

Volume 4 Issue 5

May 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 May 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 5 May 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. Alcinea 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
- **Chandrashekhar 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-Yeh 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 Elatif**
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**
AIACT&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 Fuzzy Rough Rule Based System Enhanced By Fuzzy Cellular Automata

Authors: Mona Gamal, Ahmed Abou El-Fefouh, Shereef Barakat

PAGE 1 – 11

Paper 2: Advanced Personnel Vetting Techniques in Critical Multi-Tenant Hosted Computing Environments

Authors: Farhan Hyder Sahito, Wolfgang Slany

PAGE 12 – 20

Paper 3: Fine Particulate Matter Concentration Level Prediction by using Tree-based Ensemble Classification Algorithms

Authors: Yin Zhao, Yahya Abu Hasan

PAGE 21 – 27

Paper 4: Data Flow Sequences: A Revision of Data Flow Diagrams for Modelling Applications using XML

Authors: James PH Coleman

PAGE 28 – 31

Paper 5: Comparative study of Authorship Identification Techniques for Cyber Forensics Analysis

Authors: Smita Nirkhi, Dr.R.V.Dharaskar

PAGE 32 – 35

Paper 6: Privacy Impacts of Data Encryption on the Efficiency of Digital Forensics Technology

Authors: Adedayo M. Balogun, Shao Ying Zhu

PAGE 36 – 40

Paper 7: Image Compression Using Real Fourier Transform, Its Wavelet Transform And Hybrid Wavelet With DCT

Authors: Dr. H.B.Kekre, Dr. Tanuja Sarode, Prachi Natu

PAGE 41 – 47

Paper 8: Building Low Cost Cloud Computing Systems

Authors: Carlos Antunes, Ricardo Vardasca

PAGE 48 – 52

Paper 9: Application of multi regressive linear model and neural network for wear prediction of grinding mill liners

Authors: Farzaneh Ahmadzadeh, Jan Lundberg

PAGE 53 – 58

Paper 10: DCaaS: Data Consistency as a Service for Managing Data Uncertainty on the Clouds

Authors: Islam Elgedawy

PAGE 59 – 68

Paper 11: A Computational Model of Extrastriate Visual Area MT on Motion Perception

Authors: Jiawei Xu, Shigang Yue

PAGE 69 – 78

Paper 12: A Modified clustering for LEACH algorithm in WSN

Authors: B.Brahma Reddy, K.Kishan Rao

PAGE 79 – 83

Paper 13: Jabber-based Cross-Domain Efficient and Privacy-Ensuring Context Management Framework

Authors: Zakwan Jaroucheh, Xiaodong Liu, Sally Smith

PAGE 84 – 99

Paper 14: An efficient user scheduling scheme for downlink Multiuser MIMO-OFDM systems with Block Diagonalization

Authors: Mounir Esslaoui, Mohamed Essaïdi

PAGE 100 – 106

Paper 15: QOS, Comparison of BNP Scheduling Algorithms with Expanded Fuzzy System

Authors: Amita Sharma, Harpreet Kaur

PAGE 107 – 112

Paper 16: LASyM: A Learning Analytics System for MOOCs

Authors: Yassine Tabaa, Abdellatif Medouri

PAGE 113 – 119

Paper 17: Research on Chinese University Students' Media Images

Authors: Chengliang Zhang, Haifei Yu

PAGE 120 – 129

Paper 18: Secure Medical Images Sharing over Cloud Computing environment

Authors: Fatma E.-Z. A. Elgamal, Noha A. Hikal, F.E.Z. Abou-Chadi

PAGE 130 – 137

Paper 19: Revisit of Logistic Regression

Authors: Takumi Kobayashi, Kenji Watanabe, Nobuyuki Otsu

PAGE 138 – 147

Paper 20: Study of Current Femto-Satellite Approches

Authors: Nizar Tahri, Chafaa Hamrouni, Adel M. Alimi

PAGE 148– 153

Paper 21: Neural Network based Mobility aware Prefetch Caching and Replacement Strategies in Mobile Environment

Authors: Hariram Chavan, Suneeta Sane, H. B. Kekre

PAGE 154– 160

A Fuzzy Rough Rule Based System Enhanced By Fuzzy Cellular Automata

Mona Gamal¹, Ahmed Abou El-Fetouh², Shereef Barakat³
Mansoura University, Faculty of Computer and Information Sciences
Information System Department
P.O.Box: 35516

Abstract—Handling uncertain knowledge is a very tricky problem in the current world as the data, we deal with, is uncertain, incomplete and even inconsistent. Finding an efficient intelligent framework for this kind of knowledge is a challenging task. The knowledge based framework can be represented by a rule based system that depends on a set of rules which deal with uncertainty in the data. Fuzzy rough rules are a good competitive in dealing with the uncertain cases. They are consisted of fuzzy rough variables in both the propositions and consequences. The fuzzy rough variables represent the lower and upper approximations of the subsets of a fuzzy variable. These fuzzy variables use labels (fuzzy subsets) instead of values. An efficient fuzzy rough rule based system must depend on good and accurate rules. This system needs to be enhanced to view the future recommendations or in other words the system in time sequence. This paper tries to make a rule based system for uncertain knowledge using fuzzy rough theory to generate the desired accurate rules and then use fuzzy cellular automata parallel system to enhance the rule based system developed and find out what the system would look like in time sequence so as to give good recommendations about the system in the future. The proposed model along with experimental results and simulations of the rule based systems of different data sets in time sequence is illustrated.

Keywords—fuzzy rough reduction; fuzzy rough rules; fuzzy cellular automata; Self Organized Feature Maps (SOFM).

I. INTRODUCTION

Knowledge Discovery (KD) [25] is the process of extracting valuable knowledge from concrete data sets. This process used to be accomplished manually or semi manually (part manual and part automated). The aim of soft computing techniques[15] is to completely automate the process of KD. The problem is that the data we want to extract knowledge from is uncertain, incomplete and imprecise. So we need to represent vague concepts of information. Rule based systems are composed of a set of if-then rules that represent the knowledge content of the system.

Many soft computing techniques were used for various problems in the KD process. For example, Genetic Algorithms (GAs) [5] [8] [27] [32] were used for optimization and search problems. Rough sets[31] and Artificial Neural Networks[18] are very good tools for classification and prediction problems. They are efficient in dealing with discrete data, however the real world is dealing with values like tall, short, normal, up normal and so on. Fuzzy set theory[19] [20] which deal with linguistic values of the variables are to be used here to create

rules which handle the linguistic world's problems. The hybridization between fuzzy system and various soft computing techniques is a very interesting search topic these days. Hybridizations like fuzzy rough, fuzzy neural, fuzzy genetic algorithms and many others are very powerful in dealing with uncertain knowledge in linguistic form away from the complicated mathematical calculations of probabilities. The fuzzy rough hybrid[28] is very interesting in the field of building equivalence classes with soft boundaries and degrees of membership of the objects inside these classes.

Another interesting mechanism is the fuzzy cellular automata [11]. It is a parallel processing system that is composed of a set of interconnected cells. These systems can be used efficiently to build a grid of rules in time sequence based on initial if-then rules produced from a soft computing rule generating technique. This could be very helpful for experts on the field of the data set under consideration as they need to see what the system would look like in the future. So the cellular automata can be used to enhance the rule based system to reduce the error rate or examine in which variable direction the knowledge discovered goes.

This research is concerned in producing a complete framework for KD by building a rule based system and then enhances it with the fuzzy cellular automata. This process is accomplished in three phases of a hybrid system which type is a transformation system (the output of one module is an input to the preceding module). The first module is to prepare the fuzzy variables by generating the membership function for the fuzzy subsets of the variables. This module is implemented using Self Organized Feature Maps (SOFM) [29] from a previous research. The result of this phase is passed to the fuzzy rough rule generating module which reduces the attributes first to get the reduct (the attributes which the data set fully depend on them with no redundancies) then this reduct is used to summarize the fuzzy data and produce the corresponding fuzzy rough rules. These rules are tested against the test data to measure the accuracy rates. The set of fuzzy rough rules are the core of the rule based system that represents the data set. The final module takes the rules as an initial state for the fuzzy cellular automata parallel system that iterate to generate new rules to cover the whole corresponding data space with the suitable rules that represent new data objects. This could be thought of as an enhancement of the system equation (the set of fuzzy rough rules) or the recommendations of the system in the time sequence.

The rest of this paper is organized as follows: Section II is a quick review on the previous research in generating fuzzy rules and using cellular automata in the data mining field. Section III represents the preliminaries and theories such as the declaration of the Self Organized Feature Maps (SOFM), Fuzzy rough attribute reduction, rule generation and cellular automata. Section IV gives an over view on the whole system and its modules. It goes inside the system to explain in detail the generation of the fuzzy membership functions of the subsets of the fuzzy variables, attribute reduction, designing the fuzzy rough rule set and enhancing the system using cellular automata parallel system. Experimental results and conclusion will appear in sections V and VI respectively.

II. RELATED WORK

Designing a complete framework in the KD field is a very interesting topic that many researchers cared about and tried to find the best system to accomplish the job. The system contains problems such as preparing the attributes, finding the reduct and generating the rule set that represent the core of the system. The system produced will certainly need an enhancement to its accuracy and the scope it covers on the data space. These problems many researches covered it individually in a hope to find good and efficient solutions. The reducing attribute problem is important to remove redundancy. Rough set theory is a very good way to get the reduct from crisp data but the fuzzy data may have some loss of information. A new dimensionality reduction technique that employs a hybrid variant of rough sets, fuzzy-rough sets, to avoid this information loss[30]. Genetic Algorithm[5] [8] [27] [32] [27] has been applied for the discovery of fuzzy rules which were competitive to decision tree induction rules from the perspective of predictive accuracy[32]. The individuals of the population were the fuzzy rules to be designed and the final population was the fuzzy rule set. A hybrid algorithm of two fuzzy genetics-based machine learning approaches (i.e., Michigan and Pittsburgh) for designing fuzzy rule-based classification systems has also been proposed[12] [13] [14]. A new method was also proposed to automatically learn the knowledge base (KB) by finding an appropriate database by means of a genetic algorithm while using a simple generation method to derive the rule base (RB) [24]. Also a new hybrid approach for optimization combining Particle Swarm Optimization (PSO) and Genetic Algorithms (GAs) using Fuzzy Logic for parameter adaptation and to integrate the results. Fuzzy Logic is used to combine the results of the PSO and GA in the best way possible. Also, fuzzy logic is used to adjust parameters in the FPSO and FGA[7] [8]. Other research described the use of Modular Neural Networks (MNN) for pattern recognition in parallel using a cluster of computers with a master-slave topology. Also, a parallel genetic algorithm to optimization architecture was used Fig. 8. But these researches assume that fuzzy variables are prepared or use discrimination techniques to make the membership functions of the subsets of the variables. So some researches thought of a technique for estimating the fuzzy subsets. Gene Expression Programming method uses two populations. One for Fuzzy Classification Rules which is evolved by syntax genetic programming and the other one for membership function definitions which is evolved by mutation based

evolutionary algorithm. These two populations co-evolve to better classify the underlying data set[1] [24]. A rough fuzzy hybridization to generate fuzzy if-then rules automatically from diagnoses data sets with quantitative data values based on fuzzy sets and rough set theory is accomplished in four different stages in the KD from databases[23].

There are many papers that introduced fuzzy cellular automata parallel system in the KD field to make use of its simplicity and efficiency. A fuzzy cellular traffic model is proposed which intended for detectors data fusion in traffic control system to enable utilization of complex traffic data registered by many sensors of different type's[3]. An other research introduced a cellular automata-based solution of a two-dimensional binary classification problem[2]. The proposed method is based on a two-dimensional, three-state cellular automaton (CA) with the von Neumann neighborhood. Since the number of possible CA rules (potential CA-based classifiers) is huge, searching efficient rules is conducted with use of a genetic algorithm (GA). An other paper shows a definition of a fuzzy automaton, which has the state, input, and output sets as fuzzy sets. The state transition function is defined as moving on a fuzzy relief with fuzzy peak-states and boundaries between different membership functions[17].

III. PRELIMINARIES AND THEORIES

A. Self Organized Feature Map

The Self Organized Feature Map (SOFM) [4] [29] is an unsupervised neural network that is capable of learning its weights from its input vector without supplying it with the corresponding output vector. SOFM is called self organizing or self adoption because they are able to decide what features it will use to group the input data. SOFM is usually, a two-layered network where the neurons in the output layer are organized into either a one or two-dimensional lattice structure (Bose and Liang, 1996). The SOM solves difficult high-dimensional and nonlinear problems such as feature extraction and classification of images and acoustic patterns, adaptive control of robots, and equalization, demodulation, and error-tolerant transmission of signals in telecommunications[29] and in this research it is used to find the membership function for fuzzy variables subsets. Figure 1 represents a simple structure for the SOFM where the dimension d is the number of the input neurons in the input layer for the following input data vector $x_n = [x_{n1} \ x_{n2} \ \dots \ x_{nd}]^T$ and The synaptic weight vector at neuron j in the output layer is denoted by $w_j = [w_{j1} \ w_{j2} \ \dots \ w_{jd}]^T$, $j = 1, 2, \dots, J$, where J is the total number of neurons in the output layer and w_{jk} , $k = 1, 2, \dots, d$, is the connecting weight from the j^{th} neuron in the output layer to the k^{th} neuron in the input layer.

In the learning phase, the first step is to find the best matching neuron in the output layer that is the closest to the input vector from the following equation:

$$q(x_n) = \min_{\forall j} \|x_n - w_j\| \quad (1)$$

Where,

$q(x_n)$ is the index of the winning neuron in the output layer,

x_n is the input vector,

w_j is the weight vector between the input vector and the output neuron j ,

$\| \cdot \|$ is a distance measure (usually the Euclidean norm).

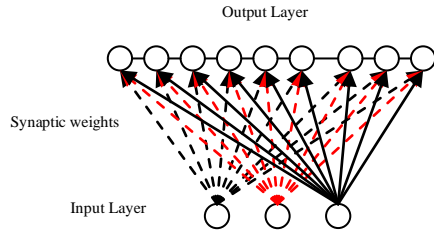


Fig. 1. a simple structure for the SOFM

The next step is to update the weight vectors associated with the winning neuron q (x_n). The learning rule for neuron $j \in N_q$ where N_q is the chosen neighborhood of winning neuron q for input vector x_n , is given by

$$w_j[t+1] = w_j[t] + \eta_{qj}[t](x_n[t] - w_j[t]) \quad (2)$$

Where

$$\eta_{qj}[t] = \begin{cases} \mu[t] & j \in N_q \\ 0 & j \notin N_q \end{cases} \quad (3)$$

Here, $\mu[t]$ is the learning rate, $0 < \mu[t] < 1$, at time index t .

In the retrieving phase, when x_n is the input vector, only the winning neuron, after convergence, will have positive response.

B. Fuzzy rough sets

Fuzzy rough set[28] is a generalization of the lower and upper approximation of the rough set[31] to allow soft boundaries. The thinking has been changed from objects which are indistinguishable (according to their attribute values) to objects similarity.

Objects are divided into fuzzy equivalence classes according to their similarities where an object could belong to more than one class with different degrees of membership. All equivalence classes are fuzzy this means that the decision values and the conditional values are all fuzzy. The lower and upper approximations are now:

$$\mu_{\underline{X}}(F_i) = \inf_x \max\{1 - \mu_{F_i}(x), \mu_X(x)\} \quad \forall i \quad (4)$$

$$\mu_{\overline{X}}(F_i) = \sup_x \min\{\mu_{F_i}(x), \mu_X(x)\} \quad \forall i \quad (5)$$

where F_i denotes a single fuzzy equivalence class. The tuple $\langle \underline{X}, \overline{X} \rangle$ is called a fuzzy-rough set. Again, it can be seen that these definitions degenerate to traditional rough sets when all equivalence classes are crisp. Additionally, if all F_i s are crisp, the result is a rough set.

C. Fuzzy rough attribute reduction

The non relevant attributes in data sets make the classification process more complicated as they take processing time and space and do not make any improvement in the classification accuracy level. The attribute reduction process aims to find those non relevant attribute and remove them from data sets before getting conducted in the classification process. The problem here is how to find those attribute carefully without affecting the final classification result. The measure of dependency between the attributes and the overall data set is reliable. For the uncertain knowledge and fuzziness issues, the fuzzy rough attribute reduction (FRAR) is an efficient and previously tested algorithm. The FRAR Algorithm is a new dimensionality reduction technique that employs a hybrid variant of rough sets (fuzzy-rough sets) in calculating the dependency between attributes and the data set avoiding information loss[30]. FRAR make use of the concepts of vagueness (for fuzzy sets) and indiscernibility (for rough sets) to measure the dependency degree between the fuzzy attributes and the uncertain data sets using the membership function of each attribute and the membership degree of the objects in each class to find the best reduct. The dependency degree is calculated over the attribute fuzzy positive region. The attribute that maximizes the overall dependency is added to the core attribute sets. The equations are illustrated in section IV.B.

D. Fuzzy rough rule generation

The fuzzy rough rules in a rule based system are if-then rules with fuzzy rough attributes in the conditional and the consequences parts. These if-then rules can be easily found by summarizing the reduced data set resulted from the attribute reduction process. This method in summarizing the data set is used previously in a research paper for rule extraction using soft computing techniques[23]. The problem in fuzzy data is that there are real values with membership degrees that may be a difficulty for the rough set technique to discretize such data and get the if-then rules. The technique here is to transform the attribute real value into a tuple that represent the membership degrees in all the attribute subsets.

For example the fuzzy attribute height is represented by the membership function illustrated in figure 2. The value 170 can be written in the format (S: 0.0; M: 0.3; T: 0.7) where S, M and T are the attribute subsets and 0.0, 0.3, 0.7 are the membership degrees in each subset respectively. In the data summarizing process, we try to type this tuple in the 0's and 1's format so the rough set theory can be used to summarize the core data set and find all the if-then rules that represent the core data set. The 1's are given to the membership degrees more than or equal to 0.5 and 0's otherwise. The previous tuple can be written as (0; 0; 1). If there are two membership degrees in the tuple of value 0.5 then both of them will be 1 (i.e. (L:0.5;M:0.5;H:0.0) will be (1;1;0)).

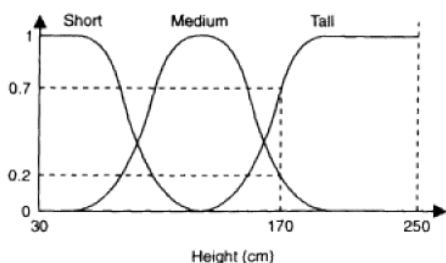


Fig.2. Membership functions representing three fuzzy sets for the variable "height."

E. Cellular automata parallel system

Cellular automata(CA)[16] is a parallel distributed processing system that aims to build a grid of cells from some initial configuration in the time sequence iterations according to some transition function (update rule). In 1d grids, there is a vector of cells where each cell has a state. The cell changes its state according to the states of its neighborhood's states and its state in the previous time step. Following [Wolfram, 1984], [Wolfram, 2002] one can represents any CA with two parameters (k; r).

Where k is a number of states, and r is a radius of neighborhood. Thus CA is defined by parameters (2; 1). There are $n = k2r+1$ different neighborhoods and kn different evolution rules. For a cellular automata vector where each cell has 2 states and 1 radius tall then the number of rules that design the cell states in the next time step will be 265 rules. These rules are called transition functions that can be written in the form

$$C_i(t+1) = F(\dots, c_{i-1}(t), c_i(t), c_{i+1}(t+1), \dots) \quad (6)$$

Where

c_i is the cell state,

$\dots, c_{i-1}(t), c_i(t), c_{i+1}(t+1), \dots$ are the cell neighbors,

F is the transition function.

In 2d and 3d cellular grids become very large, resulting in time and space consuming problems, so researchers have made some rules to generate the cells in the next time step without the conventional updating rules. It is called game of life, as it simulates human generation and existence, where the neighbors of the cells determine the cell state in the next time step. Von Neumann neighborhood and Moore neighborhood are two common neighborhood definitions for the cells in a two dimension cellular grid.

In von Neumann, each cell has neighbors to the north, south, east and west. The Moore neighborhood adds the diagonal cells to the northeast, southeast, southwest and northwest. Figure 3 shows these two neighborhood models in two dimensions. In general, in a d-dimensional space, a cell's von Neumann neighborhood will contain $(2*d)$ cells and its Moore neighborhood will contain $(3*d - 1)$ cells where d is the number of dimensions.

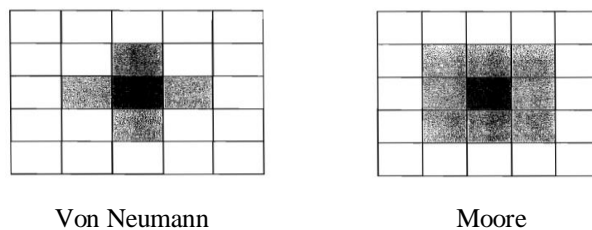


Fig. 3. Cellular automata neighborhood types

Instead of transition functions on the neighbor's states, the game of life puts two simple rules invented by the English mathematician Jon Horton Conway [16] .

These rules are:

- A dead cell becomes alive at the next generation if exactly three of its neighbors are alive.
- Alive cell at the next generation remain a live if either two or three of its neighbors are alive but otherwise it dies.

In dealing with data mining[6] , cellular automata grid represents the instance space and hence the cell state will represent the instance class. The cells are organized and connected according to attribute value ranges. The instance space will form a (multi-dimensional) grid over which the cellular automata operate. The grid will be initialized with training instances, and the CA run to convergence. The state of each cell of the cellular automata grid will represent the class assignment of that point in the instance space. The intention is that cells will organize themselves into regions of similar class assignment.

The transition function (update rule) becomes a simple voting rule that locally reduces entropy. The voting rule examines a cell's neighbors and sets the cell according the number of neighbors that are set to a given class. In this work we use the von Neumann neighborhood because it is linear in the number of dimensions of the instance space, so it scales well.

This is implemented in a transition rule called n4V1[2] . This rule examines each cell's four neighbors and sets its class to the majority class. It can be a stable rule, in that, once the cell's class has been set it will not be changed. It also can be nonstable, where the cell state can change along the iterations if the majority class of the neighborhood changed.

The global effect of the n4V1 update rule is that each cell in the grid becomes assigned with the class of the nearest initial point as measured by Manhattan distance. The Manhattan distance aspect stems from the fact that the CA uses the von Neumann neighborhood, so each cell's influence spreads outward along the two dimensions. The first neighbor of a cell that changes state, from empty to a class, will result in that its neighborhood cells changing state in the next time step[6] . If 0 means that the cell is empty and the instance's classes are 1 and 2 then the n4V1 nonsatble update rule will look like:

$$n4V1 \text{ nonstable} = \begin{cases} 0 : \text{class1 neighbors} + \text{class2 neighbors} = 0 \\ 1 : \text{class1 neighbors} > \text{class2 neighbors} \\ 2 : \text{class1 neighbors} < \text{class2 neighbors} \\ \text{rand}(\{1,2\}) : \text{class1 neighbors} = \text{class2 neighbors} \end{cases} \quad (7)$$

where rand (1, 2) selects randomly from the elements with equal probability.

This was the Boolean cellular automata where the cell either alive or dead (has a class or not). But in the case of uncertainty, the fuzzy cellular[12] is more efficient as its cells contain the state (class) and the membership degree of that class. This will help handle uncertain data where the variable are fuzzy and hence producing fuzzy classes.

Fuzzy cellular automata (FCA) [12] are continuous cellular automata where the local rule is defined as the “fuzzification” of the local rule of a corresponding Boolean cellular automaton in disjunctive normal form[11]. The “fuzzification” is accomplished by using the fuzzy extension of the Boolean operators AND, OR and NOT. Depending on which fuzzy operator is used, a different type of Fuzzy cellular automata can be defined. Among the various possible choices for the fuzzy operators, we consider the following: $(a \vee b)$ is replaced by $(\max(a, b))$; $(a \wedge b)$ by $(\min(a, b))$, and $(\neg a)$ by $(1 - a)$. The resulting local rule becomes the fuzzy real function that generalizes the original function. Then the n4V1 rule can be fuzzified by applying the fuzzy operators.

IV. THE PROPOSED FRAMEWORK

The framework proposed by this paper is to generate a fuzzy rough rule based system and enhance it using the fuzzy cellular automata. The fuzzy rough rules are simple if-then rules but with fuzzy rough variables. This research tries to build these fuzzy rough rules in three phases. The first phase is to generate the membership function for the subsets of the fuzzy variables. The second phase is to reduce the features using the fuzzy membership dependency between the features and the data set. The third phase is to design the fuzzy rough rule by summarizing the data of the reduced features basing on the rough set theory then tests these rules efficiency by the test data set. The proposed framework that outlines the main modules is represented in figure 4. The data used in the training process as well as the features data collected from experts are used as inputs to the Generating fuzzy membership function for features subsets process which outputs a data file that contains the values of the fuzzy variables and their corresponding membership degrees in the subsets of these variables. These fuzzy membership degrees and the training data again are used as inputs to the reducing features process that measures the attributes dependency and produces the core attributes. The generating fuzzy rough rules process takes the core attributes (reduced training data set) and summarizes the data set to output the corresponding fuzzy rough rules set after testing them by the test data records. The set of the fuzzy rough rules are used as an initial state for the fuzzy cellular automata parallel system to enhance the rule based system and produce what can be said as an equation of the system on time sequence. The main components are:

Generating fuzzy membership function for features subsets: This module uses the SOFM capabilities of unsupervised learning and clustering to generate the membership functions of the features subsets.

Reducing Features: This module uses the fuzzy rough attribute reduction (FRAR) algorithm to reduce the features basing on the measuring the dependency membership degree between the fuzzy variables and the training data set to produce the reduct (core attributes).

Generating fuzzy rough rules: This module summarizes the data set using the reduced data set according to the fuzzy rough theory to generate the corresponding fuzzy rough rules. These rules are applied on the testing data set to measure its accuracy rate.

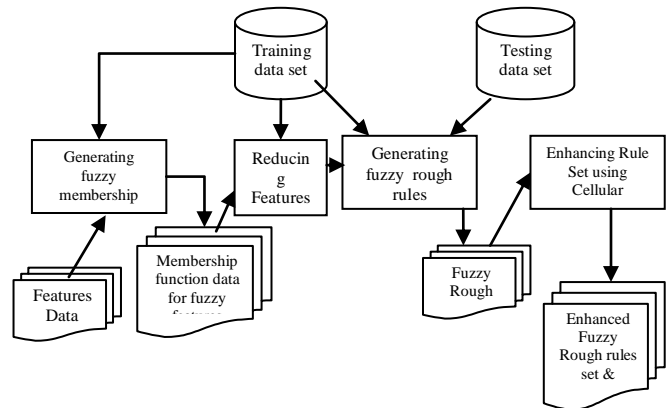


Fig. 4. the proposed framework for uncertain Knowledge

Enhancing Rule Set using Fuzzy Cellular Automata: this module takes the fuzzy rough rule set as an initial state and iterate according to the fuzzy n4V1 nonsatble update rule to produce the view of the fuzzy rough rule based system in the time sequence. These views of the system may give good recommendations for the experts in the field of the data under consideration (training data) and cover all the data space with accurate rule for more classification or prediction issues.

A. Preparing Fuzzy Features (variables)

The process of generating membership functions used to be in two phases. The first phase generates the proper clusters of the feature data. Then, the fuzzy membership function is generated according these clusters. But it is possible to generate the membership functions in one phase by combining the variable labels with the variable values in the input layer of the SOFM[4]. As mentioned the SOFM learns from its input vector so the input vector will be $X_n = (v, S_1, S_2, \dots, S_d)$ where v is the value of the feature under consideration and S_1, S_2, \dots, S_d are the subsets that the feature will be divide to. These input vectors will be the training data and can be got by asking an expert some questions about the features like do you think that the value v of feature f belongs to s_1 or s_2 or \dots, s_d ? The SOFM then goes through the learning phase and update its weights according to the learning procedure mentioned in section

III.A. After convergence the weights of the SOFM will be the values of the variable and its corresponding membership degrees of the labels (subsets) in the input layer. This technique is illustrated before in a research for Generating fuzzy membership function with self-organizing feature map by Chih-Chung Yang, N.K. Bose.

Example 1 [4] is a graphical example for illustrating the technique. Suppose that the fuzzy subsets for a variable like height would be ‘short’ and ‘tall’ and the membership function for each subset is to be found. The dataset could be collected by asking a question “Do you think a person with height 6 feet is tall or short?” After dataset is collected, the labeling information may be represented by 2-D unit vectors $[1\ 0]^T$ and $[0\ 1]^T$ for fuzzy variable ‘short’ and ‘tall’, respectively. In the training phase of SOFM, the input feature height was combined with the labeling information to form a 3-D vector, which would be the input training sample for the SOFM.

Suppose there are five neurons in the output layer as in figure 5 and the associated weights after training process are listed in Table 1.

The fuzzy membership functions for the fuzzy variables tall and short are illustrated in figure 6. From this example the feature value height=5 has a degree of membership 0 in the subset tall and a degree = 1 in the subset short.

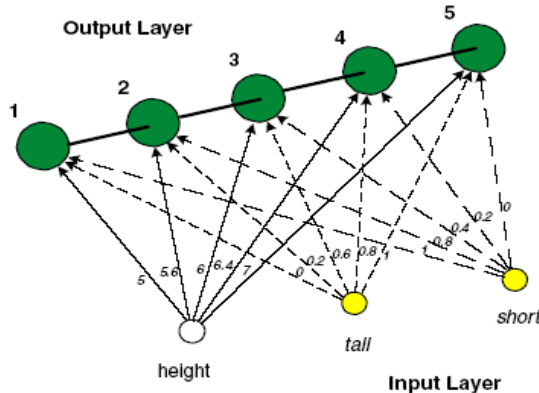


Fig. 5. [4] SOFM after training process

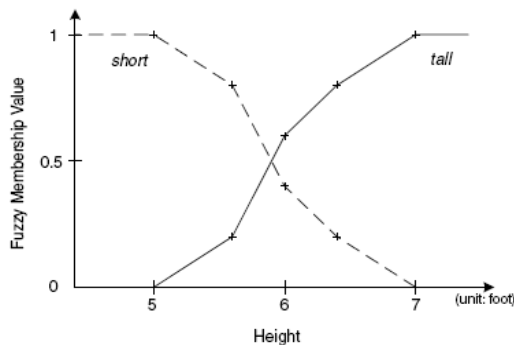


Fig. 6. [4] : Fuzzy membership function for fuzzy subsets tall and short

Neuron index	Associated weights	Feature height	Fuzzy variable	
			Tall	Short
1	$[5\ 0\ 1]^T$	5	0	1
2	$[5.6\ 0.2\ 0.7]^T$	5.6	0.2	0.7
3	$[6\ 0.6\ 0.4]^T$	6	0.6	0.4
4	$[6.4\ 0.8\ 0.2]^T$	6.4	0.8	0.2
5	$[7\ 1\ 0]^T$	7	1	0

B. Reducing Features

The reduction process aims to find the set of core attributes which the training data equivalent classes depend fully on them. Starting with the core attributes as an empty set, it can be found by calculating the dependency between all the attributes and the equivalent classes one at a time and add the attribute that maximizes the dependency degree to the core attribute set. This process is repeated until the dependency come to its highest level (usually 1). The problem, for the fuzzy attributes, is that the attributes must go through a discretization process to calculate the dependency and this causes information loss. The fuzzy Rough Set Reduction[30] solved this problem by calculating the fuzzy membership dependency degree between the fuzzy variables (attributes) and the fuzzy equivalent classes. For D is the set of equivalent classes on the universe U and A is an attribute in the fuzzy rough attributes set, the dependency can be calculated using the following equation:

$$\gamma'_{A}(D) = \frac{|\mu_{pos_A(D)}(x)|}{|U|} = \frac{\sum_{x \in U} \mu_{pos_A(D)}(x)}{|U|} \quad (8)$$

This dependency equation takes the membership degree of an object x to the fuzzy positive region of the fuzzy rough attribute A as a parameter which is:

$$\mu_{pos_A(D)}(x) = \sup_{x \in U/D} \min(\mu_{F_i}(x), \mu_{pos_A}(F_i)) \quad (9)$$

where

F_i is the fuzzy attribute subsets in D (equivalent classes);

$\mu_{F_i}(x)$ is the membership degree of object x in F_i ;

$\mu_{pos_A}(F_i)$ is the fuzzy positive region of a fuzzy equivalence class $F_i \in U/A$ which can be defined as:

$$\mu_{pos_A}(F_i) = \sup_{x \in U/D} (\mu_x(F_i)) \quad (10)$$

where

$$\mu_x(F_i) = \inf_{x \in U} \max\{1 - \mu_{F_i}(x), \mu(x)\} \quad (11)$$

where $\mu(x)$ are the membership degrees for the fuzzy attribute subsets (fuzzy membership function). The fuzzy attribute can be declared by its fuzzy membership function. Figure 2 shows a fuzzy attribute with three subsets short, medium and tall.

This process is repeated for each attribute and the one with the biggest dependency is added to the core set until no further increasing in the dependency is accomplished. In the case of uncertain knowledge, this process can never reach the absolute dependency so a threshold can be taken to define the level of certainty accepted taking in account the time and space complexity of the algorithm. The algorithm along with an illustrating numerical example can be found in[30].

C. Generating Fuzzy Rough Rules

In the Rule based system, the Rules are the system equation so finding the best accurate rules is the only way to make an accurate system. The fuzzy rough set theory is used to get the reduct fuzzy attributes and remove the unnecessary ones but it still a data set and we need to find the necessary rules to represent it. Data summarizing is an efficient way to get the if then rules but how could fuzzy attributes be summarized hence each value is represented by the corresponding membership degrees. This process of summarizing will be only used to get the if-then rules but the membership degrees will be reserved by the attribute membership function and the data set attribute values. The next tables illustrate the whole summarizing process step by step from the data (iris data set) before reduction to the if then rules.

sepal length	sepal width	petal length	petal width	class
5.7	4.4	1.5	0.4	1
5.4	3.9	1.3	0.4	1
...				
6.4	3.2	4.5	1.5	2
6.9	3.1	4.9	1.5	2
...				
6.2	2.8	4.8	1.8	3
6.1	3.0	4.9	1.8	3

After the feature reduction process the data set should look like:

petal length	petal width	Class
1.5	0.4	1
1.3	1.6	1
...		
4.5	1.5	2
4.9	1.5	2
...		
4.8	1.8	3
4.9	1.8	3

The attribute real value can be represented by the tuples [M1; M2; M3] where M1, M2 and M3 are the membership degrees in the fuzzy attributes subsets coming from the fuzzy attributes membership function defined for each attribute in the preprocessing phase.

Petal length	Petal width	class
[1; 0; 0]	[0.92;0.07;0.0003]	1
[1; 0; 0]	[0.07;0.92;0.0003]	1
[0;0.99;0.01]	[0.0033;0.8212;0.175]	2
[0.00024;0.4592;0.5405]	[0.0033;0.8212;0.175]	2
[0.0008;0.603;0.3952]	[0.003;0.070;0.929]	3
[0.00024;0.4592;0.5405]	[0.003;0.070;0.929]	3

These membership degrees can be summarized as previously illustrated in the form of 0's and 1's as in table 5.

petal length	petal width	Class
[1; 0; 0]	[1; 0; 0]	1
[1; 0; 0]	[0; 1; 0]	1
[0; 1; 0]	[0; 1; 0]	2
[0;0;1]	[0; 1; 0]	2
[0; 1; 0]	[0;0;1]	3
[0;0;1]	[0;0;1]	3

These tables can then be transformed into fuzzy rough if then rules for the entire data set as follows:

	support	confidence
petal length is short & petal width is short then Iris-setosa	0.45	0.99
petal length is short & petal width is medium then Iris-setosa	0.013	0.997
petal length is medium & petal width is medium then Iris-versicolor	0.213	0.998
petal length is tall & petal width is medium then Iris-versicolor	0.12	0.529
petal length is medium & petal width is tall then Iris-virginica	0.106	0.47
petal length is tall & petal width is tall & then Iris-virginica	0.093	1

D. Enhancing System Equation by Fuzzy cellular automata

The set of fuzzy rough rules are the system equation and can be represented graphically on a two or three dimension grid. The grid dimensions are decided by the number of conditional attributes in the core set. The grid coordinates are the core attributes and cells will represent the result (class with membership rule degree inferred) of the if-then rules. The cell state will be the c & μ where the c is the class index and μ is the membership degree of that class (calculated initially from the fuzzy rules using the simple inference methods). During the iterations of the fuzzy cellular automata parallel system, the cell state will be formed according to the fuzzy n4V1 nonstable update rule (transition function). This transition function is a fuzzification of the regular n4V1

nonstable update rule which resulted by replacing the Boolean operators AND, OR and NOT by their fuzzy extensions. The update rule decides the cell state at the next time step $[C_i(t+1) \& \mu_{ci}(t+1)]$ basing on the cell and the cell's neighbors states at the current time step.

$$\text{Fuzzyn4 V1 nonstable} = \begin{cases} [0 \& 0] : \sum_i \mu_{ci}(t) = 0 \\ [\max_{\mu_{ci}}(ci) \& \max_{\mu_{ci}}(\mu_{ci}(t)) : \mu_{ci}(t) \text{are_not_equal} \\ [\text{rand}_i(i) \& \text{rand}_{\mu_{ci}}(\mu_{ci}(t)) : \mu_{ci}(t) \text{are_equal} \end{cases} \quad (12)$$

Where $\mu_{ci}(t)$ is the class membership degree of the class i at time t .

The membership degree at time step t is the conjunction of all the neighbors' membership of the same class index:

$$\mu_{ci}(t) = \min(\mu_{\text{class}_i \text{ neighbors}}(t)) \quad (13)$$

or we can take the average instead of the min functions:

$$\mu_{ci}(t) = \text{avg}(\mu_{\text{class}_i \text{ neighbors}}(t)) \quad (14)$$

V. EXPERIMENTAL RESULTS

The proposed hybrid model is composed of four main sub modules that are implemented using soft computing techniques. The first sub model is the Generating fuzzy membership functions for features subsets which is responsible for generating the degree of membership of the values of the variables in their corresponding subsets. This process uses the SOFM, which uses its unsupervised learning and clustering ability to learn the weights of the neural net from the input training data vectors, to obtain the variables values and their corresponding membership degrees. Using these values we can draw an analogue membership function for each subset of the variables. This technique is implemented before in a previous work [4] and we just make use of it to prepare the fuzzy variables for the generating fuzzy rough rules process. The second sub module is the attribute reduction process which makes use of the fuzzy rough dimensionality reduction process to measure the membership dependency degree between the whole data set and each attribute. This process picks the set of attributes that maximizes the dependency with the minimum loss of information. This set of attributes are said to be the core attributes which is used later in the third sub module to generate the set of fuzzy rough rules which represent the system. The third process is the generating fuzzy rough rules module that applies the summarizing technique to transform the data set under the core attributes into a set of fuzzy rough if-then rules. The fourth module is the rule enhancement which makes use of the parallel distributed processing of the fuzzy cellular automata system to generate the system equation on time series. The fuzzy cellular takes the set of fuzzy rough rules generated in the third process as its initial generation and use the fuzzy update rule (fuzzy transition function) to generate new fuzzy rough rules

its corresponding membership function. These new fuzzy rough rules could be used by experts to examine the system behavior in time steps and to make the system grid visible to see which class is most probable in the direction of which attribute. This could be thought of as a prediction system as well as a classification system.

In experiments, SOFM is used to generate the membership function of each of the attributes in the data set. The SOFM is trained with 3 or 4 input neurons (one for the feature value and the rest for the subsets) depending on the number of subsets of the features and 15 output neurons that produced 225 feature values and their corresponding membership degrees. These values were used to draw an analogue function for each feature. Figure 7 shows the membership functions of the 9 conditional features (before reduction) of the breast cancer data set. These membership functions along with the data set goes through the fuzzy rough dimensionality reduction process that measures the attributes dependency and selects the core attributes that represent the data set and reduce the data set accordingly. This process depends on some threshold which defines the acceptable loss of information. In some data sets this threshold could be 0 and in others could be 0.01 or 0.001. The threshold is selected after applying more than one trial of the reduction process and selecting threshold that reduces the error rate.

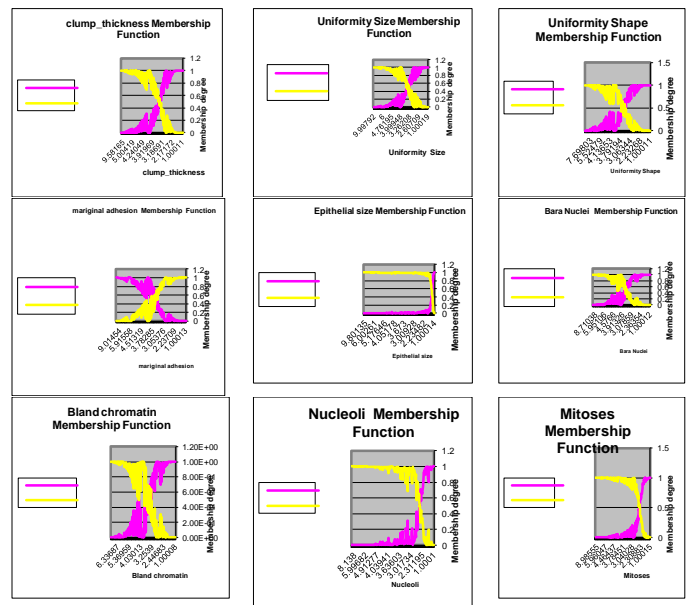


Fig. 7. Membership functions of the liver data features from the SOM process

The reduced data set (data of the core attributes only) is summarized according to the technique illustrated in section IV.C. The set of summarized fuzzy rough if-then rules are tested against a test data set to measure the overall accuracy of the system equation (if-then rules). After the whole rule generation process is finished, the if-then rules are used as the initial generation in the fuzzy cellular grid. The cell states are used as $c_i \& \mu$ where the c_i is the rule output (2 or 3 classes in experiments) and μ is the rule membership degree that

corresponds to the cell. The fuzzy transition function (fuzzy update rule) illustrated in section IV.D is used to generate fuzzy rough rule based system in time steps basing on the values of the cell's neighbors states. After convergence, the cellular grid plots the system equation which represents the system behavior.

The breast cancer fuzzy rough rule based system has a core of two attributes (the reduct is Uniformity Size and Epithelial Size). The system two dimensions grid resulting from the fuzzy cellular enhancement process along with the initial generation are illustrated in figure 9. The blue is the benign class and the color degrees are to which degree the class is accurate (membership degree of the class according to the rule membership inference) and the red is the malignant class and also the color degrees are the membership degrees. The grids are plotted by mat lab and the color bar in the figures represents the coloring memberships. The data sets used in this research to test the model are taken from the UCI machine learning repository and their properties are illustrated in table 7. The data set records are divided in two equal parts (one for the training data and one for the test data).

Table 7 : Description of the data sets properties

Name of the data set	No of attributes	No of continuous attributes	No of categorical attributes	No of data records	No of classes
weather	4	2	2	14	2
Breast Cancer	10	10	0	699	2
Wine	13	13	0	168	3
liver	6	6	0	345	2
Iris	4	4	0	150	3

The comparison between the proposed model and other techniques is listed in table 8 which shows the accuracy levels of the rule sets generated by C4.5, Neural Networks, Naïve Bays, SOFM&PGA[22] and the proposed model (SOFM + Fuzzy Rough) on five different data sets. Figure 8 shows the same comparison in graphical mode.

These comparisons show that the proposed hybrid model gave better accuracy level than the previous ones. The results indicate an average accuracy of around 87%, with accuracies above 60.1% even for quite small training sets. This compares favorably with previous systems for classifying the same data sets, whose average accuracy is 80%.

Table 8: comparison between the proposed model and other techniques found in the field of generating fuzzy rules

	c4.5	neural	naïve Bays	SOFM +PGA	SOFM + Fuzzy Rough
iris	84.5	91.2	88	81.3	89.3
Weather	68	78	64	71.4	100
liver	49	47	51	60.7	60.1
Breast Cancer	95.1	95.7	95.7	94.2	97
Wine	82	89	88	64.3	88.5

iris	84.5	91.2	88	81.3	89.3
Weather	68	78	64	71.4	100
liver	49	47	51	60.7	60.1
Breast Cancer	95.1	95.7	95.7	94.2	97
Wine	82	89	88	64.3	88.5

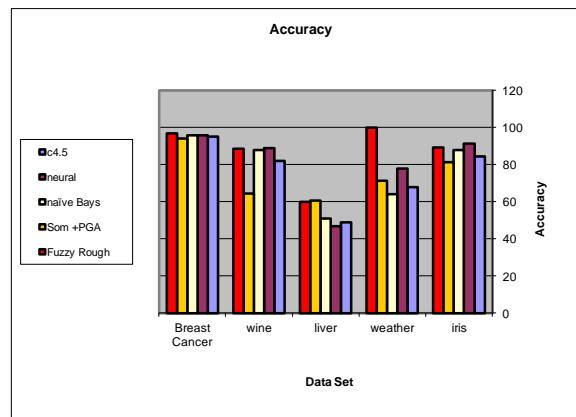
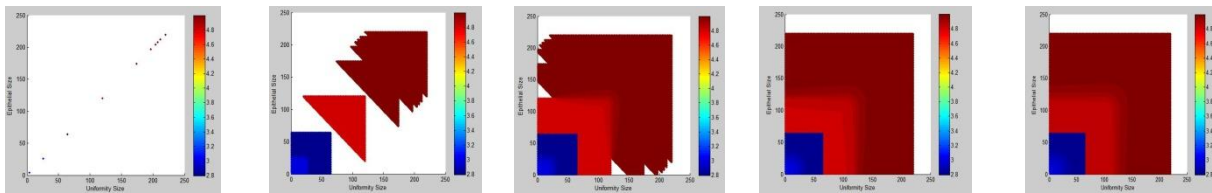


Fig. 8. The accuracy of the rule set of the proposed model and some other rule generator algorithms

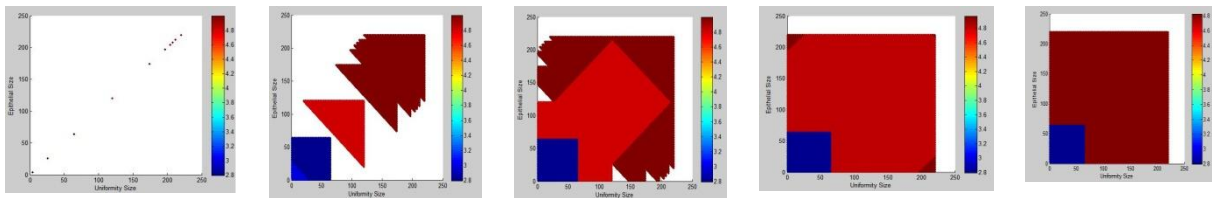
VI. CONCLUSION

Knowledge based systems are very important tools in data mining field as they extract knowledge from concrete uncertain data. Rule based systems are a branch from knowledge based systems that depend on a set of rules to extract the information (consequences) from data (propositions). The uncertainty in the data requires a special kind of rules. Fuzzy rough rules are concerned with uncertain propositions and consequences (variables). These uncertain propositions and consequences can be represented by fuzzy variables which deal with uncertainty by means of membership functions declared for the variable values. Defining good fuzzy variables is important for generating efficient fuzzy rough rules.

Designing fuzzy rough rule based system that deals with uncertainty contains a number of problems such as declaring the variables membership functions, reducing the features (variables) to get the core attributes and generating the set of fuzzy rough rules. These problems can be solved by soft computing techniques so as to automate the whole knowledge based system. The system can help in extracting knowledge but only according to the space that the rules cover. Enhancing the fuzzy rule based system is another problem that needs to be automated and the system will be suitable to cover the whole data space and give future recommendations for experts as well.



Initial grid and the fuzzy cellular output grids after 100,200,300 and 400 iterations using the equation No. 14



Initial grid and the fuzzy cellular output grids after 100,200,300 and 400 iterations using the equation No. 13

Fig. 9. The breast cancer fuzzy rough rule grids in time series.

This paper is concerned with solving the problems which face the knowledge based framework automation process by soft computing algorithms. The first problem is declaring fuzzy variables and their corresponding membership functions. SOFM are used in clustering its inputs. This property can be used efficiently in clustering the variables into the correct subsets and produce the representative membership functions. The second problem is the dimensionality problem (reducing the feature set size). The fuzzy rough attribute reduction (FRAR) algorithm measures the dependency membership degree between the fuzzy variables and the data set basing on the fuzzy-rough set theory and this will help in generating the core attributes (reduct) and solving the dimensionality reduction problem. The third problem is the fuzzy rough rule generation which is solved by the data summarizing process which find the rules from removing redundancy and repetition from the data sets. Solving these problems contributes in producing an efficient and accurate fuzzy rough rule based system which represents the knowledge based framework which this research intended to make. This system is accurate according to the accuracy measure process by calculating the error rate of the system when applied to a test data set. In an enhancement process to the fuzzy rough rule based system, fuzzy cellular automata parallel system is used to generate the system states in time series. Using the system's set of fuzzy rough rules (which can be thought of as the system equation) as the initial state and the fuzzy update rule (fuzzy transition function) which is a fuzzy equivalent to the n4v1 non stable update rule, the fuzzy cellular automata iterate to produce the system equation on time series, cover the whole data space with accurate rules and also can give suitable recommendations about the data set under consideration to the experts.

The experimental results illustrated the framework development process and the parameters used in each step. The comparisons between the proposed framework and other systems on different data sets proved that the proposed framework is accurate and stable even for small size data sets.

The fuzzy rough rule based system grid for the breast cancer data set is presented on its initial state and in time series after the cellular automata parallel system worked on it.

REFERENCES

- [1] A. A. Freitas, "Data Mining and Knowledge Discovery with Evolutionary Algorithms", Springer, Berlin, 2002.
- [2] A. Piwonska and F. Sereyński, "Solving Two-Dimensional Binary Classification Problem with Use of Cellular Automata", in *AUTOMATA the 17th International Workshop on Cellular Automata and Discrete Complex Systems Proceedings*, Santiago, Chile, 2011.
- [3] B. PLACZEK, "Fuzzy cellular model for traffic data fusion, Transport problems", volume 4 issue 4, pp. 25-35, 2009.
- [4] Chih-Chung Yang, N.K. Bose, "Generating fuzzy membership function with self-organizing feature map", *Letters Volume*, 1, Pages 356-365, April 2006.
- [5] D. E. Goldberg, "Genetic algorithms in search, optimization, and machine learning", Addison-Wesley, 412, 1989.
- [6] Fawcett, "Data mining with cellular automata", *ACM SIGKDD Explorations Newsletter*, 10(1): pp. 32-39, 2008.
- [7] Fevrier Valdez, Patricia Melin and Herman Parra, "Parallel genetic algorithms for optimization of Modular Neural Networks in pattern recognition", *IJCNN*, pp.314-319, 2011.
- [8] Fevrier Valdez, Patricia Melin and Oscar Castillo, "Evolutionary method combining Particle Swarm Optimisation and Genetic Algorithms using fuzzy logic for parameter adaptation and aggregation: the case neural network optimization for face recognition", *IJAISC*, Vol.2(1/2), pp.77-102, 2010.
- [9] Fevrier Valdez, Patricia Melin and Oscar Castillo, "An improved evolutionary method with fuzzy logic for combining Particle Swarm Optimization and Genetic Algorithms". *Appl. Soft Comput.*, Vol.11(2), pp.2625-2632, 2011.
- [10] Fevrier Valdez, Patricia Melin and Oscar Castillo, "Evolutionary method combining Particle Swarm Optimisation and Genetic Algorithms using fuzzy logic for parameter adaptation and aggregation: the case neural network optimization for face recognition", *IJAISC*, Vol.2(1/2), pp.77-102, 2010.
- [11] H. Betel and P. Flocchini, "On the Relationship between Boolean and Fuzzy Cellular Automata", 2009.
- [12] H. Ishibuchi, K. Nozaki and H. Tanaka, "Adaptive Fuzzy Rule-Based Classification Systems", *IEEE Trans. on Fuzzy Systems*, vol. 4, no. 3, pp. 238-250, 1996.
- [13] H. Ishibuchi, T. Nakashima and T. Murata, "Performance Evaluation of Fuzzy Classifier Systems for Multi-Dimensional Pattern Classification

- Problems", IEEE Trans. Syst., Man, Cybern, Part B, vol. 29, pp. 601-618, 1999.
- [14] H. Ishibuchi, T. Nakashima and T. Murata, "A Fuzzy Classifier System that Generates Fuzzy If-Then Rules for Pattern Classification Problems", Proc. of 2nd IEEE Int. Conf. Evolutionary Computation, Perth, Australia, pp. 759-764, Nov. 29-Dec. 1, 1995.
- [15] Janusz Kacprzyk, "Studies in Fuzziness and Soft Computing" ,ISBN 978-3540737223, ISSN: 1434-9922 (Print) 1860-0808 (Online), Springer Berlin / Heidelberg ,2009.
- [16] J.L. Schiff, "*Cellular Automata A Discrete view of the world*. 2008: Jhon Wiley & SONS INC. publications
- [17] J. Virant and N. Zimic , "Fuzzy automata with fuzzy relief", IEEE Trans. Fuzzy Systems, Vol. 3, No. 1, pp. 69-74, 1995.
- [18] Kenji Suzuki, "Artificial Neural Networks: Methodological Advances and Biomedical Applications", InTech, ISBN-13: 9789533072432, 2011.
- [19] Lotfi A. Zadeh, "From computing with numbers to computing with words —from manipulation of measurements to manipulation of perceptions", in International Journal of Applied Math and Computer Science, pp. 307-324, vol. 12, no. 3, 2002.
- [20] Lotfi A. Zadeh, "Fuzzy sets and systems". In: Fox J, editor. System Theory. Brooklyn, NY: Polytechnic Press, pp. 29-39, 1965.
- [21] M. Vose., "The Simple Genetic Algorithm Foundation and Theory", MIT Press, 251, 1999.
- [22] Mona Gamal, Ahmed Abo El-Fatoh, Shereef Barakat and Elsayed Radwan, "A Hybrid of Self Organized Feature Maps and Parallel Genetic Algorithms for Uncertain Knowledge", International Journal of Computer Applications (0975 – 8887) Volume 60– No.6, December 2012.
- [23] Nan-Chen Hsieh , "Rule Extraction with Rough-Fuzzy Hybridization Method", Advances in Knowledge Discovery and Data Mining; Lecture Notes in Computer Science, Vol. 5012, pp 890-895, 2008.
- [24] O. Cordon, F. A. C. Gomide, F. Herrera, F. Hoffmann and L. Magdalena, "Ten Years of Genetic Fuzzy Systems: Current Framework and New Trends", Fuzzy Sets and Systems, Pages 5-31, 2004.
- [25] Oded Maimon and Lior Rokach , "Soft Computing for Knowledge Discovery and Data Mining" , ISBN-10: 0387699341, ISBN-13: 978-0387699349, Springer; 2008 edition, November 26, 2007.
- [26] O. Cordon, F. Gomide, F. Herrera, F. Hoffmann, and L. Magdalena, "Ten years of genetic fuzzy systems: Current framework and new trends", Fuzzy Sets and Systems, pp. 5-31, 2004.
- [27] Saroj, Nishant Prabhat," A Genetic-Fuzzy Algorithm to Discover Fuzzy Classification Rules for Mixed Attributes Datasets", International Journal of Computer Applications, Vol 34– No.5, November 2011.
- [28] S. P. Tiwari and Arun K. Srivastava, "Fuzzy rough sets, fuzzy preorders and fuzzy topologies Fuzzy Sets and Systems", Vol. 210 , pp. 63-68, January 2013.
- [29] T. Kohonen, "Self-Organizing Maps", Springer Series in Information Sciences, Vol. 30, Springer, Berlin, Heidelberg, New York, ISBN 3-540-67921-9, ISSN 0720-678X 1995, 1997, 2001.
- [30] R. Jensen and Qiang Shen, "Fuzzy-Rough Sets for Descriptive Dimensionality Reduction", Proceedings of the IEEE International Conference on Fuzzy Systems, Vol: 1 ,Pp 29 – 34, 2002.
- [31] Y. Caballero, D. Alvarez, R. Bello and M. M. Garcia, "Feature Selection Algorithms Using Rough Set Theory, In Intelligent Systems Design and Applications", ISDA Seventh International Conference on, pp. 407-411 , 2007.
- [32] Z. Michalewicz, "Genetic Algorithms + Data Structures = Evolution Programs", Springer-Verlang, 252, 1992.

Advanced Personnel Vetting Techniques in Critical Multi-Tenant Hosted Computing Environments

Farhan Hyder Sahito
Institute for Software Technology
Graz University of Technology
Graz, Austria

Wolfgang Slany
Institute for Software Technology
Graz University of Technology
Graz, Austria

Abstract—The emergence of cloud computing presents a strategic direction for critical infrastructures and promises to have far-reaching effects on their systems and networks to deliver better outcomes to the nations at a lower cost. However, when considering cloud computing, government entities must address a host of security issues (such as malicious insiders) beyond those of service cost and flexibility. The scope and objective of this paper is to analyze, evaluate and investigate the insider threat in cloud security in sensitive infrastructures as well as to propose two proactive socio-technical solutions for securing commercial and governmental cloud infrastructures. Firstly, it proposes actionable framework, techniques and practices in order to ensure that such disruptions through human threats are infrequent, of minimal duration, manageable, and cause the least damage possible. Secondly, it aims for extreme security measures to analyze and evaluate human threats related assessment methods for employee screening in certain high-risk situations using cognitive analysis technology, in particular functional Magnetic Resonance Imaging (fMRI). The significance of this research is also to counter human rights and ethical dilemmas by presenting a set of ethical and professional guidelines. The main objective of this work is to analyze related risks, identify countermeasures and present recommendations to develop a security awareness culture that will allow cloud providers to utilize effectively the benefits of this advanced techniques without sacrificing system security.

Keywords—Cloud Computing; Human Threats; Multi Layered Security Strategy; Employee Screening; fMRI.

I. INTRODUCTION

Cloud computing is the crest of the wave of new and better computing possibilities – provides, efficient and cost effective way to run governmental and commercial organizations. This technology is unique in its ability to address our national defense [1]. It offers critical infrastructures and intelligence agencies an emerging platform to deliver innovative mission solutions and assist in enhanced collaboration [2]. This platform underpins the next generation of digital products and services with reducing time and decreasing cost and giving these agencies the ability to purchase a broad range of IT services in a utility- based model to increase their operational efficiencies [3]. However, despite the myriad advantages of cloud computing, several security challenges still exist, such as malicious insider [4]. Insiders, by virtue of legitimate access to their organizations' systems and networks, may pose a significant risk to critical infrastructure in cloud. It poses a significant risk and could gain complete control over the cloud

services with little or no risk of detection. This paper will focus on malicious insider as a major threat in cloud security.

Recent reports¹ expose that violent extremists are trying to obtain insider positions in critical infrastructures. Based on these reports, it is clear malicious insiders pose a significant threat to cloud computing and critical agencies and government organizations are careful in adopting this new IT approach as moving to cloud could be challenging with due consideration to the sensitivity of data [1]. Cloud providers have an extreme interest in detecting malicious insiders to assure the protection of critical infrastructures. Despite much investigation into the motivation and psychology of malicious insiders, the fact remains that it is extremely complicated to predict this threat [5]. This presents cloud vendors with a dilemma to establish an appropriate level of trust w.r.t. employees. The objective of this work is twofold and the purpose is to alleviate above threats by focusing on two proactive approaches to overcome the aforementioned issues.

The goal of the first section is to develop a framework that focuses on a multi-layered security strategy that can be used to better combat the risks of the insider threat to an organization's networks, or information systems. This paper recommends that this strategy might be a necessary tool for cloud providers, policy makers, security officers and other stakeholders to identification and assessment of risk that insiders presents to organizations. Innovations in techniques and framework will enable cloud providers to successfully execute the mission of cloud computing. The next session is aiming for extreme measures to evaluate human threats related assessment methods for employee screening and evaluations using cognitive analysis technology, in particular functional Magnetic Resonance Imaging (fMRI). The aim is to establish an appropriate level of trust at employees, effective monitoring and ensuring that insiders do not pose a foreseeable risk to critical infrastructure. The technological and non-technological outputs will ensure a better involvement of SME's as well as enhancing the competitiveness that enables users, ranging from SME's to government agencies, to experience a secure and trustworthy cloud computing experience.

II. MALICIOUS INSIDERS IN CRITICAL INFRASTRUCTURE - A THREAT TO CLOUD SECURITY

Critical infrastructures are the advanced physical and cyber-based systems essential to the state's security, economic

¹ <http://info.publicintelligence.net/DHS-InsiderThreat.pdf>

prosperity and social well-being of the nation, such as law enforcement services, power plants and Information and communication services etc. [6]. As a result of advances in technology, these critical infrastructures have become increasingly automated and interlinked. On the other side, these advances have created new vulnerabilities to physical and cyber-attacks by insiders. A report² issued by the US Department of Homeland Security revealed that violent extremists are trying to obtain insider positions in critical infrastructure in the USA and the world and may increase the impact of any attack on the government assets. It further reveals that the fall edition of AQAP (a magazine published by Al-Qaeda) encourages followers to use “specialized expertise and those who work in sensitive locations that would offer them unique opportunities” to conduct terrorist attacks in the world.

A recent study³ “2011 Cost of Data Breach Study: United States” reveals that insiders are the top cause of data breaches and 25 percent more costly than other types. Current events also expose that concerning cloud computing in the context of critical infrastructure, the threat of malicious insiders is very real. The cloud computing architecture is primarily a multi-tenant, service based architecture that necessitates the creation of certain staff positions that can be a high risk in terms of internal security threats [2]. Moving data and application in cloud environment bring with it an inherent level of risk that allows insiders to steal the confidential data (such as passwords, cryptographic keys, clear text passwords from a VM’s memory, private keys from a VM’s memory etc.) of cloud users, sabotage of information resources and various types of frauds [4]. A cloud administrator’s access to the management VM makes these attacks possible [4]. Recently, a denial of service attack launched by a malicious insider against a well know Infrastructure as a Service (IaaS) cloud by creating twenty accounts and launching virtual machine instances for each. Using those accounts, malicious insider created twenty additional accounts machine instances in an iterative fashion and thus consumed resources beyond set limits [7].

There have been several recent events that exhibit how human operators undermined cloud computing. In December 2010, Google has revealed that the accounts of Chinese human rights activists in Google’s clouds were targeted in a hacking attempt. As part of this investigation it was discovered that these accounts were not compromised through a security breach at Google, but most likely by a malicious insider. In another case⁴, David Barksdale, a 27-year-old former Google engineer, repeatedly took advantage of his position as a member of an elite technical group at the company to access users’ accounts, violating the privacy of at least four minors during his employment. Security breaches by human elements in clouds are on the increase globally and there have been similar attacks at, e.g., RSA, Lush, Play.com and Epsilon

around the world⁵. The sequence of scandals induced 2010 by the publication of classified government documents to the Wiki-Leaks website – in which volumes of sensitive documents were leaked by a trusted insider and ultimately published on an open website – has caused much embarrassment to the United States and other nations and represents the ultimate nightmare scenario for cloud vendors when considering the human aspect in cloud security [8]. It is indeed sobering to imagine that any organization could fall victim to such events and the damage malicious insider can do [8].

The exact security concerns rose from the LAPD (Los Angeles Police Department) when they decided to move to Google apps in November 2009 [7]. The security requirements included encryption of all data and background checks on employees who access the police database (e.g. criminal history records and fingerprints) could be vulnerable by malicious employees [7]. This data base is accessible to all police officials around the country. After more than two years of efforts, the city of LA has abandoned plans in January 2012 and voted to scale back the contract to migrate its 13,000 law-enforcement employees to office application platform and Google’s hosted email, saying that security needs of crucial departments were not able to be met and concluded that cloud provider could not be brought into compliance with certain security requirements of the FBI’s Criminal Justice Information Systems⁶. Instead, Google will pay up to \$350,000 per year for the LAPD for the entire term of the contract due to the delay.

Cloud providers need to face the reality that these threats are targeting critical infrastructure that could debilitating impact on state’s defense and a loss of public confidence in state’s services [3]. This could be a sign of growing pains as leading cloud providers are trying to get larger federal and government entities to go to the cloud and it also brings up questions about whether cloud service is really secure enough.

III. MALICIOUS INSIDER IN CRITICAL INFRASTRUCTURE: WHY CANNOT WE STOP IT?

The “insider” is an individual authorized to access an organization’s information system, network or data - based on trust [9]. The insider threat refers to harmful acts and malicious activities that trusted insiders might carry out such as negligent use of classified data, unauthorized access to sensitive information, fraud, illicit communications with unauthorized recipients, and any other behavior that cause harm to the organization [10]. Insiders can be system administrator, contractors, former employees, suppliers, security guards and partner employees etc. According to Noonan & Archuleta [11] malicious insiders can be labeled as three different types of actors: 1) criminals 2) ideological or religious radicals; and 3) psychologically-impaired disgruntled or alienated employees. The motivations of malicious insiders can be simple illicit financial gain, revenge for a perceived wrong, or a

² <http://info.publicintelligence.net/DHS-InsiderThreat.pdf>

³ http://www.symantec.com/about/news/resources/press_kits/detail.jsp?pkid=ponemon&om_ext_cid=biz_socmed_twitter_facebook_market_wire_linkedin_2011Mar_worldwide_costofdatabreach

⁴ <http://gawker.com/5637234/>

⁵ <http://blogs.longhaus.com.au/cloudcio/2011/05/07/cloud-computing-crisis-100-million-victims-of-cloud-privacy-breach/>

⁶ <http://blogs.longhaus.com.au/cloudcio/2011/05/07/cloud-computing-crisis-100-million-victims-of-cloud-privacy-breach/>

radicalization for the advancement of religious or ideological objectives [11].

To counter human threats, agencies have invested billions of Euros in different technical measures for years now [12]. The current security paradigms include access control and encryption to face malicious insiders and outsiders. They are implemented through passwords, physical token authentication and biometric authentication, firewalls, encrypted data transmission, data leakage prevention, behavioral-pattern threat detection, voice stress analysis and polygraphs. Various studies demonstrate that above devices and security software are normally designed to defend against external threats to secure critical infrastructure and do not protect against attacks aided by internal help in organizations [12]. An insider not only has the ability to obtain or access valuable data that resides on the internal network, but he/she can obtain this data from their workstation without causing suspicion or breaking trust. Cloud organizations are well aware of these issues, as demonstrated in a recent roundtable including senior staff as well as the sector's major companies [13]. It is of utmost importance to adopt extreme measures to secure critical infrastructure in clouds to lower the level of threats. Secondly, well thought-out policies are necessary with understanding of the potential role of a process and methods – each designed to address threats targeted at specific segments of the cloud environment. Our research aims at alleviating this critical factor with higher assurance of achieving their desired security as employees at critical positions are may be, in some instances, the first and only line of defense and vital to national security.

IV. MITIGATING INSIDER THREATS - TOWARD BEST PRACTICE ACTIONABLE FRAMEWORKS

This part focuses on malicious insiders with a variety of specialized security solutions in supporting all aspects of user trust in clouds to mitigate human threats and for making better informed decision when choosing a cloud solution. Our actionable framework aims to mitigate the potential vulnerabilities of critical infrastructure in cloud computing and key resources to ensure its protection and resilience. To achieve this goal, this paper recommends policies and actions to detect and manage insider threats and effectively fight these problems.

A. A Human Factor Vulnerability Assessment (HVA) Framework.

To mitigate malicious insider risks, this research recommends developing a human factor vulnerability assessment (HVA) framework. The purpose is to assess cloud organizational structure, its infrastructure, technical and non-technical vulnerabilities, employee roles and responsibilities, actual events of human threats incidents and existing security measures. This framework will raise awareness of the malicious insider causes, potential indicators and prevention and detection strategies, and informing cloud providers and users of their security responsibilities. This research recommends to build a data base in HVA that will consist of various incident for malicious related activates, vulnerabilities, lessons learned, and physiological profiles or statistics regarding the insider and insider misuse, or social engineering activities in cloud computing. This assessment may provide

indicators, practices, actions, policies and procedures so cloud organization can track changes in their capabilities related to human threats over time and possible mitigation strategies. The assessment's results will benefit all individuals involved in this process and enable cloud providers to gain a better understanding of vulnerabilities in their clouds. The framework must include low-cost, easily implemented policy solutions for cloud vendors that have long term effects. These security awareness and training programs will be paramount to ensure that cloud providers and users will understand their security responsibilities to mitigate social engineering threats, and properly use and protect the cloud resources entrusted to them. This security plan ought not to be static; it has to be refined and adjusted regularly as the security challenges of cloud data centers permanently will change.

B. Counter-Insider Threat Database

This database may include cloud users, security personnel, technical/non-technical staff, administrators, and all levels of organization management into a single, actionable assessment framework. Based on this database, HVA will analyze the newest approaches to topics related to human factors involved in cloud security by means of reviewing relevant publications and security surveys. It may also take into consideration the incidents knowledge and the opinion of security experts, psychologists and stakeholders, in order to find the real nature and magnitude of the malicious insider and social engineering problems to identify the best practices including individual and organizational actions and responses. It is important to compile a database in this phase with criminal cases in which current or former employees, contractors, or business partners abused the trust and access associated with their positions. This database would work on cases of insider activity from the real world documented with many insider threat cases that may provide a rich source for empirical research on real cases of insider threat. In this regard, interviews with various victim organizations as well as with perpetrator are necessary, complementing a wealth of case data with first-hand insights into the methods and motivations behind these crimes.

Secondly, the collaboration with noted psychologists and others from the law enforcement agencies to uncover key technical, social, and organizational patterns of insider behavior may severely hamper understanding of the magnitude of the problem and development of solution strategies. Insider misuse, abuse and malicious activates are yet another manifestation of betrayal of trust behavior for which cloud organization must be alert in espionage cases. The psychological profile will provide managers, security specialists and medical personnel a profile of the insider, which may become a useful tool to enable them to identify potential abusers before they cause serious damage This knowledge will investigate the objectives, data-driven examination of the motivations and behavior patterns of malicious/outsider human threats, as well as organizational issues that may influence them. Leveraging data to prevent, detect, and respond to these threats can help cloud organizations to strengthen the protection of the critical infrastructure in cloud. Potential benefits from developing an insider and outsider event database will focus on the need for improved detection, technical research priorities, and prevention through policies, education

and training. This valuation may provide reports (confidential and public) on the basis of this assessment with the observable indicators and information and guidelines that cloud providers need to develop, and a plan of action and security standards to increase their ability to prevent, detect and respond to human threats. A unified database may be built that may consist of:

- a) Incidents
- b) Statistical findings and implications regarding technical details of the incidents
- c) Insider Planning
- d) Nature of harm
- e) Communication behavior and characteristics
- f) Vulnerabilities
- g) Lessons learned
- h) Physiological profiles
- i) Statistics regarding the insider and insider misuse or abuse
- j) Case studies regarding detection and identification of the insiders
- k) Social engineering activities to develop recommendations for technology and policy solutions for future problems in cloud computing environment.

At minimum, the following activities should be engaged for a Human Factor Vulnerability Assessment (HVA) framework:

- 1) Identify categories of problems and analyze differences/similarities of cases.
- 2) Provide administrators, security specialists a physiological profile of the insider/social engineer.
- 3) To define significant characteristic types of insider misuse and assist in the development of questions for security investigations.
- 4) Develop standardized framework and material for security education, awareness and training.
- 5) Provide data for finer-grain access policies and differential access controls needed to help define what constitutes proper usage, thus facilitating the role of insider-misuse detection.
- 6) Develop recommendations for technology and policy solutions for future problems.

The next sections will define how this framework will deliver organizations the means to support business continuity by better securing their assets through well founded decision making and therefore lower the overall risks associated with cloud computing.

V. MULTI-LAYERED SECURITY STRATEGY

This task may carry out a multi layered security strategy based on the Human Factor Vulnerability Assessment (HVA).

After this initial step, evaluation of the security mechanisms must be developed to monitor compliance and effectiveness of this task to revise the specification and architecture according to the experimental results obtained with the database. Our security program identifies the four critical steps:

1. Human Factor Security Awareness and Training Program Development

2. Security Awareness and Training Material Development
3. Security Program Implementation Consultancy
4. Post-Implementation Security Program Consultancy and Periodical Evaluation

A. Human Factor Security Awareness and Training Program Development:

This step must identify:

- a) What are the plans for developing and implementing security awareness and training opportunities to mitigate the human factors in cloud that are compliant with existing directives?
- b) What awareness, training and education are needed to mitigate the human factors in clouds (i.e., what is required)?
- c) How are these needs being addressed by cloud providers?
- d) Where are the gaps between the needs?
- e) What is being done and what more needs to be done for gap analysis and targeting deficient areas for early rollout?
- f) Which needs are most critical in cloud environment?
- g) What are the roles and responsibilities of a cloud organization's personnel will be identified in design and implementation stage to maintain security standards?
- h) Who should ensure that the appropriate employee attend or view the applicable material in cloud agency?
- i) Documentation, feedback, and evidence of learning for each aspect of the program;
- j) Gap analysis and targeting deficient areas for early rollout.

B. Security Awareness and Training Material Development:

In this task we recommend to develop a standardized framework and material to identify:

- a) What skills we do want the cloud provider's management and their employees to learn and apply to mitigate human threats?
- b) What behavior to reinforce?
- c) The topics may be covered in this approach are:

- 1) Social engineering tricks
- 2) Malicious insiders
- 3) Shoulder surfing
- 4) Dumpster diving
- 5) Tailgating
- 6) Access control issues
- 7) Visitor control
- 8) Physical access to spaces as well as handling incident response.

C. Security Program Implementation:

This approach will focus on the implementation of a multi layered strategy with:

- 1) Trainings
- 2) Programs

3) *Education and awareness.*

4) *Trainings:*

This scope of this section is to ensure that cloud providers and users are appropriately trained in how to face human factors in cloud environment. It must ensure that insiders with privileged access in cloud such as administrator are only required to undergo the same vetting procedures as other insiders. It is evident that the sensitivity of functions they perform and the potential to access the most sensitive information contained in the system are much greater than those without such privileges. It is particularly important to focus on developing a strong security partnership with system managers, ensuring that these individuals receive the best security awareness training available. Furthermore, it is important to ensure that:

a) *Awareness and training material is effectively deployed to reach the intended audience.*

b) *Training material is reviewed periodically and updated when necessary and assist in establishing a tracking and reporting strategy.*

5) *Programs:*

Security programs should recommend best methods, guidelines and technologies dealing with human factor issues. Some of our guidelines are:

a) *To recommend technologies and guidelines currently available for dealing with the insider/outsider problem.*

b) *To establish an activity to evaluate on a continuing basis the effectiveness of available security methods and tools of all types that may be available to mitigate that risk.*

c) *To direct the appropriate cloud providers to accelerate the development of new tools for cloud computing to combat insider and outsider threats.*

d) *To develop robust policies to discarded floppy disks, and printouts etc.*

e) *To propose different policies that classifies the sensitive information and enforces the mandatory and discretionary access control mechanisms.*

6) *Education and Awareness*

This research identifies that many cloud operators lack an appropriate awareness of the threat insiders and social engineers pose to their operations. A strong security partnership is necessary to be developed. Education and awareness may present the biggest potential return for policy by motivating cloud operators and focusing their efforts to address the human threat. Our research attempts to enforce knee jerk policies to require all employees review the company policy after every three months, to acquaint themselves with revisions if any. Different activities are needed to be established to evaluate on a continuing basis the effectiveness of available security methods and tools of all types to mitigate that risk. Methods and techniques must be developed regarding discarded storage medium that may contain sensitive information. Appropriate awareness will help to shape the human threat policies and programs needed to address the unique insider risk profile.

Techniques for Delivering Awareness Material:

Techniques may include, but not limited to:

a) *Teleconferencing sessions*

b) *Interactive video training*

c) *Web based training*

d) *Videos, posters (“do and don’t lists” or checklists)*

e) *Screensavers*

f) *Warning banners*

g) *SMS/messages*

h) *In-person*

i) *Instructor-led sessions*

j) *Awards program*

k) *Letters of appreciation etc.*

STAKEHOLDER WORKSHOP: It is recommended to arrange workshops for stakeholders involved in this whole process for intermediate results and panel discussion for feedback.

PUBLIC WORKSHOP: Arrange workshops to recommend security personnel and cloud users a layered defense through use of guidelines and tools and how to develop an effective, comprehensive insider/outsider threat monitoring strategy.

VI. **WHO WILL WATCH THE WATCH MAN?**

In addition to above policies and framework, this paper suggests that it’s a mistake to rely just on preventive measures. Instead, it is essential to supplement them with monitoring and auditing so insider attacks can be detected and truly stopped by removing the attacker from the cloud organization. Although it’s difficult to prevent a malicious attack from a motivated insider, there are ways to spot bad behavior before it becomes a big problem. IT administrators make the cloud service and related datacenters operate, and through homogeneity and greater automation they often manage ratios of thousands of servers. The risk from these privileged cloud provider administrators must be explicitly recognized and addressed. There is a need to ensure that important location and data should not be reached and accessed by an individual and functions are not held by the same individual. Checks and balances implemented through review and approval processes are consistently applied so that gaps, even in critical situations are appropriately controlled and reviewed. Increased attention must be paid to privileged accounts, as it is mentioned earlier that how privileged administrators misused user’s data and applications by accident, for profit, or for retribution. Each employee has logical patterns of information usage, and the organization should look for abnormal usage and investigate when this occurs. For example, if an employee looks at 50 customer accounts each day and then one day looks at 100 or more, there is a potential issue that should be investigated. You always need to understand if unusual behavior is warranted or malicious. Our strategy suggests that special monitoring and double auditing should be done to monitor abnormal or suspicious behavior by administrator or the damage that can be caused.

However, we advocate that in the context of critical infrastructure, employment screening measures are necessary to implement by focusing on technical and psychological measures used in psychological research as well as in law enforcement domain to secure cloud data. This research propose that technical measure like, fMRI scanning must be

used for critical staff to detect malicious intent activities and susceptible behavior and other weaknesses (for instance, alcoholic, religious affiliation or domestic problems). This policy is needed to be implemented on pre-employment screening as well as regular employees. We propose that pre scanning measures such as reference checks can uncover prior criminal records, issues with character or credit problems.

VII. MAINTAINING SECURITY USING FUNCTIONAL MAGNETIC RESONANCE IMAGING (fMRI)

The functional MRI is widely known and accepted in the scientific community as it does have a significant amount of scientific research behind its claims and validity. This technique relies on the fact that cerebral blood flow and neuronal activation are coupled. It involves placing the subject in a donut-shaped magnetic technology, which can identify subtle changes in electromagnetic fields [14]. During scanning, when an area of the brain is in use, blood flow to that region also increases [15]. Thus, its responses associated with neuronal activation correlate with cognitive tasks and various behavioral functions [15].

An fMRI has already had a major impact on neuroscience and in clinical settings. It has been applied ranging from language comprehension to treatment of neurological impairment disease, psychiatric illness, aesthetic judgment and justification of cognitive enhancing drugs in educational settings [16]. With these rapid developments many researchers claimed this technology to be useful outside the laboratory settings. For instance, economic contexts, investing personality traits, mental illness, religious extremism, racial prejudice, suicidal thoughts aggressive or violent tendencies and truth verification [16] [17]. Proponents of this neuro-imaging technology hailed fMRI as the next “truth meter” and conclude that because of the novelty of the physiological parameters being measured, this technology may be more accurate than other traditional methods for employee screening (e.g., polygraph, see [18] [19] [20] [21]). According to researchers, this ground breaking research proves that fMRI has the capacity to address the question of guilt versus innocence. Since the first publication by [21] on deception detection by fMRI, various papers and studies (See [22] [23] [24] [25] [26] [27] [28] [29] [30] [31] [32] [33] [33]) have reported different experiments in which subjects were asked to respond deceptively in some blocks and truthfully in others. It proved that lying involve more efforts than truth and expose that specific brain areas respond strongly in generating deceptive responses. As with lying, several brain regions show significant increases and light up on during scanning when a person sees a familiar object or image or during deception compared to truth telling [23]. For instance, dorsolateral prefrontal cortex (DLPFC), anterior cingulate cortex (ACC), ventrolateral (VLPFC) and left and right cerebral hemispheres increases activity when people tell lies [31].

fMRI truth verification in individual humans has been studied in 31 original peer reviewed scientific journal articles. These studies were done by 22 different research groups that included researchers from 13 different countries (USA, UK, Canada, Australia, China, Japan, Netherlands, Switzerland, Poland, Denmark, Sweden, Germany, and Russia). Similarly,

during the employee screening phase, if any suspect employee is asked a question, the information to which is unknown then the specific regions of the brain is unusually active and it is presumed that suspect is lying; if, however, the same areas are no more active it may presumed that subject is telling the truth [18] [19] [20]. Thus, this technology has potential to reveal recognition regardless of whether the suspect speaks or attempts to conceal the recognition.

Ruben Gur, a neuropsychologist at the University of Pennsylvania, states that fMRI scans can reveal cognitive tasks when a subject recognizes a familiar situation, object, or person connected with any fraud, theft of intellectual property, and IT sabotage, no matter how hard he or she tries to conceal it [17]. This cognitive analysis technology could function as a hyper-accurate lie detector that is nearly impossible to deceive [35] [36]. For instance, an investigator could present a suspect with specific information such as of passwords, social security numbers, credit card information, other personal information, or other confidential corporate information [35]. This scientific technique allows intelligence operatives to focus their investigations on the suspects who actually commit crime and to determine if he or she has any un-authorized access to an organization's network or system before (See [22] [25] [26] [29] [33] [35]). On the other side, an information absent will provide support for the claims of innocence that employee is not guilty of committing any crime and has no knowledge specific to any data or any information [37]. This development will lead to speculations about the development of this neuro-imaging technology that could directly examine the malicious insider memories, intentions and its mind.

Interrogators will be able to confidently say that the fMRI told us this suspected employee lied about X or that he recognizes Y or fMRI picked him out as a malicious insider. This confidence that organizations will have in this neuroscience technique will be based on an aura of infallibility, scientific validity and objectivity [17]. Thus, unlike polygraph—which detects a person's emotional response to deception—fMRI measures person's decision to lie, as subjects cannot control their cerebral activity to avoid detection [38].

Thus fMRI can be used as a tool warranted in employee screening as not only has this neuro-imaging technology taken the attention of scientific communities and law enforcement agencies but it has also attracted interest of corporate world [38]. Two private firms: No Lie MRI and Cephos Corp trying to make the dream of perfect truth verification into a reality and have begun marketing since 2006. They offer high-tech lie detection services based on research comparing neuronal activation patterns [14].

The future of cloud provider may very well be under construction with this new approach that is becoming a reality for mining of knowledge from suspected employee to assess potential threats rapidly. However, there are still significant concerns must be addressed prior to moving this technology to real-world application [16]. In addition to the scientific challenges, advances in fMRI identify numerous, human rights, social, legal and general public concerns to the process of and the science behind it [39].

VIII. EVIDENCE-BASED POLICIES AND GUIDELINES: A RELIABLE RESPONSE TO EMPLOYEES CONCERNS

In this research we are proposing best practices, recommendations and guiding principles to employ fMRI brain scanning technology during the questioning of suspects in cloud environment.

1. We recommend that, members involved in screening process must be made aware of the issues raised by this technology to develop best practices and efficient internal measures. These actions should address the development of policies and procedures relating to incidental findings within the doctrine of informed consent. Informed consent should be sought before scanning as the employees should be aware of the potential dangers. He/she should read, understand and sign an informed consent disclaimer to ensure that all the necessary requirements are met. This authority will give employee confidence and more control over the construction of their identities. Human experimentation without the consent of the subject is also a violation of human rights law [40]. To assure the protection, the fMRI scan process should undergo a complete government approval process to make reasonable assurance of employee's safety.

2. Training of investigators is one of the major challenges for the implementation of this tool. This training is necessary for the evaluation of screening centers to appropriately protect employees while allowing for scanning. Thus, only trained experts will be required to evaluate subjects and conduct the scan.

3. Experts are ethically obligated to report to the appropriate authorities when they have reason to believe that employee screening is coercive and violating human rights. They must ensure that if experts do not detect any abnormal behavior, the subject is not harmed. However, if an abnormality is detected, the results of the scan should be analyzed by other highly trained neuroscientists and possibly rectified [43].

4. It is also important that professionals involved in screening will be required to acquire security clearances. This shield will make it impossible for them to share the findings with colleagues in unclassified settings [43].

5. This research recommends that "Certificate of Confidentiality and Privacy" that can make a difference in the screening context. This certificate will allow the members who have access records to refuse to disclose identifying information at the civil, criminal, legislative, federal, state, or local level if the employee is not guilty. Disclosure of sensitive information could have adverse consequences on innocent person's reputation, employability as well as financial standing [41]. This document will particularly encourage employee to participate in the scanning process.

6. Any pre-employment screening process must be compatible with all relevant legislations, for instance,

employee's right must be protected by Article 8⁷ of the European Convention on the Protection of Human Rights and Article 12⁸ of the Universal Declaration of Human Rights. The process must implement the United Nations International Labor Organization (ILO) code of practice on the Protection of Workers' Personal Data (1996)⁹ as well as European Union Guidelines 95/46 and 97/66 on data protection.

7. It is important that organizations must ensure the safety of the subjects through the systematic monitoring of the international law and human rights – including the United Nations Conventions against Torture, the International Covenant on Civil and Political Rights, and the Universal Declaration of Human Rights. The agencies must also consider the nuances of the Geneva Conventions as applied to suspected terrorists [42].

8. Question should be limited to a verification of the "real" or "personal" identity such as education, employment history, court records, credentials and other data associated with an employee. In screening process, the access to the results should be restricted for investigators in order to prevent the misuse of these preliminary data.

9. Uniformed personnel's and medical experts who are engaged in screening panel using fMRI must be held to account for their actions if they have violated human rights laws. Innocent employees or victim of this technology must be offered compensations, health care services and a formal apology to address ethical violations caused by this technology or by the professionals. A comprehensive federal investigation is required if the staff's trust in the ethical integrity of the security and medical profession being seriously compromised. If interrogators dismiss a subject for failing an fMRI scan test, they must be able to justify the action against him/her under the influence of a Human Rights Act, such as the European Convention on Human Rights (ECHR) or the UK Human Rights Act 1998. Furthermore, a policy can be introduced of only screening in case of suspicious activities. In this regard, drug testing examples of the Österreichischer Gewerkschaftsbund (ÖGB) in Austria, Deutscher Gewerkschaftsbund (DGB) in Germany, and the Confédération Générale de Travail (CGT) in France are suitable case studies for this approach [42].

10. Innovation in technology has been a key driver of change - the defense and security arenas are no exception. Similarly, members of elite unit should be well aware of current knowledge, novel literature, latest technologies, valuable processes and services about fMRI scanning for the purpose of developing image analysis to improve investigating methods. It is a major step forward in the action to our national

⁷ <http://www.echr.coe.int/NR/rdonlyres/D5CC24A7-DC13-4318-B457-5C9014916D7A/0/ENGCONV.pdf>

⁸ <http://www.un.org/en/documents/udhr/index.shtml#a12>

⁹ http://www.ilo.org/wcmsp5/groups/public/---ed_protect/---protrav/--safework/documents/normativeinstrument/wcms_107797.pdf

interests that will continue to play a key role in the effectiveness of fMRI as a counterterrorism tool. We also recommend that government must push promising research on fMRI as they could meet our defense needs through collaboration with research sectors and universities to ensure a strong research base in this area. This action must be vibrant, inventive and innovative that looks most promising in screening neuroimaging. Investigators and neuroscientists must grasp the opportunities and adapt them quickly and effectively as this benefit is critical to our security and sovereignty [42].

IX. CONCLUSION

September 11th has marked an important turning point that exposed this new type of human threat that may pose a significant risk to critical infrastructure in cloud computing. This paper has explored different countermeasures that can be taken by an organization to protect themselves against this threat. The main objective for this work is to focus on advanced security strategies, frameworks, models, multi-layered security strategies and assessment methods linked to the overall architecture of this paper. Employee screening is central to this approach. The Insider Threat Study has also revealed a surprisingly high number of malicious insiders with prior criminal convictions when hired. Having access to complete source of employee history information is the only way the interest in performing due diligence to protect key assets and the nation can be served. At one hand it is beneficial for a general improvement which ultimately leads to higher productivity, better workers, increased efficiency and will provide an acceptable level of assurance for employees who have access to protectively marked critical assets and could alleviate the burden of mistrust. Furthermore, the aim of introducing this scanning technique is also deterrence from malicious activity of any kind. Indeed this approach may deter some high risk candidates with criminal/terrorist backgrounds from applying for the job at all, which moreover may even save money and time in the recruitment process. Similarly, there is a great expectation among scientists and counterterrorism agencies that workers at critical positions will also realize the urgency of the threat and the significance of neuroimaging application to national defense.

More significant, consideration must also be given to the government's purpose in subjecting the suspect to fMRI scan. It is important that the state's interest in interrogating a potential malicious insider in a high risk position must be justifiable, appropriate and will not curtail substantial civic liberty. Whether or not policy makers or civilized society can or should allow brain scanning is a matter that will continue to be debated for years to come. However given only the choice of deciding of whether or not scan an individual using fMRI technology when it may be possible to prevent mass casualties through this scan, the state ultimately has to make sensible decisions as necessary in order to save lives.

REFERENCES

[1] Fazio, M., Paone, M., Puliafito, A., & Villari, M. (2012). HSCLOUD: Cloud Architecture For Supporting Homeland Security.
[2] <http://www.s2is.org/Issues/v5/n1/papers/paper14.pdf>
[3] Matthew, G., Allison, S., Scott, R., Jonathan, C., & Jodi C. (2012). Creating Effective Cloud Computing Contracts for the Federal

Government. A joint publication of CIO and CAOC. Federal Cloud Compliance Committee.
[4] <http://www.cio.gov/cloudbestpractices.pdf>
[5] Jackson, K. L. (2011). Implementation of Cloud Computing Solutions in Federal Agencies
[6] <http://www.njvc.com/sites/default/files/documents/NJVC%20Cloud%20computing%20for%20Government%20FINAL.pdf>
[7] Francisco, R., Salvador, A., & Miguel, C. (2011). The Final Frontier: Confidentiality and Privacy in the Cloud, IEEE Computer, vol. 44, n. 9, pp. 44-50.
[8] <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5959138>
[9] Sahito, F.; Slany, W.; Shahzad, S.K., (2011). "Search engines: The invader to our privacy — A survey," Computer Sciences and Convergence Information Technology (ICCIT), 2011 6th International Conference, vol., no., pp.640, 646.
[10] <http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=06316696>
[11] Sahito, F.; Latif, A.; Slany, W., (2011). "Weaving Twitter stream into Linked Data a proof of concept framework," Emerging Technologies (ICET), 2011 7th International Conference, vol., no., pp.1,6, 5-6
[12] http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6048497
[13] Jansen, W., & Grance, T. (2011). The NIST Guidelines on Security and Privacy in Public Cloud Computing.
[14] <http://csrc.nist.gov/publications/nistpubs/800-144/SP800-144.pdf>
[15] Maxim, M. (2012). Defending against insider threats to reduce your IT risk. white paper, security and compliance.
[16] <http://www.ca.com/~media/Files/whitepapers/insider-threat-wp-jan-2011.pdf>
[17] Greitzer, F. L., Moore, A. P., Cappelli, D. M., Andrews, D. H., Carroll, L. A., & Hull, T. D. (2008). Combating the insider cyber threat. IEEE Security and Privacy, 6(1), 61-64.
[18] <http://doi.ieeecomputersociety.org/10.1109/MSP.2008.8>
[19] Yuqing, Sun., Ninghui, Li., and Elisa, Bertino. (2011). Proactive defense of insider threats through authorization management. In Proceedings of 2011 international workshop on Ubiquitous affective awareness and intelligent interaction (UAAII '11). ACM, New York, NY, USA, 9-16.
[20] <http://dl.acm.org/citation.cfm?id=2030095>.
[21] Noonan, T. and Archuleta, E. (2008). 'The National Infrastructure Advisory Council's Final Report and Recommendations on the Insider Threat to Critical Infrastructures.
[22] http://www.dhs.gov/xlibrary/assets/niac/niac_insider_threat_to_critical_infrastructures_study.pdf
[23] Sarka, K. R. (2010). Assessing insider threats to information security using technical, behavioural and organisational measures, information security technical report, pp. 1-22.
[24] <http://www.sciencedirect.com/science/article/pii/S1363412710000488>
[25] Grosse, E., Howie, J., Ransome, J., & Schmidt, S. (2010). "Cloud Computing Roundtable," IEEE Security & Privacy, Nov./Dec. 2010, pp. 17-23.
[26] <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5655239>
[27] Simpson, JR. (2008). Functional MRI lie detection: too good to be true? J Am Acad Psychiatry Law 36:491– 8.
[28] <http://www.jaapl.org/content/36/4/491.abstract>
[29] Fenton, A., Meynell, L., and Baylis, F. 2009. Ethical challenges and interpretive difficulties with non-clinical applications of pediatric fMRI. American Journal of Bioethics (AJOB Neuroscience) 9(1): 3–13.
[30] <http://www.ncbi.nlm.nih.gov/pubmed/19132609>
[31] Garnett A, Whiteley L, Piwowar H, Rasmussen E, Illes J (2011). Neuroethics and fMRI: Mapping a Fledgling Relationship. PLoS ONE 6(4): e18537. doi:10.1371/journal.pone.0018537.
[32] <http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0018537>
[33] Marks, J.H. (2007). Interrogational Neuroimaging in Counterterrorism: A "No-Brainer" or a Human Rights Hazard? American Journal of Law and Medicine, 33, 483-500.
[34] http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1005479

- [35] Bruni, Tommaso. (2012). Cross-Cultural Variation and fMRI Lie-Detection. *Technologies On The Stand: Legal And Ethical Questions In Neuroscience And Robotics*, pp. 129-148, B. Van den Berg, L. Klaming, eds., Nijmegen: Wolf Legal Publishers, 2011. <http://ssrn.com/abstract=1983536>
- [36] Faulkes, Z. (2011). Can brain imaging replace interrogation and torture? *Global Virtue Ethics Review* 6(2): 55-78. <http://www.spaef.com/article.php?id=1266>
- [37] McCabe, D. P., Castel, A. D. and Rhodes, M. G. (2011). The influence of fMRI Lie Detection Evidence on Juror Decision-Making, *Behavioral Sciences and the Law*, 29: 566-577.
- [38] <http://castel.bol.ucla.edu/publications/McCabe%20Castel%20Rhodes%20BSL%20in%20press.pdf>
- [39] Spence, SA., Farrow, TFD., and Herford, AE. (2001). Behavioral and functional anatomical correlates of deception in humans. *Neuroreport* 12:2849–53.
- [40] <http://www.ncbi.nlm.nih.gov/pubmed/11588589>
- [41] Lee, TMC., Liu, H-L., Tan, L-H. (2002). Lie detection by functional magnetic resonance imaging. *Hum Brain Mapp* 15:157– 64.
- [42] <http://www.ncbi.nlm.nih.gov/pubmed/11835606>
- [43] Langleben, DD., Schroeder, L., Maldjian, JA. (2002). Brain activity during simulated deception: an event-related functional magnetic resonance study. *Neuroimage* 15:727–32.
- [44] http://www.med.upenn.edu/langleben/neuroimage15_2002.pdf
- [45] Ganis, G., Kosslyn, SM., Stose, S., (2003). Neural correlates of different types of deception: an fMRI investigation. *Cerebral Cortex* 13: 830–6.
- [46] <http://cercor.oxfordjournals.org/content/13/8/830.abstract>
- [47] Spence, S. A. (2004). The deceptive brain. *Journal of the Royal Society of Medicine*, 97(1), 6–9
- [48] Nunez JM, Casey BJ, Egner, T. (2005). Intentional false responding shares neural substrates with response conflict and cognitive control. *Neuroimage* 25:267–77.
- [49] <http://www.ncbi.nlm.nih.gov/pubmed/15734361>
- [50] Kozel, FA., Revell, LJ., Lorberbaum, JP. (2004a). A pilot study of functional magnetic resonance imaging brain correlates of deception in healthy young men. *J Neuropsychiatry Clin Neurosci* 16:295–305.
- [51] <http://neuro.psychiatryonline.org/article.aspx?articleid=101888>
- [52] Kozel, FA., Padgett, TM., George, MS. (2004b). A replication study of the neural correlates of deception. *Behav Neurosci* 118:852– 6.
- [53] <http://www.personal.psu.edu/krm10/PSY597SP07/Kozel%20neural%20correlates.pdf>
- [54] Lee, TMC., Liu, H-L., Chan, CCH. (2005). Neural correlates of feigned memory impairment. *Neuroimage* 28:305–13.
- [55] <http://www.ncbi.nlm.nih.gov/pubmed/16165373>
- [56] Davatzikos, C., Ruparel, K., Fan, Y. (2005). Classifying spatial patterns of brain activity with machine learning methods: application to lie detection. *Neuroimage* 28:663– 8.
- [57] http://repository.upenn.edu/cgi/viewcontent.cgi?article=1011&context=n euroethics_pubs
- [58] Langleben, DD., Loughhead, JW., Bilker, WB. (2005). Telling truth from lie in individual subjects with fast event-related fMRI. *Hum Brain Mapp* 26:262–72.
- [59] <http://www.med.upenn.edu/langleben/tellingtruth.pdf>
- [60] Phan, KL., Magalhaes, A., Ziemlewicz, TJ. (2005). Neural correlates of telling lies: a functional magnetic resonance imaging study at 4 Tesla. *Acad Radiol* 12:164–72.
- [61] <http://www.ncbi.nlm.nih.gov/pubmed/15721593>
- [62] Mohamed, FB., Faro, SH., Gordon, NJ. (2006). Brain mapping of deception and truth telling about an ecologically valid situation: functional MR imaging and polygraph investigation: initial experience. *Radiology* 238:679–88.
- [63] <http://www.ncbi.nlm.nih.gov/pubmed/16436822>
- [64] Kozel, FA., Johnson, KA., Mu, Q. (2006). Detecting deception using functional magnetic resonance imaging. *Biol Psychiatry* 58:605–13.
- [65] <http://www.musc.edu/pr/fmri.pdf>
- [66] Kozel, F. A., Johnson, K. A., Mu, Q., Grenesko, E. L., Laken, S. J., & George, M. S. (2005). Detecting deception using functional magnetic resonance imaging. *Biological Psychiatry*, 58(8), 605–613.
- [67] <http://www.ncbi.nlm.nih.gov/pubmed/16185668>
- [68] Greely, HT., and Illes, J. (2007). Neuroscience-based lie detection: The urgent need for regulation. *American Journal of Law and Medicine* 33(2 and 3):377–431.
- [69] <http://www.ncbi.nlm.nih.gov/pubmed/17910165>
- [70] Spence, S.A., Kaylor-Hughes, C.J., Brook, M.L., Lankappa, S.T., Wilkinson, I.D. (2008). Munchausen's syndrome by proxy' or a 'miscarriage of justice? An initial application of functional neuroimaging to the question of guilt versus innocence. *European Psychiatry*, 23: 309-314.
- [71] <http://www.ncbi.nlm.nih.gov/pubmed/18029153>
- [72] Tim, Bayne. (2011). *Mindreading: From Neuroimaging to the Philosophy of Mind*, Humanities Research Showcase, University of Oxford.
- [73] Edersheim, J. G., Rebecca Weintraub Brendel, & Bruce H. Price. (2012). Neuroimaging, Diminished Capacity and Mitigation, in *NEUROIMAGING IN FORENSICPSYCHIATRY*, supra note 25 at 163–64.
- [74] <http://clbb.mgh.harvard.edu/wp-content/uploads/Neuroimaging-Diminished-Capacity-and-Litigation.pdf>
- [75] Meier, B. M. (2002). International Protection of Persons Undergoing Medical Experimentation: Protecting the Right of Informed Consent, 20 *BERKELEY J INT'L L* 513, 550. <http://scholarship.law.berkeley.edu/cgi/viewcontent.cgi?article=1224&context=bjil>
- [76] Palys, T. S., and J. Lowman. 2000. Ethical and Legal Strategies for Protecting Confidential Research Information. *Canadian Journal of Law and Society*, 15(1), 39-80.
- [77] <http://www.sfu.ca/~palys/Strategies-CJLS.pdf>
- [78] [42] Sahito, Farhan.; Slany, Wolfgang. (2012), "Functional Magnetic Resonance Imaging and the Challenge of Balancing Human Security with State Security", *Human Security Perspectives 1* (European Training and Research Centre for Human Rights and Democracy (ETC), Graz, Austria) 2012 (1): 38–66 <http://www.isn.ethz.ch/isn/Digital-Library/Publications/Detail/?lng=en&id=152162>
- [79] [43] Sahito, Farhan. (2013.) "Interrogational Neuroimaging: The Missing Element in Counter-Terrorism" *International Journal of Innovation and Applied Studies*, Vol. 03, No. 02, 2013.

Fine Particulate Matter Concentration Level Prediction by using Tree-based Ensemble Classification Algorithms

Yin Zhao

School of Mathematical Sciences
Universiti Sains Malaysia (USM)
Penang, Malaysia

Yahya Abu Hasan

School of Mathematical Sciences
Universiti Sains Malaysia (USM)
Penang, Malaysia

Abstract—Pollutant forecasting is an important problem in the environmental sciences. Data mining is an approach to discover knowledge from large data. This paper tries to use data mining methods to forecast $PM_{2.5}$ concentration level, which is an important air pollutant. There are several tree-based classification algorithms available in data mining, such as CART, C4.5, Random Forest (RF) and C5.0. RF and C5.0 are popular ensemble methods, which are, RF builds on CART with Bagging and C5.0 builds on C4.5 with Boosting, respectively. This paper builds $PM_{2.5}$ concentration level predictive models based on RF and C5.0 by using R packages. The data set includes 2000-2011 period data in a new town of Hong Kong. The $PM_{2.5}$ concentration is divided into 2 levels, the critical points is $25\mu g/m^3$ (24 hours mean). According to 100 times 10-fold cross validation, the best testing accuracy is from RF model, which is around 0.845-0.854.

Keywords—Random Forest; C5.0; $PM_{2.5}$ prediction; data mining.

I. INTRODUCTION

Air pollution is a major problem for some time. Various organic and inorganic pollutants from all aspects of human activities are added daily to the air. One of the most important pollutants is particulate matter. Particulate matter (PM) can be defined as a mixture of fine particles and droplets in the air and this can be characterized by their sizes. $PM_{2.5}$ refers to particulate matter whose size is 2.5 micrometers or smaller. Due to its effect on health, it is crucial to prevent the pollution getting worse in a long run. According to WHO's report, the mortality in cities with high levels of pollution exceeds that observed in relatively cleaner cities by 15–20% [1]. Forecasting of air quality is much needed in a short term so that necessary preventive action can be taken during episodes of air pollution. WHO's Air Quality Guideline (AQG) [2] says the mean of $PM_{2.5}$ concentration in 24-hour level should be less than $25\mu g/m^3$, although Hong Kong's proposed Air Quality Objectives (AQOs) [3] is $75\mu g/m^3$ right now. Because the target data is from a new town in Hong Kong, which means there are lots of people living in this area, so it is need to be a stricter standard of air pollution in such area. As a result, we use $25\mu g/m^3$ based on 24 hours mean as our standard points. The number of particulate at a particular time is dependent on many environmental factors, especially the meteorological data and time serious factors.

Predictive models for $PM_{2.5}$ can vary from the simple to the complex; hence we have CART, C4.5, Artificial Neural Networks, Support Vector Machine among others. In this paper, we try to build models for predicting next day's $PM_{2.5}$ concentration level by using two popular tree-based classification algorithms, which are, Random Forest (RF) [4-5] and C5.0 [6-7]. CART and C4.5 are simple decision tree models because there is only one decision tree in each model. While RF and C5.0 are ensemble methods based on CART and C4.5, and each of them has a bunch of basic decision trees in the model. Some of the differences among these two algorithms are shown in Table 1.

TABLE I. BRIEF DIFFERENCE BETWEEN RF&C5.0

Algorithms	Number of Trees	Methods	Basic Classifier
RandomForest	Multiple	Bagging and Voting	CART
C5.0	Multiple	Boosting and Voting	C4.5

R [8] is an open source programming language and software environment for statistical computing and graphics. It is widely used for data analysis and statistical computing projects. In this paper, we will use some R packages as our analysis tools, namely "randomForest" package [9] and "C50" package [10]. Moreover, we also use some packages for plotting figures, such as "reshape2" package [11] and "ggplot2" package [12].

The structure of the paper is: Section 2 reviews some basic concept of tree-based classification methods, while Section 3 and 4 will describe the data and the experiments. The conclusion will be given in Section 5.

II. METHODOLOGY

A. Random Forest (RF)

RF is an effective prediction tool in data mining, which is based on CART. It employs the Bagging [13] method to produce a randomly sampled set of training data for each of the trees.

CART uses *Gini* index which is an impurity-based criterion that measures the divergences among the probability distributions of target attribute's values.

Definition 1 (*Gini Index*): Given a training set S and the target attribute takes on k different values, then the *Gini* index of S is defined as

$$Gini(S) = 1 - \sum_{i=1}^k p_i^2$$

Where p_i is the probability of S belonging to class i .

Definition 2 (*Gini Gain*): *Gini Gain* is the evaluation criterion for selecting the attribute A which is defined as

$$\begin{aligned} GiniGain(A, S) &= Gini(S) - Gini(A, S) \\ &= Gini(S) - \sum_{i=1}^n \frac{|S_i|}{|S|} Gini(S_i) \end{aligned}$$

Where S_i is the partition of S induced by the value of attribute A .

CART algorithm can deal with the case of features with nominal variables as well as continuous ranges.

Pruning a tree is the action to replace a whole sub-tree by a leaf. CART uses a pruning technique called “minimal cost-complexity pruning” which assuming that the bias in the re-substitution error of a tree increases linearly with the number of leaves. Formally, given a tree T and a real number $\alpha > 0$ which is called the “complexity parameter”, then the cost-complexity risk of T with respect to α is:

$$R_\alpha(T) = R(T) + \alpha \cdot |T|$$

where $|T|$ is the number of terminal nodes (i.e. leaves) and $R(T)$ is the re-substitution risk estimate of T .

Bagging, which stands for “bootstrap aggregating”, is an ensemble classification method. It repeatedly samples from a data set with replacement according to a uniform probability distribution. Each sample has a probability $1 - (1 - 1/N)^N$ of being selected, where N is the number of observations in the training set. If N is sufficiently large, this probability converges to $1 - 1/e \approx 0.632$, that is, a bootstrap sample contains approximately 63% of the original training data, while other data is a natural good testing dataset which is known as OOB (Out of Bag [14]) in RF. Since every sample has an equal probability of being selected (i.e. $1/N$), bagging does not focus on any particular instance of the training data. Therefore, it is less sensitive to model overfitting when applied to noisy data.

RF constructs a series of tree-based learners. At each tree node, a random sample of m features is drawn, and only those m features are considered for splitting. Typically $m = \sqrt{p}$ (as default in R “randomForest” package), where p is the number of features. The essential difference between Bagging and RF is the latter not only selecting samples randomly but also the features being selected randomly. RF will not prune the trees during the whole growing procedure.

B. C5.0

C5.0 is an advanced decision tree algorithm developed based on C4.5. It includes all functionalities of C4.5 and applies some new technologies, the most important application among them is Boosting (i.e. AdaBoost [15]), which is a technique for generating and combining multiple classifiers to improve predictive accuracy.

C4.5 uses information gain ratio which is an impurity-based criterion that employs the entropy measure as an impurity measure.

Definition 3 (*Information Entropy*): Given a training set S , the target attribute takes on k different values, and then the entropy of S is defined as

$$Entropy(S) = - \sum_{i=1}^k p_i \log_2 p_i$$

Where p_i is the probability of S belonging to class i .

Definition 4 (*Information Gain*): The information gain of an attribute A , relative to the collection of examples S , is defined as

$$\begin{aligned} InfoGain(A, S) &= Entropy(S) - Entropy(A, S) \\ &= Entropy(S) - \sum_{i=1}^n \frac{|S_i|}{|S|} Entropy(S_i) \end{aligned}$$

Where S_i is the partition of S induced by the value of attribute A .

Definition 5 (*Gain Ratio*): The gain ratio “normalizes” the information gain as follows:

$$\begin{aligned} GainRatio(A, S) &= \frac{InfoGain(A, S)}{SplitEntropy(A, S)} \\ &= \frac{InfoGain(A, S)}{- \sum_{i=1}^n \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|}} \end{aligned}$$

Similar to CART, C4.5 can also deal with both nominal and continuous variables.

Error Based Pruning (EBP) is the pruning method which is implemented in C4.5 algorithm. The idea behind EBP is to minimize the estimated number of errors that a tree would make on unseen data. The estimated number of errors of a tree is computed as the sum of the estimated number of errors of all its leaves.

AdaBoost stands for “adaptive boosting”, it increases the weights of incorrectly classified examples and decreases the ones of those classified correctly.

C5.0 is much efficient than C4.5 also on the aspect of unordered rule sets. That is, when a case is classified, all applicable rules are found and voted. This improves both the interpretability of rule sets and their predictive accuracy.

III. DATA PREPARATION

All of data for the 2000-2011 period were obtained from Hong Kong Environmental Protection Department (HKEPD)

and Hong Kong Met-online. The air monitoring station is Tung Chung Air Monitoring Station (Latitude 22°17'19"N, Longitude 113°56'35"E) which is in a new town of Hong Kong, and the meteorological monitoring station is Hong

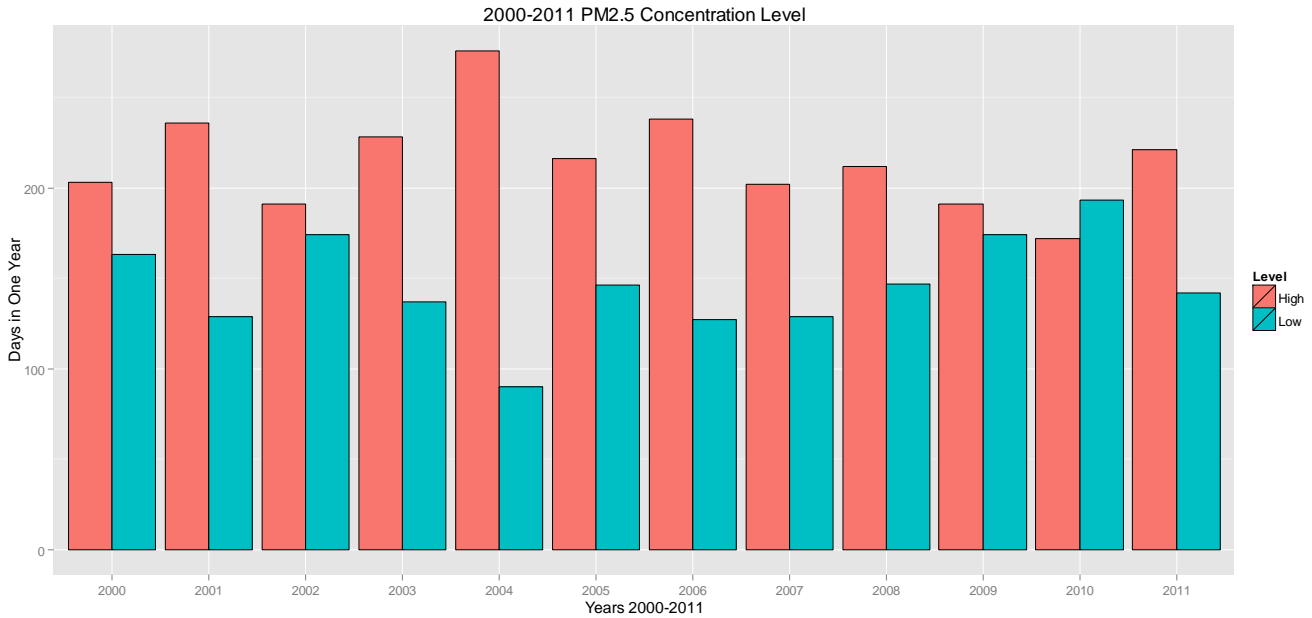


Fig.1. PM_{2.5} concentration levels in 2000-2011

Kong International Airport Weather Station (Latitude 22°18'34"N, Longitude 113°55'19"E) which is the nearest station from Tung Chung. As mentioned in Section 1, accurately predicting high PM_{2.5} concentration is of most value from a public health standpoint, thus, the response variable has two classes, which are, "Low" indicating the daily mean concentration of PM_{2.5} is below 25µg/m³ and "High" representing above it. Figure 1 shows that the days of two levels in 2000-2011.

We learn that the air quality is the best in 2010 (i.e. it has the most "Low" days), while the worst is in 2004 among these 12 years. In summary, the percentage of "Low" and "High" level is around 40.0% and 59% during 12 years in this area, respectively (around 1% missing values). Thus, if a predictive model obtains the accuracy is less than 60%, which means it approximately equals the randomly guess, that would be failure. The purpose to use data mining method is to raise the accuracy, say, at least more than 60%.

We convert all hourly PM_{2.5} data to daily mean values and the meteorological data is the original daily data. In addition, all of air data and meteorological data are numeric. We certainly cannot ignore the effects of seasonal changes and human activities; hence we add two time variables, namely the month (Figure 2) and the day of week (Figure 3).

Figure 2 clearly shows that PM_{2.5} concentration reaches a low level from May to August, during which is the rainy season in Hong Kong. But the pollutant is serious from October to next January, especially in December and January.

We should know that the rainfall may not be an important factor in the experiment as the response variable is the next day's PM_{2.5}, and it is easy to understand that rainy season includes variant meteorological factors. Figure 3 presents the trends of people's activities in some senses. We learn that the air pollution waves slightly during the week. The concentration levels are similar from Tuesday to Thursday, while the lowest level appears on Sunday. This situation can be related to Tung Chung is a living area in Hong Kong, which means the air is less influenced by factories or other pollution source (i.e. different from business area or industrial area).

At last, there are 4326 observations by deleting all NAs and 14 predictor variables (Table 2) and 1 response variable which is the next day's PM_{2.5} concentration level.

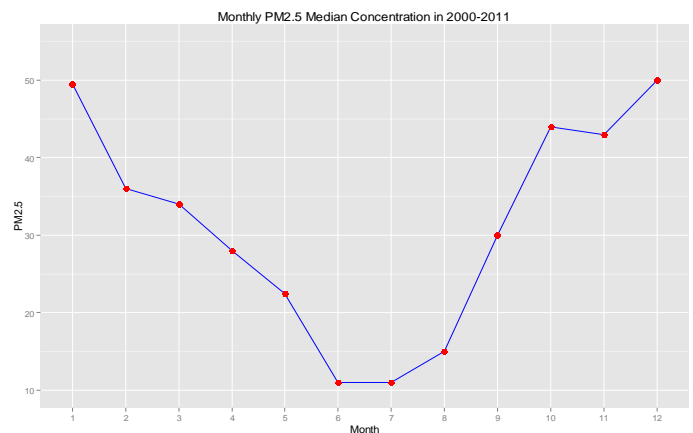


Fig.2. Monthly PM_{2.5} Concentration in 2000-2011

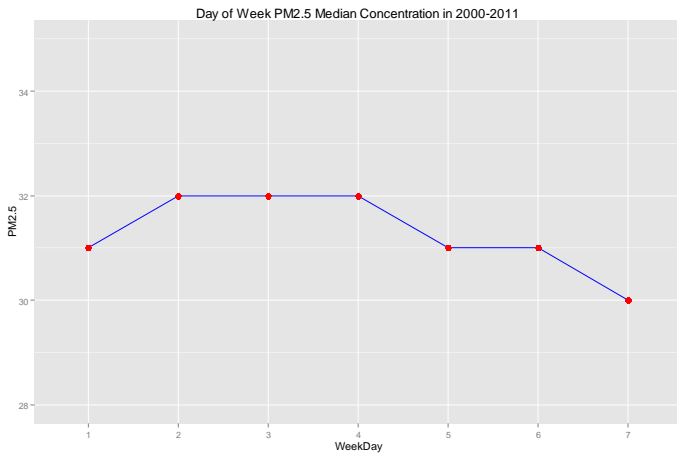


Fig.3. Daily PM_{2.5} Concentration in 2000-2011

TABLE II. VARIABLE LIST

Notation	Description	Variable Class
MP	Mean Pressure	Numeric
AT1	Max Air Temperature	Numeric
AT2	Mean Air Temperature	Numeric
AT3	Min Air Temperature	Numeric
MDPT	Mean Dew Point Temperature	Numeric
RH1	Max Relative Humidity	Numeric
RH2	Mean Relative Humidity	Numeric
RH3	Min Relative Humidity	Numeric
TR	Total Rainfall	Numeric
PWD	Prevailing Wind Direction	Numeric
MWS	Mean Wind Speed	Numeric
PM2.5	PM _{2.5} concentration	Numeric
MONTH	Month	Nominal
WEEK	Day of week	Nominal

IV. EXPERIMENTS

The experiments include three sections: the first and second experiment will test RF and C5.0, respectively. We try to train the model and select the proper parameters in each one. The third one will compare them by using 100 times 10-fold cross validation (10-fold CV) in order to understand which one is more accurate as well as stable.

A. Random Forest (RF)

We use R package “randomForest” to train and test the performance of RF model. There are two important parameters in RF model, that is, the number of splitting feathers (i.e. “mtry”) and the number of trees (i.e. “ntree”). We try to select

a proper number in order to obtain the best testing accuracy by 10-fold CV. Firstly, we set “ntree” from 1 to 500 and “mtry” from 2 to 5. The result is shown in Figure 4. We learn that when the number of splitting feathers is 2 or 3 is somewhat better than other two values as the accuracy of both are tightness. The best accuracy of each splitting number is shown in Table 3. According to this result we choose *mtry* = 3 and *ntree* = 98 in the following experiments. Note that the default value in R package is *mtry* = \sqrt{p} (as we mentioned in Section 2) and *ntree* = 500, generally speaking, we can use “mtry” as the default value and select “ntree” by using 10-fold CV. Alternatively, one can choose the function *tuneRF* for selecting parameters in RF model and the details can be checked in the help file of randomForest package. Figure 5 shows that the importance of variables in RF model, we can see the most important predictor is the previous PM_{2.5}, and then MDPT, MP, and MONTH. The criterion of this list is according to the mean decreasing *Gini* gain of each predictor. Why the variable WEEK is not important in RF model? A reasonable explanation is that WEEK waves slightly on each day, moreover, all the medians of PM_{2.5} concentration are higher than 25 $\mu\text{g}/\text{m}^3$ (see Figure 3) which is the boundary between response variable.

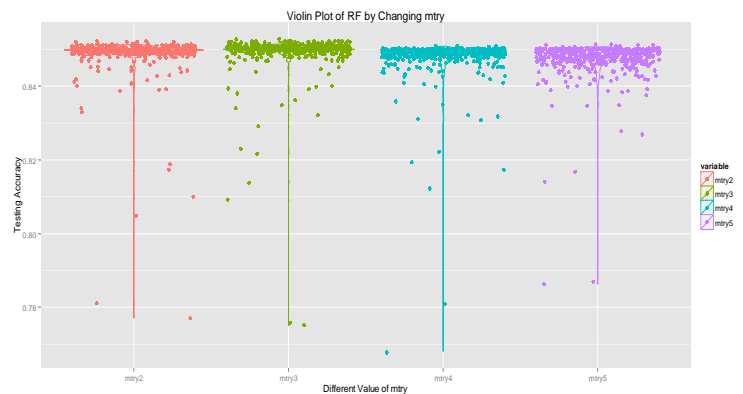


Fig.4. Accuracy on Different Number of Splitting Feathers

TABLE III. RESULT OF RF

mtry	ntree	Testing Accuracy
2	95	0.852
3	98	0.853
4	246	0.851
5	349	0.851

B. C5.0

We will use “C50” package for building C5.0 model in this paper. Similar as RF, we try to obtain the best number of trees at first. Note that the maximum number of trees in “C50” package is 100, which is much less than RF (i.e. *ntree* = 500). Some of the results are shown in Table 4. We find that the highest accuracy is at the 32nd tree, whose testing accuracy is 0.852. Figure 6 indicates the trends of accuracy by changing number of trees in RF and C5.0. The training accuracy of both models increases steadily and stays at a stable level at last. The testing accuracy waves little serious at the beginning,

especially, C5.0 is much higher than RF when there are only a few trees. But both of them float in a moderate level later, RF is higher than C5.0 at this phase. Figure 7 shows the variables importance of C5.0 model. According to this result, we learn that the most important predictor is the previous PM_{2.5}, too. And MONTH, PWD are also important variables, but it is different from the result of RF. C5.0 calculates the percentage of splits associated with each predictor. We think this result is more accurate than RF algorithm, because *Gini* gain will be in favor of those variables having more values and thus offering more splits [16]. But C5.0 uses gain ratio which avoids this problem. Variable WEEK is still not important in this model.

TABLE IV. RESULT OF C5.0

Trees	Training	Testing
1	0.868	0.843
2	0.868	0.843
3	0.877	0.842
.....
31	0.943	0.847
32	0.945	0.852
33	0.945	0.849
.....

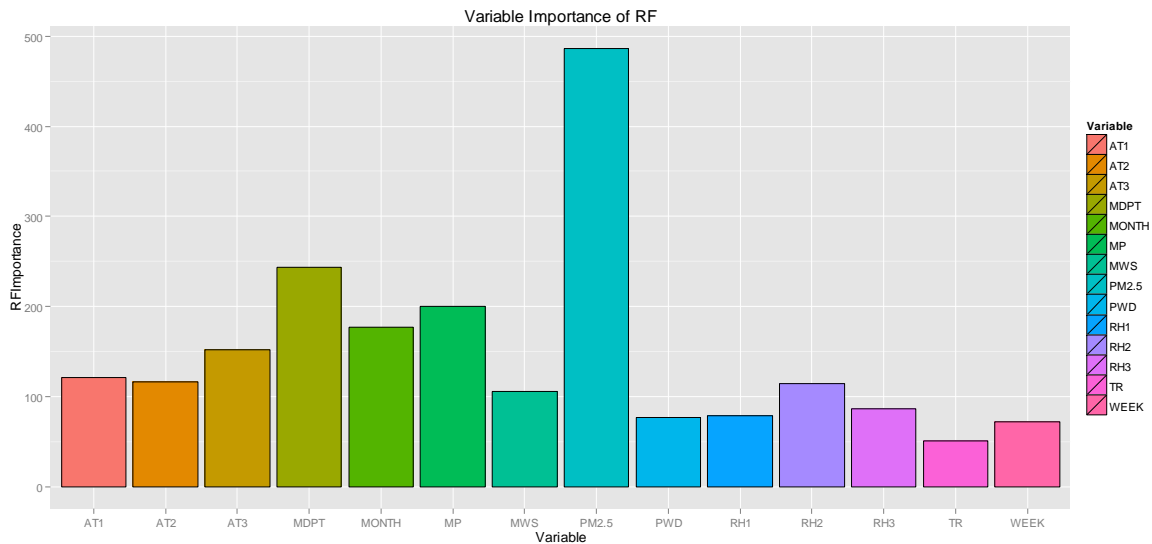


Fig.5. Variable importance of RF

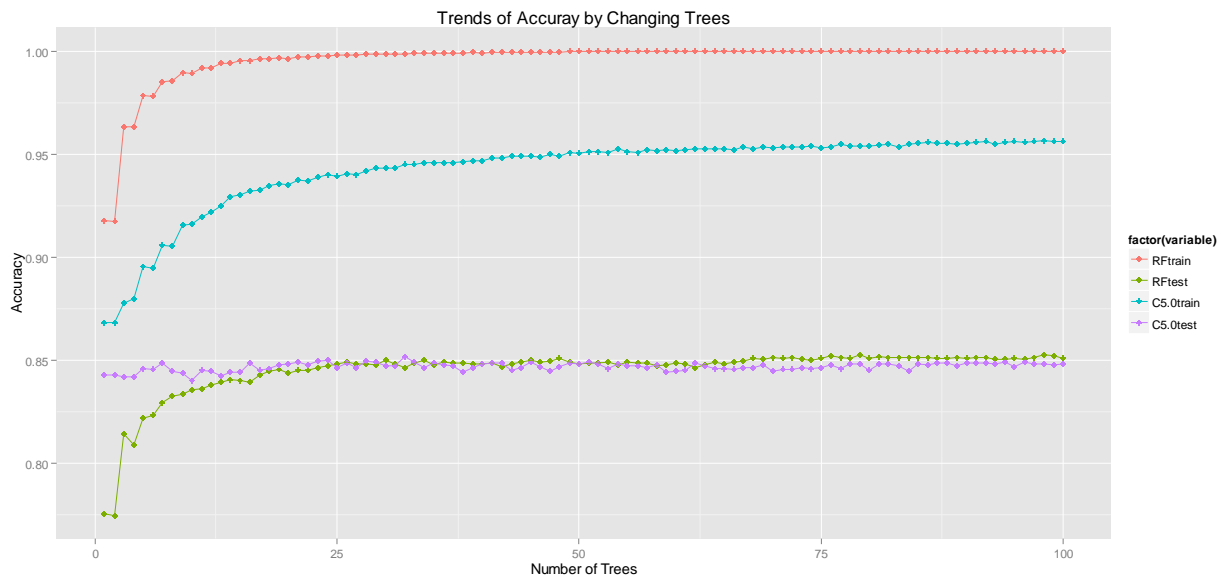


Fig.6. Trends of Testing Accuracy by Changing Number of Trees in RF & C5.0

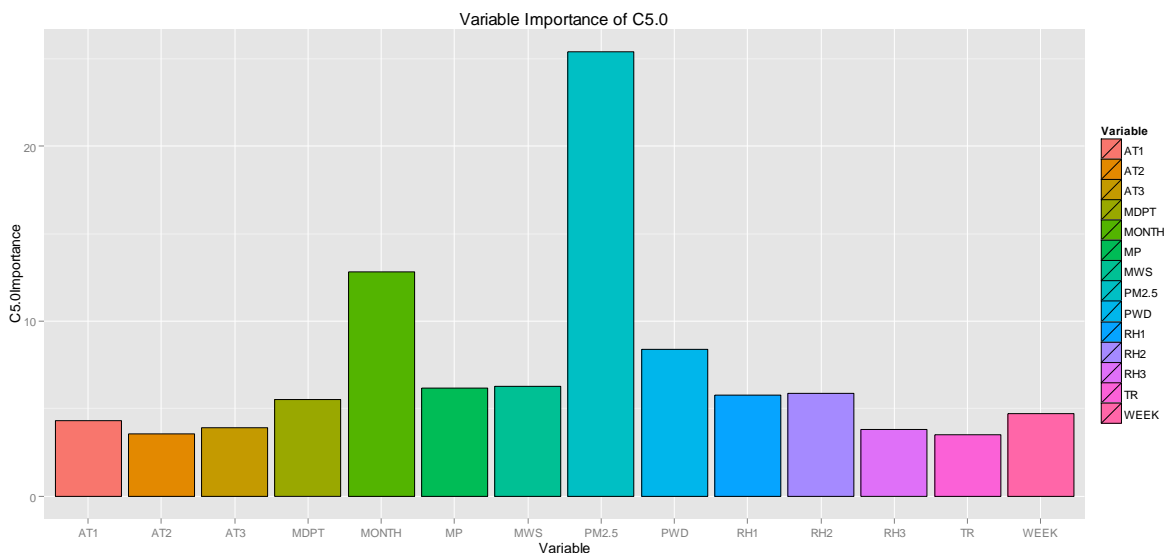


Fig.7. Variable importance of C5.0

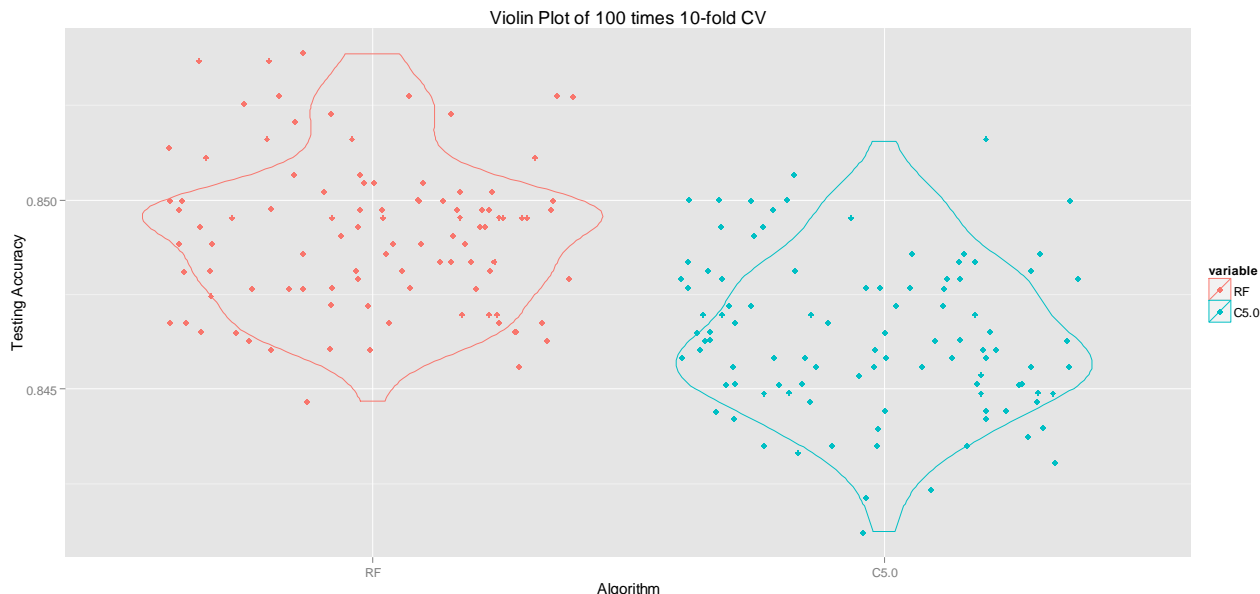


Fig.8. Violin Plot of 100 Times 10-fold CV

C. Comparison

We compare two algorithms by using 100 times 10-fold CV with the result shown in Table 5.

We learn that RF obtains the best result, and its accuracy is around 0.845~0.854. While C5.0 also gets a moderate accuracy, or say, only a little bit worse than RF. Another issue is the stability of these two algorithms during repeated times.

Figure 8 shows the violin plot of 100 times 10-fold CV. We can see that RF is more stable than C5.0 during 100 times and its accuracy is also better than C5.0.

TABLE V. COMPARISON BETWEEN RF&C5.0

	Maximum	Minimum	Median
RF	0.854	0.845	0.849
C5.0	0.852	0.841	0.846

V. CONCLUSION

In this paper, we build $PM_{2.5}$ concentration levels predictive models by using two popular data mining algorithms, which are RF and C5.0. The dataset, which is from a new town in Hong Kong, includes 4326 rows and 15 columns by deleting all missing values. Based on all experiments, we have our conclusions as below.

1) *Selecting the best parameters in each model based on the testing accuracy by using 10-fold CV. For RF model, the number of trees is 98 and the number of splitting features is 3. For C5.0 model, the best number of trees is 32. We prefer to use default value of mtry in RF model and select ntree by 10-fold CV.*

2) *According to 100 times 10-fold CV, the best result is from RF which is around 0.845~0.854. It not only obtains the highest accuracy but also performs more stable than C5.0.*

3) *Another issue between them is the importance of variables, and we prefer the result of C5.0 as it is unbiased.*

4) *The advice of using RF or C5.0 in practice is to select the number of iterations at first. 10-fold CV is the selecting method in this paper, while researchers can repeat this process many times, for instance, 10 times 10-fold CV should be better than once. In summary, the selecting process has to maximum limit reducing the random error.*

ACKNOWLEDGMENT

The authors wish to thank Hong Kong Environmental Protection Department (HKEPD) and Hong Kong Met-online allowing us to use all the data in this paper.

REFERENCES

- [1] Air quality and health, <http://www.who.int/mediacentre/factsheets/fs313/en/index.html>
- [2] WHO Air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide (Global update 2005), World Health Organization, whqlibdoc.who.int/hq/2006/WHO_SDE_PHE_OEH_06.02_eng.pdf
- [3] Proposed New AQOs for Hong Kong, http://www.epd.gov.hk/epd/english/environmentinhk/air/air_quality_objectives/files/proposed_newAQOs_eng.pdf
- [4] Leo Breiman, "Random Forests", Machine Learning, Volume 45, Issue 1, pp. 5-32, 2001.
- [5] Leo Breiman, "Statistical Modeling: The Two Cultures", Statistical Science, Volume 16, No. 3, pp. 199-231, 2001.
- [6] J. R. Quinlan, "C4.5: Programs for Machine Learning", Morgan Kaufmann, Los Altos, 1993.
- [7] C5.0: An Informal Tutorial, www.rulequest.com/see5-unix.html
- [8] R-project: <http://www.r-project.org/>
- [9] Andy Liaw, Matthew Wiener, "randomForest" package, Version 4.6.7, <http://cran.r-project.org/web/packages/randomForest/randomForest.pdf>
- [10] Max Kuhn, Steve Weston, Nathan Coulter, "C50" package, Version 0.1.0 14, <http://cran.r-project.org/web/packages/C50/C50.pdf>
- [11] Hadley Wickham, "reshape2" package, Version 1.2.2, <http://cran.r-project.org/web/packages/reshape2/reshape2.pdf>
- [12] Hadley Wickham, "ggplot2" package, Version 0.9.3.1, <http://cran.r-project.org/web/packages/ggplot2/ggplot2.pdf>
- [13] Leo Breiman, "Bagging Predictors", Machine Learning, volume 24 issue 2, pp. 123-140, 1996.
- [14] Leo Breiman, "Out-of-bag estimation", <http://www.stat.berkeley.edu/~breiman/OOBestimation.pdf>
- [15] Yoav Freund, Robert Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting", European Conference on Computational Learning Theory, pp. 23-37, 1995.
- [16] Leo Breiman, Jerome Friedman, Richard Olshen, Charles Stone, "Classification and Regression Trees", Wadsworth Int. Group, pp.42, 1984.

Data Flow Sequences: A Revision of Data Flow Diagrams for Modelling Applications using XML.

James PH Coleman
Edge Hill University
St Helen's Rd, Ormskirk, UK, L39 3LG.

Abstract—Data Flow Diagrams were developed in the 1970's as a method of modelling data flow when developing information systems. While DFDs are still being used, the modern web-based which is client-server based means that DFDs are not as useful. This paper proposes a modified form of DFD that incorporates, amongst other features sequences. The proposed system, called Data Flow Sequences (DFS) is better able to model real world systems in a way that simplifies application development. The paper also proposes an XML implementation for DFS which allows analytical tools to be used to analyse the DFS diagrams. The paper discusses a tool that is able to detect orphan data flow sequences and other potential problems.

Keywords—Data Flow Diagrams; Modelling diagrams; XML; Data Flow Sequence Diagrams

I. INTRODUCTION

Data Flow Diagrams (DFDs) [1] were developed in the late 1970's as a method of modelling the flow of data through an information system. According to Bruza [2] they are often used in the preliminary design stages to provide an overview of the system. Today there are a number of advanced modelling tools (including UML [3] – which was developed by Grady Booch, Ivar Jacobson and Jim Rumbaugh at Rational Software in the 1990s) and Business Activity Models [4] and other tools that not only describe the data flow, but also specify the processing steps involved. These tools can then be (in some cases) to automatically develop the code.

Data flow diagrams are one of essential perspectives of the structured-systems analysis and design method SSADM [5]. SSADM is one particular implementation and builds on the work of different schools of structured analysis and development methods.

Kolhatkar [6] proposed the development of an XML representation of DFDs to overcome a number of identified weaknesses with the graphical DFDs used. These included: the amount of time it takes to actually “draw” the DFDs given that DFDs are usually developed iteratively and ambiguity in understanding given that there are a number of different models in drawing DFDs. There exists at least 2 major versions (Yourdon & Coad [7] and Gane & Sarson [8]).

II. DATA FLOW SEQUENCE DIAGRAMS

In this paper we will consider a revised and modernised form of DFD that is better suited to modern applications, particularly web-based applications. Web based applications

are characterised by the client-server nature of the relationship – where the main entity (the User) communicates with a client system (usually called a web browser), and the web browser then communicates with one/more servers (called web servers) which may themselves communicate with other processes using system systems as SOAP [9] HTTP-based systems.

The main difference between an application and a web-based application stems from the fact that web-based applications exists within a context of a web-page that is displayed by the web browser. This web-page is downloaded from the web server, which is again in the context of a web-page (the application). This means that all data flows communicate with processes that are sub-components of a page (or group of processes), and these pages are *downloaded* from pages from the Server.

In order to support this extended definition, DFS diagrams include the concept of sequence – that is, dataflows are sequenced. This indicates the sequence in which dataflows, and processes/entities receive data, process data and then produce output.

Processes run on either the Client system, the Server system or on a separate system detached from the client or server. An example of this would be a DB server. Even though a DB system (or datastore) may actually be running on the same networking device as the web server, by putting it as being separate from the Web Server, this indicates that the DB server may be physically separate.

Kolhatkar [6], in his proposal for representing DFDs in XML, established a number of XML tags, including: <process>, <entity>, <dataflow> and <datastore>, each with a number of attributes. Processes, entities and data stores have an id attribute that is used in the dataflow to identify the source and destination tags.

This article introduces a number of new concepts to the DFD, forming the Data Flow Sequence Diagrams (DFS). These changes are:

The Introduction of a *Client* and *Server* as sites for executing processes, and

On the Client and Server, there are *Process Groups* (called a *procgroup*) which conceptually form a *page equivalent* for web-based applications.

Clients and proggroups allows the Designer to introduce the concept of a web page cookie – where a cookie is datastore on a client and inside a proggroup, so that if the proggroup closes then the datastore is lost. This mimics the behaviour of page-bound cookies which are only accessible from the current web page, and when the web page is replace, then the cookie is removed. At the same time, the cookie is also accessible to the same web-page (proggroup) on the server.

Similarly, processes, datastores and the server allows DFS to mimic the PHP *session variable* – which is a variable that only exists on the server. It has a wider scope than the proggroup, but always only exists on the server.

Processes exist on both clients, and servers. Processes can be executed inside a proggroup, if it is a process created by a web page, or outside the proggroup environment as would happen for instance with a PHP DB request, which comes from a process in a proggroup on the server and is sent to a DB server for processing, and then returned.

Irrespective of whether a process is in a proggroup or not, dataflows connect entities to/from processes, datastores to/from processes or process to/from process.

Figure 1 illustrates a DFS diagram that represents a sub-set of the Guest Book system where the User has supplied his personal information, this data is stored in a DB, and then all guest book entries are displayed. In this example, there is a proggroup that is labelled “cgbook.php” which indicates that the included processes are a part of the gbook.php file on the Client. Similarly, there is a proggroup called sgbook.php which includes the processes that are executed on the Server.

III. XML REPRESENTATION

Figure 2a illustrates how a portion of this DFS is represented in XML. For clarity, the Context diagram has been removed, as has a number of the context-level dataflows.

```

<?xml version="1.0" encoding="UTF-8"?>
<!-- New document created with EditiX at Wed Dec 05 15:16:12 GMT 2012 -->
<dfs>
...
<level id="1">
<entity id="E1"><label>User</label></entity>
<client>
<process id="CP1"><label>request web page</label></process>
<proggroup id="gbook.php">
<process id="CP2"><label>Display Web Page</label></process>
<process id="CP3"><label>Obtain User Data</label></process>
<process id="CP4"><label>Display Guest Book and prepare for next entry</label></process>
</proggroup>

```

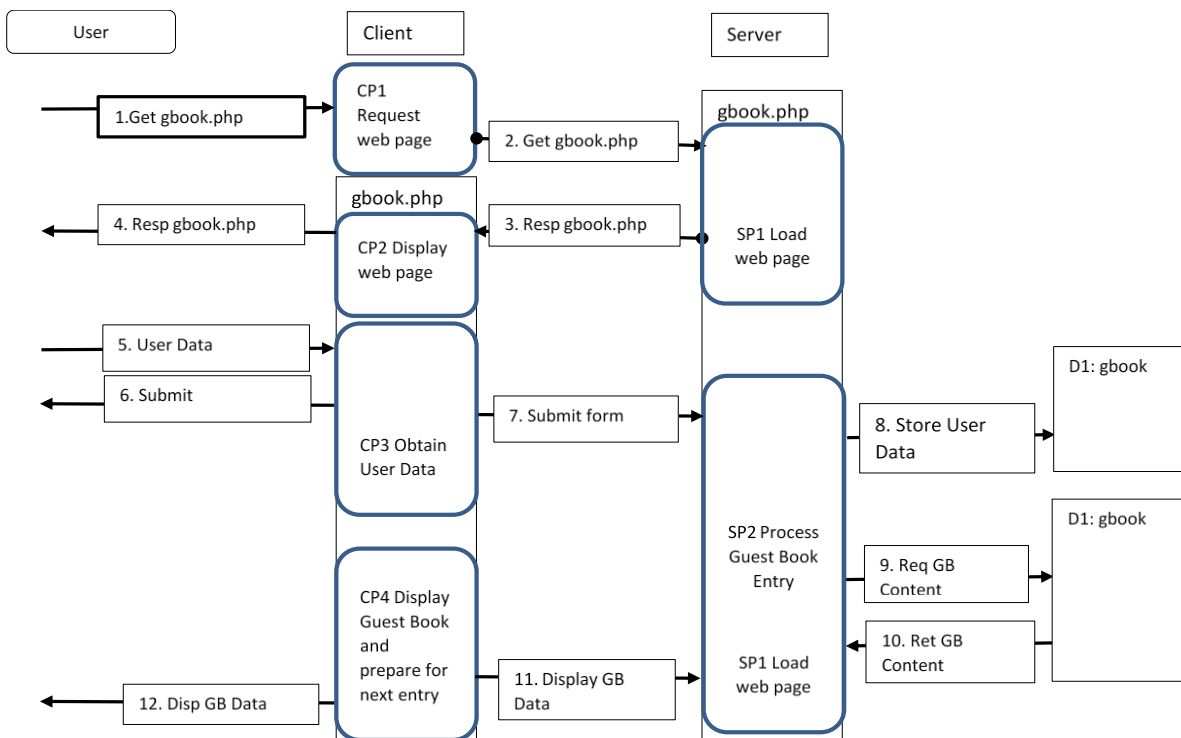


Fig. 1 DFS diagram for Guest Book Example system

```
</client>
<server>
  <process id="SP1"><label>load          web
page</label></process>
  <procgroup id="gbook.php">
    <process id="SP2"><label> Process Guest Book
Entry</label></process>
  </procgroup>
</server>
<datastore id="DB1"><label>Guest      Book
DB</label></datastore>
  <dataflow from="E1" to="CP3" seq="5"><label>Enter
user data</label></dataflow>
  <dataflow from="E1" to="CP3" seq="6"><label>Submit
Form</label></dataflow>
  <dataflow from="CP3" to="SP2" seq="7"><label>Submit
Form Data</label></dataflow>
  <dataflow from="SP2" to="DB1" seq="8"><label>Store
User Data</label></dataflow>
  <dataflow from="SP2" to="DB1"
seq="9"><label>Request GB Data</label></dataflow>
  <dataflow from="DB1" to="SP2"
seq="10"><label>Retrieve GB Data</label></dataflow>
  <dataflow from="CP4" to="E1" seq="11"><label>Display
GB Data</label></dataflow>
  ...
</level>
</dfs>
```

Fig 2a: Figure 2a illustrates a simplified implementation of a web-based guest-book application.

Each process, entity, dataflow, datastore, procgroup has a unique identifier. The uniqueness of these different objects enable the dataflows to specify the relevant process, entity or datastore without needing to distinguish between local and global identifiers, and also allows dataflows to link between levels. That is, a Level 2 DFS diagram is able to reference an entity at a higher (or lower) level.

```
<!ELEMENT dfs ( context , level )>
<!ELEMENT level ( entity+ , client ,
server , datastore+ , dataflow+ )>
<!ATTLIST level
id CDATA #REQUIRED
>
<!ELEMENT dataflow ( label )>
<!ATTLIST dataflow
from CDATA #REQUIRED
to CDATA #REQUIRED
```

```
seq CDATA #REQUIRED
>
<!ELEMENT label ( #PCDATA )>
<!ELEMENT datastore ( label )>
<!ATTLIST datastore
id CDATA #REQUIRED
>
<!ELEMENT label ( #PCDATA )>
<!ELEMENT server ( process+ , procgroup+
)>
<!ELEMENT procgroup ( process+ )>
<!ATTLIST procgroup
id CDATA #REQUIRED
>
<!ELEMENT process ( label )>
<!ATTLIST process
id CDATA #REQUIRED
>
<!ELEMENT label ( #PCDATA )>
<!ELEMENT process ( label )>
<!ATTLIST process
id CDATA #REQUIRED
>
<!ELEMENT label ( #PCDATA )>
<!ELEMENT client ( process+ , procgroup+
)>
<!ELEMENT procgroup ( process+ )>
<!ATTLIST procgroup
id CDATA #REQUIRED
>
<!ELEMENT process ( label )>
<!ATTLIST process
id CDATA #REQUIRED
>
<!ELEMENT label ( #PCDATA )>
<!ELEMENT process ( label )>
<!ATTLIST process
id CDATA #REQUIRED
>
<!ELEMENT label ( #PCDATA )>
<!ELEMENT entity ( label )>
<!ATTLIST entity
id CDATA #REQUIRED
>
<!ELEMENT label ( #PCDATA )>
<!ELEMENT context ( process , entity+ ,
dataflow+ )>
<!ELEMENT dataflow ( label )>
<!ATTLIST dataflow
from CDATA #REQUIRED
to CDATA #REQUIRED
seq CDATA #REQUIRED
>
<!ELEMENT label ( #PCDATA )>
<!ELEMENT entity ( label )>
<!ATTLIST entity
id CDATA #REQUIRED
>
```



```
<!ELEMENT label ( #PCDATA )>
<!ELEMENT process ( label )>
<!ATTLIST process
id CDATA #REQUIRED
>
<!ELEMENT label ( #PCDATA )>
```

Fig2b: provides the DTD for the DFS XML representation.

IV. DFS VALIDATION AND ANALYSIS

Having the DFS being represented using XML enables automatic validation and analysis of the DFS diagram to ensure the diagram truly represents the real world. As an example, an analyser tool has been developed that is able to follow the flow of data from process to process. It is also possible to start at a dataflow and back-track to find out where the data came from.

Figure 3 illustrates part of the track of a dataflow in the system described in Figure 2 above – the Guest Book system. Here the tool shows the different paths. By holding the mouse over a dataflow, you are able to see the destination process/entity/datastore.

Follow DataFlow

CDF1 ▾ Forward ▾ Go

CDF1(E0 to P0) CDF2(P0 to E0)

Symbol Table

ID	Level	Type	Description	Seq	From	To
P0	-1	process	Guest Book System			
E0	-1	entity	User			
CDF1	-1	dataflow	Send Details	1	E0	P0
CDF2	-1	dataflow	Return Guest Book	2	P0	E0
E1	1	entity	User			
client	1	datastore	null		client	
CP1	1	process	request web page		client	
Cgbook.php	1	datastore	null		client	Cgbook.php
CP2	1	process	Display Web Page		client	Cgbook.php
CP3	1	process	Obtain User Data		client	Cgbook.php
CP4	1	process	Display Guest Book and prepare for next entry		client	Cgbook.php
server	1	datastore	null		server	
SP1	1	process	load web page		server	
Sgbook.php	1	datastore	null		server	Sgbook.php
SP2	1	process	Process Guest Book Entry		server	Sgbook.php
DB1	1	datastore	Guest Book DB			

Fig. 2 DFS Analyser

Analysers can also search for a number of anomalous conditions such as:

- Where there exists a process for which there output dataflows but no input data flows
- Where there exist a process that has an input dataflow (or dataflows) but no out dataflow.
- Processes where there are no connecting dataflows.

- Datastores where there are input but no output dataflows
- Datastores has output dataflows but no input dataflows.
- Dataflows between entities without an intervening process
- Dataflows between datastores without an intervening process

The Analyser can also provide a definitive list of the elements that make up the DFS diagram in the form of a symbol table that lists all the key information about the entities, processes and datastores showing whether they are client/server/other based and whether the a process is part of a proggroup. Similarly the system shows all dataflows, and indicates which element it is connected to. In this way the tool aids the designer is locating data flows and process orphans.

V. CONCLUSIONS

DFS Diagrams enable developers to model real world applications with a much richer diagrammatic system than the traditional Data Flow Diagram. DFS Diagrams are specifically designed to support web-based applications with the concept of a client and server being an integral part of the DFS system.

With the ability to specify the DFS Diagram using XML, then the diagram can be analysed using XML processing tools such as XPath and XSL. Further, the XML representation can also be used to analyse the DFS diagram looking for fundamental errors in design, as well as the ability to follow a dataflow in sequential order from any starting point to the logical end of the dataflow.

The XML representation can be expanded to include code specification, and in this way can be sued to automatically create applications.

The Analyser currently provides a limited set of validation and analyser tools. Given the flexibility of XML and its efficiency in processing, other analytical tests can be incorporated into the system to aid finding logical and practical problems with the DFS design.

REFERENCES

- [1] W. Stevens, G. Myers, L. Constantine, "Structured Design", IBM Systems Journal, 13 (2), 115-139, 1974.
- [2] Bruza, P. D., Van der Weide, Th. P., "The Semantics of Data Flow Diagrams", University of Nijmegen, 1993
- [3] Marc Hamilton, "Software Development: A Guide to Building Reliable Systems" p.48; Prentice Hall, 1999.
- [4] Cadle, J, Eva M, Hindle K, Paul D, Rollaston C, Tudor D, Yeates D: "Business Analysis" 2nd Edition; British Informatics Society Limited; 2010
- [5] SSADM. "Business Systems Development with SSADM". The Stationery Office. 2000. p. v. ISBN 0-11-330870-1.
- [6] S S Kolhatkar "XML Based Representation of DFD: Removal of Diagramming Ambiguity" International Journal of Advanced Computer Science and Applications, Vol. 2, No. 8, 2011
- [7] Coad, P., Yourdon E, "Object-Oriented Analysis", 2nd Edition, Englewood Cliffs, NJ: Prentice-Hall, 1991.
- [8] C. Gane and T. Sarson. "Structured Systems Analysis: Tools and Techniques", New York: IST, Inc., 1977
- [9] Simple Object Access Protocol (SOAP) 1.1, W3C Note 08 May 2000; <http://www.immagic.com/eLibrary/ARCHIVES/SUPRSDED/W3C/W000520N.pdf> Accessed 8/1/2013.

Comparative study of Authorship Identification Techniques for Cyber Forensics Analysis

Smita Nirkhii

Department of Computer Science & Engg
G.H.Raisoni College of Engineering
Nagpur, India

Dr.R.V.Dharaskar

Director
MPGI
Nanded, India

Abstract—Authorship Identification techniques are used to identify the most appropriate author from group of potential suspects of online messages and find evidences to support the conclusion. Cybercriminals make misuse of online communication for sending blackmail or a spam email and then attempt to hide their true identities to void detection. Authorship Identification of online messages is the contemporary research issue for identity tracing in cyber forensics. This is highly interdisciplinary area as it takes advantage of machine learning, information retrieval, and natural language processing. In this paper, a study of recent techniques and automated approaches to attributing authorship of online messages is presented. The focus of this review study is to summarize all existing authorship identification techniques used in literature to identify authors of online messages. Also it discusses evaluation criteria and parameters for authorship attribution studies and list open questions that will attract future work in this area.

Keywords—cyber crime; Author Identification; SVM

I. INTRODUCTION

Cyber crime is also known as computer crime, the use of a computer to further illegal ends, such as committing fraud, trafficking in child pornography and intellectual property, stealing identities, or violating privacy.

Cybercrime, especially through the Internet, has grown in importance as the computer has become central to commerce, entertainment, and government. Senders can hide their identities by forging sender's address; Routed through an anonymous server and by using multiple usernames to distribute online messages via different anonymous channel.

Author Identification study is useful to identify the most plausible authors and to find evidences to support the conclusion.

Authorship analysis problem is categorized as [13]

1) *Authorship identification (authorship attribution): It determines the likelihood of a piece of writing to be produced by a particular author by examining other writings by that author.*

2) *Authorship characterization: It summarizes the characteristics of an author and generates the author profile based on his/her writings like Gender, educational, cultural background, and writing style*

3) *Similarity detection: It compares multiple pieces of writing and determines whether they were produced by a single author without actually identifying the author like Plagiarism detection. To extract unique writing style from the number of online messages various features need to be considered are Lexical features, content-free features, Syntactic features, Structure features, Content-specific features*

Although authorship attribution problem has been studied in the history but in the last few decades, authorship attribution of online messages has become a forthcoming research area as it is confluence of various research areas like machine learning, information Retrieval and Natural Language Processing. Initially this problem started as the most basic problem of author identification of anonymous texts (taken from Bacon, Marlowe and Shakespeare) [1], now has been grown for forensic analysis, electronic commerce etc. This extended version of author attribution problem has been defined as *needle-in-a-haystack* problem in [2]

When an author writes they use certain words unconsciously and we should able to find some underlying pattern for an authors style. The fundamental assumption of authorship attribution is that each author has habit of using specific words that make their writing unique Extraction of features from text that distinguish one author from another includes use of some statistical or machine learning techniques.

Rest of the Paper is organized as follows. Section 2 Reviews existing techniques used for Authorship Analysis along with their classification. Section 3 explains basic procedure for authorship analysis. Section 4 summarizes Comparisons of various techniques since year 2006 till 2012. Section 5 Reviews performance evaluation parameters required for Authorship Analysis Techniques followed by section 6 which is conclusion.

II. STATE OF THE ART OF CURRENT TECHNIQUES

This section gives fundamental idea on existing Authorship Attribution Techniques followed by their comparison in next section. In literature, this problem was solved using statistical Analysis and Machine learning techniques. These are mainly categorized as shown in Figure 1.

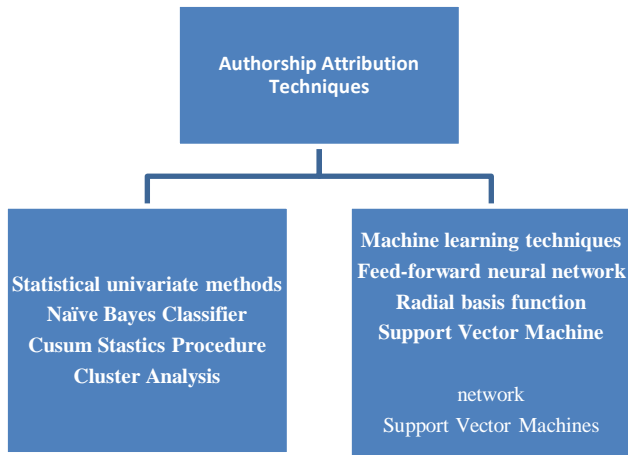


Fig. 1. Authorship Attribution Techniques

STATISTICAL UNIVARIATE METHODS

- A) *Naive Bayes classifier*: In this Classifier Learning and classification methods based on probability theory. In Literature it is found that Bayes theorem plays a critical role in probabilistic learning and classification. It uses prior probability of each category given no information about an item.
- B) *B.CUSUM statistics procedure*: In stastical analysis the cusum called cumulative sum control chart, the CUSUM is a sequential Analysis technique used for onitoring change detection. As its name implies, CUSUM involves the calculation of a cumulative sum.
- C) *Cluster Analysis*: Cluster analysis is an exploratory data analysis tool for solving classification problems. Its purpose is to sort cases (people, things, events, etc) into groups, or clusters, so that the degree of association is strong between members of the same cluster and weak between members of different clusters.

III. MACHINE LEARNING TECHNIQUES

A. Feed-forward neural network :

A feed forward neural network is an artificial neural network where connections between the units do *not* form a directed cycle. This is different from networks. The feed forward neural network was the first and arguably simplest type of artificial neural network devised. In this network, the information moves in only one direction, forward, from the input nodes, through the hidden nodes (if any) and to the output nodes. There are no cycles or loops in the network.

B. Radial basis function network:

A radial basis function network is an artificial neural network that uses radial basis functions as activation functions. The output of the network is a linear combination of radial basis functions of the inputs and neuron parameters.

Radial basis function networks are used for function approximation, time series prediction, and system control.

C. Support Vector Machines:

In machine learning, support vector machines (SVMs, also support vector networks are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. The basic SVM takes a set of input data and predicts, for each given input, which of two possible classes forms the output, making it a non-probabilistic binary linear classifier.

IV. CLASSIC PROCEDURE FOR AUTHORSHIP IDENTIFICATION

Figure 2 shows classic approach to model authorship identification problem.

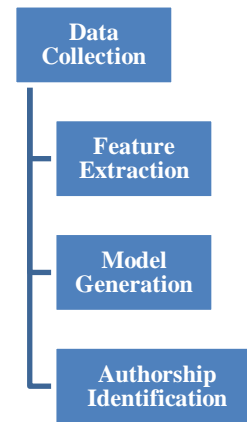


Fig. 2. Typical Procedure for Authorship Identification

Step1: Data collection:-Collect online messages written by potential authors from online communication.

Step2: Feature Extraction:-After extraction, each unstructured text is represented as a vector of writing-style features

Step3: Model Generation:-Dataset should be divided into training and testing set. Classification techniques should be applied. An iterative training and testing process may be needed

Step4: Author Identification:-Developed model can be used to predict the authorship of unknown online messages

V. COMPARISON OF VARIOUS TECHNIQUES

This section compares the various techniques used for authorship identification research forum since 2006 to 2012. History of studies on authorship attribution problems presented in tabular format and year wise. For each method, we identify the corpus on which methods were tested, the feature types used and the categorization method used, size of Training set. Table 1 represented the comparative study of all authorship techniques.[5][6][7][8][9][10].

YEAR/AUTHORS	FEATURES	TECHNIQUES	CORPUS	NUMBER OF AUTHORS	TRAINING SET
(2006) Rong Zheng, Jiexun Li, Hsinchun Chen, Zan Huang	Lexical, syntactic, structural, content Specific	SVM	English Internet newsgroup messages & Chinese Bulletin Board System (BBS) messages.	20	48 for English 37 (Chinese)
2006 Ahmed Abbasi and Hsinchun Chen	Lexical, syntactic, structural, content Specific	PCA	USENET forum, Yahoo group forum , website forum for the White Knights	10	30 msgs per forum
2007 cyran	Lexical, syntactic,	ANN	Novels of two famous Polish writers, Henryk Sienkiewicz and Boleslaw Prus	2	168
2007 Daniel Pavelec, Edson Justino, and Luiz S. Oliveira	Linguistic Features	SVM	Our sources were two dif- ferent Brazilian newspapers, Gazeta do Povo (http://www.gazet adopovo.com.br) and Tribuna do Paran´	10	150
2008 EFSTATHIOS STAMATATOS	Stylistic Features	SVM	Corpus Volume 1 (RCV1) Arabic Corpus:	10	1000
Kim Luyckx and Walter Daelemans	Syntactic Features	Memory based learning approac	Personae corpus	145	1400 words
2008 Chun Wei	Email features	clusterin g	Email dataset	42	4200
2008(Hamilton)	Syntactic Features	Stylomet roy		145	2000
2008 Farkhund Iqbal, Rachid Hadjidj, Benjamin C.M. Fung, Mourad Debbabi	Stylometric Features	Frequent Pattern	Enron Dataset	158	200399
2008(M.Connor)	Syntactic	Decision Trees/KN N.	Emails collected from users	12	120
2009 Rachid Hadjidj, Mourad Debbabi, Hakim Lounis, Farkhund Iqbal,Adam Szporer, Djamel Benredjem	Stylometry Features	Stastical Analysis, Machine Learning	Enron Dataset	158	200399
2011 George K. Mikros1 and Kostas Perifanos	Linguistic features	Regulariz ed Logistic Regressio n (RLR) SVM	Dataset	-	-
2012 Ludovic Tanguy, Franck Sajous, Basilio Calderone,	Linguistic Features	Machine Learning Tool	Dataset	10	100 words

VI. CONCLUSION

The complexity level of aforementioned problem is determined by the various parameters like the number of authors and size of training set. This both the parameters play vital role to determine prediction accuracy. Although these parameters are considered critical to the complexity of the problem and therefore the prediction accuracy, there are no studies examining their impact on the authorship-identification performance in a systematic way. The problem of authorship attribution is explored well in the area of literature, newspapers etc but limited work has been done for the authorship identification of online messages like blogs, emails and chat. This comparative study concluded that if number of author's increases and size of training sets decreases then performance degrades. Thus, by considering all these parameters further research direction is to improve prediction accuracy.

REFERENCES

- [1] Estival 2008] [Abbasi et. al. 2008] [Koppel et. al. 2003] [De Vel et. al. 2001].
- [2] Li, J., Chen, H., & Huang, Z. "A Framework for Authorship Identification of Online Messages: Writing-Style Features and classification Technique", *Journal of the American Society for Information Science*, 57(3), 378–393. doi:10.1002/asi.2006.
- [3] Abbasi, A., & Chen, H. "Visualizing Authorship for Identification", *English*, 60–71, (2006).
- [4] Stańczyk, U., & Cyran, K. A. "Machine learning approach to authorship attribution of literary texts", *Journal of Applied Mathematics*, 1(4), 151–158, (2007).
- [5] Pavelec, D., Justino, E., & Oliveira, L. S. "Author Identification using Stylometric Features", *Inteligencia Artificial*, 11(36), 59–65. doi:10.4114/ia.v11i36.892, (2007).
- [6] Stamatasos, E. "Author identification: Using text sampling to handle the class imbalance problem", *English*, 44, 790–799. doi:10.1016/j.ipm.2007.05.012, (2008).
- [7] Iqbal, F., Hadjidj, R., Fung, B. C. M., & Debbabi, M. "A novel approach of mining write-prints for authorship attribution in e-mail forensics", *Information Systems*, 5, 42–51. doi:10.1016/j.diin.2008.05.001, (2008).
- [8] Iqbal, F., Binsalleh, H., Fung, B. C. M., & Debbabi, M. "Mining writeprints from anonymous e-mails for forensic investigation", *Digital Investigation*, 1–9. doi:10.1016/j.diin.2010.03.003, (2010).
- [9] Mikros, G. K., & Perifanos, K. "Authorship identification in large email collections: Experiments using features that belong to different linguistic levels, (2011).
- [10] Tanguy, L., Sajous, F., Calderone, B., & Hathout, N. "Authorship attribution: using rich linguistic features when training data is scarce", (2012).

Privacy Impacts of Data Encryption on the Efficiency of Digital Forensics Technology

Adedayo M. Balogun
School of Computing and Mathematics,
University of Derby,
Derby, United Kingdom.

Shao Ying Zhu
School of Computing and Mathematics,
University of Derby,
Derby, United Kingdom.

Abstract—Owing to a number of reasons, the deployment of encryption solutions are beginning to be ubiquitous at both organizational and individual levels. The most emphasized reason is the necessity to ensure confidentiality of privileged information. Unfortunately, it is also popular as cyber-criminals' escape route from the grasp of digital forensic investigations. The direct encryption of data or indirect encryption of storage devices, more often than not, prevents access to such information contained therein. This consequently leaves the forensics investigation team, and subsequently the prosecution, little or no evidence to work with, in sixty percent of such cases. However, it is unthinkable to jeopardize the successes brought by encryption technology to information security, in favour of digital forensics technology. This paper examines what data encryption contributes to information security, and then highlights its contributions to digital forensics of disk drives. The paper also discusses the available ways and tools, in digital forensics, to get around the problems constituted by encryption. A particular attention is paid to the *Truecrypt* encryption solution to illustrate ideas being discussed. It then compares encryption's contributions in both realms, to justify the need for introduction of new technologies to forensically defeat data encryption as the only solution, whilst maintaining the privacy goal of users.

Keywords—Encryption; Information Security; Digital Forensics; Anti-Forensics; Cryptography; TrueCrypt

I. INTRODUCTION

Data is becoming largely existent in today's world than they were anticipated some three decades ago [1]. Individuals are keeping lot more amount of information than organizations kept in the yesteryears. Significant amounts of such information are valued and consequently preferred to be known to them alone. Such valued information includes their financial details, medical records, locations, as well as professional and network information. Businesses and organizations possess larger amounts of information than individuals. A good amount of such information is critical to their sustained existence and growth. Their intellectual properties and trade secrets are kept away from potential exploits, thus, considered very private. Governments and agencies keep sensitive information that may affect the stability of their jurisdictions, politically or economically, if divulged.

The necessity to keep such information within the required confines describes a component purpose of Information Security, which involves the totality of activities to ensure the protection of information assets that use, store, or

transmit information from risk through the application of policies, education, training, awareness, and technology [2]. Data security involves the consideration of potential confidentiality, integrity, and availability threats to data services, using functions such as identification, authentication, authorization and audit [3].

An important and popular methodology for enforcing information security is encryption, which is itself an element of cryptography. Cryptography provides a secret communication mechanism between two or more parties. Symmetric and Public Key Cryptography employ various algorithms to ensure the security of data items 'at rest', 'in use', and 'in motion'.

Data encryption may not be an explicit solution to information security problems, as organizations remain increasingly vulnerable to data breach incidents, but it is still the most efficient fix when deployed adequately [4]. This has led to the growing availability of full disk encryption tools. Disk manufacturers are embedding full encryption tools into their products, making encryption more available for use [5].

The study conducted by [4] showed the increased usage of full disk, virtual volume, native disk, and flash drive encryptions over two years. However, for reasons other than the cost of deployment and managing an encryption solution, some organizations have shunned or still undecided about adopting encryption solutions. They insisted that "availability is more important than confidentiality" [6]. The time the encryption and decryption processes take before data is made accessible to potential users may cause delay in organizations' operations, depending on how complex the base algorithm is. Such delays may escalate to a sort of denial-of-service situation, which may be adverse to organizations' businesses. On the electronic discovery front, unavailability problem prevents anticipated investigation of cyber-incidents [7]

The rest of this paper is organized thus: Section 2 presents the reason for this work. Section 3 highlights the contributions of data encryption to information security and digital forensics. Section 4 discusses the effects of data encryption on digital forensics processes, as well as the currently improvised digital forensics methods to defeat data encryption issues. In section 5, justifications for new technologies to help forensic investigation of encrypted data containers are discussed. Conclusions are given in Section 6.

II. NECESSITY & SCOPE

Surveys revealed the continuously increasing adoption of cryptographic solutions by organizations for various data security platforms within the last five years [6]. The report of the surveys infers the anticipation of non-users to adopt partial or holistic cryptographic solutions in the nearest future. This suggests the impending domination by cryptographic procedures, to protect information in the computer world. There are ways for investigators to outmaneuver the use of cryptography as a provocation to digital forensics processes. These methods are either by legally obtaining appropriate 'search and seize' authorizations or tactically planning to catch the offender unawares and hence, access live – running and unencrypted – systems [5]. However, only a handful of encryption incidents encountered by investigators have been solved using those methods. The larger lot of about 60% often does not get prosecuted, not because they were missed, but because nothing could be done to access the potential evidence [8][9]. An instance is the case of Brazilian banker Daniel Dantas, whose strong truecrypt passphrase has foiled all attempts by Brazilian police and FBI to access his encrypted potential evidential hard drives [10][11].

The inconsistency of legal systems across boundaries does not make the process easier, as laws may or may not enjoin perpetrators to help the investigators access the encrypted medium [12]. This was evident in the Dantas' suspected money laundering case, where Brazil had no legislation to make him reveal his passphrase or encryption type, unlike the United Kingdom [11]. Therefore, researchers and developers need to be reminded of privacy-enforcement threats to forensic investigations, and pestered about the need for technologies to help deal with accessing encrypted storage devices.

III. ENCRYPTION CONTRIBUTIONS

A. Data Encryption for Information Security

In order to examine threats contributed by a technology, the solutions it offers should be considered too [13]. Encryption, as an element of cryptography, is a methodology for achieving information security, through secretive communications [14].

The United Kingdom's Data Protection Act 1998 most suitably describes the confidentiality element of information security. It seeks to ensure that the information held by organizations of their customers and employees are safeguarded from other uses than they were obtained [15]. This is meant to avert incidents such as identity crimes, and protect such potential victims from damages and embarrassment that unauthorized use of their data may cause [16]. The powers conferred on the Information Commissioner's Office (ICO) and the Financial Services Authority (FSA) to spot check and fine defaulting organizations, as well as the necessity for card-accepting organizations to comply with industry standards, like the Payment Card Industry Data Security Standard (PCI DSS), has led to the increasing adoption of encryption solutions [6].

There is also a huge necessity to ensure the confidentiality of data items, at rest, in use, or in motion [17]. Financial organizations, where transactions are regularly performed on

data, have to ensure that such data are not subject to unauthorized access or modifications. The combination of the encryption and hash technologies to create digital signatures and certificates, which are used to ensure data confidentiality and integrity, is a laudable approach [18].

As far as information security is concerned, data encryption technology has been of invaluable success on the confidentiality and integrity fronts. Whereas on the availability front, it is known for delays on sparse occasions. Serious availability issues caused by the deployment of encryption solutions are not unheard of, although they are usually addressable by providers [19]. In an overall sense, it is hence, agreeable to regard data encryption as a massive solution for information security challenges.

B. Data Encryption for Digital Forensics

During their hard disk sanitization study at Massachusetts Institute of Technology's Laboratory, [14] found out the ease with which data can be retrieved from disk drives. The ability to recover deleted data and locate hidden data was not a challenge, because there were forensic tools that require little or no specialized user training in existence. The success of those tools was attributed to the “widespread failure of the market to adopt encryption technology” [14]. Some eight years after his work at the MIT laboratory, [20] admitted that the current forensic tools are struggling to be useful to digital forensics investigators when certain data are concerned. He stressed the increasing occurrence of such data and identified format incompatibilities, encryption and lack of training as the reasons. [21] Highlighted data scalability and encryption as some of the unaddressed issues too.

Encryption of data on disk drives is implemented at the file system encryption and full disk encryption (bitstream) levels. At the file system encryption level, individual files are encrypted with separate keys. Although the file system encryption protects virtually all the files in a disk drive, other data outside the file system are omitted. Full disk encryption secures data on disk drives with a single symmetric key. Full disk encryption protects data in all areas of the disk drive, including areas outside the file system. Such data are the hidden files, swap files, file metadata, temporary files and caches, registry files, and boot sector data [22][23][24][25].

The preservation and acquisition of an encrypted disk drive can be tricky, depending on the power state, level, and type of encryption used – hardware-based or software-based [5]. It may be easier to preserve and acquire a file system-encrypted disk drive than a fully-encrypted disk drive in the powered off state. Likewise, the acquisition of a software-implemented and fully-encrypted disk drive may be easier than a hardware-implemented and file system-encrypted disk drive. Acquisition may be totally impossible in cases where disk is not accessible [8].

The examination and analysis phases of digital forensics investigation suffer the most from encryption technology. Reference [7] explained that there may be a possibility to recover an encrypted data, but it is often impossible to process the data. An examiner's tool needs access to read the contents of the encrypted data to be processed. However, they

downplayed the threat posed to digital forensics by stating the possibility to circumvent encryption technology, even though it may be time-consuming and luck-dependent.

IV. DATA ENCRYPTION TOOLS AND KNOWN FORENSICS MANOEUVRING METHODS

There are numerous data encryption solutions for disk drives. Each solution addresses data protection and privacy requirements using different methods. Some encryption solutions are compatible with particular operating systems, unlike others who are portable. They protect data at different levels and employ different key management and authentication methods. Most encryption solutions are implemented as software, but hardware-based solutions are preferred by some organizations [25]. Here is a list of popular disk drive encryption solutions: Microsoft's BitLocker, Symantec's PGP, Apple's FileVault, WinMagic's SecureDoc, IronKey's D200, RSA Data Security's RSA SecurPC, and McAfee's Endpoint [25][26][27]. However, the TrueCrypt solution has been briefly examined in this paper because of its immensely controversial reputation and gross utilization [8][9][22].

Although investigators are not entirely incapacitated by data encryption, sometimes it is down to the legal system to help get access to evidential data [7]. They stated that perpetrators used to employ the file-system encryption because they are concerned only about the data that are incriminating in their own opinions. The other areas left unencrypted usually contain sufficient evidence to prosecute them.

However, the threat posed by a full disk encryption solution is more detrimental to digital forensics processes. But as the saying goes that "no machine is 100% efficient", these encryption solutions have some exploitable vulnerabilities.

The encryption status seizes to hold for all data on the entire disk from the point the symmetric key has been accepted by the system until it is shut down. Deductively, data becomes accessible and inaccessible when the system is powered on and off respectively [7][25]. Digital forensics investigators need to execute an authorized and well-planned "search and seizure", with the aim of catching the perpetrator unawares while his system is running.

Another way to circumvent the threat is the traditional search for the encryption key [25]. There is a possibility, no matter how unlikely, that the key to decrypt the disk drive is written on a notepad or stored in USB drive somewhere at the scene.

Advanced memory-based procedures are also used to overcome the encryption threat. The concept of the full disk encryption that decrypts and makes data available for use is a memory function. The encryption key is stored in the memory the first time it was supplied. It remains in the memory and is used to automatically decrypt required data until the system is powered off. Various techniques can be used to retrieve the encryption key from the memory. RAMs hold data for few seconds to minutes – extendable by keeping the RAM cooled – without power. They can then be accessed through dedicated tools, such as MoonSol's Windows Memory, GMG Systems'

KnTList, Passware or F-Response. However, tools such as Wiebetech's HotPlug can be used to transfer the running system to a back-up power supply in case required expertise is not available and seized system needs to be moved to a laboratory [5].

A. The TrueCrypt Scenario

TrueCrypt was first released in 2004, as a privacy-enforcing solution by the TrueCrypt Foundation. It is a very sturdy, free and open-source software solution, which allows intentional or accidental (sudden power outage), partial or full encryption and decryption processes, without compromising its data security ability. Several encryption algorithms used by TrueCrypt to encrypt data include Advanced Encryption Standard (AES), TwoFish, Serpent, AES-Twofish, Serpent-AES, AES-Twofish-Serpent, Twofish-Serpent, and Serpent-Twofish-AES. The PIPEMD-160, SHA-512 or Whirlpool hash algorithms, with a Random Number Generator, are used to create key-files for stronger security. TrueCrypt is capable of running in a portable mode, without being installed on the target disk drive. It supports smart cards and tokens for added security level and provides hot-keys to perform encryption tasks swiftly. It encrypts data in a partition/drive as a whole, or in a file-hosted container [5][9][22][24][27][28][29].

1) Challenges Posed by TrueCrypt

TrueCrypt employs both the file-system and full-disk encryption methods. This ensures that all data, including registry entries, swap space, file metadata, temporary files, and hidden data are protected from unauthorized access. This leaves investigators with the options to plan a surprise access to a running system, search for possibly exposed encryption key, or use advanced memory-fetching techniques for keys or data [9][24][29].

TrueCrypt is an on-the-fly-encryption (OTFE)-based software. This concept ensures that only the data required by the user is available for access. All data are encrypted from the onset. The required data is copied into the memory and decrypted for use there. As such, the data on the disk drive remain encrypted. Once the data being processed by the user has changed and need to be saved to disk, TrueCrypt encrypts and saves it to the disk. Further, restricting investigators' access methods to those highlighted earlier [24].

Plausible deniability is a feature confidently boasted by TrueCrypt. The encrypted volume/drive appears as a drive with random data. A suspicious reaction to the randomness can be quashed with an explanation of a normal wipe process that fills drive with random data. This plausible deniability is possible because TrueCrypt does not have any signature that would be otherwise found in the drive's partition table [5][24][28][29].

It also allows volumes and operating systems to be hidden inside a visible TrueCrypt volume. The hidden volume is encrypted by a different key than the volume which contains it. It resides within the illusional random data created by the encryption of the visible volume. The user chooses the volume to mount for use by the encryption key he supplies. If the legal system forces him to disclose his encryption key, he can disclose the key for the encrypted volume alone. The hidden

volume, which will usually contain evidential information, remains oblivious to investigators [24][30].

A case example was a suspected terrorist, whose laptop was turned on and had the TrueCrypt mount window displayed. After initial refusal to disclose the encryption key, a High Court order requested he did. The case was dismissed later on, as no evidence could be gathered by the investigation/prosecution team against him (5). It is suffice to say that the suspected terrorist might have supplied the outer encrypted volume's key rather than the hidden volume's key. The plausible deniability feature of TrueCrypt proves to be a higher defeat to the three get-around digital forensics methods currently being used. Since, the TrueCrypt encryption solution is a widely available free software, experienced cyber-criminals will continue to use the features to avoid being prosecuted.

The following section looks at possible counter-actions to the robustness offered by encryption solutions for disk drives.

V. MUTED SOLUTIONS TO DATA ENCRYPTION THREATS

Reference [8] highlighted two possible solutions to the forensics threats posed by inaccessible information containers. The first was the collaboration of digital forensics experts and storage device manufacturers to develop and implement a standard back-door across all storage devices.

This would seem the perfect solution, but Forte thought it was an excessive ask and "very unlikely to succeed". Three years on, there has not been any such collaboration or development. Forte pointed that the other solution involves a cooperation of the parties involved in an incident towards the adjustment of evidence preservation and analysis methodologies. He however admitted that the latter is only possible in e-discovery processes where both parties may be willing to cooperate rather than cyber-crime incidents whose subjects are often reluctant. Undecided about what solution was more feasible, even though the former seemed more effective, [8] believed there was enough time to find solutions before encrypted storage containers become pervasive, and as sort lightened the push for the former solution.

Reference [24] itself, highlighted that physical or remote (by malware or rootkit) access to such truecrypt-protected computer can compromise the encryption. The access may be used to install keyloggers and memory capturing software or hardware devices. These devices can obtain the encryption keys and passwords, as well as the unencrypted data which could be decrypted using acquired keys. Kleissner's "stoned" bootkit is a particular example of rootkit that can give such access to a target computer. It infects and controls the master boot record of the truecrypt program, and consequently allows the user to bypass the full volume encryption feature of truecrypt to access data resident on the computer [31]. However, the law enforcement agencies would be breaching the privacy rights of such individuals if physical or remote access to their computer is gained without their consent. A situation where law enforcement agencies have legal right to physically and remotely access a target computer is the sex offender monitoring case, in which the offender has actually been convicted of the crime. Unfortunately, a yet-to-be-

convicted suspect that has actually committed a cybercrime will ordinarily not give such approval. He may also protect himself with the self-incrimination legal clause. Yet again, privacy needs have rendered another potentially-viable solution illegal.

VI. CONCLUSION

The effectiveness of data encryption as a mechanism for enforcing information privacy is massive. This is evident by the reported widespread use of various data encryption solutions at the organizational and individual levels. However, its huge success for data access restriction has been a threat for digital forensics processes over the years. Cyber-criminals have been exploiting the information confidentiality ability of data encryption solutions, to restrict digital forensics investigators' accesses to potential evidence. The ubiquitous availability, inexpensive cost and easy implementation of encryption solutions enhance the threats posed to digital forensics processes. Investigators sometimes get around the encryption challenge through careful and thoughtful planning of search and seizure, thorough search for exposed encryption keys, and advanced in-memory data retrieval techniques. Yet, a minimum of 60% of computer incidents involving data-encryption end up unprosecutable.

The TrueCrypt software went even further by providing users with plausible deniability and non-repudiation abilities. This makes digital forensics investigations of encrypted disk drives harder and less feasible. Consequently, this undesired situation constitutes an indirect reason for the rise in occurrence of computer incidents. As much as data encryption helps offenders get away from being caught, the necessity for data privacy and security cannot be sacrificed for digital forensics. Unfortunately, the only digital forensics solution to a threatening information security solution will have to be unanimously considered by disk drive manufacturers. There should be a technology that will provide a backdoor for digital forensics investigators to gain access to the most securely encrypted disk drives. However, there will have to be a restriction to the distribution of such technology when it comes to existence. This is to avoid its abuse by non-law enforcement practitioners (and potential computer criminals) to illegally access target data.

ACKNOWLEDGMENT

The encouragement received from Dr. Shao Ying Zhu was quite appreciated by the author.

REFERENCES

- [1] M. Sheetz, (2007). Computer Forensics: An Essential Guide for Accountants, Lawyers and Managers. Florida: John Wiley & Sons.
- [2] M. Whitman and H. Mattord, (2012). Principles of Information Security 4th ed. Boston: Cengage Learning.
- [3] M. Stamp and P. Stavroulakis, (2010). Handbook of Information and Communication Security. Heidelberg, Berlin: Springer-Verlag.
- [4] Ponemon Institute, LLC., (2010). Annual Study: UK Enterprise Encryption Trends. Available at http://www.symantec.com/content/en/us/about/media/pdfs/Symc_Ponem_on_Encryption_Trends_UK.pdf. Retrieved 2nd February, 2013.
- [5] E. Casey, G. Fellows, M. Geiger and G. Stellatos, (2011). The growing impact of full disk encryption on digital forensics. Digital Forensics. Vol. 8. pp. 129–134.

- [6] K. Getgen, (2009). Encryption and Key Management Industry Benchmark Report. Available at http://beepdf.com/doc/153430/2009_encryption_and_key_management_industry_benchmark_report.html. Retrieved 19th February, 2013.
- [7] E. Casey and G. Stellatos, (2008). The impact of full disk encryption on digital forensics. Digital Forensics. ACM SIGOPS Operating Systems Review. 42(3). pp. 93–98.
- [8] D. Forte, (2009). Do encrypted disks spell the end of forensics? Computer Fraud and Security. 2009(2). pp. 18-20.
- [9] D. Behr, (2008). Anti-Forensics: What it is, What it Does and Why You Need to Know. New Jersey Lawyer. Issue No. 255. Available at <http://www.njsba.com/images/content/1/0/1002013/Dec2008.pdf#page=4>. Retrieved 20th February, 2013.
- [10] J. Leyden, (2008). Fall of an Opportunist. The Economist. http://www.economist.com/world/americas/displaystory.cfm?story_id=1272516. Retrieved 16th February, 2013.
- [11] J. Leyden, (2010). Brazilian banker's crypto baffles FBI. The Register. http://www.theregister.co.uk/2010/06/28/brazil_banker_crypto_lock_out/. Retrieved 14th February, 2013.
- [12] J. Anastasi, (2003). The New Forensics: Investigating Corporate Fraud and the Theft of Intellectual Property. Hoboken, New Jersey:John Wiley & Sons.
- [13] L. Hildner, (2006). Defusing the Threat of RFID: Protecting Consumer Privacy through Technology-Specific Legislation at the State Level. <http://heinonline.org/HOL/LandingPage?collection=journals&handle=hein.journals/hcrcl41&div=9&id=&page=> Retrieved 3rd March, 2013.
- [14] S. Garfinkel and A. Shelat, (2003). Remembrance of data passed: A study of disk sanitization practices. IEEE. 1(1). pp. 17–27.
- [15] Information Commissioner's Office, (2009). The Guide to Data Protection. Available at http://www.ico.gov.uk/for_organisations/data_protection/~media/documents/library/Data_Protection/Practical_application/the_guide_to_data_protection.ashx Retrieved 2nd March, 2013.
- [16] Data Protection Act (1998). <http://www.legislation.gov.uk/ukpga/1998/29/contents>. Retrieved 4th February, 2013.
- [17] M. Stamp, (2006). Information Security: Principles and Practice. 2nd Edition. Hoboken, New Jersey:John Wiley & Sons.
- [18] S. Bosworth, M. Kabay and E. Whyne, (2009). Computer Security Handbook. 5th Ed. Hoboken, New Jersey:John Wiley & Sons.
- [19] K. Mandia, P. Proise, and M. Pepe, (2003). Incident Response & Computer Forensics. 2nd Ed. Emeryville, California:McGraw-Hill.
- [20] S. Garfinkel, (2010). Digital forensics research: The next 10 years. Digital Investigation vol. 7, pp. S64–S73.
- [21] N. Beebe, (2009). The Good, The Bad, and the Unaddressed. IFIP International Conference of Digital Forensics. pp. 17-36.
- [22] Q. Miao, (2010). Research and Analysis on Encryption Principle of TrueCrypt Software System. 2nd International Conference on Information Science & Engineering. pp. 1409-1412.
- [23] B. Carrier, (2005). File System Forensic Analysis. Upper Saddle River, NJ:Addison-Wesley.
- [24] TrueCrypt Foundation, (2012). TrueCrypt Documentation. Available at <http://www.truecrypt.org/docs/> Retrieved 15th February, 2013.
- [25] E-security Planet, (2012). Buyer's Guide to Full Disk Encryption. Available at <http://www.esecurityplanet.com/mobile-security/buyers-guide-to-full-disk-encryption.html> Retrieved 18th January, 2013.
- [26] R. Snyder, (2006). Some security alternatives for encrypting information on storage devices. Proceedings of the 3rd annual conference on Information Security Curriculum Development. pp. 79–84.
- [27] S. Dean, (2010). On-The-Fly Encryption: A Comparison. Available at <http://otfedb.sdean12.org> Retrieved 16th February, 2013.
- [28] B. Oler and I. Fray, (2007). Deniable file system: Application of Deniable Storage to Protection of Private Keys. International Conference on Computer Information Systems and Industrial Management Applications (CISIM) 2007. pp. 225-229.
- [29] Czeskis, D. Hilaire, K. Koscher, S. Gribble, T. Kohno and B. Schneier, (2008). Defeating Encrypted and Deniable File Systems: TrueCrypt v5.1a and the Case of the Tattling OS and Applications. Available at http://static.usenix.org/events/hotsec08/tech/full_papers/czeskis/czeskis.html/ Retrieved 28th February, 2013.
- [30] S. Garfinkel, (2007). Anti-Forensics: Techniques, Detection and Countermeasures. The 2nd International Conference on i-Warfare and Security. Monterey, CA. Available at simson.net/ref/2007/slides-ICIW.pdf. Retrieved 21st February, 2013.
- [31] P. Kleissner, (2009). Stoned Bootkit White Paper. Black Hat Technical Security Conference. USA. Available at <http://www.blackhat.com/presentations/bh-usa-09/KLEISSNER/BHUSA09-Kleissner-StonedBootkit-PAPER.pdf>. Retrieved 13th February, 2013.

Image Compression Using Real Fourier Transform, Its Wavelet Transform And Hybrid Wavelet With DCT

Dr. H.B.Kekre
Sr. Professor, MPSTME,
Department of Computer
Engineering
NMIMS University
Mumbai, India

Dr. Tanuja Sarode
Associate Professor
Department of Computer
Engineering, TSEC
Mumbai University
India

Prachi Natu
Ph. D. Research Scholar
MPSTME, NMIMS University
Mumbai, India

Abstract—This paper proposes new image compression technique that uses Real Fourier Transform. Discrete Fourier Transform (DFT) contains complex exponentials. It contains both cosine and sine functions. It gives complex values in the output of Fourier Transform. To avoid these complex values in the output, complex terms in Fourier Transform are eliminated. This can be done by using coefficients of Discrete Cosine Transform (DCT) and Discrete Sine Transform (DST). DCT as well as DST are orthogonal even after sampling and both are equivalent to FFT of data sequence of twice the length. DCT uses real and even functions and DST uses real and odd functions which are equivalent to imaginary part in Fourier Transform. Since coefficients of both DCT and DST contain only real values, Fourier Transform obtained using DCT and DST coefficients also contain only real values. This transform called Real Fourier Transform is applied on colour images. RMSE values are computed for column, Row and Full Real Fourier Transform. Wavelet transform of size $N_2 \times N_2$ is generated using $N \times N$ Real Fourier Transform. Also Hybrid Wavelet Transform is generated by combining Real Fourier transform with Discrete Cosine Transform. Performance of these three transforms is compared using RMSE as a performance measure. It has been observed that full hybrid wavelet transform obtained by combining Real Fourier Transform and DCT gives best performance of all. It is compared with DCT Full Wavelet Transform. It beats the performance of Full DCT Wavelet transform. Reconstructed image quality obtained in Real Fourier-DCT Full Hybrid Wavelet Transform is superior to one obtained in DCT, DCT Wavelet and DCT Hybrid Wavelet Transform.

Keywords—Real Fourier Transform; Hybrid Wavelet Transform; DCT

I. INTRODUCTION

Image compression is storing images using lesser number of bits than its original size. Image compression leads to less storage space and less bandwidth for transmission. Hence in this world of internet and multimedia applications image compression is of utmost important and interesting area to work on. It is used to store images in medical image database, to generate image database in biometrics and many other applications. Image compression is divided into two categories: lossy and lossless [1]. Compression ratio and image quality of decompressed image, these are two major things to be considered in image compression. As compression ratio increases, quality of reconstructed image starts degrading. Many compression techniques like vector quantization,

predictive coding, differential image coding, transform coding have been introduced. Transform based techniques are popular for image compression especially at low bit rate. In transform domain, many researchers have worked on image compression and still this area is equally popular to work on. Discrete Cosine Transform is widely used. It separates an image into different frequency components. Low frequency components are located at top left corner giving high energy compaction. High frequencies are located in bottom right corner. Elimination of these high frequency elements gives transformed image with few low frequency components. If image is reconstructed from such lesser number of transformed, low frequency elements, it gives compressed image without losing much data contents in original image. Wavelet transform coding is preferred over simple orthogonal transform in image compression due to its multi-resolution property. It provides enhanced image quality even at higher compression ratios [2]. Recently hybrid transformation techniques have come into picture which combines properties of two different transforms [3]. It gives compressed image with visually perceptible image quality. This paper focuses on Real Fourier Transform which is an orthogonal transform, obtained by combining cosine and sine coefficients of Fourier transform. Wavelet transform is generated from orthogonal Real Fourier Transform by using the algorithm in [4] and also hybrid wavelet is generated by following the procedure in [5] and compression of image is studied using all these three transforms. Remaining paper is organised as follows: Section II contains brief introduction to related work. Section III discusses proposed technique. In proposed technique, orthogonal Real Fourier Transform, Real Fourier Wavelet Transform and Real Fourier Hybrid Wavelet Transform are discussed. In section IV results obtained from experimental work are discussed and finally conclusion is drawn in section V.

II. RELATED WORK

A lot of work has been done on image compression and still it is going on. Image compression using biorthogonal wavelet transform is proposed by liu in [6]. A lifting scheme wavelet based transform with a modified entropy coding algorithm is proposed in [7]. It discusses effect of block sub-band coding on compression factor and quality of an image. Hybrid DCT-VQ approach is presented in [8] where high frequency elements are first removed by applying 2D-DCT on colour image and then VQ algorithm is applied on this reduced 2D vector to speed up the process. Use of wavelet transform started with Haar

transform. In [9] Adman Khashman et.al. Proposed an image compression technique using neural network and Haar wavelet transform. Their paper used Haar wavelet compression with nine compression ratios and a supervised neural network that learns to correlate the gray image intensity with a single optimum compression ratio. Two neural networks getting different input image sizes are implemented in this technique and a comparison between their performances was presented. A combined approach of image compression based on vector quantization and wavelet transform is proposed using RBF neural network in [10] which works on grayscale images.

III. PROPOSED TECHNIQUE

This paper proposes a new transform called real Fourier transform. We know that Fourier transform is actually implemented using complex numbers, where the real part is weight of cosine and imaginary part is weight of sine. DCT is similar to Fourier Transform but uses only cosine functions i.e. using only real and even functions. Discrete Sine Transform (DST) uses only sine functions i.e. real and odd functions. DCT and DST contain only real data sequence. Both DCT and DST can be computed by using two FFT's of original data sequence of length N. i.e. they are equivalent to FFT of twice the length[11]. To avoid complex values in the output like Discrete Fourier transform, combination of DCT and DST coefficients are considered to form transformation matrix. It gives Real Fourier Transform. Thus transform matrix contains only real values and no complex exponentials. Using this transform matrix column transform, row transform and full transform of image is obtained [12]. In column transformed image, energy compaction occurs towards upper portion of the transformed image. In row transformed image energy compaction takes place towards left portion of an image whereas in full transformed image high energy coefficients are present at top left corner. After applying transform, low energy coefficients are eliminated to represent the image in lesser number of bits. Hence in column transform we eliminate rows in lower portion and in row transform, columns at the right side of transformed image are eliminated. In full transform low energy coefficients in both, rows and columns are eliminated. Same numbers of coefficients are eliminated in above three cases and results are compared. Steps to apply transform on an image are as follows:

A. Column, Row and Full Transform

- Separate R, G, and B plane of an image.
- Generate Real Fourier Transform matrix 'T' of size 256x256, since image size is 256x256.
- Apply transform on each plane of image separately.
- For column transform use $F = [T] * [f]$, where f is original image and F is transformed image.
For row transform use $F = [f] * [T]^T$ where $[T]^T$ is transpose of transform matrix.
For full transform use $F = [T] * [f] * [T]^T$

- From column/row/full transformed image eliminate specific number of low energy coefficients. It is done by eliminating columns/rows/combination of columns and rows respectively. Apply inverse transform to reconstruct the image from compressed image.
- Calculate Root mean Square Error between original image and reconstructed image.
- Compare RMSE values in column/ row/ full transform to compare their performance

B. Real Fourier Wavelet Transform

Use of wavelets is beneficial over simple orthogonal transform because it concentrates on local characteristics of image. Using algorithm in [5], $N^2 \times N^2$ Real Fourier Wavelet Transform matrix is generated from $N \times N$ Real Fourier Transform matrix. Using the steps 1 to 8 in case (A) of proposed technique, column, Row and full Real Fourier Wavelet Transform of image is obtained and performance of each is compared.

C. Real Fourier Hybrid Wavelet Transform

Orthogonal transforms are used to analyse global properties. Each transform has its own characteristics. In wavelet transform, some orthogonal transform focuses on global properties of data whereas some exhibit local properties in better manner. Hybrid wavelet combines properties of two different orthogonal transforms giving strengths of both transforms. Here two different transforms selected are: Real Fourier Transform and Discrete Cosine Transform. Following the methodology in [6], Real Fourier-Cosine Hybrid wavelet Transform matrix is generated. 8x8 Real Fourier transform matrix and 32x32 size Cosine matrix is used to generate 256x256 size Real Fourier-Cosine Hybrid wavelet Transform matrix. Column, Row and Full Hybrid wavelet Transform of images is computed using steps 1 to 4 in section (A). In hybrid wavelet transform, energy of each row/ column is calculated and rows/columns are sorted in descending order of their energy. Rows/columns having lowest energy are then eliminated to compress the image, i.e. now steps 5 to 8 in section (A) are used to compute RMSE. Results in each case are computed using RMSE as a measuring parameter. Also, the performance of DCT column transform[12], DCT column wavelet transform[13] with different sizes of two Cosine Transform matrices, Real Fourier Column transform, Real Fourier column wavelet transform and Real Fourier-Cosine column hybrid wavelet transform is compared.

Similarly, DCT row transform, DCT row wavelet transform with different matrix sizes, Real Fourier row transform, Real Fourier row wavelet transform and Real Fourier-Cosine row hybrid wavelet transform are compared.

All these six cases are considered for full transform and compared.

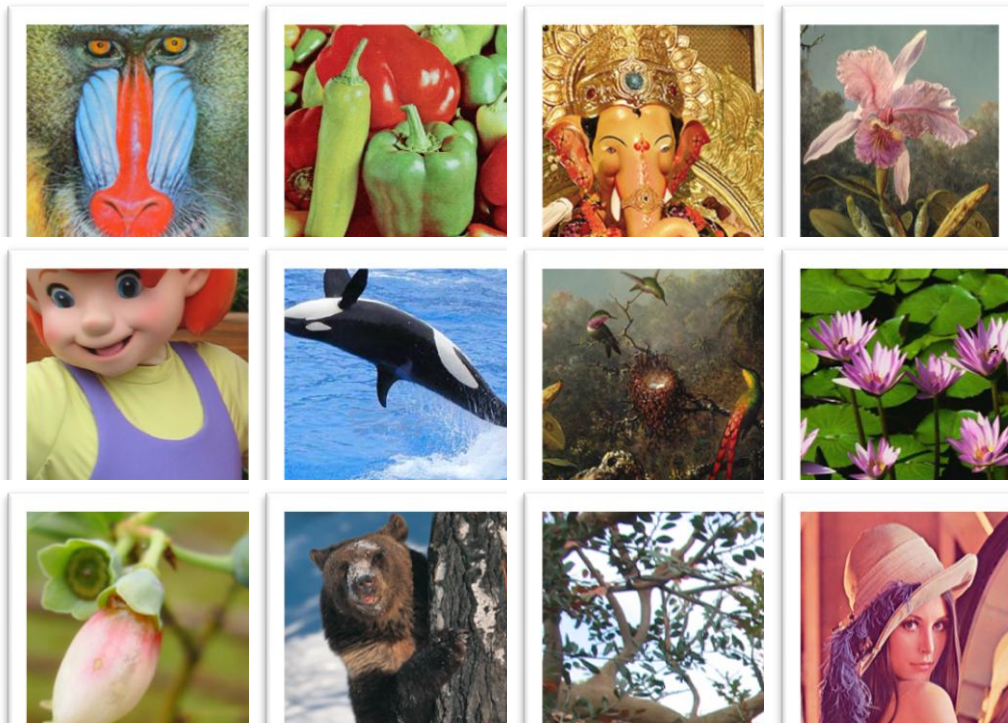


Fig.1. Set of twelve test images of different classes used for experimental purpose namely (from left to right and top to bottom) Mandrill, Peppers, Lord Ganesha, Flower, Cartoon, dolphin, Birds, Waterlily, Bud, Bear, Leaves and Lenna

IV. RESULTS AND DISCUSSIONS

Proposed technique is applied on set of colour images. Twelve different colour images are used for experimental purpose. Each image is of size 256x256. Matlab 7.2 on AMD dual core processor is used for experimentation. Fig.1 shows sample colour images used for experimental purpose.

Graph in Fig.2 compares the results of Column, Row and Full Real Fourier Transform. It has been observed that approximately same results are obtained for Column and Row transform. When more numbers of coefficients are eliminated i.e. higher compression ratio, there is considerable difference between RMSE values of Full transform, column transform and row transforms. Elimination of 240 rows from 256 rows of transformed image gives compression ratio 16. At this compression ratio, RMSE value is 19.734 in Full transform. In column and row transform it is nearly 25.

Fig. 3 shows graph of RMSE values for Real Fourier Wavelet transform with three different cases: Column wavelet, Row wavelet and Full wavelet transform. Here 256x256 size transform matrix is obtained from 16x16 size Real Fourier Transform matrix.

From figure, it can be observed that in Real Fourier Full Wavelet Transform, at higher compression ratio, RMSE values are decreased by more than half as compared to column wavelet and row wavelet transform. It means image quality of reconstructed image obtained by Real Fourier Full Wavelet Transform is much better than obtained in Column Wavelet and Row Wavelet Transform.

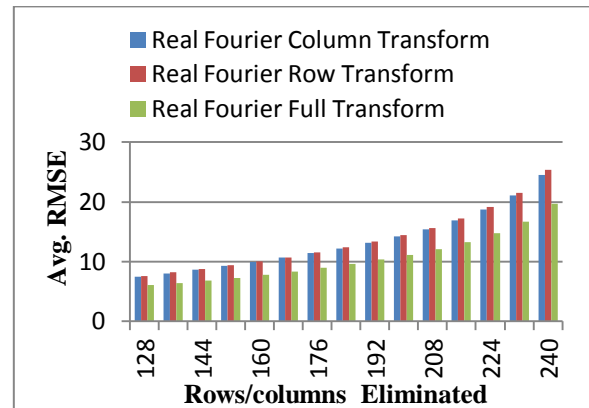


Fig.2. Comparison of RMSE in Real Fourier Column Transform, Real Fourier row Transform and Real Fourier Full Transform

Fig. 4 compares the performance of Real Fourier Hybrid Wavelet Transform. Here two different transforms are used to generate hybrid transform matrix. Hybrid Transform matrix size is same as that of image size. First, Real Fourier Transform of 8x8 is considered and second Discrete Cosine Transform of 32x32 is selected 256x256 Hybrid Transform matrix is generated using algorithm in [6]. Computed RMSE values for Column, Row and Full Real Fourier-Cosine Hybrid Transform are compared in following figure. This graph clearly indicates that, RMSE values are reduced to one third at higher compression ratios when Full Hybrid wavelet Transform is applied on an image. It assures better image quality even when more coefficients are removed from transformed image.

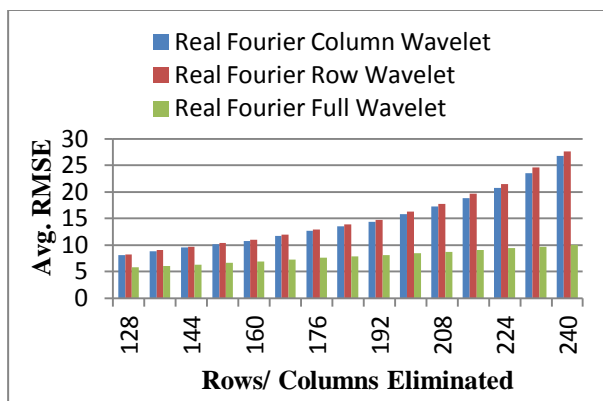


Fig.3. Comparison of RMSE in Real Fourier Column Wavelet Transform, Real Fourier Row Wavelet Transform and Real Fourier Full Wavelet Transform

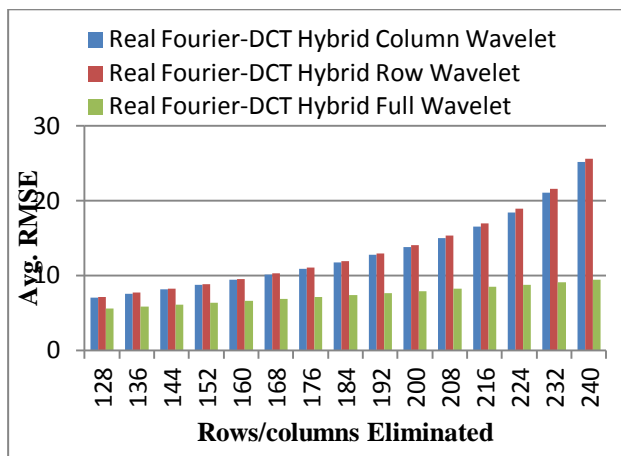


Fig.4. Comparison of RMSE in Real Fourier Hybrid Column Wavelet Transform, Real Fourier Hybrid Row Wavelet Transform and Real Fourier Hybrid Full Wavelet Transform

Fig. 5 compares RMSE values for column transforms of DCT, Real Fourier Transform and their combination. When we keep on removing more number of rows, DCT column wavelet transform (with $m=8$, $n=32$) and Real Fourier-DCT column Hybrid Wavelet Transform gives same results. It is because Cosine matrix of 32×32 is used in both cases to extract local properties of image. These results are slightly better as compared to DCT column, DCT column wavelet, Real Fourier column and Real Fourier column wavelet transform. Elimination of 240 rows from 256 rows gives compression ratio 16. At this high compression ratio DCT column Transform gives slight better performance.

In Fig. 6, performance for different cases of row transform is shown. It is nearly same as different cases of column transform. As we eliminate more number of columns, RMSE increases. Comparison of Full Transforms of Cosine, Real Fourier and their combination is given in Fig.7. From figure; it has been observed that Full Hybrid Wavelet Transform gives better performance than orthogonal Full Wavelet Transform. Among DCT Full and Real Fourier-DCT Full Hybrid Transform, second combination gives better results.

Fig. 7 shows results of Real Fourier- DCT Full Wavelet Transform. It gives the best results among all. Even after eliminating 240 rows, RMSE value of 9.4 is obtained. It indicates that perceptible image quality is obtained using this transform by retaining small amount of data.

Fig. 8 shows reconstructed images using different six cases of Column transform mentioned in Fig. 5. Well known ‘pepper’ image is selected to demonstrate the quality of reconstructed image. Fig. 9 shows reconstructed images for different Row transforms. For Full transforms, reconstructed images are shown in fig. 10. Comparison of Fig. 8,9,10 clearly indicates that Real Fourier-DCT Full hybrid Wavelet Transform gives best quality of reconstructed image among all.

V. CONCLUSION

In this paper Real Fourier Transform is studied. It is applied on the image in three ways: Column transform, Row transform and Full transform. Performance of these three is compared and it is found that Full transform gives better results than row and column Real Fourier transform. From $N \times N$ Real Fourier transform wavelet of $N^2 \times N^2$ is generated. In this case also Real Fourier Full Wavelet transform gives better compression than Column and Row wavelet Transform. Among Real Fourier Full wavelet Transform and Real Fourier Full Transform, Full Wavelet transform gives much better results when more number of coefficients are eliminated to get high compression ratio.

Also Real Fourier-DCT Hybrid Wavelet Transform matrix is generated and results for column, row and full hybrid transform are considered. Full hybrid wavelet transform gives better results than full transform as well as full wavelet transform. All results of Real Fourier Transform are also compared with and results of Cosine Transform, Cosine Wavelet transform and Cosine-Cosine Hybrid Wavelet Transform. It has been observed that, Real Fourier-Cosine Full Hybrid Wavelet Transform outperforms all other transforms giving lowest value of RMSE 9.407 when 240 rows from 256 rows of transformed image are eliminated. It gives perceptible image quality at compression ratio 16.

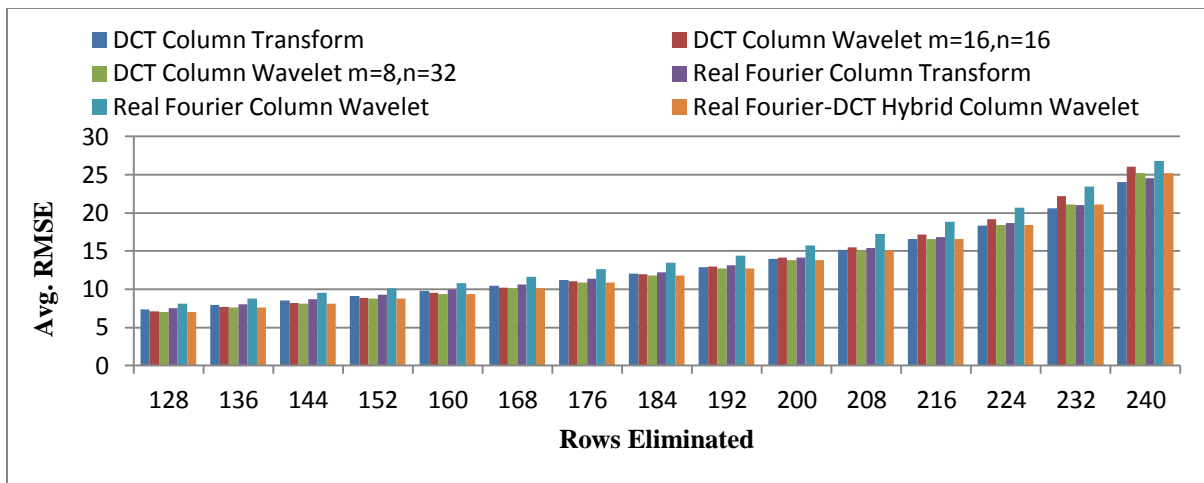


Fig.5. Comparison of RMSE for different cases of Column Transforms of DCT and Real Fourier Transform: DCT Column Transform, DCT Column Wavelet Transform (m=16, n=16), DCT Column Wavelet Transform (m=8, n=32), Real Fourier Column Transform, Real Fourier Column Wavelet Transform Real Fourier-DCT Hybrid Column Wavelet Transform (m=8, n=32).

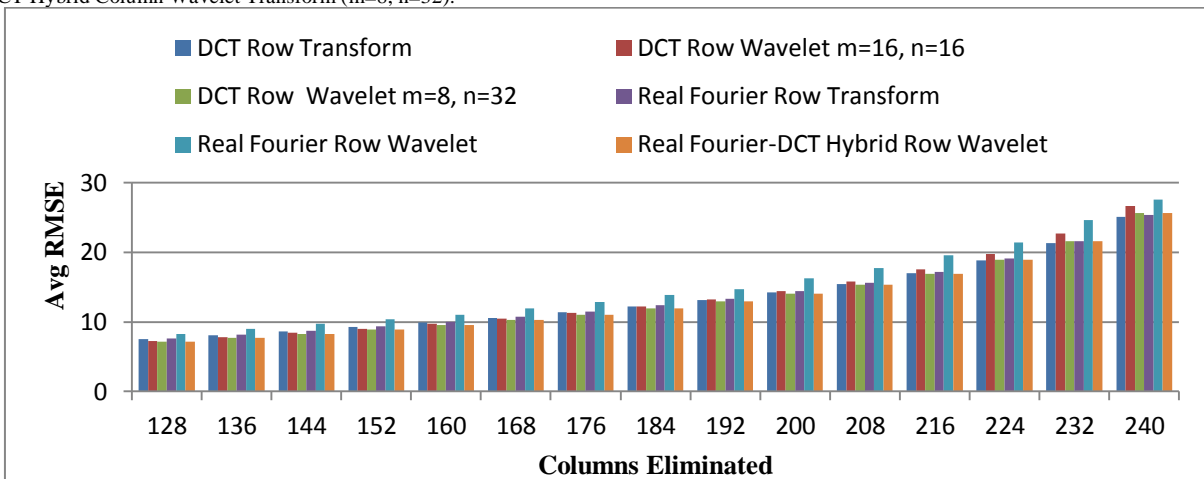


Fig.6. Comparison of RMSE for different cases of Row Transforms of DCT and Real Fourier Transform: DCT Row Transform, DCT Row Wavelet Transform (m=16, n=16), DCT Row Wavelet Transform (m=8, n=32), Real Fourier Row Transform, Real Fourier Row Wavelet Transform Real Fourier-DCT Hybrid Row Wavelet Transform

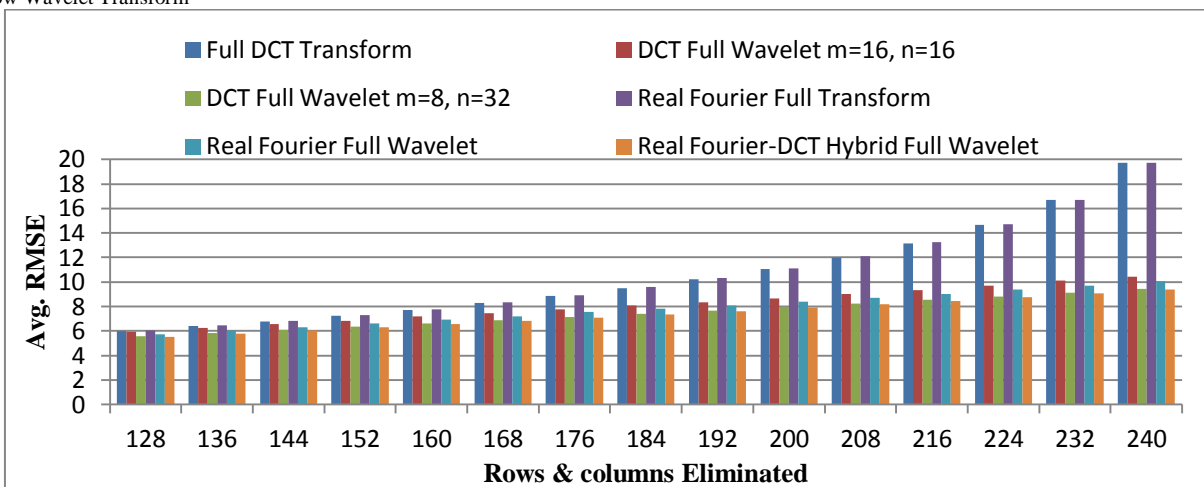


Fig.7. Comparison of RMSE for different cases of Full Transforms of DCT and Real Fourier Transform: DCT Full Transform, DCT Full Wavelet Transform (m=16, n=16), DCT Full Wavelet Transform (m=8, n=32), Real Fourier Full Transform, Real Fourier Full Wavelet Transform, Real Fourier-DCT Hybrid Full Wavelet Transform

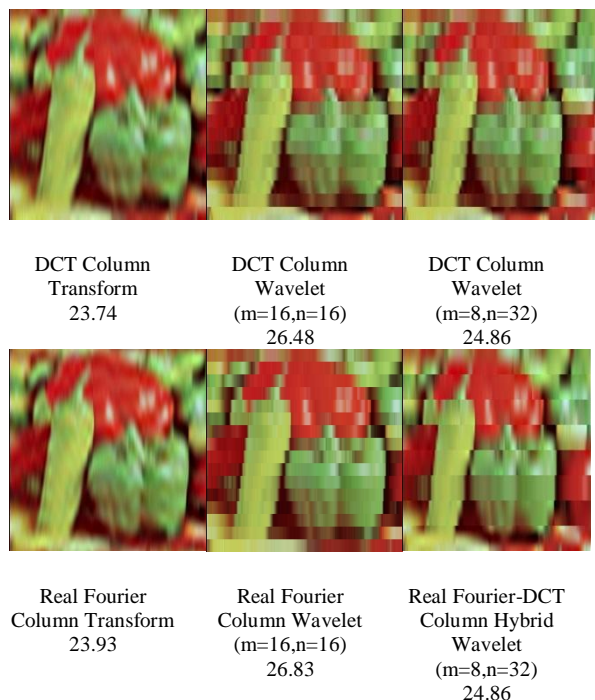


Fig.8. Reconstructed images with RMSE values using six different cases of column transform when 240 rows of 256x256 images are eliminated

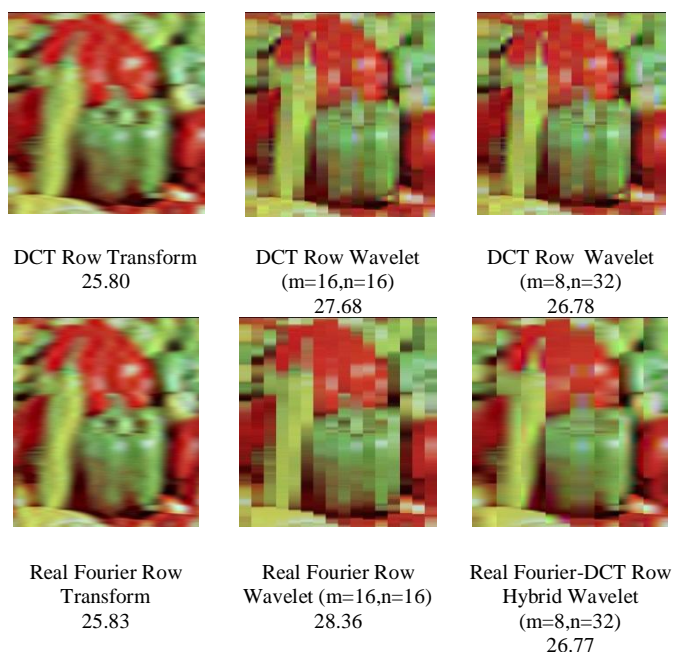


Fig.9. Reconstructed images with RMSE values using six different cases of Row transform when 240 columns of 256x256 images are eliminated

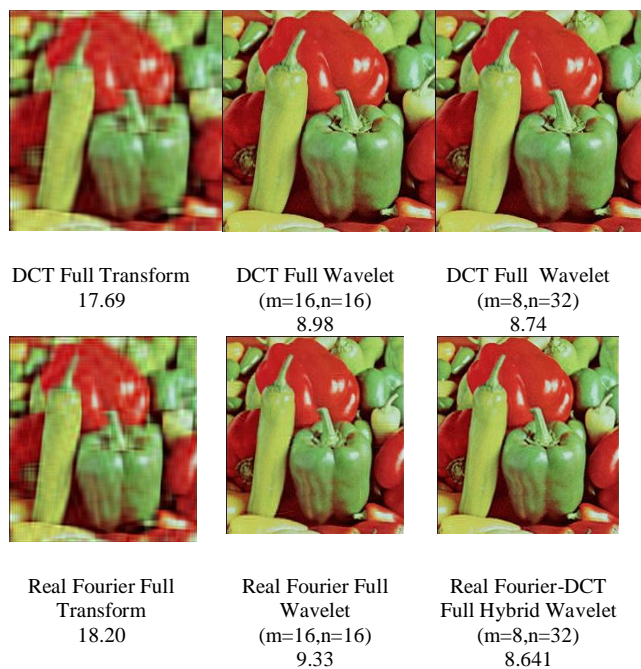


Fig.10. Reconstructed images with RMSE values using six different cases of full transform when 240 rows/columns of 256x256 images are eliminated

REFERENCES

- [1] Prabhakar.Telagarapu, V.Jagan Naveen, A.Lakshmi Prasanthi, G.Vijaya Santhi, "Image Compression using DCT and wavelet ansformation", IJSPIPPR, vol 4, issue 3, pp. 61-74, Sept 2011.
- [2] M. J. Nadenau, J. Reichel, and M. Kunt, "Wavelet Based Color Image Compression: Exploiting the Contrast Sensitivity Function," IEEE Transactions Image Processing, Vol.12, no.1, pp. 58-70, 2003.
- [3] Rachana Dhannawat, Tanuja Sarode, H. B. Kekre, "Kekre's Hybrid Wavelet TransformTechnique with DCT, Walsh, Hartley and Kekre's Transform for Image Fusion", International Journal of Computer Engineering and Technology(IJCET), Vol. 4, Issue 1, Feb 2013, pp. 195-202.
- [4] H.B.Kekre, Tanuja Sarode, sudeep Thepade, Sonal Shroff, " Instigation of Orthogonal Wavelet Transforms using Walsh, Cosine, Hartley, KekreTransforms and their use in Image Compression", International Journal of Computer Science and Information Security (IJCSIS), Vol 9, No. 6, pp. 125-133, 2011.
- [5] H.B.Kekre, Tanuja Sarode, Sudeep Thepade, "Inception of hybrid wavelet TransformUsing Two Orthogonal Transforms and its use for Image compression", IJCSIS, vol 9, no. 6, 2011.
- [6] Hong Liu, Lin Pai Zai, Ying Gao, Wen Ming Li, Jiu-Fei Zohu, "Image compressionBased on Biorthogonal Wavelet Transform", in Proc of ISCT, 2005.
- [7] Loay A. George, Aree A. Mhammad, "Intra Frame Compression Using Lifting Scheme Wavelet-Based Transformation (9/7-Tap Cdf Filter)", International Conference on Multimedia Systems And Applications, MSA 2007, June 25-28, 2007, Las Vegas Nevada, USA.
- [8] Arup Kumar Pal, G.P. Biswas, S. Mukhopadhyay, "A Hybrid DCT-VQ Based Approach for Efficient compression of Colour Images", International Conference on Computer and Communication Technology, (ICCT), 2010, pp. 177-181.
- [9] Adnan Khashman, and Kamildimililer, "Image Compression using Neural Networks and Haar Wavelet," WSEA Transactions on Image Processing, Vol. 4, no. 5, pp. 330- 339, 2008.
- [10] G. Boopathi, Dr. S. Arockiasamy, "Image Compression: Wavelet Transform using Radial Basis function (RBF) Neural Network", IEEE Transactions, 2012, pp. 340-344.

- [11] H. B. Kekre, J. K. Solanki, "Comparative Performance of Various Trigonometric Transforms For Transform Image Coding" *International Journal of Electronics*, Vol.44, Issue 3, pp.305- 315.
- [12] H.B.Kekre, Tanuja Sarode, Prachi Natu, "Efficient Image Compression Technique Using Full, Column and Row Transforms on Colour Image", *International Journal of Advances in Engineering and Technology*, Vol.6, Issue 1, March 2013, pp. 88-100.
- [13] H. B. Kekre, Tanuja Sarode, Prachi Natu, " Image Compression Using Column, Row and Full Wavelet Transforms Of Walsh, Cosine, Haar, Kekre, Slant and Sine and Their Comparison With Corresponding Orthogonal Transforms", *International Journal of Engineering Research and development (IJERD)*, Vol. 6, Issue 4, March 2013, pp.102-113.

Building Low Cost Cloud Computing Systems

Carlos Antunes

Informatics Engineering Department, School of Technology
and Management, Campus 2
Polytechnic Institute of Leiria
Leiria, Portugal

Ricardo Vardasca

Faculty of Engineering / Faculty of Advanced Technology
University of Porto / University of South Wales
Porto, Portugal / Cardiff, United Kingdom

Abstract— The actual models of cloud computing are based in megalomaniac hardware solutions, being its implementation and maintenance unaffordable to the majority of service providers. The use of jail services is an alternative to current models of cloud computing based on virtualization. Models based in utilization of jail environments instead of the used virtualization systems will provide huge gains in terms of optimization of hardware resources at computation level and in terms of storage and energy consumption. In this paper it will be addressed the practical implementation of jail environments in real scenarios, which allows the visualization of areas where its application will be relevant and will make inevitable the redefinition of the models that are currently defined for cloud computing. In addition it will bring new opportunities in the development of support features for jail environments in the majority of operating systems.

Keywords—cloud computing; IAAS; jail environments; optimization; PAAS.

I. INTRODUCTION

The concept of cloud computing is intrinsic to the provision of services and resources over the internet facilitating access to customers regardless the location of where they are and at a lower cost than if they had these resources and services locally[1].

High availability, high performance and load balancing are terms that become more common in daily practice. Any organization, independently of its size, is dependent of an IT infrastructure and services. In the great majority any failure can mean a loss of business profitability[1].

The last decades have strengthened the notion that data processing can be done more efficiently in large computational clusters and storage systems accessible via the Internet[1].

Since then the applicational models of services in cluster present several variants, presenting even these days, as a huge challenge for service delivery in cloud computing[2].

The growth in usage of cloud computing solutions may be seen frenetic, and physical means applied to cloud computing has grown enormously, to the point that only large companies bearing support this type of implementations.

The current common implementation of clouds is through implementing virtual environments, which can be defined as a emulated software independent computer systems, providing the launch of several devices in parallel, as if they were separate physical machines within one physical machine [3]. Other possible implementation is through jails, which can be intended as being the execution of a parallel instance of the

operating system itself, isolating a particular area of files, and separating in its environment the respective users and associated processes [4].

This article aims to present a valid alternative to virtualized solutions, allowing to optimize the performance of cloud solutions in terms of resource consumption, enforcing flexibility that certainly in a future will boost redefinition of the implementation models of cloud computing.

This document is organized as follows: sections were the problems of cloud computing are highlighted; the concepts of virtualization, virtualized systems, jail services and jail environment are presented; the different resources made available are attended with a comparison between implementations with virtual and jail environments; the tested applicational models with suggested solution are defined; and final the conclusions and future research directions are pointed.

II. CLOUD COMPUTING PROBLEMS

There are several attempts to define cloud computing but none of them is consensual, some are very vague and abstract. Even the definition of the National Institute of Standards and Technology (NIST) defines it as an evolutionary paradigm [5]. With the emergence of cloud computing a lot of discussion has been raised about how to define it as a computational model. Models clearly matured have been published and debated and service providers hold a clearly defined model for its products.

Gartner [6] says that tension between the short-term risk and the long-term imperative strategic will define the market development in the next 2 to 3 years.

Thus, it can be considered that the current implementations of cloud computing are a business model based on the distribution of services with high availability with load balancing. Three layers support this model, being the layer being made available to users of complete abstraction of the technologies used. Any technology component associated with this type of solution is then hidden, placed in the protocol and infrastructure layers designated protocol as a service (PAAS) and infrastructure as a service (IAAS) respectively.

Offering IAAS requires platforms that can manage the infrastructure shared and stored dynamically, the current models based on the usage of virtual machines by that reason cannot be the future of cloud computing.

III. MAIN CONCEPTS

A. Virtualization

Virtualization appeared in the 60s, in order to discharge the large hardware solutions that were the *ex-libris* at the time of any institution. "Mainframe war" was a typical term at the time, between different groups of those solutions providers, struggling with who had the most powerful machine. IBM and other competitors as BESM and Strela, roamed the markets in search for dominance [3].

Nowadays, the computational capabilities of any common computer, face the same problems of rigidity and underutilization of the 60's mainframes. Achieving the total advantage of the full capacity provided by the hardware is an holistic view, being impossible to obtain with a single system [3].

Since the 90's, several software providers, in particular VMware has developed virtualization solutions for high performance, supporting several platforms [3].

The emerging of large-scale virtualization solutions has overshadowed the development of parallel solutions, which mostly ends up not moving from prototypes. Therefore, it is pertinent to make some reflections on virtualization alternatives, and the reasons why they are forgotten.

With the need of high availability required by the increasing deployment of cloud solutions, it is pertinent to assess whether commercial solutions on the market go against the full needs of customers.

B. Virtualized systems

Virtualized systems are attractive, versatile in some contexts, offering tremendous flexibility of configurations and solutions. However, there are well-known problems which are reflected in the performance and reliability of most of the implemented solutions [3].

Virtualization is not a free concept. Consumption of resources and all the mechanisms for managing them by the provider of virtualized environments is inevitable. Joining these consumed resources with the need of resources required by the operating system for supporting the deployed applications, for most implementations is a difficult scenario to support and the performance of existing solutions will be affected.

Moreover, the concept of independent operating systems, levels of management and maintenance is difficult to obtain, and is often a barrier to the concept of IAAS, which requires isolation.

C. Jail Services

The jail services base their operation on a shared kernel that generates process isolation through concurrent access to devices and resources of the machine, avoiding any kind of virtualization [7].

This concept has originated with the need to implement security requirements such as integrity, privacy and data protection, as well as the need to delegate management functions or administration of certain areas/services implemented [7].

Jail services arised in 1982 by Mr. Bill Joy, with the implementation of the chroot functionality. This solution was developed for open environments and implemented on the concept of virtualization based in processes [4].

The chroot implements jail services in an isolation concept of file structure associated with certain processes that are running on the machine, allowing interaction with other system processes [8].

However the implementation of chroot has some weaknesses, although the visibility is limited to finding its single file system subtree, and the concept of isolation does not extend beyond the file structure, sharing all other elements such as users, and other network devices.

D. Jail Environment

Aiming to fill the gaps in the mechanism chroot, the jail environments arise, which are available from the FreeBSD operating system version 4.2. This new feature is a mix between jail services and virtualized environment.

Poul-HenningKamp, father of the jail environment, based his development on the absorption of the advantages of the two previously mentioned mechanisms, creating a service trapped with insulation concepts of the directory structure, and at the same time doubling the structure of the base operating system, creating a different concept for virtualization and network services [7].

The Jails were developed based on concepts from chroot, analogously implementing the structure of independent directory features but adding insulation at various levels. Each environment jail owns and operates autonomously their services and local settings, regardless of the host where it is implemented.

Oppositely to what may be thought, in terms of administration, the jails do not increase the weight in the administration and management of the system administrator, since despite having local security policies, these can be imported security policies already implemented in the base operating system. On the other hand, maintaining the concept of service imprisoned, administration and management of jails can be made from the base machine, without for that being necessary to access the jail environments.

Two requirements need to be attended when considering jails:

- The mechanism of discretionary access control to retain any service;
- To each jail is allowed to have a superuser which can locally manage files and processes within the environment.

Each jail has an associated address and a file system independent. The child processes generated in jail environments inherit its structure, lying immediately limited to the environment where they are generated.

To define constraints for each environment the main system administrator can use a configuration set (flags) and control variables (sysctl), which allow changing the working format,

access permissions and security associated with each environment. However, there are options that require careful planning in order to prevent access to administrative functions, such as, among others, the ability to load or disable devices, or interfere with the firewall rules. It is clear that, although the flexibility of the jail environments, there are limitations on access to certain features which must be guaranteed to be secured solution.

The Jail environments are defined by 4 elements:

- A root directory - the starting point from which a jail is set;
- A hostname - the hostname that will be used within the jail;
- An IP address - that is assigned to the jail and cannot be altered in any way after setting up the same;
- A command - an executable that allows to provide the jail environment.

There are situations where the usage of imprisoned environments present great asset, especially when referring to the optimization of resources, which quickly translate into solutions for more profitable investments due to the saving in computation load, storage and energy consumption.

IV. RESOURCES AVAILABLE

The mad rush of the leading telecommunications operators to build megalomaniac infrastructures with a view to supporting the solutions marketed has accentuated in recent years, pushed with easy access to the latest hardware at attractive prices, it can become an illusion, and at long term can lead to serious problems of maintenance and administration.

According to David Strauss the models used in current implementations of cloud solutions are outdated, and there is a need to evolve the models to resource optimization, and improved performance of the proposed solutions. The resources required and expended to implement reliable models are huge, and there is a realistic need to reduce that [4].

Since its conception, that the consumption of resources, was always a point to consider in the development of high-availability solutions. Having regard to the use of jails passes to manage processes rather than independent virtual machines, will occur significant differences with regard to this point.

System Virtualization - it is a mandatory feature when it comes to virtual machines, non-existent when it comes to solutions based on jails. However, it may resemble the base operating system (FreeBSD), which is a mandatory requirement for the provision of the jail environment. However the differences are huge, can immediately be noted that the dependence of virtualized systems from the hardware emulated software performance as well as the use of additional drivers, causes bottlenecks in the access to real devices. Meanwhile the use of jail environments in concurrent access to the existing devices on the machine where they are configured prevents the existence of any intermediary in accessing them.

Operating System - this must be one of the highest points earned with systems based in jails, because the concept of

using the jail environment is the sharing of the kernel resources from the base machine, removing the need for additional resources related to the implementation of a new operating system instance. Thus there has been a saving of significant resources such as: storage, memory and processing load.

File structure - another relevant point is the space required by the file structure of a jail, which is relatively lower than any real operating system installation, as a simplified copy of the basic structure of a system based on FreeBSD, requiring less storage resources

Flexibility - The ability to adapt to different situations is important when implementing a solution, and as the virtualized solutions, jail environments provide enormous flexibility, increasing some important features such as the ability to share binary data sharing location via mount points.

The usage of virtual machines and jails environments for virtualization has huge similarities in concepts of isolation, but nevertheless profound differences to the operational level [5]. For better understanding, an analogy to the processes and threads running on an existing system is presented. A virtualized environment can be compared to a process, being holder of a wide range of information, but heavy when it is released, at the beginning, at the end and when restored, as well as the amount resources expended. Threads can be compared to jail environments, which are lighter, more flexible and have a faster execution.

Briefly it can be noted that the jail environments are lighter, based from the outset on the fact of sharing the kernel with the base machine, excluding the need of installing a new operating system, having a simplified file structure, with the possibility of sharing binary files, allowing the release of several isolated instances of the same service, such as the launch and the end of such processes at times almost imperceptible and insignificant consumption [5].

Table I presents a comparative summary of the requirements and implementation solutions required by virtual environments.

TABLE I. REQUIREMENTS COMPARISON OF VIRTUAL ENVIRONMENTS

Resources	Virtual machines	Jail environment
System virtualization	Yes	Operating system base
Operating system	Yes	No
Files structures	Yes	Simplified
Virtualized drivers	Yes	No
Data sharing	Network	Network or local
Starting time	Operating system startup	Minimal (> 3s)

The consumption of resources continues to be a constant point of discussion within the development teams of the current models, and there is a considerable increase in resources when the start of virtualized environments for supporting cloud computing, causing temporarily instability in the support system, unlike the jail environments, that while having a minimum starting duration not exceeding two seconds and a

residual resource consumption. In Table II is shown the resources consumption at the system startup. The use of jail environment increments important advantages in terms of time and disturbance in the base system.

TABLE II. RESOURCE CONSUMPTION AT STARTING BASE SYSTEMS

N° of Machines	Virtual machines			Jail environment		
	CPU	RAM	TIME	CPU	RAM	TIME
1	~20%	~50MB	~30s	~5%	~1MB	~2s
2	~20%	~50MB	~30s	~5%	~1MB	~2s
3	~20%	~50MB	~30s	~5%	~1MB	~2s

A further study is being conducted, and the results and analysis will be presented in the future in another publication.

Note that when the startup of virtualized environments, the resource consumption increases exponentially and may temporarily cause instability in the support system, unlike that jail environments, while having a minimum starting duration (not exceeding two seconds), has a consumption resource imperceptible. With the use of environments imprisoned instead of virtualized environments, it is obtained a drastic reduction in both storage space and in terms of computational resources required, which could reach values close to 50%.

V. APPLICATION MODELS

The Resulting of its adaptability, jail environments are liable to be used in virtually any situation. Fruit of experience in handling some solutions based on jails, and observing the most common types of implementation, there are presented the three groups that tended to be the most advantageous application of such solutions. These models are based on computational models used in the distribution of virtualized services, and applicational models for high availability solutions.

A. Per Service

One of the critical points in systems management is the need of logical and physical structure organization of services to implement, and the related decentralization of management and administrative, presenting itself as a concern the necessity of the existence of multiple users with high credentials for access to the base system, making it a real jigsaw at the implementation level of the safety policies.

Using imprisoned environments, it will facilitate this process by providing the isolation process and responsibilities, allowing to define restricted areas of access, as well as limitations of services and features, providing improvements substances in the implementation of security policy (fig. 1).

These models are conceived for large and medium companies, which, holding a considerable amount of services, intend to decentralize the delegation of responsibilities by implementing improvements in safety of the available systems.



Fig. 1. Implementation of jail environments per service.

B. Per User

Similarly to the distribution service, it is possible to envision a host server environment in which every jail is associated space of each user, allowing the implementation or the respective mapping services and data in isolated and secure environment. This approach allows to implement a higher-level control security features assigned, which translates into a more cost effective solution and affordable than the currently existing solutions, also enabling independent settings tailored to the needs of each client concerned.

This type of solution allows each user to maintain its placeholder, contracted services and respective settings properly insulated, giving you a greater sense of control and freedom over the system that acquired (fig. 2).

It is the ideal solution for many hosting service providers, taking into account the limitations, the level of systems administration, some customers who require proprietary solutions and high flexibility.



Fig. 2. Implementation of jail environments per user.

With this model of service delivery, it is possible to distribute to the user a powerful environment, flexible and insulated, allowing to delegate to him the tasks of monitoring and security management to the administrator of the base system.

C. High availability warranty

Major focus of current solutions include the need to provide high availability of services, and, in most situations, not always present themselves as affordable. The use of jails can make a huge transformation in the concept, allowing the replication of existing jails, or coexisting with other jails environments by configuring sharing binaries or providing access to common areas, across common mount points (fig. 3).

The application of jail environments are susceptible to being applied in all high availability scenarios developed so far, making them a huge competitor to the virtualized environments. The authors of this publication intend to do another on the subject of implementing high availability solutions using jail environments.

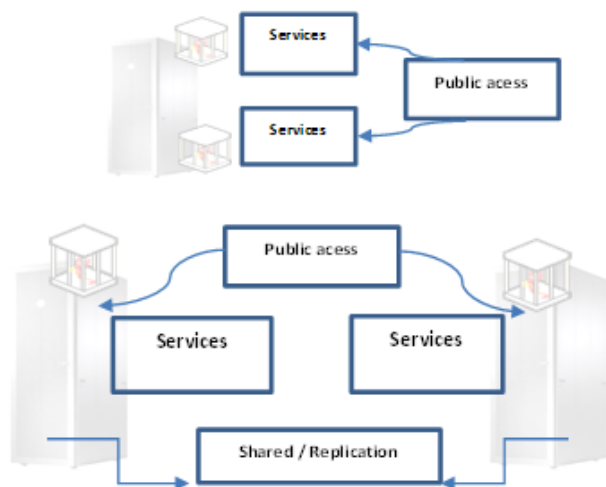


Fig. 3. Implementation of jail environments for high availability warranty.

This can be done through jails being replicated redundantly and using protocols that guarantee high availability, such as Common Address Redundancy Protocol (CARP) or solutions such as Heartbeat, aiming to provide high availability solutions at a reduced price and without huge requirements hardware.

VI. DISCUSSION

The Summarizing it can be noted that all models of cloud, have their basis of a structure already established for some time, but it still presents some gaps, despite the many developments in the methodologies of the cloud.

The current jail environments, yet not been considered as substitutes of virtual systems, but certainly will emerge as a major competitor, and these with regard to the optimization of new models.

Carefully analyzing the basic concepts of the cloud model and bearing in mind the features provided by jail environments, it can be believed that a new conception of architecture will be developed, being based on the use of jail environments and taking into consideration its advantages. That matches the

required standards of cloud solutions, specifically in the IAAS layer.

There is a vast area in this field of work and development that is still unexplored, as it happened with virtualization, enhancing their use as there appear new developed applicational models, because its concepts fit perfectly with the new trends of the models the cloud.

Similar evolution happened in distributed computing between processes and threads, with time threads based implementations proved to be more effective than process based ones and for high performance systems replaced them.

As future research direction is proposed the optimization of resources in the IAAS layer, since it is an area that requires immediate analysis, there much work ahead in defining new models and applicational scenarios to relief the maintenance of cloud computing solutions. It is also proposed a study on the feasibility of high availability models with a view of supporting the 2 top layers of cloud solution model implementing a jail environment.

REFERENCES

- [1] R. Buyya, C. S. Yeo, S. Venugopal, J. Broberg, I. Brandic. "Cloud computing and emerging IT platforms: vision, hype, and reality for delivering computing as the 5th utility". *Future Generation Comput. Syst.* 2009,25, p. 599–616.
- [2] M. Armbrust et al. "Above the clouds: a Berkeley view of cloud computing". UC Berkeley Technical Report. 2009.
- [3] F. Galan, D. Fernandez, J. Ruiz, O. Walid, T. de Miguel. "A Virtualization Tool in Computer Network Laboratories", 5th International Conference on Information Technology Based Higher Education and Training (ITHET'04), Istanbul, June 2004.
- [4] J. Sacha, J. Napper, S. Mullender, and J. McKie. *Osprey: Operating system for predictable clouds*. In DSN Workshops, p. 1-6. IEEE, 2012.
- [5] P. Mell and T. Grance. "NIST definition of cloud computing". National Institute of Standards and Technology. October 7, 2009.
- [6] D.M. Smith, "Hype cycle for cloud computing", Gartner Research G00214915 (2011).
- [7] D. Strauss, "Containers - Not Virtual Machines, Are the Future Cloud", *The Linux Journal*, 228, p. 118-123, April 2013.
- [8] P.-H. Kamp and R. N. M. Watson. "Jails: Confining the Omnipotent Root". In Proceedings of the 2nd International SANE Conference, Maastricht, The Netherlands, May 2000.

Application of multi regressive linear model and neural network for wear prediction of grinding mill liners

Farzaneh Ahmadzadeh

Division of operation and maintenance
Luleå University of Technology
Luleå, Sweden

Jan Lundberg

Division of operation and maintenance
Luleå University of Technology
Luleå, Sweden

Abstract—The liner of an ore grinding mill is a critical component in the grinding process, necessary for both high metal recovery and shell protection. From an economic point of view, it is important to keep mill liners in operation as long as possible, minimising the downtime for maintenance or repair. Therefore, predicting their wear is crucial. This paper tests different methods of predicting wear in the context of remaining height and remaining life of the liners. The key concern is to make decisions on replacement and maintenance without stopping the mill for extra inspection as this leads to financial savings. The paper applies linear multiple regression and artificial neural networks (ANN) techniques to determine the most suitable methodology for predicting wear. The advantages of the ANN model over the traditional approach of multiple regression analysis include its high accuracy.

Keywords—Wear prediction; Remaining useful life; Artificial neural network; Principal Component Analysis; Maintenance scheduling ; Condition Monitoring.

I. INTRODUCTION

The development of a maintenance system for mechanical structures that has both intelligent features in fault detection and knowledge accumulation is an academic goal for researchers as it can greatly assist industry where it is now almost impossible to manually analyse the rapidly growing data to extract valuable decision-making information.

Engineering prognostics is used by industry to manage business risks that result from equipment failing unexpectedly; reliability estimation of equipment and estimation of its remaining useful life (RUL) are mandatory [1].

In practice, such estimations remain predominantly intuitive and are based on the experience of personnel familiar with the equipment. However, due to improved asset reliability and an ageing engineering workforce, it is increasingly difficult to rely on experience. Furthermore, human decision making is not always sufficiently reliable or accurate when dealing with complex equipment with a multitude of interrelated failure modes.

Therefore, developing methods to reduce industry's dependence on human experience is desirable. Appropriate model selection to ensure successful practical implementation requires a mathematical understanding of each model type and an appreciation of how a particular business intends to utilise

the models and their outputs. In reality, industry sites will not be able to use every prognostic modelling option with equal efficacy. The models' ability to perform the modelling is highly dependent on the availability of required data, skilled personnel and computing infrastructure. Consequently, model requirements must be clearly understood [2].

For example, autogenous mills used in the mining industry and in ore dressing plants can cause major bottlenecks in downtime and negatively influence production economics. The rubber liners inside these mills are critical for protecting the mill shell and grinding ore. The replacement and inspection of these mill liners are major factors in mill stoppages and lead to production losses. Therefore, the wear prediction of mill liners is critical for making replacement decisions, as is prediction accuracy.

Selecting the appropriate wear prediction method can lead to a significant reduction in the overall costs [3, 4]. This paper examines two methodologies, linear multiple regression and artificial neural networks (ANN), to determine which is best for prediction [5].

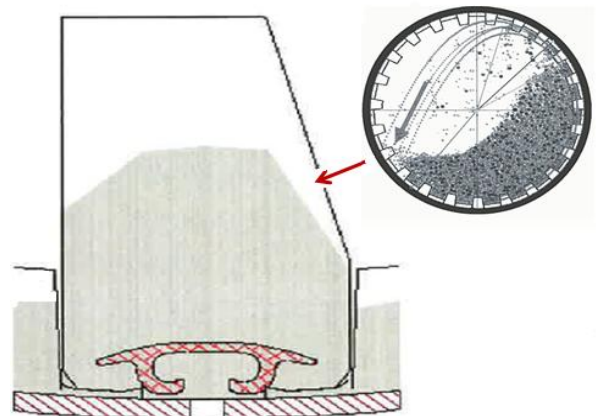


Fig. 1. Cross sectional view of grinding mill and Shell feed lifter bar (by Metso Mineral)

II. DATA COLLECTION AND PREPARATION

Assessing wear using life cycle data is hampered because of the unavailability of operating information, particularly in the wear out phase of liner measurement. The data requirements

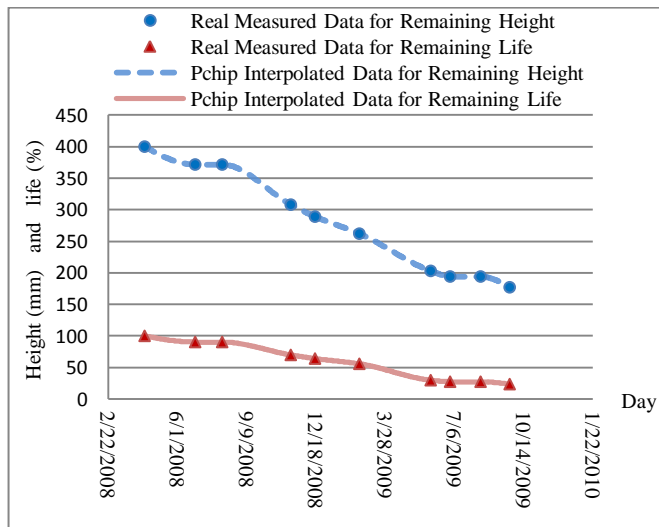
for this research were met by selecting the shell feed lifter bar (LB) of the liner of grinding mill machines see Fig. 1. The following sources were used to gather life cycle and condition monitoring (CM) data.

- Metso mineral AB for wear measurement data during different life cycles.
- Boliden mineral AB for process data during the same life cycle.

Process data for five years were obtained from the Boliden mining company. The process data include the ore type, ore feed (ton/h), power (kW), angular speed (% of centrifugal critical speed), torque (% of the max torque), water addition (m³/h), grinding energy (kWh/ton), load (ton). The mill in this case study processes ore types which come from different mines and have different physical characteristics in grade values (% of metal content), densities, hardness indexes, rock size etc.

A. Interpolation Technique

Because total value of lost production during any mill stoppage is extremely high, it is not economical to stop the mill at intervals and measure liner wear, except for maintenance, inspection, installation, and replacement. As a result, few wear measurement data were available from Metso. The solid circles and triangle in Fig. 2 show real measured remaining height and remaining life of LB during one life cycle; other data shown in the figure were generated by piecewise cubic Hermit interpolating polynomial (PCHIP). PCHIP preserves the shape of the data and respects monotonicity. It is the best interpolating method for this study because of the monotonically decreasing characteristics of CM data in the context of the remaining height and remaining life of the liners.



Real and interpolated data using PCHIP method

Methodologies

Making reliable decisions on maintenance and replacement

of the liners' components requires a thorough analysis of their behaviour during their life cycle. As every product has a different operating environment, individual assessment of the capacity of each component is necessary.

CM and process data must be analysed to identify which parameters can be used to estimate the wear of components. As recommended in [5, 6], the following methods were selected for analysing and investigating these data to determine the wear of components under given conditions of use.

B. Multiple Regression Analysis

Regression analysis is one of the most useful tools to estimate and forecast future trends of variables by analysing historical data. It is an extremely flexible procedure that can aid decision making in many areas, such as sales, medicine, weather forecasting etc. Regression is a technique used to predict the value of a dependent variable using one or more independent variables. Mathematically,

$$Y = a + b_1x_1 + b_2x_2 + \dots + b_kx_k \quad (1)$$

Where Y is the response variable, a and b_i i=1...k are regression coefficients, x₁, x₂ x_k represent the explanatory variables, and K is the total number of explanatory variables.

The method of least squares estimation as [7] is used to calculate the regression coefficients. R-square, the coefficient of determination is used as the performance measure to determine how successfully the method explains the variation of the data [8]. However, the purpose of carrying out regression analysis is to know how the explanatory variables, also called predictors or independent variables, such as ore type, ore feed, power, speed, torque, water addition, grinding energy, and load, are related to the response (dependent) variables in this case, the remaining height and remaining life of the LB. In other words, the primary goal is to estimate or predict the LB's wear given current and past values of the explanatory variables.

The regression analysis is done using Microsoft Excel. It can perform stepwise regression analysis that helps to determine the impact of each of the explanatory variables in the system. The multiple regression equation can be expressed as (2). Applying a multiple regression technique to the data yields the following regression equation:

$$\text{Wear} = a + b_1(\text{Ore type}) + b_2(\text{Ore feed}) + b_3(\text{Power}) + b_4(\text{Speed}) + \dots + b_5(\text{Torque}) + b_6(\text{Water}) + b_7(\text{energy}) + b_8(\text{Load}) \quad (2)$$

Where, wear in (2) shows the remaining height or remaining life of the liner. The multiple regression coefficients' p-values and the correlation results for when the dependent variable is either the remaining life or the remaining height of the LB for shell feed are shown in Tables I and II respectively.

TABLE I. MULTIPLE REGRESSION ANALYSIS RESULTS WHEN DEPENDENT VARIABLE (WEAR) IS REMAINING LIFE

Wear	Intercept	Ore type	Ore feed	Power	Speed	Torque	Load	Grinding energy	Water addition	R ²
<i>Coefficients</i>	298,77	1,44	-0,04	0,17	-4,12	-3,32	0,01	-0,43	3,13	0,40
<i>P-value</i>	0%	2%	0%	3%	3%	1%	98%	82%	0%	

TABLE II. MULTIPLE REGRESSION ANALYSIS RESULTS WHEN DEPENDENT VARIABLE (WEAR) IS REMAINING HEIGHT

Wear	Intercept	Ore type	Ore feed	Power	Speed	Torque	Load	Grinding energy	Water addition	R ²
<i>Coefficients</i>	990,63	4,23	-0,09	0,50	-11,81	-10,29	0,11	-1,59	8,89	0,42
<i>P-value</i>	0%	1%	0%	3%	3%	1%	86%	76%	0%	

TABLE III. MULTIPLE REGRESSION ANALYSIS RESULTS WHEN DEPENDENT VARIABLE (WEAR) IS REMAINING LIFE AFTER APPLYING STEPWISE REGRESSION

Wear	Intercept	Ore type	Ore feed	Power	Speed	Torque	Water addition	R ²
<i>Coefficients</i>	308,29	1,42	-0,03	0,18	-4,37	-3,47	3,10	0,40
<i>P-value</i>	0%	2%	0%	2%	1%	0%	0%	

TABLE IV. MULTIPLE REGRESSION ANALYSIS RESULTS WHEN DEPENDENT VARIABLE (WEAR) IS REMAINING LIFE AFTER APPLYING STEPWISE REGRESSION

Wear	Intercept	Ore type	Ore feed	Power	Speed	Torque	Water addition	R ²
<i>Coefficients</i>	1039,04	4,19	-0,09	0,53	-12,84	-10,84	8,78	0,42
<i>P-value</i>	0%	1%	0%	1%	1%	0%	0%	

Because we want an explanatory model, we only keep variables where the error (p-value) is less than 0.05, giving us a 95% confidence level. Stepwise regression (backward elimination) is used; this involves starting with all candidate variables, testing the deletion of each variable (those with a p-value greater than 5%) to determine whether this improves the model, and repeating this process until no further improvement is possible.

In Tables I and II, the load and grinding energy which have p-values greater than 5% are deleted from (1); this does not improve the R² but the rest of variables are significant at 95% confidence level; see results in Tables III and IV. Thus, the remaining height and remaining life of the shell feed LB can be predicted, following (3), (4), as

$$Remain_life = a + b_1(Oretype) + b_2(Orefeed) + b_3(Speed) + \dots + b_4(Torque) + b_5(Wate) \quad (3)$$

$$Remain_hight = a + b_1(Oretype) + b_2(Orefeed) + \dots + b_3(Power) + b_4(Torque) + b_5(Wate) \quad (4)$$

Despite this remarkable improvement in establishing a good correlation between input and output variables, the multiple regression results are still not acceptable for decision making on life assessment because of the low value of the coefficient of determination, (R²). In the best case scenario, fitted models for the remaining height and remaining life explain 40 and 42% respectively of the total variation in the data.

C. Artificial Neural Network

Artificial neural networks (ANNs) are a special case of adaptive networks; they have been extensively explored in the literature because they can perform nonlinear modelling without a priori knowledge and are able to learn complex relationships among inputs and outputs. Moreover, from a computational point of view, ANNs are quick processes. The idea of using ANNs for forecasting is not new. For example, [9] used the windrow's adaptive linear network to forecast the weather. However, due to the lack of a training algorithm at the time, the research was limited and ANNs were left aside. Since the 1980s, research has expanded. One of the first successful applications of ANNs in forecasting is reported by [10] who designed a feed forward ANN to accurately mimic a chaotic series. In general, feed forward ANNs (PMC, RBF) trained with the back propagation algorithm have been found to perform better than classical autoregressive models for trend prediction in nonlinear time series [11, 12]. Many factors can affect the performance of ANNs (number of inputs and outputs nodes, number of layers, activation functions, learning algorithm, training sample etc.). Thus, building a neural network predictor is a non-trivial task. Since the 1990s, many studies have sought to improve the accuracy of predictions while reducing the time required for processing. ANNs have successfully been used to support the prediction process, and research work emphasizes its importance. Nevertheless, some authors remain skeptical, feeling that the design of an ANN is more of an art than a science, or calling ANNs black boxes, implying that an ANN has no explicit form to explain or analyze the relationships between inputs and outputs. However,

ANNs have been used to find nonlinear or linear relationships between input (process data) and output (CM) variables, thereby predicting the values for CM data which show wear. The proposed ANN model has the following characteristics:

1) *Architecture:* The proposed neural network model is a multilayer feed-forward back-propagation neural network introduced in [13]; see Fig. 3. The back-propagation neural network model has the advantages of handling non-linear problems with a learning capability [14]. The architecture of the proposed network consists of two hidden layers of sigmoid (tansig) neurons followed by an output layer of a linear neuron (purelin). Hidden layers with nonlinear transfer functions allow the network to learn nonlinear and linear relationships between input and output variables, while the linear transfer function of the output layer lets the network produce outputs outside the range [-1, 1].

The number of inputs to the proposed network is given by the number of available inputs or process data (ore type, ore feed (ton/h), power (kW), speed (%), torque (%), water addition (m³/h), grinding energy (kWh/ton), and load (ton)); the number of neurons in the output layer is constrained to two, remaining height and remaining life.

The number and size of layers between network inputs and the output layer were determined by testing several combinations of numbers of layers and various numbers of neurons in each layer. Each of the selected combinations was tested with several different initial conditions to guarantee that the proposed model could provide the best solution. The resulting network consists of two hidden layers of 25 and 50 neurons respectively. See Fig. 3 for its architecture.

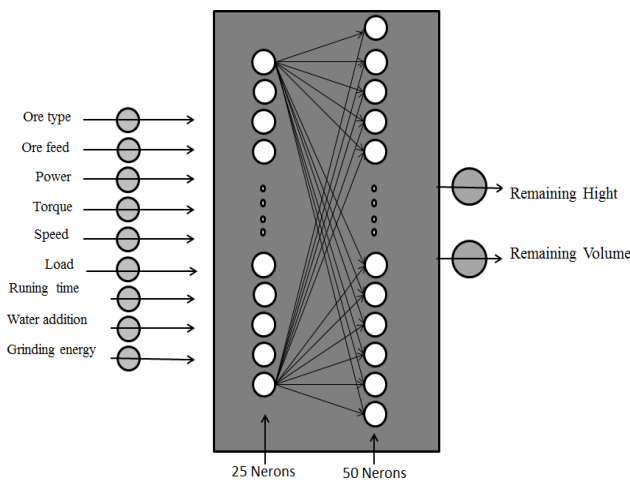


Fig. 2. Architecture of the proposed ANN.

2) *Data preparation:* 886 data sets were collected. Among these data sets, 80% of the data (708 sets) were used for the neural networks' training phase, while the remaining 20% (177

sets) were used to test the network. The testing data were grouped in multiples of 6: 6, 12, 18, and so on.

3) *Training the network:* The training style was supervised learning which provides a set of examples (the training set) of proper network *behaviour*. A training set consists of inputs and the corresponding correct outputs (targets). One of the most powerful learning algorithms, the Levenberg-Marquardt algorithm [13], was used to train the network. In function approximation problems, this algorithm is considered to have the fastest convergence.

4) *Learning and generalization:* After the training was completed, the network was tested for its learning and generalization capabilities. The test of its learning ability was conducted by testing its ability to produce outputs for the set of inputs (seen data) used in the training phase. The test for the network's generalization ability was carried out by investigating its ability to respond to the input sets (unseen data) that were not included in the training process.

Figs. 4 to 7 show relative error for predicted remaining height and remaining life in the training and testing phase (seen and unseen data). These indicate the proposed network's performance. As shown in Figs. 4 and 5 the maximum relative error was less than 6% and 10% for remaining height and life respectively for seen data during the training phase and less than 4% and 10% for remaining height and life respectively for unseen data during the testing phase.

In short, the network predicts the height and remaining life of the liners with accuracy greater than 90%. Therefore, the proposed model can approximate the input-output function with high accuracy.

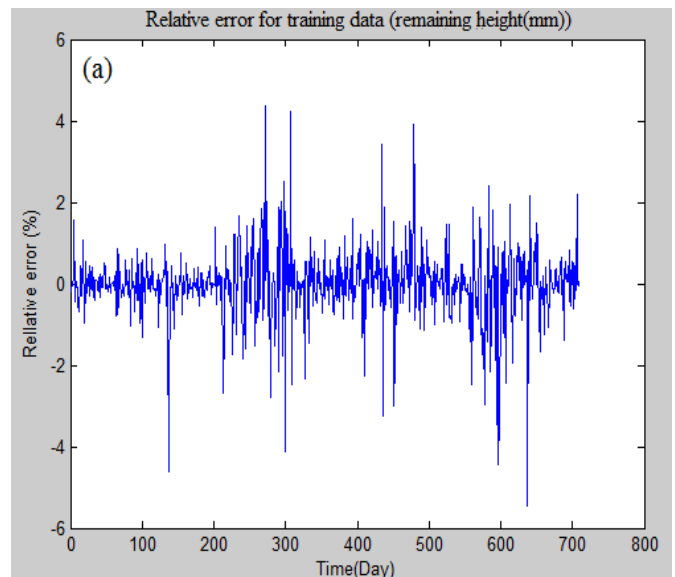


Fig. 3. Relative error for predicted remaining height for training phase

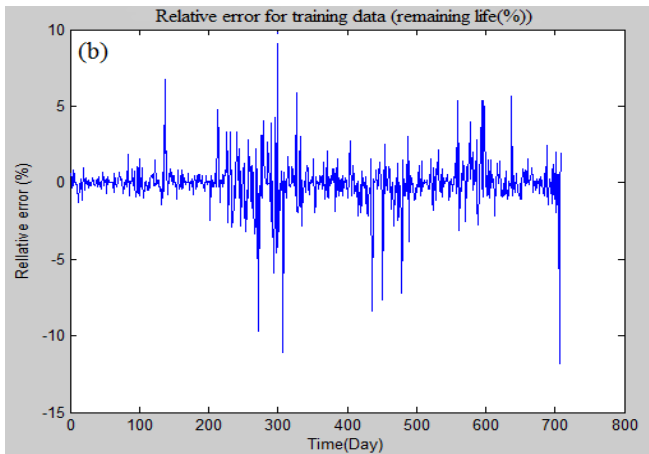


Fig. 4. Relative error for predicted remaining life for training phase

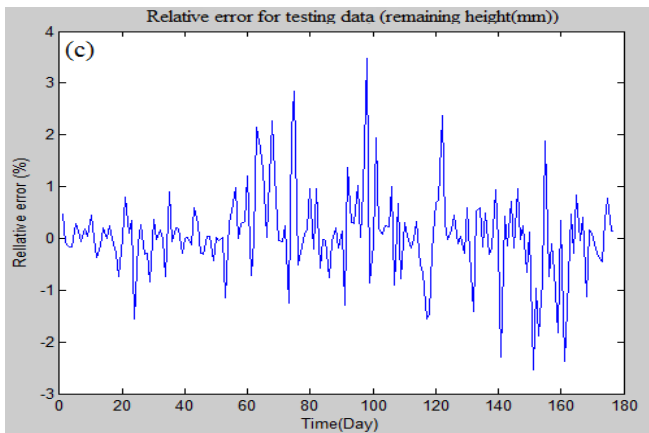


Fig. 5. Relative error for predicted remaining height for testing phase

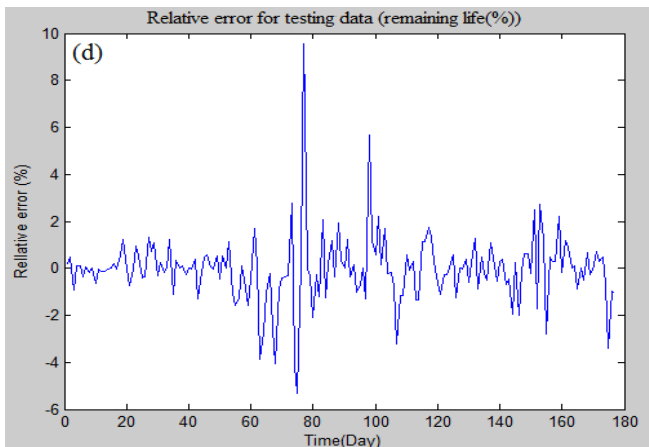


Fig. 6. Relative error for predicted remaining life for testing phase

5) *Network performance*: The performance of the neural network model is very consistent for both the training and testing data. The network's outputs have a correlation coefficient of about 0.99987 with the desired (actual) outputs, as shown in Fig. 8. Clearly, the neural network model is capable of handling the complex nonlinear interrelationships between variables. In addition, there was no substantial

difference in network's output when it was trained with seen or unseen data.

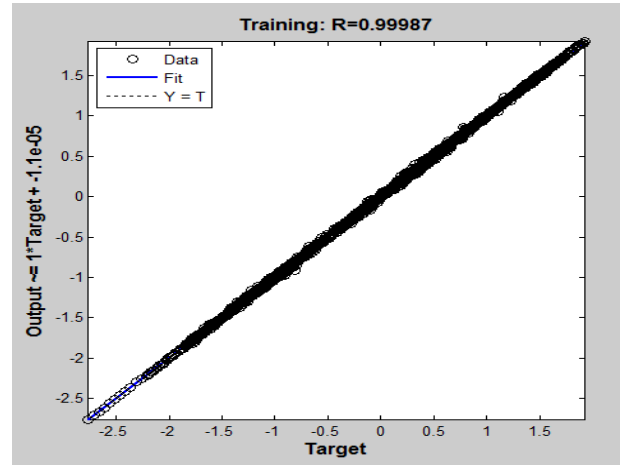


Fig. 7. Neural Network's Performance

III. RESULTS AND DISCUSSION

In this study, regression analysis and artificial neural networks were employed to develop input-output relationships for wear prediction. When we consider two life cycles of data (around five years), it is evident that based on the coefficients of determination (R^2 values), the artificial neural network model is the best of the two methods.

In one hand, regression analysis methods produce reasonable results in situations where the input variables follow a well-defined trend over the age of the component, but this method has been found struggling to maintain its estimation accuracy when the input variables exhibit a complex trend. This limits the application of regression procedures to restricted ranges of data sets making them unsuitable for the whole range of life cycle data. On the other hand, ANN which has been widely used for various prediction and forecasting problems is predominantly useful for many complex real-world problems because of their flexible nonlinear modeling ability.

A summary of the results appears in Table V. The results reveal that the classical procedures of regression analysis fails to produce acceptable results, as the correlation between input and output variables is very low. Further, the multiple regression analysis lacks the required level of accuracy, as the R^2 is around 40% and 42% for remaining height and remaining life respectively, so it is not trustworthy for making decision for maintenance scheduling or replacement. But the proposed neural network can adapt to the data presented to it in the form of input-output patterns with high accuracy more than 99%.

The model shown in Table V has very high values of R^2 for ANN. A comparison summary given in the table shows that the classical methods are no longer capable of producing reasonable results in situations where input-output relationships are nonlinear and complex. Once trained, however, the neural network model yields outputs very close to the desired targets. Therefore, the artificial neural network is the best methodology for wear prediction of grinding mill liners.

TABLE V. COMPARISON OF ESTIMATION ACCURACY FOR ANN AND REGRESSION MODEL

Wear prediction accuracy base on method	R^2	
	ANN	Multiple Regression
Height	0.99987	0.42
Life (Volume)		0.40

IV. CONCLUSION

Wear prediction with nonlinear inputs is far more complicated than with linear inputs, especially in the case of an intricate mixture of fluctuating and unpredictable trends. This study has analysed life cycle data (inputs) by employing regression analysis and ANN to predict the wear of grinding mill liners. It finds the classical procedures of regression analysis inadequate to handle nonlinear and complex input-output relationships in life cycle data, as the correlations between input and output variables established by these techniques are low. The neural network, on the other hand, is very effective in this respect. Furthermore, ANNs are found to have 90% accuracy, and the performance of the proposed model has consistent results for both training and testing data. The study further reveals that employing the proposed ANN as well as condition monitoring data analysis tools is the key factor in securing remaining life estimates associated with higher levels of certainty because maximum relative error was less than 6% for remaining height and less than 10% for remaining life in both training and testing phases.

The findings represent a critical advance in sustainable management of maintenance procedures in industry, especially for heavy duty equipment like grinding mill liners which must work constantly, since it allows for a better understanding of not only service requirements of mill liners but the remaining life of each liners component.

The other advantages of using ANN includes: It doesn't require very expensive and sophisticated equipment for data recording and analysis. The accuracy of the results are very high. The ANN model can accommodate a wide range of input variables with complex and nonlinear input trends/patterns. The proposed methodology does not require disassembly of the liners or stoppage of grinding mills to make decisions on replacement, inspection, installation and maintenance scheduling. The dynamic nature of the proposed methodology

opens up future studies of wear prediction for different categories of liner component such as lifter shell, inner lifter bar, shell plate, feed end discharge end, Etc.

ACKNOWLEDGMENT

The authors would like to thank Boliden, Metso Mineral for supporting this research and permission to publish this article. Special appreciation is extended to the operating maintenance engineers Jan Burstedt (Boliden), Lars Furtenbach and Magnus Eriksson (Metso) for sharing their valuable experience, knowledge, and data to improve the paper.

REFERENCES

- [1] M. El-Koujok, R. Gouriveau and N. Zerhouni, "From monitoring data to remaining useful life: an evolving approach including uncertainty," 34th SReDA Seminar and 2nd Joint ESReDA/ESRA Seminar, Spain, , 2008.
- [2] J.Z. Sikorska, M. Hodkiewicz, L. Ma, "Prognostic modeling options for remaining useful life estimation by industry". Mech. Syst. Signal. Pr. , vol. 25, pp. 1803-1836, 2011.
- [3] A. R. Wijaya, "Multivariate analysis to predict the lifetime of liner due to wear," Annual Rep., Lulea University of technology, Sweden. 2010.
- [4] R. Dandotiya, J. Lundberg, "Replacement decision model for mill liners". J Quality In Maint.Eng. In press. 2011.
- [5] Lucifredi, C. and M. R. Mazziari. "Application of Multiregressive Linear Models, Dynamic Kriging Models and Neural Network Models to Predictive Maintenance of Hydroelectric Power Systems," Mech. Syst. Signal Pr. , vol. 14, pp. 471 – 494, 2000.
- [6] D.C. Montgomery, Design And Analysis of Experiments: Response Surface Method and Designs. John Wiley and Sons: New Jersey, 2005.
- [7] H. J. Motulsky, A. Christopoulos, Fitting Models to Biological Data Using Linear and Nonlinear Regression: A Practical Guide to Curve Fitting. 1st ed., Oxford: Newyork , 2004.
- [8] M. J. C. Hu, E. Halbert, "An adaptive data processing system for weather forecasting". J Appl. Meteor., vol. 3, pp. 513–523, 1964.
- [9] A.Lapedes, R. Farber, "Nonlinear signal processing using neural networks: prediction and system modeling," Tech . Rep. LA-UR87-2662, Los Alamos National Laboratory. 1987
- [10] G. Zhang, B.E. Patuwo and M.Y. Hu. "Forecasting with artificial neural networks: the state of the art," Int. J Forecasting, vol. 14, pp. 35-62, 1998.
- [11] R.C.M Yam, P.W. Tse, L. Li and P. Tu, "Intelligent predictive decision support system for condition-based maintenance", Int. J Adv. Manuf. Tech., vol. 17, pp. 383-391, 2001.
- [12] S. Haykin, Neural networks: a comprehensive foundation. 2nd ed., Upper Saddle River, New Jersey: Prentice Hall, 1999.
- [13] G. P. Zhang, M. Qi., "Neural network forecasting for seasonal and trend time series," Eur. J Oper. Res., vol. 160, pp. 501 – 514, 2005.
- [14] M.I. Mazhar, S. Kara, H. Kaebernick, "Remaining life estimation of used components in consumer products: life cycle data analysis by Weibull and artificial neural networks," J Oper. Manag., vol. 25, pp. 1184–1193. 2007.

DCaaS: Data Consistency as a Service for Managing Data Uncertainty on the Clouds

Islam Elgedawy

Computer Engineering Department,
Middle East Technical University,
Northern Cyprus Campus,
Guzelyurt, Mersin 10, Turkey.

Abstract—Ensuring data correctness over partitioned distributed database systems is a classical problem. Classical solutions proposed to solve this problem are mainly adopting locking or blocking techniques. These techniques are not suitable for cloud environments as they produce terrible response times; due to the long latency and faultiness of wide area network connections among cloud datacenters. One way to improve performance is to restrict access of users-bases to specific datacenters and avoid data sharing between datacenters. However, conflicts might appear when data is replicated between datacenters; nevertheless change propagation timeliness is not guaranteed. Such problems created data uncertainty on cloud environments. Managing data uncertainty is one of the main obstacles for supporting global distributed transactions on the clouds. To overcome this problem, this paper proposes an quota-based approach for managing data uncertainty on the clouds that guarantees global data correctness without global locking or blocking. To decouple service developers from the hassles of managing data uncertainty, we propose to use a new platform service (i.e. Data Consistency as a Service (DCaaS)) to encapsulate the proposed approach. DCaaS service also ensures SaaS services cloud portability, as it works as a cloud adapter between SaaS service instances. Experiments show that proposed approach realized by the DCaaS service provides much better response time when compared with classical locking and blocking techniques.

Keywords—clouds; cloudlet; cloud adapter; data uncertainty; DCaaS; SaaS; PaaS

I. INTRODUCTION

Clouds are the next-generation datacenters virtualized through hypervisor technologies, where cloud-vendors can dynamically provision their virtualized nodes on demand to their customers according to the specified service level agreements [3]. Cloud computing is the computing paradigm that enables the whole solution stack (from hardware to software) to be delivered as services over the internet. Such services are classified into three basic classes: Software as a service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS) [3]. *SaaS services* are applications that customers need. *PaaS services* are services needed to deploy and deliver SaaS services such as database and middleware services. *IaaS services* are services needed to specify the required virtualized computer infrastructure such as disk and memory requirements.

Real-life cloud environments usually constituted from a collection of datacenters connected via a Wide Area Network (WAN). A datacenter is constituted from thousands of machines connected via a LAN (i.e. local area network) forming what is known as a *cloudlet* (i.e. a small cloud). Latency in WANs is much bigger than latency in LANs. This difference in latency distribution inside cloud environments created a non-homogenous timing model for the cloud. For example, latency between two machines inside a datacenter is in the range of 100 msec; while latency between two machines connected via WAN is in the range of 1000 msec (i.e. when machines are in different continents). Such latency difference makes the WAN connections as the main bottleneck in cloud environments. Hence, existing classical concurrency control and transaction management approaches (such as ones discussed in [7] [9]) are not suitable for cloud environments, as they opt to accommodate the slowest latency inside the cloud environment, which badly hurts services performance. To overcome this problem, many approaches have appeared [4][5][6][8][20][22][24] proposing a restricted version of cloud computing, in which requests of users with similar latency values (known as a user-base) are directed to the closest datacenter such that no data sharing between datacenters is allowed. However, data could be replicated later between datacenters storages as a background process to keep databases eventually synchronized and to create multiple copies of the database for backup purposes [12]. We define such computing model as “cloudlet computing”. A cloudlet is a small cloud, so it is similar to cloud in terms of offered services; however it differs in restricting its physical scope into only one datacenter. Cloudlet computing only supports what is known as a mini-transactions [20], which are transactions restricted to a single datacenter to guarantee good performance [4][5][6][8][20][24]. On the other hand, cloudlet computing cannot support global transactions (such as in flight reservation and banking), as global data correctness is not guaranteed due to lack of global control. In other words, cloudlet computing paradigm ensures local correctness of the data within the cloudlet but cannot ensure global correctness of the data (among all cloudlets), as conflicts might appear when data is replicated between cloudlets due to lack of global control. Furthermore, there is no guarantee for change propagation timeliness as updates propagation depends on many different factors such WAN latency and replication schedule of cloud vendor. Hence, data uncertainty becomes a very important characteristic on the clouds and must be

managed by applications as cloud data stores management systems cannot overcome such problem [13]. However, managing data uncertainty in service code is not an easy task, as it requires services to be designed in a different way to deal with missing or multiple values of data objects. Actually, managing data uncertainty is one of the main obstacles for writing transactional applications on the clouds [3]. We argue that service developers should be totally decoupled from managing data uncertainty in their code to improve service maintainability, portability and reusability, nevertheless reducing service development efforts and time. We argue that we need to create a new breed of database management services (such as transaction management, data access, and data replication) that take into consideration the data uncertainty resulting from the cloud non-homogenous timing model. Unlike the classical centralized database management systems, such new breed of database management services must be totally decoupled and must ensure good performance, high availability, and high scalability of services as well as global data correctness, which enables us to easily support global distributed transactions. This paper proposes our first initiative towards achieving these goals. Hence we summarize the contributions of the paper as follows:

1) *First, we propose to use a new middleware platform service for handling data consistency and data uncertainty issues (i.e. Data Consistency as a Service (DCaaS)) on behalf of service developers, hence service code will be totally decoupled from data uncertainty management code leading to faster maintainable SaaS service development. SaaS developers will not write SQL statements in their SaaS service code to access data; instead they will write invocations for the DCaaS service APIs operations to access their data. Furthermore, DCaaS service ensures service cloud portability, as it also decouples SaaS services from directly accessing PaaS services operations; hence no SaaS service code will change if the cloud vendor is changed. The only required change is in the interface between the DCaaS and the PaaS services, which could be handled easily using service adapters [28].*

2) *Second, we propose to use a multi-level data consistency approach for handling SaaS services data objects to enhance service performance. As maintaining strong data consistency is a costly process [15], we argue that it should be only used for objects that their correctness is crucial for services correctness, while for less important data we could go for weaker consistency notions such as eventual or session consistency [23]. Service developers will dynamically define their consistency requirements according to their business logic in a form of a Data Consistency Plan (DCP), and then submit such plan to the DCaaS service, which will make sure such consistency requirements are fulfilled during data access operations. Currently, we support three levels of data consistency strong, eventual, and session that service providers choose from to define the required DCP; more details are given in Section 4.*

3) *Third, we propose a quota-based approach for ensuring global data correctness among cloudlets. The proposed approach applies inventory management principles to ensure fulfillment of users requests, that it requires service providers to divide crucial objects capacity among cloudlets by specifying a quota for each cloudlet such that DCaaS services makes sure no cloudlet user request consume more than the allocated quota. Hence, when data is replicated between cloudlets no conflicts could arise. When a given DCaaS service instance requires more than assigned quota due to high volume of requests, it could contact other DCaaS instances to borrow extra quota. If quota borrowing process fails the request is rejected. To achieve such goals, we provide different protocols for quota borrowing, object stabilization, and DCaaS fault tolerance to ensure protocols liveness and safety properties, more details are given in Sections 4, 5 and 6.*

Experiments show that proposed DCaaS service adopting the proposed data consistency approach provides much better response time when compared with classical locking and blocking techniques. The rest of the paper is organized as follows. Section 2 provides a brief background and discusses related work. Section 3 provides solution model and assumptions. Section 4 introduces the quota-based approach proposed for ensuring data global correctness. Section 5 discusses different management issues of data consistency plan and proposes the adopted object stabilization protocol. Section 6 discusses various design aspects of the proposed DCaaS service such as required APIs and DCaaS recovery. Section 7 provides some basic comparative simulation experiments for proposed approaches, and finally Section 8 concludes the paper. This paper is the extended version of the paper proposed in [27].

II. BACKGROUND AND RELATED WORK

Ensuring data consistency over partitioned distributed database system is a classical problem that attracted many researchers. Data replication is one of the methods used for sharing data between database instances; in which multiple copies of the shared data are stored with the SaaS service instances [7][9][12]. Such copies (replicas) are frequently updated by broadcasting changes to all instances. However, this is not an easy step, as correctness of the data must be maintained. One important aspect of replicated data correctness is mutual consistency, in which all copies of the same logical data must agree on exactly one current value for the data items without violating the logic of the executed transaction. Furthermore, the problem becomes more complicated, when a failure occurs (e.g. due to network failure or server failure) as the correctness of the shared replicated data could be compromised via uncoordinated updates. Classical solutions proposed to solve this problem are mainly adopting locking or blocking techniques to ensure data correctness. Good surveys for such approaches could be found in [7] [9]. Such classical approaches adopt a pessimistic strategy that assumes conflicts occur frequently. Hence, they suspend all other instances from working (via locking or blocking) when a given instance needs to do some updates for the shared data. These techniques provide very bad

performance when applied on cloud environments [5] [6] [11] [15] [22], as they tend to create considerably high overhead over the slow faulty WAN connections due to exchanged synchronization messages and performed reconciliation transactions, which of course badly hurts services availability and customers' response times. CAP theorem [11] clearly states that there is a tradeoff between Data consistency, service availability and partition tolerance. This means asking for high availability and consistent data would imply that we cannot tolerate network partitioning. In cloud environments, data is partitioned over multiple machines to provide high scalability, and as networks between these machines could simply fail, this means partitioning (data and networks) is crucial for cloud environments. Hence, cloud vendors opt to choose between availability and consistency. Recent approaches (such as Google's BigTable [4], Yahoo PUNTS[5], Amazon's Dynamo [8], G-store [6] and Apache Cassandra [24]) proposed to go for weaker forms of consistency on the clouds such as eventual consistency [20], in which they trade consistency for availability, that all service instances are allowed to work normally without any suspension and process their transactions locally. This is known as the optimistic strategy; as it assumes conflict occur rarely. However, when a conflict is detected undo transactions and/or compensating transactions should be performed by the services, also in some cases some data versions could be lost. Going for weaker forms of consistency requires developers to design programs in new ways that can tolerate such data inconsistencies according to business logic. This could be done via writing correcting transactions or using data time stamps to decide between multiple versions of the data as in [21]. Solution proposed in this paper compromise between the optimistic and pessimistic approaches such that it maintains local correctness within a cloudlet using pessimistic approaches and ensures data global correctness between cloudlets by using a quota-based approach that adopts lazy replication approaches as in optimistic approaches to ensure availability and scalability.

III. SOLUTION MODEL AND ASSUMPTIONS

In this paper, we assume cloud vendors provide a PaaS service for accessing the SaaS database (that is a tenant in the physical cloud database). Objects of the SaaS database are stored as simple key-value data format. SaaS database could be partitioned among different cloudlets; hence we require cloudlets PaaS services to provide a lazy replication mechanism (as a background process) to replicate their data changes. A PaaS service could be accessed by one or more DCaaS services simultaneously; hence we require a PaaS service to provide a local concurrency protocol mechanism between DCaaS instances accessing it. Each SaaS instance handles a given user-base of SaaS customers. Each cloudlet could create multiple SaaS, DCaaS, and PaaS service instances to increase availability, throughput, and enhance response times, as depicted in Figure 1. Hence we require each DCaaS instance to keep reference to other DCaaS instances created inside and outside its cloudlet. We model DCaaS service instances as peers and they can communicate with each other in a P2P manner. We require all the communications between SaaS, DCaaS, PaaS services to be done in an asynchronous mode, as the clouds timing model is

non-homogeneous. Hence, fast services will not wait for slow services responses and could process other requests. We require a state machine to implementation be installed at each DCaaS instance to realize proposed protocols, the exchanged messages between state-machines are calls for DCaaS API operations.

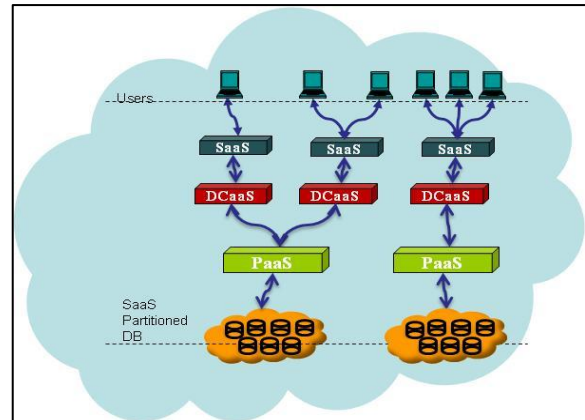


Fig. 1. SaaS, DCaaS, PaaS single cloudlet deployment

IV. QUOTA-BASED DATA CONSISTENCY APPROACH

Work in [13] clearly indicates that data uncertainty must be managed in distributed transactions in order to meet real life requirements. Management of data uncertainty is not a new problem. Actually, in business, handling data uncertainty is a fact of life and many solutions have been adopted by businesses for managing such uncertainty such as reserved inventory, allocations against credit lines, and budgeting. We propose to handle uncertainty for data objects using similar business strategies. For example, inventory management is primarily about specifying the shape and percentage of stocked goods required at different locations within a facility or within many locations of a supply network. Inventory management is the process of efficiently overseeing the constant flow of units into and out of an existing inventory. This process usually involves controlling the transfer in of units in order to prevent the inventory from becoming too high, or dwindling to levels that could put the operation of the company into jeopardy. Hence, we argue we could use the same process for managing transactions accessing objects on distributed data stores such that the data store act as inventories, the objects act as the goods, and the users' requests act as the consuming demand. For example, an airline reservation service could have its database partitioned among many cloudlets (i.e. inventories). Instead of globally locking flight data object whenever a booking operation is made, we will allocate a quota of seats (i.e. goods) for each cloudlet such that each cloudlet locally handles its incoming users requests (i.e. demand) and its DCaaS service instances make sure it does not exceed the allocated quota. When such condition is fulfilled, no data conflicts (i.e. different bookings for the same seat) could appear when replication occurs between cloudlets. A cloudlet ensures the correctness of its transactions using its own concurrency control approach using any locking or blocking technique. This approach will not hurt performance as latency inside cloudlets is small (i.e. within the range of 100ms), which still provide acceptable response

time [20]. Another example, in banking, if we need to access a given account, instead of locking the account object, we could allocate a budget for each cloudlet to manage its incoming withdrawal and deposit requests.

To ensure global data consistency, we require each SaaS service provider to define a capacity quota for its strong consistency data objects for each cloudlet; then provides such information to the corresponding DCaaS service instances via a DCP. DCaaS makes sure that none of incoming users requests to consume more than the allocated objects quotas. In case of one request requires capacity more than allocated quota, the involved DCaaS service instance tries to borrow quota from other DCaaS instances. In case of success, it accepts the request and processes it, otherwise it rejects it. A DCaaS service instance could borrow from instances located in its cloudlet, or from instances in other cloudlets. As borrowing from outside cloudlets requires communications via WAN connection, hence only requests requiring extra quota will be affected. We argue that quota should be distributed between cloudlets in a manner that minimizes the borrowing rate, that quotas should be proportional to the volume of the cloudlets users-bases such that cloudlet with bigger volume should take a bigger quota. We perform the quota borrowing process adopting a simple protocol depicted in Figure 2.

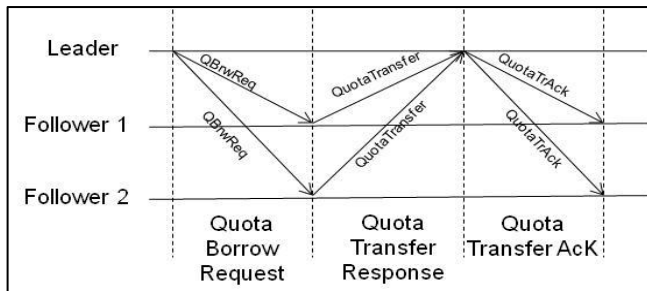


Fig. 2. Stages of Quota Borrowing Protocol

The quota borrowing protocol works as follows. The DCaaS service instance requesting the quota sends its request first to DCaaS instances in its cloudlet with the required of borrow amount. Each DCaaS service instance received the quota borrow request replies back with the quota amount it can transfer. This amount ranges from zero to the required amount. The leader collects all quota transfer responses and acknowledges other DCaaS instances with the amounts it will take. Once DCaaS instances receive such acknowledgment it updates its share of quota with the acknowledged amount. Such protocol requires a state machine to be installed at each DCaaS instance, exchanged messages between state-machines are calls for DCaaS API operations; more details about DCaaS APIs will be given in Section 6. The easiest quota distribution strategy is to equally divide the object capacity among cloudlets. However, the proper quota distribution strategy should be based on thorough demand forecast analysis. In case a service provider makes a mistake in allocating the quotas, DCaaS service instances will automatically redistribute the quotas among themselves when requests arrives via the quota borrowing process. The price of wrong quota allocation is longer response times due to the slow quota borrowing process (if WAN connections are used). However, once quota

borrowing process is finished, response times dramatically improve, as all incoming requests will be handled locally inside the cloudlet, as shown in Section 7.

V. SERVICE DCP MANAGEMENT

Our solution divides the responsibility of managing data uncertainty between the service provider and the DCaaS service. It decouples the definition of the data management strategy from its implementation. The proposed solution requires SaaS service providers to specify the required strategy, while the DCaaS service work on implementing and executing this strategy. A SaaS service provider specifies its strategy by defining a Data Consistency Plan (DCP) for its SaaS service then submits such DCP to the DCaaS service to implement it. DCP specifies the required consistency level for each data object and its corresponding object stabilization method. Service providers could change their DCP at run time without changing their service code. For each data access, DCaaS service checks the required consistency level defined in the DCP; then invokes the corresponding data access procedure. This section discusses different management aspects of SaaS data consistency plan. First it introduces the supported data consistency levels and provides a formal definition for a DCP. Then it describes DCP change management process. Finally, it shows the adopted object stabilization protocol as well as the supported stabilization methods in case of conflicts.

A. DCP Definition and Creation

Currently, we support three levels of data consistency (i.e. Strong, Eventual, and Session). Strong consistency implies that the global correctness of the data object is maintained such that any SaaS instance accessing the object is actually reading its up-to-date correct value. Eventual consistency implies that object correctness is locally maintained (i.e. within a cloudlet) but not globally (i.e. between all cloudlets). However, if there are no global conflicts between cloudlets, and no more new updates are made to the object, eventually all database accesses will return the same last updated value, as cloud vendors perform a lazy replication process between cloudlets to synchronize their DBs [12]. Session consistency implies that the SaaS instances read its own writes only. This means object data will be maintained only at the DCaaS service instance cache and does not go to the PaaS service for storage. Hence, those data will be lost after the session terminates. We require each SaaS provider to define a DCP for its service; hence each SaaS service instance will follow the same service DCP. A DCP indicates the required consistency level for each data object. Also it indicates the required stabilization method to be applied in case of object values divergence. DCP also specifies the cloudlet quota for strong consistency objects. We formally define a DCP as a set of Object Access Patterns (OAP) that $DCP = \{OAP(i)\}$, where an $OAP(i) = \langle i, c, s, q \rangle$, i is the data object reference, c is the required consistency level, s is the required stabilization method, and q is the cloudlet quota distribution plan and it is defined as a set of cloudlet quota allocations, that $q = \{\langle \text{Cloudlet reference}, \text{CloudletQuota} \rangle\}$. As the number of cloudlets is always small, the size of such quota list is not a problem. We support different stabilization methods, more

details about object stabilization will be provided later. We require each data object to have only one OAP. For example, a SaaS provider for a flight reservation service X , which require access for two data objects *Customer* and *Flight*, The corresponding DCP could be defined as $DCP(X) = \{ \langle \text{Customer}, \text{Eventual}, \text{Thomas}, \{ \} \rangle, \langle \text{Flight}, \text{Strong}, \text{Exact}, \{ \langle 1, 50 \rangle, \langle 2, 200 \rangle \} \rangle \}$. This means the consistency of the customer object is eventual and Thomas write rule (i.e. last write wins) will be applied in case of conflicts, while the consistency of the flight object is strong, and history method will applied in case of conflicts. It also shows that we have two cloudlets, the first cloudlet has a quota of 50, and the second one has a quota of 200. As we can see the given DCP definition is working at the object level, however we could extend the definition to work on the attribute level, by defining DCP as a set of Attributes Access Patterns (AAP) that $DCP = \{ AAP(i, j) \}$, where an $AAP(i, j) = \langle i, j, c, s, q \rangle$, i is the data object reference, j is the attribute reference, c is the required consistency level, s is the required stabilization method, and q is a set of allocated cloudlets quota. For example, a DCP over flight attributes could be defined as $\{ \langle \text{Flight}, \text{PlaneModel}, \text{Eventual}, \text{Thomas}, \{ \} \rangle, \langle \text{Flight}, \text{Capacity}, \text{Strong}, \text{Max}, \{ \langle 1, 50 \rangle, \langle 2, 200 \rangle \} \rangle \}$. We do not require specific granularity level for the DCP definition; we leave this choice to SaaS providers to decide. If the SaaS providers choose an object level or a higher level, DCP size could be small and fits nicely in memory but performance could be affected due to local concurrency control locks. However, if they choose the attribute level, the DCP size could be big; hence DCP could not fit into memory and require storage. Of course, this is a classical optimization problem a SaaS provider has to solve. Once quick solution is to compress DCPs using any query-aware compression technique (such as one in [26]) to avoid DCP storage. Another approach to minimize the DCP size is to assume default values for unspecified objects and attributes. We use eventual consistency as the default consistency level, and Thomas rule as the default stabilization method. It is important to note that in this paper, we require DCP to have only one access pattern for each object/attribute. However, in future work we are planning to relax this condition to allow a given object to have different access patterns that DCaaS could choose from in a context-based manner (i.e. choice could be based on the executed SaaS operation, PaaS response time, Users SLAs).

B. DCP Change Management

To provide flexibility for SaaS providers, we provide them with the option to change their DCPs at run time whenever they like and the DCaaS service will do the necessary adjustments to fulfill the new requirements. The DCaaS service contains different components to handle different consistency requirements (refer to Figure 4). It is important to note that change in DCP does not require change in the SaaS service data access code, as DCP change occurs through a specific DCP APIs, while the data access occurs via invocations for different API operations, more details about DCaaS APIs will be given in section 6.

As we do not allow different access patterns for the same data object, whenever DCaaS service instance receives a request for DCP change, it automatically becomes the DCaaS instances leader and notifies the other DCaaS service instances

with the DCP change and make sure it is executed at all instances. For consistency level upgrade request from session to eventual, the DCaaS instance leader updates the corresponding DCP entry, then stores the object value written in its cache into the data store via the PaaS service, and then notifies other instances and waits for their acknowledgments. If all instances replied, it considers the request is fulfilled. In case of missing or slow acknowledgment, the leader tries back after certain timeout window, if an instance still not replying, the leader consider it as a failed node and store the change request for later when it recovers, more details about DCaaS recovery will be given later. For consistency level upgrade request from session/eventual to strong, the DCaaS instance leader updates the corresponding DCP entry, and then starts to stabilize the object values in all cloudlets as correctness of such values were not maintained before the upgrade request. This is done by broadcasting a stabilization request for all DCaaS instances. We have different strategies for stabilizing different object values that differ in their costs, more details are given later in Section 5.3. Once the leader stabilizes the object value, it computes the DCaaS instances quotas then sends for each DCaaS instance the new object value and its allocated quota, more details about quota computation are given in Section 6. For consistency level downgrade request from strong to eventual, the DCaaS instance leader updates the corresponding DCP entry to stop quota checks, as now only local correctness is required. For consistency level downgrade request strong/ eventual to session, the DCaaS instance leader updates the corresponding DCP entry, and then creates an entry in its cache for the object and stop storing object updates into the data store as all updates has to in the cache only. In both cases, DCaaS leader notifies other DCaaS instances with the change and waits for their acknowledgement.

C. Objects Stabilization

When a given DCaaS instance receives a request for consistency upgrade to strong consistency, it requires stabilizing the object value, as every DCaaS instance could have a different value. Our stabilization protocol is very simple. First, we assign the DCaaS instance receiving the change request as the leader who will orchestrate the change. Other DCaaS instances will be the followers. The leader sends a stabilization request messages to all DCaaS instances and waits for their response.

Each DCaaS instance must reply back to the leader with the current value of the object using a stabilization response message. The leader collects all values and computes the new object value by applying the stabilization method defined in the DCP. Finally, the leader sends to each DCaaS instance a stabilization command message to propagate the computed value and monitor instances acknowledgments. Once a DCaaS instance receives a stabilization command, it updates the object value and its DCP and replies with an update acknowledgement. Of course, implementation of such protocol requires a state machine to be installed at each DCaaS instance, exchanged messages between state-machines are calls for DCaaS API operations; more details about DCaaS APIs will be given in Section 6. Figure 3 summarizes the steps of the stabilization protocol.

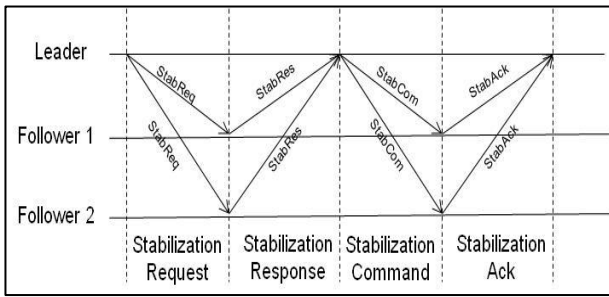


Fig. 3. Stages of Object Stabilization Protocol

Figure 3 shows the different stages of the protocol. To simplify the diagram we assumed propagation delays between DCaaS instants are constant. However, in real life, propagation delays are different. Such stabilization process is a very costly process as it involves communication over WAN connections. Hence, we advise SaaS providers to avoid frequent consistency upgrades. We assume all exchanged messages are asynchronous. We propose different types of stabilization methods in order to provide SaaS providers with the flexibility to choose the most suitable ones for their business logic. Each data object will have its own stabilization method defined in the DCP. To stabilize an object, we propose to use different methods for stabilization varying in complexity, cost, and correctness. 1) Exact method. 2) Thomas write rule. 3) Basic uncertainty filters: Min, Max, Avg, and Sum. 4) Customized uncertainty filter. The exact method guarantees the correctness of the data object value. However, it is the most expensive method and we do not recommend it for SaaS providers due to the huge amount of communication and computation involved. The exact method requires keeping track of transaction history at each DCaaS instances, then sending these histories to the leader to find a global order for all transactions. Then the leader has to execute these transactions to compute the new value, and then distribute the new history and new value to other DCaaS instances. Of course, finding such global transaction order is a very expensive task as it could require many transactions rollbacks over all DCaaS instances. Hence, SaaS providers should use this method only for objects that is extremely crucial for their business logic. The Thomas write rule is one of the most famous methods in conflict resolution. It simply returns the value with the most recent time stamp. Hence, each DCaaS instance should send the leader the object value with its corresponding time stamp. The leader simply chooses the most recent one. The problem with this approach if real time is used is to have global clock synchronization, which is not feasible. However, there are many solutions proposed in distributed computing area for this problem such as use of lamport clock [16]. To avoid the headache of global clock synchronization, we provide the option to use basic uncertainty filter that are used in the area of probabilistic databases [1] [2]. That DCaaS follower sends only the values, and the leader applies one of the basic probabilistic basic functions (such as Min, Max, Avg, Median, and Sum) to get the new object value. Finally, we provide the SaaS providers to provide their own customized uncertainty filters if they did not like to use basic ones.

VI. DATA CONSISTENCY AS A SERVICE

DCaaS service is basically proposed to decouple SaaS developers from managing data uncertainty aspects in their services code. SaaS developers will not write SQL statements in their SaaS service code to access data; instead they will write invocations for the DCaaS service APIs operations to access their data. Also DCaaS service decouples SaaS developers from PaaS services, hence it ensures SaaS clouds portability as no changes will occur to the SaaS service code if we change cloud vendors, the only change will be in the DCaaS service interface with the PaaS service, which could be managed by service adapters [28]. DCaaS takes the consistency requirements of a given SaaS service as a DCP, and then automatically implements and executes the given DCP for each data access. SaaS developers have the flexibility to change their consistency requirements on run time without changing their SaaS service code. This section briefly discusses various design aspects of DCaaS service. First, it discusses the DCaaS structure and configuration, and then it describes the different DCaaS APIs, and finally it illustrates adopted protocols for DCaaS service recovery.

A. DCaaS Structure and Configuration

As we support three different levels of data consistency, DCaaS service should have implementations for approaches realizing the adopted data consistency levels. To decouple DCaaS service code from the realizing approaches implementations, we encapsulate each data consistency approach as a *component service* to be invoked by the DCaaS service. We can think of the DCaaS service as the orchestrator for these service components. For example, a developer could invoke a DCaaS write operation to update a value of an object (e.g. *DCaaS.Write (X,1)*). Listing 1 depicts a sketch for a DCaaS service write operation. DCaaS should invoke the write operation version corresponding to the required data consistency level.

Listing 1: A sketch for DCaaS Write operation

```

int Write (DataObject X, ObjectValue V)
{
    Consistency Level L = GetConsistencyLevel(X);
    Switch (L)
    {
        Case Strong : status= Strong-Write (X, V);
        Case Eventual: status= Eventual -Write (X, V);
        Case Session: status= Session -Write (X, V);
        //...
    }
    return (Status);
}

```

Each component service communicates with the PaaS service to perform the required operations on the data store, as in Figure 4. The DCaaS service is not necessary to be located on the same machine of its service components or the PaaS service. We require that each cloudlet to have at least one SaaS service instance, one DCaaS service instance and one PaaS service instance.

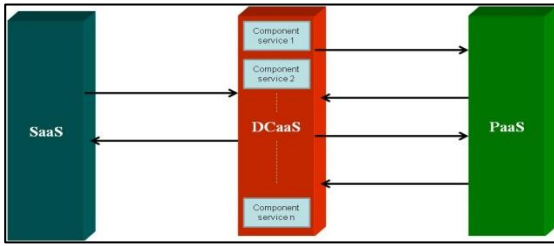
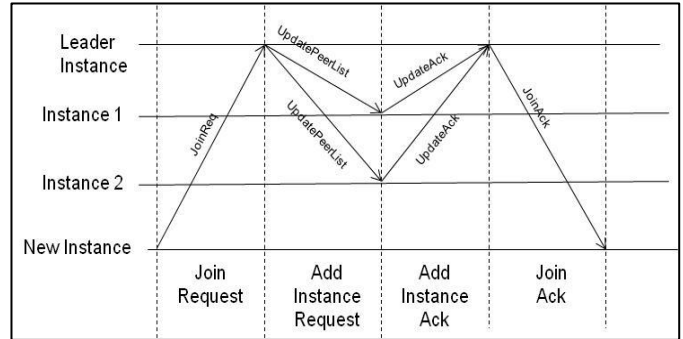


Fig. 4. SaaS, DCaaS, PaaS interactions

A cloudlet management system could clone SaaS and DCaaS service instances to improve performance, however, we require a certain configuration protocol to be followed whenever a new DCaaS service instance is created in order to ensure DCaaS functional correctness. When a DCaaS service instance is created, it will not be in the active state unless the configuration process is finalized. This configuration process could be done manually or automatically. In the manual mode, after a SaaS provider creates a DCaaS instance at each cloudlet, it provides each DCaaS instance with the corresponding DCP as well as a list of other DCaaS service instances (i.e. we define it as the peer list). Once a DCP is loaded, the DCaaS service instance creates a list of Strong Consistency Objects Quotas (i.e. $SCOQ = \{ \langle Object\ Reference, InstanceObjectQuota \rangle \}$) to keep track of its quota. This is done by copying DCP strong object entries into the list and setting *InstanceObjectQuota* to the quota of its cloudlet (i.e. *CloudletQuota* defined in the DCP). In case a cloudlet has only one DCaaS service instance, *InstanceObjectQuota* will be equal to *CloudletQuota*. However, if the cloudlet has multiple DCaaS instances, The *CloudletQuota* should be equal to the sum of all *InstanceObjectQuota* belonging to its DCaaS service instances. In the automated mode, the SaaS providers provides only one DCaaS instance with the DCP and the peer list, and this DCaaS service automatically contacts the other DCaaS Service instances in the peer list to upload the required DCP and the given peer list by invoking specific DCaaS APIs. Again, once each DCaaS loads its DCP, it creates its local SCOQ list. Once a DCaaS instance has its DCP, peer list, and SCOQ list ready, it becomes now in the ready state. When a new DCaaS instance is added to the DCaaS peer network, the cloudlet quota of strong consistency objects specified in the DCP has to be redistributed among all DCaaS service instances inside this cloudlet, and then each DCaaS service instance should update its SCOQ list with the new quota values. This is done via a join DCaaS instance protocol, in which a new DCaaS service instance sends to the current cloudlet leader a join request. If there is no current leader, the new instance sends the join request to any existing DCaaS service instance, which will become the leader. We adopted this simple leader selection approach to avoid doing leader election process. Once a leader receives the join request and makes sure the new instance is authentic and not malicious (security aspects are out of the scope of this paper), it adds it to its peer list and updates its SCOQ list by dividing the cloudlet quota of each object by the number of instances in the new peer list, then sends add instance request to all DCaaS

instance in its old peer list. Once a DCaaS instance receives the update peer list request, it adds the new node to its peer list and updates its SCOQ list by dividing the cloudlet quota of each object by the number of instances in the new peer list, and then acknowledges the leader. Once the leader receives all acknowledgments, it replies back to the new instance with the join accepted message and provides it with the peer list and the DCP, from which the new instance will compute its SCOQ



list. Figure 5 summarizes the steps of the join protocol.

Fig. 5. Stages of DCaaS Instance Join Protocol

Once each DCaaS instance computes its new SCOQ list, it becomes in the ready state and could process user requests. It is important to note that DCaaS instances check first if the required instance to be added is not in their peer list before they do the SCOQ list computation, otherwise they keep the old SCOQ list, as no changes are occurred. This is important issue to make sure the join requests from the recovered instances or new instances are idempotent. DCaaS instance recovery is discussed in the later in this section.

B. DCaaS APIs

DCaaS service API should support operations required for different service interactions. For example, it should support data access operations, operations for objects stabilization, operations for DCP management and quota redistribution, operations for peer list management, and operations for quota borrow and transfer. For data access APIs, DCaaS service implements a simple API interface for reading and writing operations. The read operation API is *Read(DataObject X)*, while the write operation API is *write(DataObject X, ObjectValue V)*. For managing data consistency plans, DCaaS service should provide DCP management APIs such as *LoadDCP(DCP p)*, *ModifyConsistencyLevel(DataObject X, ConsistencyLevel L)*, *ModifyCloudletQuota(DataObject X, CloudletReference R, Quota Q)*. *LoadDCP* is used to load a DCP when a DCaaS service is created, while *ModifyConsistencyLevel*, is used to modify a given data object consistency level. *ModifyCloudletQuota* is used to modify a given cloudlet quota. For peer list management, we should have APIs such as *LoadPeerList(List L)* to upload a peer list when DCaaS is created, *UpdatePeerList(UpdateType T, UpdateDetails D)* to update peer list contents. For quota redistribution protocol, we should have APIs such as *JoinRequest(Instance I)* to request to join the current DCaaS peer network, *UpdateAck(Instance I)* to inform leader with updates confirmation, and

joinAck(DCP P, PeerList R) to acknowledge the acceptance of a new DCaaS instance. For stabilization Protocol, we should have APIs such as *StabReq(Object x)* to request stabilization of a given object, *StabRes(Object x)* to respond with the object value, and *StabCom(Object x)* to enforce a common object value, and finally *StabAck(Object x)* to acknowledge object stabilization. For quota borrowing protocol, we should have APIs such as *QBrwReq(Object x, Amount y)* to request borrowing a certain amount of quota, *QuotaTransfer(Object x, Amount y)* to transfer certain amount of quota to another DCaaS instance, and *QuotaTrAck(Object x, Amount y)* to acknowledge the quota transfer process. For leader election protocol, we should have APIs such as *LeaderReq(Instance I)* to nominate a leader, *LeaderAck(DCP P, PeerList R)* to accept a leader nomination, *Synch(DCP P, PeerList R)* to synchronize DCaaS instances, and *SynchAck* to acknowledge the success of the synchronization process. Of course all DCaaS APIs will be under proper security management; however security is out of the scope of this paper.

C. DCaaS Recovery

In case of a given DCaaS instance failure. We will use classical DB recovery approaches using data logs for recovering eventual consistency objects to rollback any uncommitted transactions, while for session consistency objects, we will just fetch the last values from DB. The problem will be in the strong consistency objects, as the allocated quota for strong consistency objects has to be redistributed among remaining DCaaS instances. Quota redistribution is done when the leader or any other DCaaS instances noticed the failure of such DCaaS instance. Hence, it sends *UpdatePeerList* request to all the DCaaS instances in the peer list, so they can remove such instance from their peer list and update their SCOQ list. When a DCaaS instance recovers from failure, it follows the join protocol in Figure 5 to rejoin the DCaaS peer network. It is important to note that join request is idempotent, hence if multiple copies of the same join request are somehow created, they will have the same effect and no problems could occur. It is also important to note that when a DCaaS instance receives a request for adding a new instance, it checks its log to see if it has a previous history with this instance that if there exist any unfinished communications or acknowledgments so that they can pursue it. The recovery problem becomes more complicated in case of a leader failure during a given protocol execution. In this case, DCaaS instances who still alive could need to elect a different leader to accomplish the required tasks. For example, in case of join protocol, DCaaS instances will send their update acknowledgments to the new instance directly if they notice leader failure. In this case, the new instance will receive multiple join acknowledgments, which will not cause a problem as the join acknowledgments operation is idempotent. However, if the new DCaaS instance times out for not receiving any join acknowledgement, it could resubmit its request to another DCaaS instance. In case of the leader failure during stabilization protocol (see Figure 3), if it failed before receiving stabilization responses, we will have no problems as no DCPs have changed, however if it failed before receiving all stabilization acknowledgments this means we could have a problem. As some DCaaS instances could

have successfully received the stabilization command and updated their DCPs while other instances could not do such updates, if the leader recovers back it could pursue the stabilization process with the remaining DCaaS instances, however if the leader could not recover, this means we have a DCP inconsistency problem as different DCaaS instances will have different versions of the DCP. This problem will be solved after the election of a new leader that will make sure all DCaaS instances are using a common DCP configuration. Leader election occurs when one or more of the follower instances notice the leader failure, and broadcast the election request. Leader election process occurs as depicted in Figure 6. First, a DCaaS instance broadcasts to other instances a request for being the leader, other instances could accept and respond by their (compressed) DCPs and peer lists or reject the request. If majority of instances accepted, this means a new leader is elected, otherwise elections has to be repeated. Once a new leader is elected, the new leader starts synchronizing other DCaaS instances to have a common DCP and peer list. Once a DCaaS instance receives a synchronization request, it updates its information and computes its new SCOQ. Of course we could adopt different strategies for choosing a common DCP and a common peer list. The simplest strategy is choosing the most recent ones. Other strategies is to choose the most restrictive ones, the least restrictive ones, ones leading to least cost updates, etc.

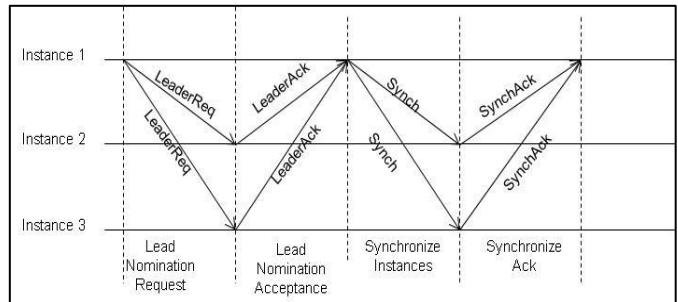


Fig. 6. Stages of Leader Election Protocol.

In our implementation, we let the choice of the DCP selection strategy as one of the configuration parameters for DCaaS services. Of course, there is no optimal strategy as each one has pros and cons. Comparison between strategies is out of the scope of this paper due to space limitation. In case of the leader failure during the quota borrowing protocol (see Figure 2), we will have no problems if failure occurs before receiving the *QuotaTransfer* messages as no quota has actually transferred. However, if failure occurs after sending *QuotaTrack* messages, this means transferred quotas are lost. Again, if the leader recovered before an election request generation, the quotas will be recovered; otherwise quotas will be redistributed during the new leader election process. It is important to note that there is impossibility of distributed consensus with one faulty process [10]. Hence, fault tolerance algorithms should be based on majority rather than complete consensus. Therefore, implementations of proposed approaches adopt a majority of instances (as in Paxos algorithms [17-19]), where f is the number of expected failed instances. We are also extending this approach to handle SaaS requests in order to ensure Byzantine fault tolerance [14] for

SaaS services. This is achieved by allowing the a SaaS service to send its requests to multiple DCaaS service instances, and then accept a response only when it is returned by majority of DCaaS instances, details of this approach is out of the scope of this paper.

VII. EXPERIMENTS

We performed basic simulation experiments using the cloudsim tool [25] that enables us to simulate cloud environments. For simplicity, we assumed that we have only one DCaaS instance per cloudlet; we have two identical cloudlets (i.e. datacenters) with WAN connection of 500ms, and one user-base accessing the first cloudlet with latency 50ms. This user-base generates 1000 request per hour, and each request contains one read and one write operations for a strong consistency data object. We simulated both cloudlets with 5 virtual machines each. Each virtual machine contains 512 MB and 1KB bandwidth. Each cloudlet is build using two 4-core processors identical servers with 10000 MIPS, 200GB RAM, 10 TB storage, and 1MB bandwidth. We run the simulation for period of 1 day and computed the average, minimum and maximum response time for the whole user-base. In our experiments, we compare between the pure locking approach (that locks record on both cloudlets for every request), against the proposed DCaaS service approach. We assumed all objects in the DCP require strong data consistency. However, to show the borrowing effect, we repeated the experiment, by adopting different global quota borrowing rates from outside the cloudlet, which are 0%, 10%, and 50%, which means are 0%, 10%, and 50% of the data accesses will require quota borrow operation, respectively. We choose to compare global quota borrowing (among cloudlets) rather than local quota borrowing as global quota borrowing is the main process that could negatively affect response time due to access of WAN connections. However, local quota borrowing process occurs internally inside the cloudlet where latency is small, hence it will not have a huge impact of response time. Experiments results are listed in Table 1.

TABLE I. EXPERIMENTS RESULTS

Approach	Avg (ms)	Min (ms)	Max (ms)
Locking Approach	1600	1038	2242
DCaaS with 0% Quota Borrow	200	150	258
DCaaS with 10% Quota Borrow	350	249	465
DCaaS with 50% Quota Borrow	600	414	815

As we can see, when the quota is enough (i.e. 0% quota borrow), the DCaaS instance does not need to communicate with the other DCaaS instance through the WAN, as all requests are fulfilled within the cloudlet; hence response time is drastically improved (i.e. from 1600ms to 200 ms). However, when DCaaS needs to borrow quota from the other cloudlet response time starts to increase as WAN connection is used, for example when 50% quota borrow is required response time becomes 3 times worse (i.e. 600 ms). Based on results in Table 1, we conclude that response time increases when the global quota borrowing percentage increases. Hence,

to minimize such response time, we argue that the initial quota distribution among cloudlets should be based on their user-based demand rates, that cloudlet with higher demand should get higher percentages of the quota.

To show the effect of DCP adoption on performance, we conducted a similar experiment when the percentage of strong data objects in the DCP is 0%, 10%, 50%, and 100% respectively. We assumed that the global quota borrow percentage is 50% to have comparable results with Table 1. Experiments results are listed in Table 2. As we can see, when we have no strong data objects (i.e. the 0% DCP case) the response time improves as no need for borrowing operations at all. We achieved much better performance (i.e. 50 ms) when compared with the 0% global quota borrow case in Table 1 (i.e. 200 ms). This is because there is no locking is required to maintain local correctness. However, when we started to increase the percentage of the strong data objects in the DCP, response time starts to increase as quota borrowing operations are required, which require access for WAN connections. Hence, we conclude that to improve performance, we should minimize the percentage of the strong data objects in the DCP. However, in case of having strong data objects in the DCP, the initial quota distribution between cloudlets should be distributed in a manner that minimizes the global quota borrowing rate. We argue that the quota should be distributed according to the cloudlet user-base demand rate.

TABLE II. EXPERIMENTS RESULTS

Approach	Avg (ms)	Min (ms)	Max (ms)
DCP with 0% Strong Consistency Objects	50	36	65
DCP with 10% Strong Consistency Objects	66	48	129
DCP with 50% Strong Consistency Objects	200	141	260
DCP with 100% Strong Consistency Objects	600	414	815

VIII. CONCLUSION AND FUTURE WORK

In this paper, we argued that strong consistency requirements should be adopted only for data objects crucial for application correctness, otherwise weaker forms of data consistency should be adopted. Therefore, we proposed to use the concept of data consistency plan (DCP) to define the consistency requirements for SaaS services, and proposed to use a new platform service (i.e. Data Consistency as a Service (DCaaS)) for executing such DCP plan to decouple SaaS developers from managing data uncertainty issues in their code. We also proposed a quota-based approach for managing data uncertainty on eventually consistent cloud data stores. The proposed approach ensures global data consistency by distributing the capacity of strong consistency data objects among datacenters, and then adopts a lazy replication approach for synchronizing the data stores. Experiments show that proposed quota-based approach realized by the DCaaS service provides much better response time when compared with locking and blocking techniques.

Future work will be mainly focused on providing SaaS developers more flexibility for defining the service DCP, by

allowing an object to have different consistency levels at the same time; depending on the performed SaaS operations and customers' SLAs. This will be achieved by having a new object model that adopts different uncertainty modeling and analysis techniques.

REFERENCES

- [1] D. Barbar'a, H. Garcia-Molina, and D. Porter. The management of probabilistic data. *IEEE TKDE*, 4(5), 1992.
- [2] O. Benjelloun, A.D. Sarma, A. Halevy, and J. Widom. ULDB: Databases with Uncertainty and Lineage. In *VLDB*, 2006.
- [3] R. Buyya, J. Broberg, and A. Goscinski (eds), "Cloud Computing: Principles and Paradigms", ISBN-13: 978-0470887998, Wiley Press, New York, USA, March 2011.
- [4] F. Chang, J. Dean, S. Ghemawat, W. C. Hsieh, D. A. Wallach, M. Burrows, T. Chandra, A. Fikes, and R. E. Gruber. Bigtable: A Distributed Storage System for Structured Data. In *OSDI*, pages 205–218, 2006.
- [5] B. F. Cooper, R. Ramakrishnan, U. Srivastava, A. Silberstein, P. Bohannon, H.-A. Jacobsen, N. Puz, D. Weaver, and R. Yerneni. PNUTS: Yahoo!'s hosted data serving platform. *Proc. VLDB Endow.*, 1(2):1277–1288, 2008.
- [6] S. Das, D. Agrawal, and A. E. Abbadi. G-Store: A Scalable Data Store for Transactional Multi-Key Access in the Cloud. In *SoCC*, 2010.
- [7] S.B. Davidson, H. Garcia-Molina, and D. Skeen, "Consistency in partitioned networks", *ACM Comput. Surv.*, vol. 17, no. 3, pp.341–370, 1985.
- [8] G. DeCandia, D. Hastorun, M. Jampani, G. Kakulapati, A. Lakshman, A. Pilchin, S. Sivasubramanian, P. Vosshall, and W. Vogels. Dynamo: Amazon's highly available key-value store. In *SOSP*, pages 205–220, 2007.
- [9] A. Demers, D. Greene, C. Hauser, W. Irish, J. Larson, S. Shenker, H. Sturgis, D. Swinehart, and D. Terry, "Epidemic algorithms for replicated database maintenance", in *Proc. of ACM Conference on Principles of Distributed Computing (PODC'87)*, 1987.
- [10] M. Fischer, N. Lynch, and M. Paterson. Impossibility of Distributed Consensus With One Faulty Process. *Journal of the ACM*, 32(2), 1985.
- [11] S. Gilbert and N. Lynch, "Brewer's Conjecture and the Feasibility of Consistent, Available, Partition-Tolerant Web Services", *SIGACT News*, vol. 33, no. 2, 2002.
- [12] J. Gray, P. Helland, P. O'Neil, and D. Shasha, "The dangers of replication and a solution", in *Proc. of ACM SIGMOD International Conference on Management of Data*, pp. 173–182, 1996.
- [13] P. Helland. Life beyond distributed transactions: an apostate's opinion. In *CIDR*, pages 132–141, 2007.
- [14] R. Kotla, L. Alvisi, M. Dahlin, A. Clement, and E. Wong, "Zyzyva: Speculative byzantine fault tolerance", In *Symposium on Operating Systems Principles (SOSP)*, 2007.
- [15] T. Kraska, M. Hentschel, G. Alonso, and D. Kossmann, "Consistency Rationing in the Cloud: Pay only when it matters", in *Proc. of the international Conference on VLDB*, 2009.
- [16] L. Lamport. Time, Clocks, and the Ordering of Events in a Distributed System. *Commun. ACM*, 21(7), 1978.
- [17] L. Lamport. The part-time parliament. *ACM Transactions on Computer Systems*, 16(2):133–169, May 1998.
- [18] L. Lamport. Fast Paxos. *Distributed Computing*, 19(2):79–103, Oct. 2006
- [19] L. Lamport. Lower bounds for asynchronous consensus. *Distributed Computing*, 19(2):104–125, Oct.2006.
- [20] D. B. Lomet, A. Fekete, G. Weikum, and M. J. Zwillig. Unbundling transaction services in the cloud. In *CIDR Perspectives*, 2009.
- [21] K. Manassiev and C. Amza, "Scalable database replication through dynamic multiversioning", in *Proc. Centre for Advanced Studies on Collaborative research*, 2005.
- [22] H. Wada, A. Fekete, L. Zhao, K. Lee and A. Liu, "Data Consistency Properties and the Trade-offs in Commercial Cloud Storages: the Consumers' Perspective", in *Proc. of the 5th biennial Conference on Innovative Data Systems Research*, 2011.
- [23] W. Vogels, "Eventually Consistent", *ACM Queue* vol. 6, no. 6, December, 2008.
- [24] Cassandra, available: <http://cassandra.apache.org/>
- [25] CloudSim, available : <http://www.cloudbus.org/cloudsim/>.
- [26] I. Elgedawy, B. Srivastava, and S.Mital, "Exploring Queriability of Encrypted and Compressed XML Data", In *proceedings of the 24th of the International Symposium on Computer and Information Sciences*, Northern Cyprus, 2009.
- [27] I. Elgedawy, "Data Consistency as a Service (DCaaS)", submitted to the 27th of the International Symposium on Computer and Information Sciences, France, 2012. Submission number 11.
- [28] I. Elgedawy, "On-demand conversation customization for services in large smart environments", *IBM Journal of Research and Development*, Special issue on Smart Cities, Vol. 55, No. 1/2, January 2011.

A Computational Model of Extrastriate Visual Area MT on Motion Perception

Jiawei Xu

School of Computer Science, University of Lincoln, Lincoln,
LN6 7TS, United Kingdom

Shigang Yue

School of Computer Science, University of Lincoln, Lincoln,
LN6 7TS, United Kingdom

Abstract—Human vision system are sensitive to motion perception under complex scenes. Building motion attention models similar to human visual attention system should be very beneficial to computer vision and machine intelligence; meanwhile, it has been a challenging task due to the complexity of human brain and limited understanding of the mechanisms underlying the human vision system. This paper models the motion perception mechanisms in human extrastriate visual middle temporal area (MT) computationally. MT is middle temporal area which is sensitive on motion perception. This model can explain the attention selection mechanism and visual motion perception to some extent. With the proposed model, we analysis the motion perception under day time with single or multiple moving objects, we then mimic the visual attention process consisting of attention shifts and eye fixations against motion- feature-map. The model produced similar gist perception outputs in our experiments, when day-time images and nocturnal images from the same scene are processed. At last, we mentioned the future direction of this research.

Keywords—Motion perception; daytime and nocturnal scenes; spatio-temporal phase

I. INTRODUCTION

The current research established three criterions on human visual perception. They are sparse criteria, temporal slowness criteria and independent criteria [1]. This paper researches the motion cues based on the sparse standard. The meaning behind the sparse criteria indicates that most neurons show a relatively low response to external stimuli, includes visual, auditory and olfactory signal, etc. Only a few of them yields a distinct activity. The response distribution of one neuron to the stimuli inputs has a property of sparse and discrete. These characters are of paramount importance and lead the dimensional deduction and feature extraction to the visual system research.

Temporal slowness criteria is described as following, the signal and environment are rapid change with the time, however, the features are slowly change with the time. Then, if we can extract slowly-changed features from the visual inputs, such as random motion, angular transformation or spin, the computational algorithm will be robust to the bio-inspired model.

The third criteria means the neuron are independent to external stimuli. The combination of independent feature subspace and multi-dimensional independent component analysis explain this criterion effectively.

Motion is a vector defined by direction and speed. In the primate visual system, motion is represented in a specialized pathway that begins in striate cortex (V1), extends through extrastriate areas MT (V5) and MST, and terminates in higher areas of the parietal and temporal lobes [2]. While the neural representation of direction in this pathway, and its relationship to perception, have been studied extensively. With the motion feature integrated into the saliency map, the proposed attention model will be able to respond to motion feature naturally. Motion feature is often a dominant factor in complex dynamic scenes. The model can mimic the visual process after adopting the motion cues into the model.

The middle temporal area (MT) is sensitive to visual motion, as discovered by neurobiologists using electroencephalogram (EEG) and gamma-aminobutyric acid (GABA) [3]. It links the bridge between LGN (lateral geniculate nucleus), V1 (Primary visual cortex) and MST (medial superior temporal area), the feedback between these area are parallel and circular [4]. Nearly all neurons in MT area show their preference on the specific motion direction and angle. The instantaneous firing rate at the specific phase is 10 times higher than other phases at a certain neuron [5]. The neurons that react similar response to a certain kind of features can compose a neuron cluster and work synchronously [6]. These information may lead to attention shifts, eye fixations although the underlying neuronal mechanism has not been fully understood.

In this paper, we proposed a computational model which can mimic the human visual selection instantaneously. In order to represent the complex and irregular neuron activities, we model the visual motion via the topological way to cluster same response neurons in a cluster way.

This paper is organized as the following. Section 2 will briefly mention the previous work. The mathematical part and algorithm will be discussed in section 3. Experiments and evaluation combines section 4. And last section is conclusion and future work.

II. RELATED WORK AND MODEL FLOWCHART

The current research on MT is mostly based on the electrophysiological recording and micro stimulation experiments. In 2005, Jing Liu [7] studied Correlation between speed perception and neural activity in the middle temporal visual area. They trained rhesus monkeys on a speed discrimination task in which monkeys chose the faster speed of two moving random dot patterns presented simultaneously in spatially segregated apertures. Evidence from these

experiments suggests that MT neurons play a direct role in the perception of visual speed. Comparison of psychometric and neurogenetic thresholds revealed that single and multi-neuronal signals were, on average, considerably less sensitive than were the monkeys perceptually, suggesting that signals must be pooled across neurons to account for performance. The initial research on MT can be traced back to 1988, W T.

Newsome [8] found the selective impairment of motion perception following lesions of MT. The injection of the ibotenic acid into MT caused striking elevations in motion thresholds; however, had little or no effect on contrast thresholds. The results indicate that neural activity in MT contributes selectively to the perception of motion.

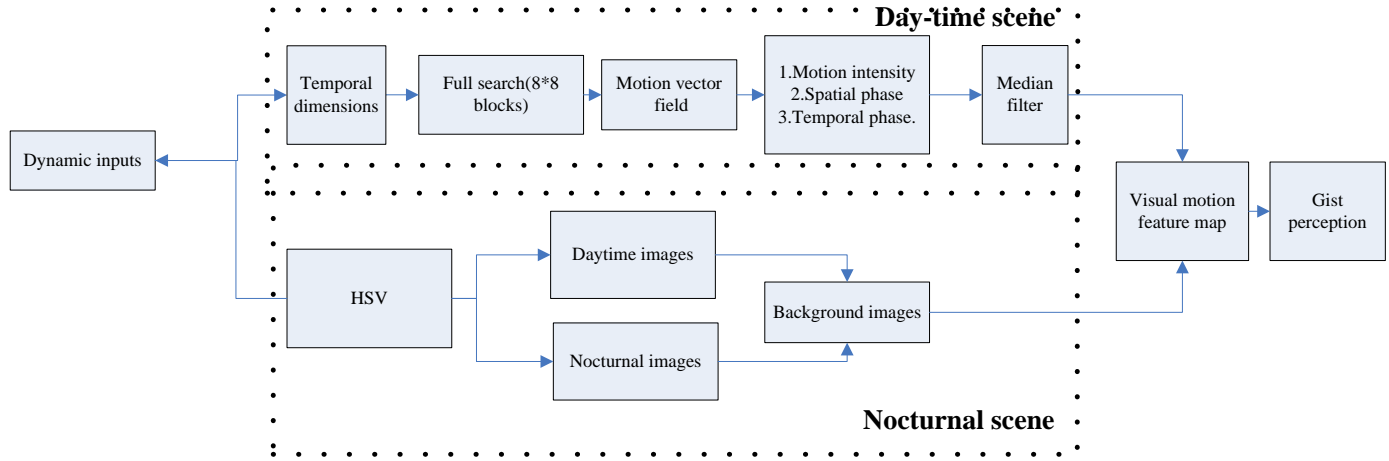


Fig.1. The sketch map of our designed model.

We divide our model into daytime and nocturnal directions, respectively. As many state-of-the art models focus on the daytime images and neglect the night scenes currently, this paper also compare the experimental results with the daytime scenes on the same situation, which further improve the robustness and rationality of proposed model. The framework of our model is represented in figure 1. The system consists of four parts, (1) Motion cues extraction under daytime, (2) objects based segmentation under nocturnal vision, (3) motion perception map, (4) gist perception. In the following section, we will explain the model and algorithm in detail.

III. COMPUTATIONAL MODEL AND ALGORITHM

The model designing concept is described as the following. The motion intensity cue reveals the highly moving objects. The spatial cues indicate the different motion objects in spatial, while the temporal cues donates the variability of one object in the temporal dimensional. Also, the motion orientation weights the motion saliency map and affects the results on a critical extent. For example, when we capture a 135 degree motion on a motion saliency map consisted by most of 45 degree motion vector. This is quite singular and obvious to our human vision system, which means a high tuning weight on the next stage.

A. Motion perception under Daytime video clips

In this section, we introduce the architecture of motion attention model under daytime scenes. We integrated this element into our model as previous approaches [9] [10] are not well considered or simplified this part. Here, we start our research based on AVI video stream. However, we only select the uncompressed video clips to keep the information fidelity.

TABLE I. Motion perception map

```

%% %% motion perception map %% %%
for i=1:f
    if i==f
        break
    end
    frame1=C(i).data;
    frame2=C(i+1).data;
    [px,py]=FullSearch(frame1,frame2);
    PxPy(i).Mx=px;
    PxPy(i).My=py;
    MotAtt(i).Intensity=MotionIntensity(PxPy);
    MotAtt(i).SpatialPhase=SpatialPhase(PxPy);
end
MotAtt(i-1).TemporalPhase=TemporalPhase(PxPy);
FeatureMap=MedianFilter(MotAtt);
for j=1:i
    if j==i
        break;
    end
    SaliencyMap(j).data=FeatureMap(j).Intensity+FeatureMap(j).
    TemporalPhase+FeatureMap(i-1).TemporalPhase;
    img = 255*mat2gray(SaliencyMap(j).data)
    figure(1)
    set(gcf, 'position', [0 0 1366 768]);
    subplot(1,1,1)
    image(img);
    colormap(gray(255));
    axis image off;
end
    
```


In each frame, the spatial layout of motion vectors would compose a field called Motion Vector Field (MVF) [11]. If we consider MVF as the retina of eyes, the motion vectors will be the perceptual response of optic nerves. We set 3 types of feature cues, motion intensity cues, spatial phase, and temporal phase, when the motion vectors in MVF go through such cues, they will be transformed into three kinds of feature maps. We fuse the normalized output of cues into a saliency map by linear combination, and it will be tuned by the weight. Finally, the image processing methods are adopted to detect attended regions in saliency map image, where the motion magnitude and the texture are normalized to [0, 255]. The selection of texture as value, which follows the intuition that a high-textured region produces a more reliable motion vector, provides this method a significant advantage that when the motion vector is not reliable for camera motion, the V component can still provide a good presentation of the frame.

After transforming the RGB to HSV color space, motion saliency can be calculated using the segmentation result of section. An example of saliency map and motion attention is illustrated in Figure 3. Figure 3(a) is the corresponding motion saliency map based on 9 dimensional MVF, while figure 3(b) is the result provided on 2 dimensional MVF. According to our assumption, there will be three cues at each location of macro block $MB_{i,j}$. Marco block is a basic unit of motion estimation in video encoder and it is consisted by an intensity pixel and two chromatic pixel blocks. Hereby we adopt 16*16 Marco block due to the computational burden. Then the intensity cues can be obtained by computing the magnitude of motion vector

$$I(i, j) = \sqrt{dx_{i,j}^2 + dy_{i,j}^2} / MaxMag \quad (2)$$

here (dx_{ij}, dy_{ij}) indicate two components of motion vector, and $MaxMag$ is the maximum magnitude in MVF. The spatial coherence cues induces the spatial phase consistency of motion vectors has high probability to be in a motion object. By contraries, the area with inconsistent motion vectors is possible to be located near the edges of objects or in the still condition. First, we calculate a phase histogram in spatial window with the size of $m*m$ pixels at each location of Marco block. The bin size of each is 10 degree, as we segment the 360 degree into 36 intervals, which means from 0 degree to 10 degree we regard it as a same angle. Then, we measure the phase distribution by entropy as following:

$$C_s(i, j) = -\sum_{t=1}^n p_s(t) \log(p_s(t)) \quad (3)$$

and
$$p_s(t) = SH_{i,j}^m(t) / \sum_{k=1}^n H_{i,j}^m(k) \quad (4)$$

Where C_s donates spatial coherence, $SH_{i,j}^m(t)$ is the spatial phase histogram whose probability distribution function is $p_s(t)$, and n is the number of histogram bins. Similarly, we define temporal phase coherence within a sliding window with the size of $W(frames)$. It will be the output of temporal coherence cues as expressed below:

$$C_t(i, j) = -\sum_{t=1}^n p_t(t) \log(p_t(t)) \quad (5)$$

and
$$p_t(t) = TH_{i,j}^W(t) / \sum_{k=1}^n TH \sum_{i,j}^W(k) \quad (6)$$

Where C_t denotes temporal coherence, $TH \sum_{i,j}^W(t)$ is the temporal phase histogram whose probability distribution function is $p_t(t)$ and n is still the number of histogram bins. Moreover, we increase the frame number as a temporal dimension and the output is easier to distinguish the difference. The result indicates the attended region can be more precise if we elongate the frame number as shown in figure 5.

The Laplacian filter is to remove the impulse noise generated by the input frames. Hereby we adopt the median filter can also preserve the edge information and sharpen the image details. We adopt 3*3, 7*7..., 25*25 window slides at the experiment stage, but finally we utilize 3*3 window as the convenience of later computation. The detail code is given as the following.

TABLE II. Temporal phase cues

```
function TemporalPhaseVect=TemporalPhase(PxPy)
[p,q]=size(PxPy(1).Mx);
Timesize=length(PxPy);
for m=1:p
    for n=1:q

        for s=1:36
            Num(s)=0;
        end

        for i=1:Timesize
            TimeWinx(i)=PxPy(i).Mx(m,n);
            TimeWiny(i)=PxPy(i).My(m,n);
            x=TimeWinx(i);
            y=TimeWiny(i);
            if x>0&&y>0
                Phase=pi/2*0+asin(abs(y)/sqrt(x^2+y^2));
            elseif x<0&&y>0
                Phase=pi/2*2-asin(abs(y)/sqrt(x^2+y^2));
            elseif x<0&&y<0
                Phase=pi/2*2+asin(abs(y)/sqrt(x^2+y^2));
            elseif x>0&&y<0
```

```

Phase=pi/2*4-asin(abs(y)/sqrt(x^2+y^2));
elseif x>0&&y==0
    Phase=0;
elseif x<0&&y==0
    Phase=pi;
elseif x==0&&y>0
    Phase=pi/2;
elseif x==0&&y<0
    Phase=pi*3/2;
else Phase=0;
    end
    Angle=Phase/2/pi*360;
Data(i)=Angle;
for t=0:35
    if (Data(i)-t*10)<10&&(Data(i)-t*10)>=0
        Num(t+1)=Num(t+1)+1;
    end
    if Data(i)==360
        Num(1)=Mum(1)+1;
    end
    end
    TemporalPhaseVect(m,n)=EntropyMethod(Num);
    end
end
end

```

TABLE III. Motion intensity

```

%%%% motion intensity & spatial phase %%%%%%%
function Intensity=MotionIntensity(PxPy)
MotionVectX=PxPy.Mx;
MotionVectY=PxPy.My;
[m,n]=size(MotionVectX);
[p,q]=size(MotionVectY);
for i=1:m
    for j=1:n
        a=MotionVectX(i,j);
        b=MotionVectY(i,j);
        Vect(i,j)=sqrt(a^2+b^2);
    end
end
LargestVect=max(max(abs(Vect)));
for i=1:p
    for j=1:q
        c=MotionVectX(i,j);
        d=MotionVectY(i,j);
        Intensity(i,j)=sqrt(c^2+d^2)/LargestVect;
    end
end
end

```

TABLE IV. Spatial phase cues

```

%%%% spatial pahse%%%%
function TemporalPhaseVect=TemporalPhase(PxPy)
[p,q]=size(PxPy(1).Mx);
Timesize=length(PxPy);
for m=1:p
    for n=1:q

        for s=1:36
            Num(s)=0;
        end

        for i=1:Timesize
            TimeWinx(i)=PxPy(i).Mx(m,n);
            TimeWiny(i)=PxPy(i).My(m,n);

            x=TimeWinx(i);
            y=TimeWiny(i);

            if x>0&&y>0
                Phase=pi/2*0+asin(abs(y)/sqrt(x^2+y^2));
            elseif x<0&&y>0
                Phase=pi/2*2-asin(abs(y)/sqrt(x^2+y^2));
            elseif x<0&&y<0
                Phase=pi/2*2+asin(abs(y)/sqrt(x^2+y^2));
            elseif x>0&&y<0
                Phase=pi/2*4-asin(abs(y)/sqrt(x^2+y^2));
            elseif x>0&&y==0
                Phase=0;
            elseif x<0&&y==0
                Phase=pi;
            elseif x==0&&y>0
                Phase=pi/2;
            elseif x==0&&y<0
                Phase=pi*3/2;
            else Phase=0;
            end

```

B. Motion perception model under nocturnal video clips

The previous survey confirmed these facts. Cone-shaped and rod cells contain $6 * 10^6$ and $1.2 * 10^6$ on human retina, respectively [12]. The former one distributed on the center of retina, however, the later one are located on the periphery of retina. On the day time, human vision and motion perception are completed by the cone-shaped cells. However, rod cells activate its function under night vision. Cone-shaped cells, conversely, need high light intensities to respond and have high visual acuity. Different cone cells respond to different colors (wavelengths of light), which allows an organism to see color [13].

Rod cells are highly sensitive to light, allowing them to respond in dim light and dark conditions. These are the cells that allow humans and other animals to see by moonlight, or with very little available light (as in a dark room). However, they do not distinguish between colors, and have low visual acuity (measure of detail) [14].

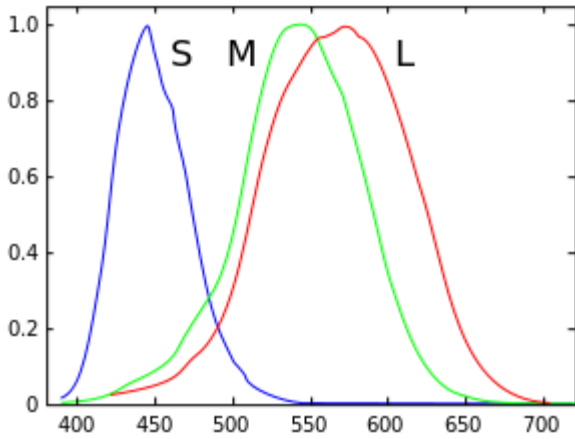


Fig.2. Normalized response spectra of human cone cells, S, M, and L types. Vertical axis: Response [15]. Horizontal axis: Wavelength in nanometers.

Generally, the difficulties of night image problem mainly contain two aspects. The first is that the obtained night image appears much noise, due to reasons of sensor noises or very low luminance. The second is the high light or dark areas in which the scene information cannot be seen clearly by the observers.

To mimic the biological process, we convert the videos from RGB to HSV color space for the convenience of process, and enhance the contrast of video inputs, thus lead to motion estimation at the later stage.

The enhancement of contrast can be classified into 3 steps. The first is calculate contrast c , then using the nonlinear transformation to get c' , which means x_i to x , then last step is compute the pixel grayscale value using c' . The mathematical equation is:

$$c = \frac{|x - x_i|}{x + x_i} \quad (7)$$

$$c' = \psi(c) \quad (8)$$

$$x' = \begin{cases} \frac{x_i(1-c')}{(1+c')}, & x < x_i \\ L_{\max} - \frac{(L_{\max} - x_i)(1-c')}{(1+c')}, & x \geq x_i \end{cases} \quad (9)$$

where x_i is the average gray-scale value of attended pixel, L_{\max} is the maximum gray-scale value, while ψ is convex transformation as $\psi(0) = 0, \psi(1) = 1, \psi(c) \geq c$.

Considering the background images of daytime and night are the images of the same scene captured under different illumination. Both objects, such as road, building, cars and

player are extracted and the remaining part is classified into background. To distinguish the night vision and daytime vision, we assume if the luminance values of night images background are larger than the luminance of daytime images background, we classify the videos into night videos, vice versa.

After background model estimate, the background image of day and night (DB and NB) are transformed from RGB color space to HSV (Hue-Saturation-Value) color space. An illumination segmentation map $L_{(x,y)}$ can be computed as (10),

$$L_{(x,y)} = \begin{cases} 1 & (NB_{(x,y)}(V) - DB_{(x,y)}(V)) > 0 \\ 0 & (NB_{(x,y)}(V) - DB_{(x,y)}(V)) < 0 \end{cases} \quad (10)$$

Where $DB_{(x,y)}(V)$ and $NB_{(x,y)}(V)$ denote the luminance value of background image DB and NB separate at position (x,y) .

To achieve real-time and accurate moving objects segmentation, we first use illumination histogram equalization in the night video $N_{(x,y)}(V)$. Pixels will be classified into M levels according to their illuminance. After that different thresholds will be assigned for different classes in the background subtraction. Let $p(i)$ denotes the ratio of pixels, which luminance equals to i in $N_{(x,y)}(V)$, G denotes the equalized images, and it can be computed through the equation (11):

$$G_{(x,y)} = M * f(m), m = 1, \dots, M \quad (11)$$

Where $f(m) = \sum_{x=0}^{x=m} p(i)$ and $G_{(x,y)}$ will be modified to nearest integral number. For the high light area has already

Been exacted. The motion map M can be computed by

$$M_{(x,y)} = \begin{cases} \left\{ \begin{array}{l} |N_{(x,y)}(R) - NB_{(x,y)}(R)| \\ |N_{(x,y)}(G) - NB_{(x,y)}(G)| \\ |N_{(x,y)}(B) - NB_{(x,y)}(B)| \end{array} \right\} > T(m) \}, \text{ or} \\ 0 \end{cases} \quad (10)$$

where $T(m)$ represents the threshold at luminance level m and $m = G_{(x,y)}$.

The final fusion rule we used is choosing the maximum value of the coefficients of the night input image and daytime reference background image for the high frequency band. For the low frequency band, the coefficients of the images are weighted according to the motion and illumination map.

C. Gist perception under dynamic scene

Recently, situation awareness (SA) [16] has been developed as a theoretical mental model for the gist perception under dynamic scenes.

It includes three levels: perception with focalized attention, comprehension of the current situation, and projection of future status. One interesting point of SA is that it proposes a goal-directed task analysis method to determine what aspects of the situation are important for perception

From the biological review, psychophysical experiments first demonstrated that humans are sensitive to average or centroid position. More recent work by Alvarez and Oliva [17][18] suggests that selective attention may play a minimal role in this process.

Using a multiple object tracking task found that even when observers were unable to identify individual unattended objects, they could localize the centroid of salient objects.

While Chong and Treisman [19] demonstrated that distributed attention could improve an estimate of the mean, this work showed that a summary might be derived even in the absence of attention. Consistent with this, Demeyere and colleagues found that a patient with simultanagnosia could perceive ensemble color in an array of stimuli despite being unaware of the array.

After obtaining the motion cues maps and fixation points, we selected the most gathering fixation points than other regions. After we get these points on each frame, if the points occupy on a relatively concentrated area, we then assumed it as the regions of attention. To indicate the region of interests, we will add a red circle with the radius of 64 pixels to indicate gist perception on visual scenes.

The computational results are elaborated in Section IV. We implement 4 groups of experiments and made the performance evaluation to compare our model's effectiveness with other standard models.

IV. EXPERIMENTS AND RESULTS

To demonstrate effectiveness of the propose attention model, we have extensively applied the method on several types of video sequences from the benchmarks. The detail of the testing results is given in table 3.

D. Benchmark Datasets

We applied our model on different types of videos to verify its feasibility and generality. The dataset are from [20][21][22][23], as detailed in Table 2, includes surfing player, parachute landing, outdoor, traffic artery and other video sequences with high or low motion features.

By implementing two kinds of experiments, we are intended to verify two predictions. The first one is to measure the motion effects on the judgment of human visual attention selection. We prove this predication by comparing the static attention selection model and the results generated by our model is more close to the ground-truth results, pointed out by the participants with normal or corrected vision. The second one is the potential eye fixations on video clips, we are trying to verify the predication that eye saccade yields simultaneous fixations in a millisecond time; however, human eyes are inclined to select the most dense regions with the fixation numbers. This predication matches the result as we can see from experiment 2.

E. Experiments on the motion perception under daytime

The first group experiments are based on the single object moving on the video clip. The tests are short movies with AVI format and 1366*768 frame sizes, 15 fps.

TABLE V. Benchmark testing datasets of daytime vision

No.	Video Subjects	Temporal dimension
1	Surfing player	22
2	Glider	18
3	Moto cyclist	25
4	Traffic artery	109

Testing videos

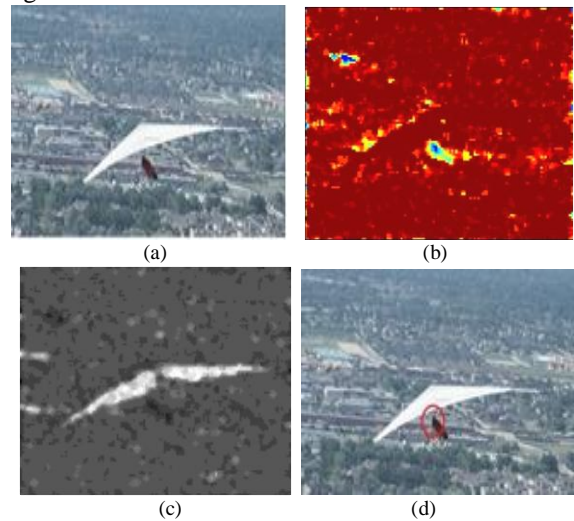


Fig.3. From top to bottom row, we illustrate a sample frame of one testing video clips shown on (a), (b) is the entropy map obtained by computing 22 temporal dimensions, (c) is the visual motion-feature-map, (d) is the gist perception.

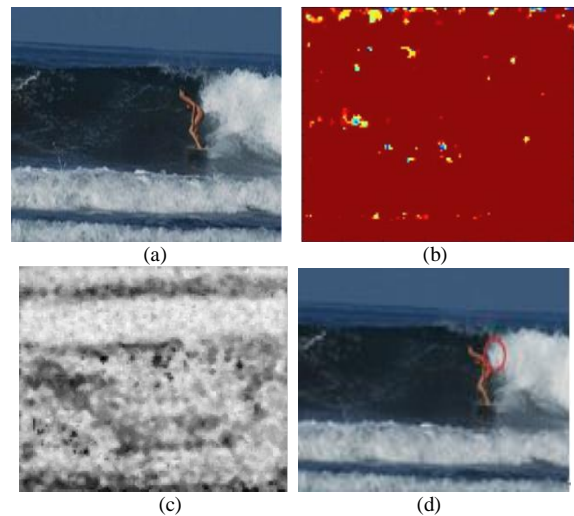


Fig.4. From top to bottom row, we illustrate a sample frame of one testing video clips shown on (a), (b) is the entropy map obtained by computing 18 temporal dimensions, (c) is the visual motion-feature-map, (d) is the gist perception.

The following figures emphasis multiple moving objects on the testing videos, we need to verify the model's robustness under more complex background.

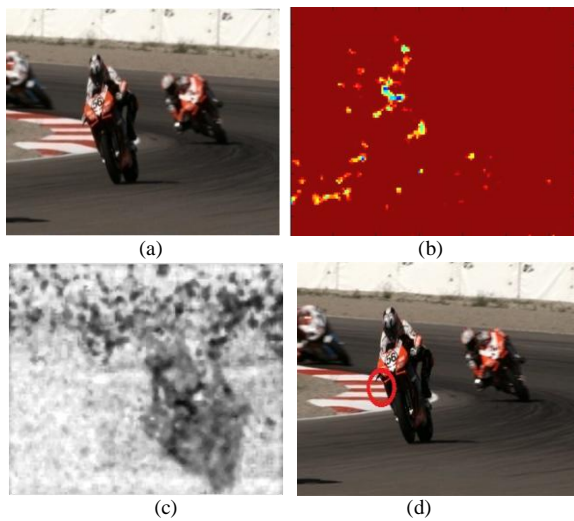


Fig.5. From top to bottom row, we illustrate a sample frame of one testing video clips shown on (a), (b) is the entropy map obtained by computing 25 temporal dimensions, (c) is the visual motion-feature-map, (d) is the gist perception.

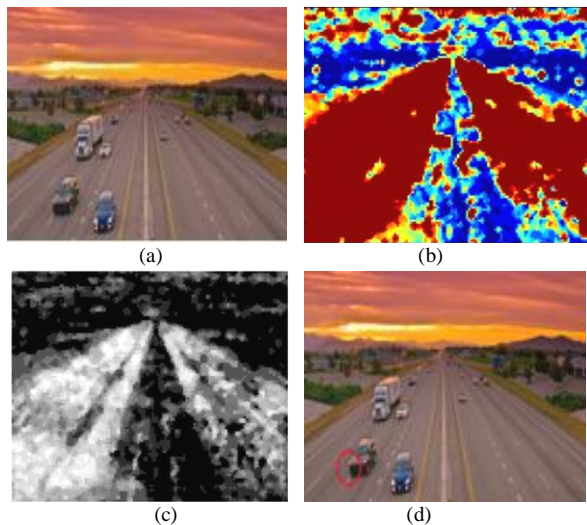


Fig.6. From top to bottom row, we illustrate a sample frame of one testing video clips shown on (a), (b) is the entropy map obtained by computing 109 temporal dimensions, (c) is the visual motion-feature-map, (d) is the gist perception.

From these experiments in figure 5 and figure 6, we conclude these common features as the following. First, the computational burden increase exponentially with the temporal dimension, our testing platform is based on a Windows 7 Intel Core i5 laptop using Matlab 2010b software. The shortest time is 5.73s; the longest time is 24.62s, respectively. Second, visual motion-feature-map in figure 5 (c) and 6 (c) indicate the dynamic motion vectors by computing the pre-setting temporal dimensions, the whiten area indicates higher entropy and motion activity area; however the darker area is relatively low-motion area. Third, the gist perception is based on the weight competition based on the maximum motion cues. Every weight competition computes for one fixation and the maximum value will be selected as the gist perception which represented by red circle for the saliency output. This is discussing in experiment 2.

F. Experiments on the eye fixations and motion perception under daytime

In this experiment, we analysis the relationship between the eye fixation and motion cues map. As we can find in figure 3, the potential eye fixations are representing by the symbol “+”. We test on a new video clip with the genre “parachutes landing”, each frame corresponds to a motion cues map as we show in figure 4.

Frame 30:

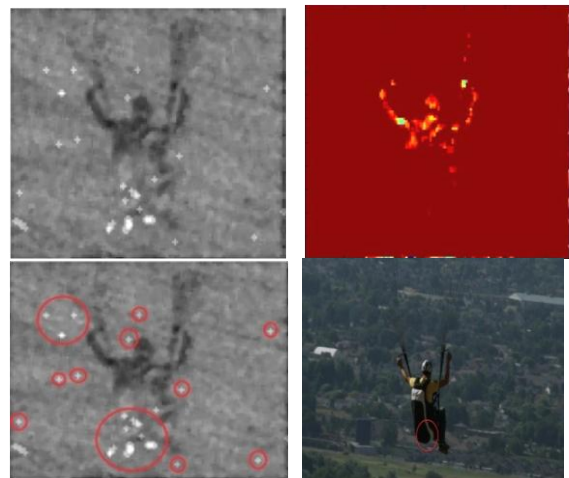


Fig.7. From left to right on first row, the left image is the motion cues map composed by 33 frames, while the right one is the corresponding entropy response with red setting background.

Frame 11:

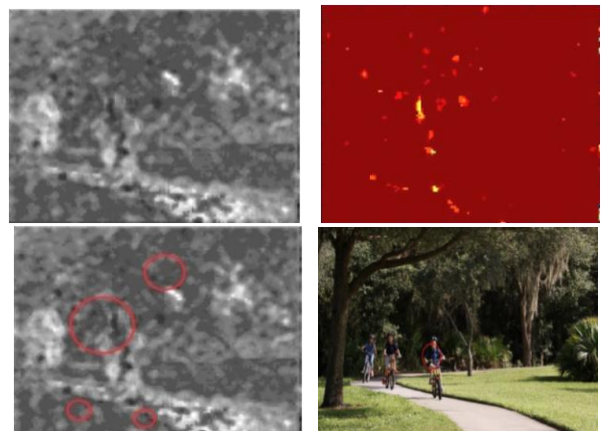


Fig.8. Another testing video with same methods in figure 4, only difference is 19 frames in total.

As shown in figure 7 and figure 8, in this group, we detected the salient regions on the center of the map and white “+” symbols are mostly scattered on the middle-bottom and left-center parts of the image. The white “+” symbols indicate the eye fixation regions; we can find the distinct result that most eye fixation regions are located on the parachutes with a larger circle. We also find other fixations with relatively small circle on the other part of images; however, these points will be selected as the sub-salient region according to WTA

(Winning-take-all) and IOR (Inhibition of return) mechanisms. The right image of bottom row indicates the saliency.

G. Experiments on the nocturnal motion perception

In this part, we illustrated the results by using the algorithm from part B of section 3. The experiments are based on the capture the same position scenes during daytime and night, using the high illumination to get the motion maps.

The figure 9 represents the images after contrast enhancement. Figure 10 shows the images under daytime and night background, then we computes the motion perception map by using the equation (7) (8) (9).

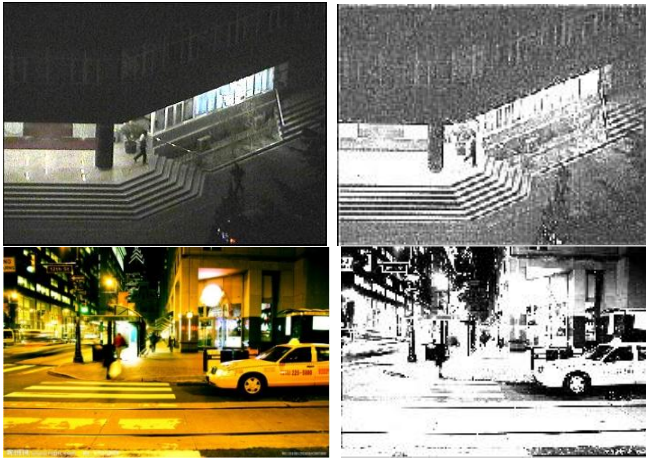


Fig.9. Frames enhancement examples by using the histogram equalization

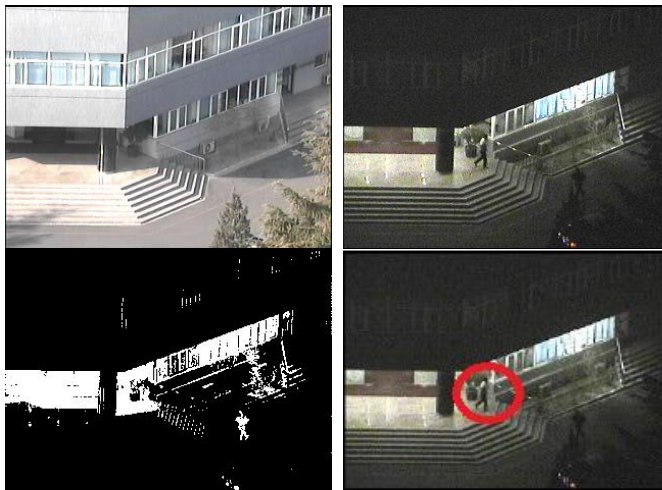


Fig.10. motion perception under nocturnal scenes. Top row, from left to right, daytime input video and night input video. Bottom row, motion perception map and gist perception of scenes.

H. Performance Evaluation

In order to further verify the proposed method, we compared our approach with several state-of-the-art methods.

A lot of measure standard have been proposed since the attention models pop out.

Generally, there are 2 criteria adopted in the evaluation, the salient information is well displayed, quantify the attention models to sticking out the salient region. We measured the overall performance of the proposed method with respect to precision, recall, and F-measure, and compared them with the performance of existing competitive automatic salient object segmentation methods, such as Itti & Koch's method [24] [25], AIM [26] and Achanta's method [27].

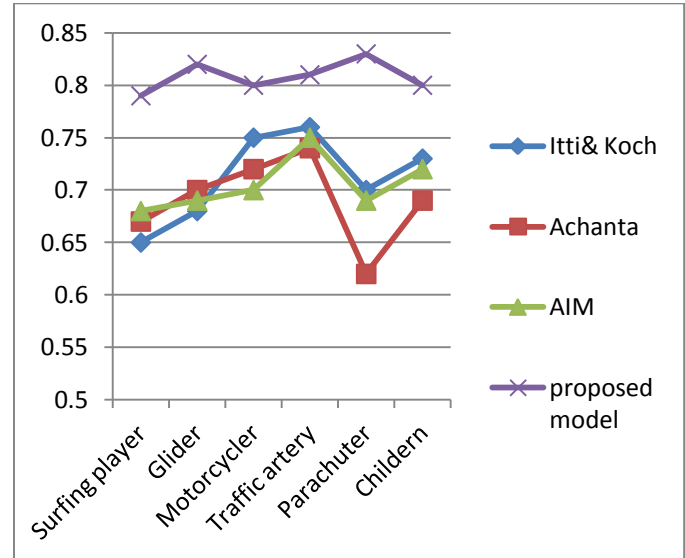


Fig.11. Evaluation of our proposed method under daytime

According to the standard evaluation methods, precision is the percentage that the detected saliency map divided on the non-ground-truth saliency map as been predicted. Recall is a measure of the percentage provided that the detected saliency map divided on the ground-truth saliency map as been predicted. The highest percentage of precision indicates the real attention region as the test participants assumes them as the attention region. The recall is similar as the false positive. F-measure is a special method which predicts the overall performance of the model. Precision (P), recall (R), F-measure used in this study is calculated from:

$$P = \sum(S * A) / \sum(S) \quad (11)$$

$$R = \sum(S * A) / \sum(A) \quad (12)$$

$$F = 2 * P * R / (P + R) \quad (13)$$

Here S donates the proposed attention regions, A is the ground truth attention regions, $S * A$ indicates the gray-scale image by the gray value of pixel wise multiplication. $\sum(\dots)$ is the summation of the gray value of each pixel. Obviously, a larger value F means a better effect result.

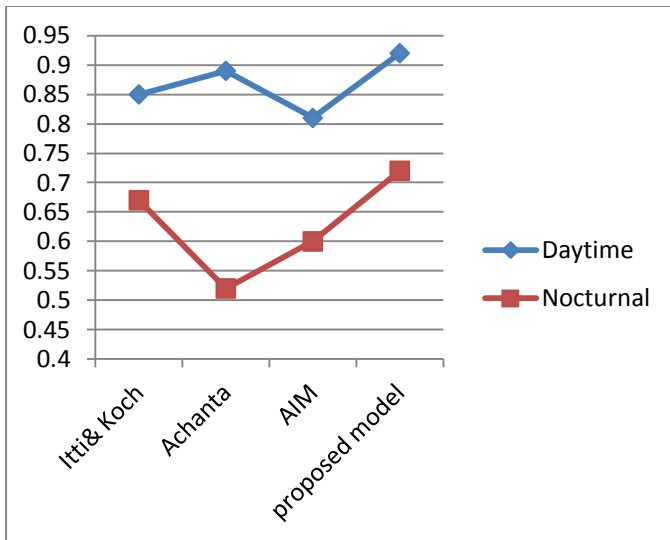


Fig.12. Evaluation of our proposed method under daytime and nocturnal scenes.

In figure 13, the horizontal axes are the proposed model by our model and other state-of-the-art models. We proposed three kinds of performance standards as the motion perception, eye fixations and nocturnal vision were compared with the ground-truth data (best result as 1), the vertical axes shows our results improved overall performance on these evaluation standards.

V. DISCUSSION AND CONCLUSION

In this paper, we proposed a new method to estimate the visual motion process on human visual attention and eye

fixations by constructing a computational model. This is a novel and state-of-the-art way. Besides, a serial of comparisons are implemented to test the robustness on the model via the day-time and nocturnal scenes. Unlike psychological methods, the technique using computer vision explains the human attention selection more vividly. This model can explain the attention selection mechanism and visual motion perception to some extent. With the proposed model, we analysis the motion perception under day time with single or multiple moving objects, we then mimic the visual attention process consisting of attention shifts and eye fixations against motion- feature-map. The model produced similar gist perception outputs in our experiments, when day-time images and nocturnal images from the same scene are processed. At last, we mentioned the future direction of this research.

We focus on the motion cues and the effects on the human visual system. Generally, the results are satisfactory and we are trying to simulate the motion effects in the top-down and bottom-up pathway. As they will leads to different outputs if we consider individual agents in the real world. The daytime and nocturnal vision is also compared via different approaches.

This paper has addressed the motion cues into the human visual model, however, in real life, motion perception are mostly irregular and abrupt. The video clips are selected from benchmark and normalized before the experiments. The robustness of algorithm needs improvement in next stage. Also, it is also believed that the visual neurons to respond to motion cues is vital for not only low level animals such as insects, but also import in the emergence of complex human brains [28][29][30][31][32].

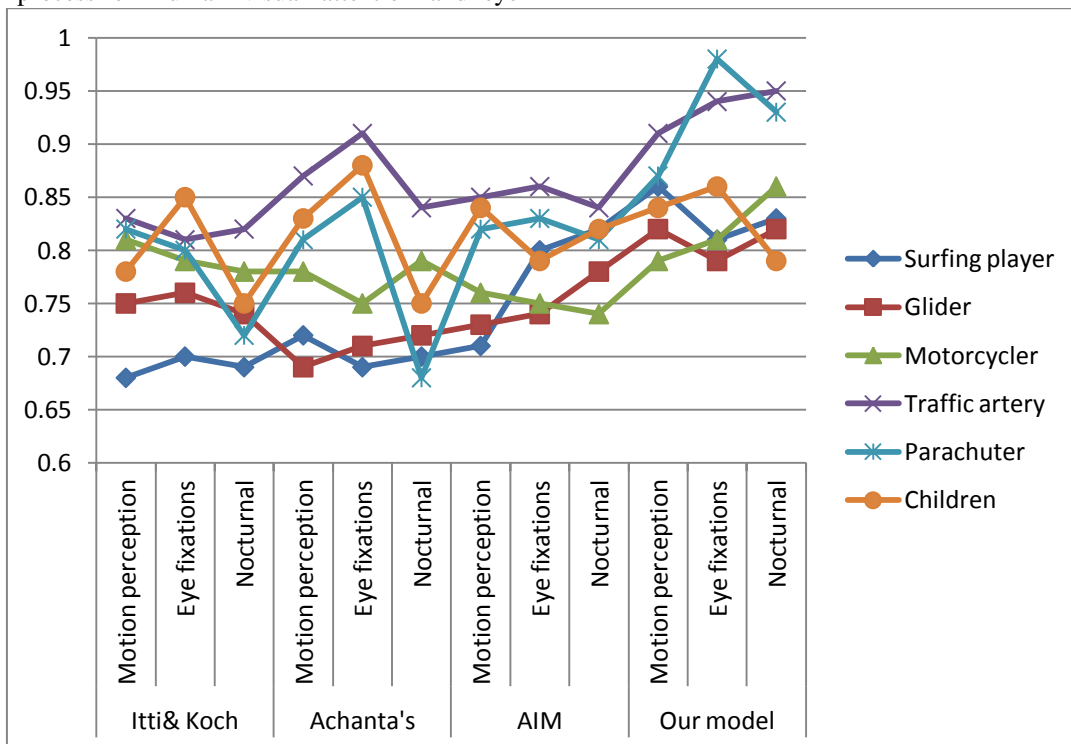


Fig.13. This figure indicates the precision (P), recall (R), F-measure comparisons between the proposed method and other state-of-the-art methods under various testing standards, such as motion perception eye fixations and nocturnal vision.

We will further integrate more motion cues into the attention model, and will implement these models to robots for efficient human robot interaction in the future. Another important factor is the top-down cues will affect our visual decision largely during the daily life, this issue has been proved by Yang [33] and other scholars [34]. The later stage is to intergate motion cues and top-downs cues together which can reflect the visual processing and enhance the model's robustness in the future work.

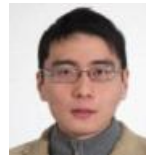
VI. ACKNOWLEDGEMENT

Thanks to all of the collaborators whose modeling work is reviewed here, and to the members of school of computer science, at the University of Lincoln, for discussion and feedback on this research. This work was supported by the grants of EU FP7-IRSES Project EYE2E (269118), LIVCODE (295151) and HAZCEPT (318907).

References

- [1] C. Koch. "The quest for consciousness", Roberts & Company Publishers, 2004.
- [2] J. Maunsell, D. Essen "Functional properties of neurons in middle temporal visual area of the macaque monkey. I. Selectivity for stimulus direction, speed, and orientation." J Neurophysiol 49 (5): 1127-47, 1983.
- [3] C. Rodman, "Afferent basis of visual response properties in area MT of the macaque. I. Effects of striate cortex removal". J Neurosci 9 (6): 2033-50, 1988.
- [4] W.Desimone, "Selective Attention Gates Visual Processing in the Extrastriate Cortex". Science 229(4715), 1985.
- [5] M.Mercier,Sophie, "Motion direction tuning in human visual cortex", European Journal of Neuroscience,Vol29,pp424-434,2009
- [6] YS Bonneh, , Motion-induced blindness and microsaccades: cause and effect, Journal of Vision, 10(14):22, 1-15, 2010.
- [7] J. Liu and Newsome, WT. Correlation between speed perception and neural activity in the middle temporal visual area. J. Neurosci. 25(3):711-722, 2005.
- [8] W. Newsome and EB Paré,. A selective impairment of motion perception following lesions of the middle temporal visual area (MT). J. Neurosci. 8: 2201-2211, 1988.
- [9] J. Tsotsos and A Rothenstein ,"Computational models of visual attention", Scholarpedia, 6(1):6201. doi:10.4249/scholarpedia.6201, 2011.
- [10] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in IEEE Conf. Computer Vision and Pattern Recognition,2007.
- [11] YF Ma, L Lu, HJ Zhang, M Li, "A user attention model for video summarization" ACM Multimedia, 2002..
- [12] C.W Oyster,"The human eye: structure and function". Sinauer Associates, 1999.
- [13] E.Strettoi, "Complexity of retinal cone bipolar cells". Progress in Retinal and Eye Research, 29 (4), pg. 272-283, 2010.
- [14] A.Roorda, D.R.Williams, "The arrangement of the three cone classes in the living human eye". Nature 397 (6719): 520-522, 1999.
- [15] R. W. G. Hunt . "The Reproduction of Colour" (6th ed.). Chichester UK: Wiley-IS&T Series in Imaging Science and Technology. pp. 11-12, 2004.
- [16] L. Itti, C. Koch, Feature Combination Strategies for Saliency-Based Visual Attention Systems, Journal of Electronic Imaging, Vol. 10, No. 1, pp. 161-169, Jan 2001.
- [17] M. Z. Aziz and B. Mertsching, "Fast and robust generation of feature maps for region-based visual attention," IEEE Trans. Image Process., vol. 17, no. 5, pp. 633-644, May 2008.
- [18] Alvarez, G.A., & Oliva, A. Spatial Ensemble Statistics: Efficient Codes that Can be Represented with Reduced Attention. *Proceedings of the National Academy of Sciences*, 106, 7345-7350, 2009..
- [19] SC Chong, A Treisman, Statistical processing: computing the average size in perceptual groups. Vision Research 45, 891-900, 2005.
- [20] ftp://ftp.cs.rdg.ac.uk/pub/PETS2001/
- [21] http://cim.mcgill.ca/~lijian/database.htm
- [22] http://vision.ucsd.edu/~bbabenko/project_miltrack.shtml
- [23] http://people.csail.mit.edu/tjudd/research.html
- [24] L. Itti, Quantitative Modeling of Perceptual Saliency at Human Eye Position, Visual Cognition, Vol. 14, No. 4-8, pp. 959-984, Aug-Dec, 2006.
- [25] R. J. Peters, L. Itti, Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention, In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jun 2007.
- [26] L.Bruce, N.D.B., Tsotsos, J.K., Attention based on information maximization. Journal of Vision, 7(9):950a, 2007.
- [27] R. Achanta and S. Süsstrunk, Saliency Detection for Content-aware Image Resizing, IEEE International Conference on Image Processing, 2009.
- [28] S. Yue and F.C. Rind, "Redundant neural vision systems - competing for collision recognition roles," *IEEE Transactions on Autonomous Mental Developments*, 2013 (in press).
- [29] S Yue and Rind F. Claire, "Postsynaptic organizations of directional selective visual neural networks for collision detection," Neurocomputing (in press), DOI: 10.1016/j.neucom.2012.
- [30] S Yue and Rind F. Claire, "Visually stimulated motor control for a robot with a pair of LGMD visual neural networks," IJAMEchS, 2012.
- [31] HY Meng , A Kofi, S Yue, H Andrew, H Mervyn, P Nigel, H Peter "Modified Model for the Lobula Giant Movement Detector and Its FPGA Implementation," Computer Vision and Image Understanding, vol.114(11), pp.1238-1247, 2010.
- [32] S. Yue and Rind F. Claire, "Collision detection in complex dynamic scenes using a LGMD based visual neural network with feature enhancement," IEEE Transactions on Neural Networks, vol.17(3), pp.705-716, 2006.
- [33] J.Yang, M.Yang: Top-down saliency via joint CRF and dictionary learning. CVPR 2012: 2296-2303
- [34] C.Kanan, M. H., Zhang, G. W. SUN: Top-down saliency using natural statistics. Visual Cognition, 17: 979-1003, 2009.

AUTHORS PROFILE



Jiawei Xu received the B.S. and M.S. degrees in computer engineering from Shanghai University of Engineering Science and Technology, Shanghai, China, 2007 and Hallym University, Korea, 2010, respectively. Now he is a PhD student in the School of Computer Science, University of Lincoln, UK. His research interests include computer vision, human attention models, and visual cortex modeling. He was a pattern classification engineer in JTV Co.Ltd, Beijing during the year of 2011.



Shigang YUE is a Professor of Computer Science in the Lincoln School of Computer Science, University of Lincoln, United Kingdom. He received his PhD and MSc degrees from Beijing University of Technology (BJUT) in 1996 and 1993, and his BEng degree from Qingdao Technological University (1988). He worked in BJUT as a Lecturer (1996-1998) and an Associate Professor (1998-1999). He was an Alexander von Humboldt Research Fellow (2000, 2001) at University of Kaiserslautern, Germany. Before joining the University of Lincoln as a Senior Lecturer (2007) and promoted to Reader (2010) and Professor (2012), he held research positions in the University of Cambridge, Newcastle University and the University College London(UCL) respectively. His research interests are mainly within the field of artificial intelligence, computer vision, robotics, brains and neuroscience. He is particularly interested in biological visual neural systems, evolution of neuronal subsystems and their applications – e.g., in collision detection for vehicles, interactive systems and robotics. He is the founding director of Computational Intelligence Laboratory (CIL) in Lincoln. He is the coordinator for several EU FP7 projects. He is a member of IEEE, INNS, ISAL and ISBE.

A Modified clustering for LEACH algorithm in WSN

B.Brahma Reddy
ECE, VBIT, Hyderabad,
India

K.Kishan Rao
Vaagdevi College of Engineering,
Warangal, India

Abstract—Node clustering and data aggregation are popular techniques to reduce energy consumption in large Wireless Sensor Networks (WSN). Cluster based routing is always a hot research area in wireless sensor networks. Classical LEACH protocol has many advantages in energy efficiency, data aggregation and so on. However, determining number of clusters present in a network is an important problem. Conventional clustering techniques generally assume this parameter to be user supplied. There exist very few techniques that can solve the problem of automatic detection of number of clusters satisfactorily. Some of these techniques rely on user supplied information, while others use cluster validity indices. In this paper, we proposed a rather simple method to identify the number of clusters that can give satisfactory results. Proposed method is compared with classical LEACH protocol and found to be giving better results.

Keywords—Clustering Index; LEACH; Wireless Sensor Networks; Energy optimization; Network lifetime

I. INTRODUCTION

Node clustering and data aggregation are popular to reduce energy consumption in large Wireless Sensor Networks (WSN). Clustering in WSN is the process of dividing the nodes of WSN into groups, where each group agrees on a central node, called the Cluster Head (CH), which is responsible for gathering the sensory data of all group members, aggregating it and sending to Base Station (BS). Cluster based routing is always a hot research area in wireless sensor networks. Classical LEACH protocol has many advantages in energy efficiency, data aggregation and so on. However, determining number of clusters present in a network is an important problem. Conventional clustering techniques generally assume this parameter to be user supplied. There exist very few techniques that can solve the problem of automatic detection of number of clusters satisfactorily. Some of these techniques rely on user supplied information, while others use cluster validity indices which need additional computation time. There are several indexes such as Dunn's, PBM, Davis-Bouldin, Global and Mahalanobis distances, SVD entropy, Krzanowski and Lai, Hartigan, Silhouett, Gap Statistic proposed by earlier authors for cluster validity. We define the optimal clustering as the one which gives data transmission from the cluster members to CH and subsequently from CH to BS incurs the minimal energy or maximize total transmissions. In this paper, we proposed a simple method to identify the number of clusters that can give increased number of transmissions.

Rest of the paper is organized as follows: Section II explains a few cluster validity indexes; Section III highlights related work done by other authors; Section IV describes the

network model used; Section V presents simulation results and analyses; finally Section VI concludes observations.

II. CLUSTER VALIDITY INDEXES

Assignment to clusters relies on a distance measure; in the case of genetic algorithms, the criterion function of the optimization is called the fitness function. Here we present some measures which we tested with our algorithm. As a general guideline, these measures should favor for minimal differences between points within the cluster (intra-cluster, DCH) and maximal differences between points of different clusters (Inter-cluster, DBS).

In the original LEACH protocol, the probability corresponds to the number of desired CHs in the network. Additional metrics such as remaining node energy can also be used to change the clustering properties. LEACH divides the whole network into several clusters, and the run time of network is broken into many rounds. In each round, the nodes in a cluster contend to be cluster head according to a predefined criterion. In LEACH protocol, all the sensor nodes have the same probability to be a cluster head, which makes the nodes in the network consume energy in a relatively balanced way so as to prolong network lifetime. However, number of clusters may vary in each round. Because of this reason, network lifetime can not be defined in terms of rounds. Better definition for network lifetime is in terms of number of transmissions. Each sensor node n decides independently of other sensor nodes whether it will claim to be a CH or not, by picking a random s between 0 and 1 and comparing s with a threshold function value $T(n)$ based on a user-specified probability p . If $s \leq T(n)$ then the node claims to become CH. The threshold is defined as follows [1]:

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

Where G is the set of nodes that have not been CHs in the last $1/p$ rounds.

In the proposed algorithm, number of clusters is fixed and predetermined using cluster validation criteria. And each round one of the cluster members become cluster head. Thus, there is one packet transmission per round from each CH. To understand the components of the energy consumed, we separate the cost into intra and inter cluster energy consumption. Intuitively, as cluster size grows, energy consumption inside the cluster also grows. At the same time energy consumption from cluster heads to the base station drops significantly, since fewer CHs are present. For finding

optimal cluster combination we validate certain characteristics given below:

$$\text{Average Cluster distance } d_{CH} = \frac{\sum_{i=1}^n d(i,CH)}{n}$$

$$\text{Average BS distance } d_{BS} = \frac{\sum_{i=1}^k d(i,BS)}{k}$$

Conditions

1. Distance to CH must be less than distance to BS

$$\frac{d_{CH}}{d_{BS}} \leq 1 \quad (1)$$

2. Energy consumed must be less for route via CH compared to direct to BS

$$d_{CH}^2 + d_{BS}^2 \leq d_{DirBS}^2 \quad (2)$$

3. None of cluster 'i' must be closer to CH of cluster 'j'
 $d_{BSi} - d_{BSj} \geq (d_{CHi} | d_{CHj})$ (3)

4. Energy consumed by CH for a transmission
 $\left(\frac{n}{k} - 1\right) \cdot E_{RX} + \left(\frac{n}{k} - 1\right) \cdot E_{DA} + E_{elec} +$

$$\varepsilon_{fs} \cdot d_{BS}^2 = E_{BS}$$

5. Energy consumed by a cluster member to forward data to CH

$$E_{elec} + \varepsilon_{fs} \cdot d_{CH}^2 = E_{CH}$$

Where n in number cluster members nodes and k is number of clusters. E_{RX} , E_{elec} , E_{DA} , ε_{fs} are Receive energy, Transmit energy, Data aggregation energy and free space loss respectively.

$$\sigma_{BSi} = \sqrt{E[BS_i^2] - E^2[BS_j]} \text{ where } j \neq i \text{ (Cluster } i) (i = 1 \dots k) \text{ ---Standard deviation in remaining energy of different Clusters. It must be close to zero. Global best (4)}$$

$$\sigma_{CHi} = \sqrt{E[CH_i^2] - E^2[CH_j]} \text{ where } j \neq i \text{ (Cluster } i) (i = 1 \dots \text{cluster-size) ----- Standard deviation in remaining energy of nodes in a cluster. It must be close to zero. Local best (5)}$$

Equations 1, 2 and 3 ensures clustering saves energy. Equations 4 and 5 give Global best and local best position of CHs.

Automatic determination of number of clusters present in a wireless sensor network has been a challenging problem to the researchers. There are two aspects of this problem: i) finding number of clusters and ii) finding the clusters themselves. Majority of the existing techniques assume the number of clusters as an input parameter to be supplied by the user. One of the most common techniques is k-means algorithm [1]. The k-means algorithm is a simple partitioned clustering algorithm. The objective of this algorithm is to partition the given data set S containing N data elements $\{x_1, x_2, \dots, x_N\}$ into k clusters, $\{C_1, C_2, \dots, C_N\}$ such that

$$C_i = \emptyset \text{ for all } i$$

$$C_i \cap C_j = \emptyset \text{ if } i \neq j$$

$$C_1 \cup C_2 \cup \dots \cup C_k = S$$

However, the unanswered question is which partition represents the best clustering solution. This question may be answered if we perform some test for the tendency of clustering of the concerned data set before clustering it.

In this paper, we used the cluster area as input to determine the number of clusters. Group of nodes within radius of clustering area are formed as a separate cluster. If any node is falling in more than one group, then the node will be retained in the cluster where the node is closer. Thus a node is ensured to present only in one cluster. In order to ascertain the characteristics of the clusters, we used Dunn's Index, PBM Index and Davis-Bouldin Index.

- A. **Dunn's index** [2]. For any k -partition Dunn defined the following index:

$$V_D = \min_{1 \leq i \leq K} \left\{ \min_{1 \leq j \leq K, j \neq i} \left\{ \frac{\delta(C_i, C_j)}{\max_{1 \leq k \leq K} \Delta(C_k)} \right\} \right\}$$

Where

$$\Delta(C_k) = \max_{x,y \in C_k} \{d(x,y)\}$$

$$\delta(C_i, C_j) = \min_{x \in C_i, y \in C_j} \{d(x,y)\}$$

Larger values of V_D corresponds to good clusters and the number of clusters that maximize V_D is taken as the optimal number of clusters. $d(x,y)$ is the distance between node x and node y .

- B. **PBM Index**: It is defined as

$$PBM(K) = \left(\frac{1}{K} \frac{E_1}{E_K} D_K \right)^2$$

Where K is the number of clusters, E_K and D_K represent sum of within cluster dispersions and maximum between cluster separations respectively.

- C. **Davis-Bouldin (DB) Index**: This index is a function of the ration of sum of within-cluster scatter to between-cluster separation.

$$DB = \frac{1}{K} \sum_{i=1}^K R_{i,qt}$$

$$R_{i,qt} = \max_{j,j \neq i} \left\{ \frac{S_{i,q} + S_{j,q}}{d_{ij,t}} \right\}$$

Where $S_{i,q}$ is the q th moment of the points in cluster C_i with respect to their mean? And it is a measure of the dispersion of the points in that cluster. $d_{ij,t}$ is the Minkowski distance of order t between the centroids that characterize clusters C_i, C_j . The objective is to minimize the DB index for achieving proper clustering.

III. RELATED WORK

This paper is not the first to analytically evaluate clustering techniques. In a very recent [3], the author addresses problem in terms of number of sensors in an optimal cluster. Jamshid Shanbehzadeh, Saeed Mehrjoo, Abdolhossein Sarrafzadeh [4] proposed a hybrid GA-PSO based clustering algorithm that improved the lifetime of WSN effectively. In [5] O.A.Mohamed Jafar and R.Sivakumar carried a survey on ant-based clustering algorithms. Akramul Azim, Mohammad Mahfuzul Islam [6] introduced relay nodes to act as cluster heads and decrease the probability of premature death of original nodes. Jin Wang, Xiaoqin Yang, Yuhui Zheng, Jianwei Zhang, Jeong-Uk Kim [7] proposed energy-based clustering algorithm and demonstrated improved energy efficiency. GRASP based algorithm was proposed for cluster formation problem by Victor de Oliveira, Matos, Jose Elias C.Arroyo, Andre Gustavo dos Santos, Luciana B.Goncalves [8]. In [9] S.Rao Rayaoudi proposed a novel approach based on intelligent water drops algorithm to solve economic load dispatch problem. Liu Ban-teng, Chen You-rong, Zhou Kai, Jingyu Hua [10] proposed Boolean sensing model based on Poisson point process to identify the function of the rate of coverage and the node density in unit area, and then calculates the total number of nodes in the region, next uses the greedy strategy of the Prim algorithm to find a spanning tree with the maximum weight, and constructs a approximate solution for the minimum connected dominating set. Malay K Pakhira [11] proposed Visuval Assessment of Tendency based algorithm for automatic determination of number of clusters identification. Anna Forster, Alexander Forster, Amy L. Murphy [12] presented an experimental analysis for optimal cluster sizes. Benjamin Auffarth [13] presented a genetic algorithm that is fast and able to converge on meaningful clusters for real-world data sets and discussed cluster validity criteria. Jianguo SHAN, Lei DONG, Xiaozhong LIAO, Liwei SHAO, Zhigang GAO, Yang GAO [14] presented another improved version of LEACH protocol to extend life cycle of the network.

IV. SIMULATION MODEL

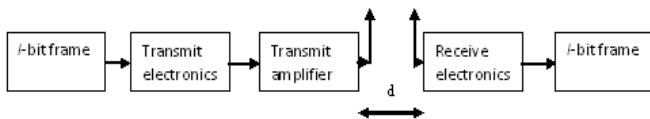


Fig.1. Radio energy model

We define WSN to be a two-dimensional network graph $G = (V, E)$ where V and E are set of Vertices (Wireless sensor nodes) and Edges (Transmission links) respectively. There is a link from sensor u to sensor v if and only if v is located in u 's transmission range, then v is called neighbor of u . All nodes are equipped with adjustable transmit power. Initially, transmit power level is fixed to transmission range R . However, when acting as CH, transmit power can be adjusted to communicate with BS. We assume clusters are circles with a transmission range of R . Nodes can communicate to all their neighbors, defined as those nodes whose distance is less than R . Energy is spent when a node sends as well as receives a packet. Energy

model of the network is shown in Fig1. Two-tier model is considered in this paper. Cluster members communicate with CH and CH aggregates the data and communicates with BS. We compute clusters and its members as given in Algorithm below.

Algorithm to computer clusters

```

Input :  $G = (V, E), R$ 
 $C = (C_1, C_2, \dots, C_k) = \emptyset$ 
repeat till all  $V_i$  are covered
  for  $i = 1$  to  $k$ 
    Select  $V_i$  next highest node – degree
     $C_i = C_i \cup V_i \cup$  all one – hop neighbors of  $V_i$ 
  endfor
endrepeat
repeat for all  $C_i$   $i = 1..k$ 
  for all members of  $C_i$ 
    if  $d(\text{member}, C_i)$  is minimum retain
    else remove from  $C_i$ 
  endrepeat
endrepeat
Output:  $C$ , cluster groups and members in each group
    
```

V. SIMULATION RESULTS

In this section, we evaluate the performance of our proposed algorithm through the simulations with respect to classical LEACH algorithm. We compare our proposed algorithm with LEACH based on three performance metrics: i) spread of the dead/alive nodes at each round; ii) Number of packets transmitted at FND(First-node dead), HND (Half nodes dead), AND (All Nodes Dead); iii) Cluster indices of Dunn's, PBM and DB. The reference network of our simulations consists of 100 nodes distributed randomly in an area of $100 \text{ m} \times 100 \text{ m}$. The BS is located at position (50, 50). Here we use the typical values $E_{elec} = 50 \text{ nJ/bit}$, $\epsilon_{fs} = 10 \text{ pJ/bit/m}^2$ and $\epsilon_{mp} = 0.0013 \text{ pJ/bit/m}^4$. As noted previously, the cluster heads are responsible for aggregating their cluster members' data. The energy for data aggregation is set as $E_{DA} = 5 \text{ nJ/bit/signal}$. The initial energy of all nodes set to 0.1 J. Every node transmits a 4000-bit message per round to its cluster head. Here we assume every node having knowledge of other node position and compute distance from BS. We normalize the distance of each node to longest node distance from BS. We computed Cluster indices for R value ranging from 0.1 to 0.9 and found to be optimum for 0.3. With this we assumed $R=0.3$ for comparison of the proposed algorithm with LEACH. p is set to 0.1 (about 10% of nodes per round become cluster heads) for LEACH. For the proposed algorithm, one of the live cluster members with highest remaining energy becomes CH in each round. Thus there is one packets transmission for every round from each cluster in the proposed algorithm. In LEACH, there can be a round without CH and thereby no packet transmission. Hence, we use number packet transmissions to BS instead of number of rounds as a measure of network lifetime. Table1 gives the comparison of Indexes values for both LEACH and proposed algorithm. As mentioned earlier in this paper, numbers of cluster heads vary

in each round for LEACH. For comparison purpose we have taken number of cluster heads of first round in LEACH protocol. But in case of proposed algorithm, number of cluster heads is same in each round. So, all Indexes computed for LEACH are related to first round only. Fig 2 gives the comparison of Packets transmitted to BS at various stages. Fig 3 gives the comparison of live nodes with respect to rounds. Here, it may be observed that though the number of rounds increases in case of LEACH, total number packets transmitted do not increase. This happens because in some rounds there may not be any CH. But in case of proposed algorithm, there will be packet transmission in every round from every cluster unless all nodes of a cluster die.

TABLE I. Comparison of Cluster Validity Indexes

Index	LEACH	Proposed algorithm
Dunn's	0.1148	0.4253
PBM	0.00007	0.0006
DBI	1.8644	2.3431
Number of clusters	9	12

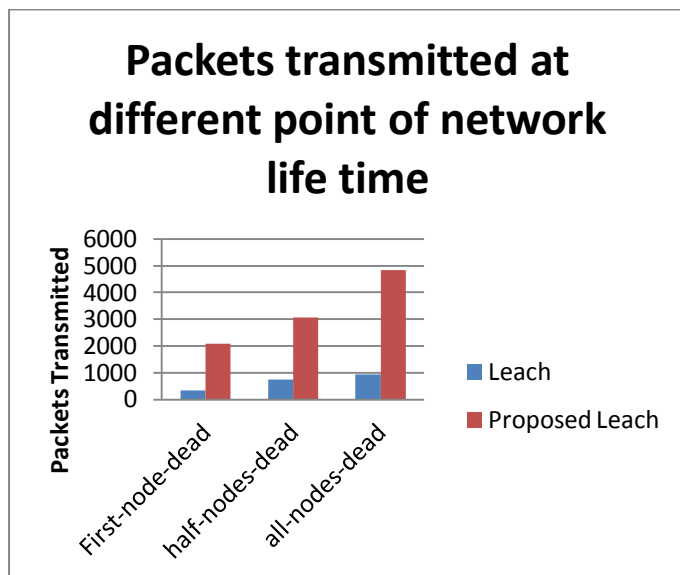


Fig.2. Number of packets transmitted

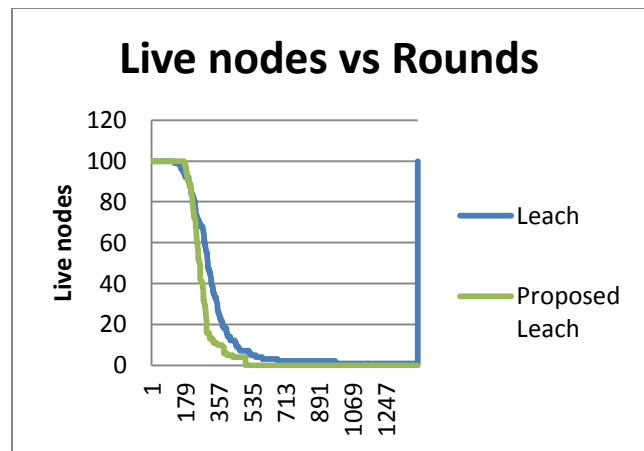


Fig.3. Number of live nodes vs Rounds

VI. CONCLUSION

In this paper, we considered a well known energy efficient clustering algorithm for WSNs called LEACH algorithm and proposed a new clustering algorithm for improved network lifetime. In this new approach, we used fixed number of clusters and cluster number is computed based on validity tests. This makes the optimum number of clusters than in LEACH algorithm and hence increases network lifetime. The simulation results show that the proposed algorithm can make the remaining energy more uniform throughout the network at every round. In the present work we assumed E_{DA} , E_{elec} to be constant. However, they can also influence the energy consumption per round. In future work, we can improve our algorithm for including variation due to E_{DA} , E_{elec} condition.

References

- [1] J.B. Ball and D Hall, "Some methods of classification and analysis in multivariate observations", Proc. of fifth Berkeley symposium on mathematical and probability, pp-281-297, 1967
- [2] T.Havens, J.C. Bezdek, J.M.Keller, and M.Popescu, "Dunn's cluster validity index as a contrast measure of VAT images", Proc. ICPR, Tampa, Florida, 2008
- [3] Wang,D, "An energy-efficient cluster head assignment scheme for hierarchical wireless sensor networks", International journal of Wireless Information Networks, 15(2), 2008, pp-61-71
- [4] Jamshid Shanbehzadeh, Saeed Mehrjoo, Abdolhossein Sarrafzadeh, "An Intelligent energy efficient clustering in wireless sensor networks", Proc. of International multi conference of engineers and computer scientists, Vol-1, March 16-18, 2011, Hongkong.
- [5] O.A.Mohamed Jafar and R.Sivakumar, "Ant-Based clustering algorithms: A brief survey", International journal of computer theory and engineering, Vol-2, No-5, October, 2010, pp-787-796
- [6] Akramul Azim, Mohammad Mahfuzul Islam, " A relay node based hybrid low energy adaptive clustering hierarchy for wireless sensor networks", International journal of energy, information and communications, vol-3, Issue-3, August, 2012, pp41-53
- [7] Jin Wang, Xiaoqin Yang, Yuhui Zheng, Jianwei Zhang, Jeong-Uk Kim, "An Energy-based Clustering Algorithm for wireless sensor networks", International Journal of Future Generation Communication and Networking Vol. 5, No. 4, December, 2012, pp-89-98

- [8] Victor de Oliveira, Matos, Jose Elias C. Arroyo, Andre Gustavo dos Santos, Luciana B. Goncalves, "AN Energy Efficient Clustering algorithm for wireless sensor networks", IJCSNS International journal of computer science and network security, Vol-12, No-10, October 2012, pp-6-15
- [9] Rao Rayaoudi, "An Intelligent Water Drop Algorithm for Solving Economic Load Dispatch Problem", International journal of Electrical and Electronics Engineering 5:1, 2011, pp-43-49
- [10] Liu Ban-teng, Chen You-rong, Zhou Kai, Jingyu Hua, "The research of wireless sensor networks optimization algorithm based on control", Third International Symposium on Information Processing, 2010, pp-420-422
- [11] Malay K Pakhira, "Finding number of clusters before finding clusters", Procedia Technology 4 (2012) 27-37, published by Elsevier Ltd.
- [12] Anna Forster, Alexander Forster, Amy L. Murphy, "Optimal cluster sizes for wireless sensor networks: An experimental analysis", ADHOCNETS 2009 LNICST 28 pp-49-63, 2010
- [13] Benjamin Auffarth, "Clustering by a genetic algorithm with biased mutation operator", 2010 IEEE WCCI CEC. Barcelona, Spain, July 18-23, 2010. doi:10.1109/CEC.2010.5586090.
- [14] Jianguo SHAN, Lei DONG, Xiaozhong LIAO, Liwei SHAO, Zhigang GAO, Yang GAO, "Research on Improved LEACH Protocol of wireless sensor networks", PRZEGLĄD ELEKTROTECHNICZNY Vol 2013, No 1b

AUTHORS PROFILE



VBIT.



B. Brahma Reddy obtained his B.Tech from JNTU and M.Tech from IIT, Madras in 1980 & 1982 respectively. He has worked for Indian Institute of Science, Indian Telephone Industries, National Informatics Centre, DishnetDSL, Reliance Infocomm for nearly 25 years. Past 6 years he is working as Professor in JNTU affiliated college. Currently he is pursuing his Doctoral programme and working for

Prof. K. Kishan Rao obtained his B.E., M.E., degrees both in Distinction from O.U. College of Engineering, Osmania University in 1965 and 1967 respectively and Ph.D. from IIT Kanpur in the year 1973. Joined Regional Engineering College, Warangal in 1972 and Retired as Principal in 2002. He has Guided 3 Ph.D. Candidates, 76 M.Tech. Dissertations, 5 Research Projects and published about 68 Technical Papers in National and international Journals and Conferences. He is a Senior Member of IEEE, Life Member of IETE, Life Member of ISTE and Life Member of APSA. His Research Interest are Wireless & Mobile Communications, Adaptive Digital Signal Processing and OFDMA Networks. Presently working as Director and Adviser to Vaagdevi Group of Technical Institutions, Warangal.

Jabber-based Cross-Domain Efficient and Privacy-Ensuring Context Management Framework

Zakwan Jaroucheh, Xiaodong Liu, Sally Smith

School of Computing
Edinburgh Napier University
10 Colinton Road, EH10 5DT, Edinburgh, UK

Abstract—in pervasive environments, context-aware applications require a global knowledge of the context information distributed in different spatial domains in order to establish context-based interactions. Therefore, the design of distributed storage, retrieval, and dissemination mechanisms of context information across domains becomes vital. In such environments, we envision the necessity of collaboration between different context servers distributed in different domains; thus, the need for generic APIs and protocol allowing context information exchange between different entities: context servers, context providers, and context consumers. As a solution this paper proposes *ubique*, a distributed middleware for context-aware computing that allows applications to maintain domain-based context interests to access context information about users, places, events, and things - all made available by or brokered through the home domain server. This paper proposes also a new cross-domain protocol for context management which ensures the privacy and the efficiency of context information dissemination. It has been robustly built upon the Jabber protocol which is a widely adopted open protocol for instant messaging and is designed for near real-time communication. Simulation and experimentation results show that *ubique* framework well supports robust cross-domain context management and collaboration.

Keywords—pervasive computing; cross-domain context management; context modeling; Jabber protocol; privacy.

I. INTRODUCTION

In the emerging and challenging pervasive environments, users will wear smart clothes that will monitor their bio signals; they will carry smart cards that will handle automatically their transactions; invisible chips will be embedded everywhere in the smart homes and offices to assist them in their daily life tasks; more sophisticated control navigation and control will be embedded into their vehicles. All these devices will cooperate together to create a context-aware pervasive environment that supports humans in everyday activities, e.g., business, health care, or education. In this respect, the user will enjoy a new experience in a non-obtrusive way as the existing infrastructures will be more proactive and dynamically adaptable to current situations; user preferences; and environmental context in a less intrusive way [1]. Context-awareness is the cornerstone to achieve the vision of such a pervasive environment. It helps to support non-intrusive adaptability of applications to new situations and to turn a static computing environment into a dynamic ecology of smart and proactive applications [2].

In this paper, we base our context management framework on the notion of context domain explained in [3] which organizes the pervasive environment hierarchically and establishes a context management scope. A context domain is defined as an abstraction of a spatial area which has a clear boundary and it is built on top of the traditional notion of network domain. Essentially, context domain establishes (i) the place and responsibility of context instances storage; (ii) the responsibility for managing context providers and consumers inside the domain; and (iii) a set of sub-domains.

Although users are more interested in context information related to their location, other context information from other domains may also be relevant to the current task at hand. For instance, a dynamic recalculation of the quickest routes for a trip involves acquiring the latest contextual information such as traffic congestion from remote sources. In this respect, we can imagine a domain-based context management system where the context information available in each domain is managed by a separate context server. While moving, the user roams across domains. In addition, each domain may maintain its own sensors and mechanisms for inferring context related to this user. Consequently, collaborative context management across domains is needed.

In particular, an efficient cross-domain context management middleware system for such a setting needs to fulfil key requirements that include: (i) domains of context perception, (ii) uniform API interface for accessing context servers, (iii) efficient context information dissemination, (iv) support of cross-domain reasoning, (v) dynamic matching between context providers and consumers, and (vi) support for privacy. In this paper, we propose *ubique*, a new domain-based context management infrastructure for disseminating context information between context providers, context consumers and context servers, and a set of APIs for interfacing between these entities. *ubique* fulfils the above mentioned key requirements and it forms an underlying robust and generic infrastructure for context management, which significantly simplifies the development of context-aware pervasive applications.

This paper is structured as follows. In Section 2, we outline the different requirements that should be fulfilled by a cross-domain context management system. Then in Section 3, we critically review the advantages and weaknesses of existing solutions with respect to the defined criteria. Section 4 describes the context dissemination problem. In Section 5, we detail our new proposed approach for context management.

Finally, we evaluate the proposed approach by mean of real experimentation and simulation and we draw conclusions.

II. REQUIREMENTS AND CHALLENGES

Hereafter, we refer to the computational entity responsible for transparently binding the context consumers (CCs) (i.e. applications) with corresponding context providers (CPs) a context server (CS). The context management in each domain is done by the context server available in that domain. The complexity of developing context-aware applications that require context information available in different CSs makes the use of a cross-domain context management middleware crucial. From our pilot experiments and literature analysis, we identify that a middleware for such a setting must fulfil the key requirements such as:

Domains of context perception: Since the context information is naturally distributed, the context management must be distributed in order to allow efficient and scalable dissemination of context. However, the task of context-aware developers becomes more difficult as it requires a priori knowledge of the computational entities responsible for providing the context information they are interested in. Their task becomes even more complex when context providers dynamically enter and leave the pervasive environment. Thus, there is a need for a dynamic discovery mechanism of context providers.

Furthermore, the middleware scalability could be increased by restricting the access and perception of the context to some domains [3]. Moreover, as we will see later, the notion of home domain server reduces the number of CSs that may be involved in the resolution of context interests. This requirement conforms to the principle of system boundary [4] of pervasive applications.

Uniform API interface and protocol: In order to enable every party to become a context provider and implement its own CS, every CS should: (i) obey a certain protocol with which context information can be federated between different CSs; and (ii) implement a standard API which allows context providers to register and publish context information in it, and context consumers to acquire context information they are interested in. This way, for instance, an organization can operate a CS for its members, and an individual can run a CS as a context provider for a single user or family members. Therefore, similar to the Next Generation Service Interfaces (NGSI) [5], providing a standard API for accessing such information, allows third party application developers to build new services based on the context made available to them.

Efficient context information dissemination: With regard to situations involving mobile users roaming across domains, additional restrictions may arise (e.g. concerning limited connectivity and bandwidth, unknown network conditions, etc.), thus exchanging context information between domains should be fast and only the required information should be transferred when users roam across domains. This requirement calls for a federation protocol between CSs. Furthermore, the middleware should support the “publish on demand” mode of operation. That is, usually context providers publish context constantly and independently of existing consumers. In this

case if a context provider publishes at a higher rate the context information is more accurate in terms of freshness. However, this is a costly operation in terms of the network bandwidth usage (i.e. increase of the number of messages sent through network), processing power, and energy consumption (e.g. battery usage of WiFi scanners). Thus, the middleware should enable providers to publish when there is a corresponding consumer.

Cross-domain reasoning: As the context information is originated from different domains, a cross-domain context management system should facilitate the context information reasoning that spans multiple domains. That is, in order to track user’s behaviour there is a need to consider the context information available in the different domains the user visits [6]. Hence, understanding the user’s current situation may require considering the different states the user experienced in these domains. For example, to identify if the current day was busy for the user there is a need to consider the different activities and states the user has experienced in work, shopping, on the road, etc.

Dynamic matching between context providers and consumers: Typically developers define context interests which should be transparently kept across distributed CSs. The main challenge in such dynamic environment is therefore to accommodate changes on the environment without infringing active context interests. The middleware should allow the context consumers (applications) to register their interests in context information; and the context providers to register their capabilities. Then, for any change in either the context consumers or providers, a matching function should be triggered so that applications asynchronously receive notifications of context information that match their interests. In addition, the application should be able to specify its context interests on the basis of context types and meta-attributes such as precision and accuracy and to indicate additional restrictions based on properties of the provider or the context publication. In this case, the middleware has to be responsible for choosing the most adequate context providers among a dynamic set of available ones.

Support for privacy: The flow of context information between different distributed domains obviously raises user privacy issues. A cross-domain system should protect user’s information and guarantee privacy across domains. As we will see later the usage of a home domain server provides an interesting approach for control privacy of context access, since it is a central point of access for a given entity’s context. A user can control the context dissemination for some consumers through modifying its privacy policy published in his home domain server.

III. LIMITATIONS OF CURRENT APPROACHES

Classical work in context-aware computing has developed centralized and application-specific solutions such as Context Toolkit [7] which provides a set of abstractions that can be used to implement reusable software components for context sensing and interpretation. The context information is directly acquired from a sensor by means of the context widget component. Widgets can be combined with interpreters, which transform low-level information into higher-level information

that is more useful to applications, and aggregators, which group related context information together in a single component. Finally, context-aware applications can invoke actions using actuators, and locate suitable widgets, interpreters, and aggregators using discoverers. Another interesting work is Gaia [8] which adopts the concept of active spaces, which are physical spaces where devices in a heterogeneous network, such as PDAs and printers, can discover each other, auto-configure and dynamically start a context-aware interaction. It provides a framework to develop user-centric, resource-aware, multi-device, context-sensitive and mobile applications. However, these approaches offer solutions for restricted and small-size smart spaces environments, with localized scalability.

GLOSS [9] composes heterogeneous context management systems through hierarchical or peer-to-peer interconnection methods. By introducing the notion of Global Smart Spaces, GLOSS supports interaction amongst people, artifacts and places while taking account of both context and movement on a global scale that facilitates the implementation of location-aware services. It allows users to pick up small notes left for them in the environment. GLOSS uses the idea of home nodes, however, it has been designed to manage location context only.

More recent middleware offers access to context information in distributed repositories. For example, the Context Fabric (Confab) [10] provides architecture for privacy-sensitive systems, as well as a set of privacy mechanisms that can be used by application developers. It maintains context information in distributed tuple-spaces called infospaces. Each infospace is a repository responsible for storing one or more context types. An application interested in a certain context, builds a context query using the address of the responsible infospace. In order to handle queries over distributed infospaces, Confab offers a query processing service, which distributes queries over distributed infospaces and composes the query results. Privacy is supported by adding operators to an infospace to carry out actions when tuples enter or leave the space. However, as Confab focuses so heavily on privacy, it does not adequately address the other middleware requirements such as mobility or context information dissemination across domains.

The scalability issue is considered in PACE [11], which is another distributed middleware focusing on offering a flexible context model called CML (Context Modeling Language) and advanced context-based programming abstractions for distributed context-aware applications. PACE is organized in layers that provide, in addition to context management, an interface to execute distributed context queries, and an adaptation layer, which maintains a reusable repository of adaptation abstractions. Applications use a catalog and meta-attributes to discover which repository satisfies their context requirements. However, when a user roams across domains, this discovery mechanism does not allow developers to identify the repositories existing in the domains visited by the roaming user which contain his context information.

CAMUS [12] is another distributed middleware where context-aware system federation is composed by environments based on CAMUS services, which disseminate context

information as tuples, in order to increase dissemination efficiency. Each service of an environment must be registered in a Jini discovery service. A CAMUS context domain is an environment that supports a minimum set of CAMUS services. The set of Jini services responsible for each CAMUS domain composes a federation. In order to access context information or to use a service of a specific domain, a client must query the Jini federation, using parameters such as the name and localization of the domain. CAMUS, however, does not address cross-domain context dissemination and how to ensure user's privacy.

Another interesting approach to allowing distributed context management based on federating context-aware services is Nexus [13]. Nexus supports heterogeneity among context management systems' context models, i.e. each context management system can adopt a particular context model and must implement an abstract interface and register itself at an Area Service Register. Thus, it focuses on the data management aspect of large-scale pervasive computing systems. A client may access context information provided by the federation, by using a query language. However, there is no concept such as domain or environment: each context server is a repository of a specific context type [3].

The Context Management Framework (CMF) proposed in MobiLife project [14][15] is designed for the discovery of, exchange of, and reasoning on context information. It is a set of components, which are connected at run time, that together provide the relevant context information for the service or application, using sensing and interpretation mechanisms. The main tasks for the CMF are to enabling the discovery of context providers, to provide a published agreement or interface contract between context providers and context consumers, and binding context consumers with the matched context providers in order to use their context service functions through the use of context broker. Therefore, in CMF there is no concept such as domain so that the application is able to specify the domain(s) from which the context information is originated. In addition, the infrastructure needed for setting and enforcing privacy of user-controlled data available through context providers is controlled by the Trust Engine. However, we believe that this setting weakens enforcing the privacy since a malicious context provider can skip contacting the trust engine to verify if the context consumer is eligible to access the context information; thus a centralized trusted entity responsible to enforce the privacy is needed.

ICE [16] is a scalable context management middleware for Next Generation Networks. It is based on the concepts of context sessions and context flows. The idea is to separate signaling data from content exchange, as in IP Multimedia Subsystem, to establish context sessions for more scalable and adaptive management of context information. The Context Access Language (CALA) has been designed to support context queries and subscriptions. However, ICE focuses heavily on efficient context information dissemination between context sources and sinks. Thus, it ignores in its designed protocols ensuring entities privacy. In addition, context sources' descriptions and context sinks' queries/subscriptions must be registered in a centralized entity - the context broker. Thus, as the user roams between domains, this adds complexity

to the developers as they must know in advance which context broker they have to contact to get the context information they are interested in.

From the perspective of globally connecting sensors, the Open Geospatial Consortium provided the Sensor Web Enablement (SWE) initiative [17] to building a framework of open standards for exploiting Web-connected sensors and sensor systems of all types such as flood gauges, air pollution monitors, Webcams, etc. SWE provides the opportunity for adding a real-time sensor dimension to the Internet and the Web. It focuses on developing standards to enable the discovery, exchange, and processing of sensor observations, as well as the tasking of sensor systems in order to achieve a "plug-and-play" Web-based sensor networks. Thus, SWE cannot be directly applied to achieve context-awareness because, for example, Sensor Model Language (SensorML) describes sensors systems; provides information needed for discovery of sensors, location of sensor observations, etc. but it does not consider modelling the entities about which the sensor is able to provide information.

Compared to this solution, Chen et al. [18] propose a data-centric infrastructure based on Context Fusion Networks (CFNs) to support context-aware pervasive-computing applications. CFNs are based on an operator graph model, in which context processing is specified by application developers in terms of sources, sinks and channels. In this model, sensors are represented by sources, and applications by sinks. Operators, which are responsible for data processing, act as both sources and sinks. At runtime, the implemented peer-to-peer (P2P) infrastructure instantiates the operator graphs on behalf of context-aware applications. Solar consists of a set of functionally equivalent hosts named Planets. The components messages will be delivered to a Planet with the numerically closest ID; therefore, unlike our proposed approach, Solar services focuses on the data objects instead of on where they live i.e. from which domain they are originated. In addition, Solar does not address privacy enforcement. Another hybrid approach to modeling contextual information that incorporates the advantages of object-oriented and ontology-based modeling techniques is introduced by Lee and Meier [19]. The objective is to support a specific large-scale pervasive domain, namely the transportation domain. Their notion of Primary-Context Model and the Primary-Context Ontology is used to share context between different domains. Although their approach is interesting, it does not address other issues such as mobility and cross-domain context dissemination.

Zebedee et al. [20] introduced ACMF, an adaptable context management system by adopting autonomic computing paradigm. This system is implemented by using the Web services and the Web Services Distributed Management (WSDM) standards. ACMF views each device in terms of the roles it plays with respect to context management which includes client, server, and context proxy. ACMF defines a context model and a set of context exchange protocols between devices. ACMF models the pervasive computing environment as a collection of domains where each domain contains a set of regions and a set of device types. A domain is a logical representation of a physical space, such as a building or campus, containing regions and device-types. In this respect,

their domain concept is very similar to the domain concept used in our approach. However, because the focus is on exchanging context information between devices available on a local area (one region) ACMF does not address cross-domain context dissemination, which is a requirement in pervasive environment. Therefore, querying context information available in distributed domains is not possible in their approach.

Closely integrated with an application domain of e-health, Pung and Gu et.al proposed a Context-Aware Middleware for Pervasive Homecare (CAMP) [21]. The middleware offers several key-enabling system services that consist of P2P-based context query processing, context reasoning for activity recognition and context-aware service management. The key contribution of CAMP is physical context data collection and reasoning, however, it lacks innovation in the architecture of context management.

Most of the previous work focussed on the software engineering perspective of the distributed context management. From a knowledge management perspective, Castelli and Zambonelli [22] addressed the distributed management of context information from a knowledge management perspective. They propose a self-organized agent-based approach to autonomously organize distributed contextual data items into knowledge networks. These data atoms as well as any higher-level piece of contextual knowledge represents a fact which can be expressed by means of a four-fields tuples (Who, What, Where, When); they call it W4 Data Model. This model is able to represent data coming from heterogeneous sources and to promote ease of management and processing. These knowledge atoms are linked via general-purpose mechanisms and policies to form W4 knowledge networks which can facilitate services in extracting useful information out of a large amount of distributed contextual items. The usage of tuple-space like repositories supports heterogeneity and facilitates building the knowledge network; however, because the focus is on the knowledge management perspective other requirements e.g. mobility between domains has been partially addressed. In addition, despite the efficiency in retrieving tuples during query resolution phase, using the spidering approach to create the knowledge networks may be inefficient when considering the rapidly changing context information such as entities location.

If we look at the aforementioned requirements and at the approaches described above, it reveals that research in the area of context management is well established and many ideas have been developed for addressing most of the above requirements individually. However, none of the examined approaches supports all of our requirements to a sufficient extent. Therefore, there is a need to design a new context management framework that takes into consideration the distribution of context in different domains and the necessity to protecting user's privacy.

IV. CONTEXT DISSEMINATION PROBLEM

Consider a simple context federation scenario: a user is subscribed to a CS located in domain A; namely CSA. This server maintains the profile information of its subscribed users and maintains a sensor infrastructure for domain A. We call

this server the *home domain server* (HDS) of its subscribed users. Likewise, the context server CSB maintains users' profiles and physical context information of domain B. Obviously as long as the user is still in the domain A the scenario is rather simple; all the context information needed by the application about this user exists in CSA. However, when the user move from A to B (we call the user a *foreign entity* in domain B), the context information related to the users maintained by CSA and CSB (such as location or environment context information) may become relevant to the applications interested in the user's context. In this case, we call the CSB the *visited domain server* (VDS). Thus, there is a need for a mechanism which allows applications to know which domains are visited by the user at any point of time and the context information gathered about the user in these visited domains.

One possible solution is to use tuple space (e.g., Confab [10]). Confab architecture structures context information into distributed tuple-spaces called *infospaces*, which store tuples about a given entity. An application interested in a certain context, builds a context query using the address of the responsible *infospace*. Although distributed *infospaces* contribute to decrease the context management overhead in a distributed environment, this distribution is not kept transparent to applications, which must know what *infospace* contains the desired context information. Another possible solution is to maintain in the HDSs "links" to the VDSs. In this case, in order to handle the application's queries about the users (or entities) over distributed domains, the HDS may have to distribute queries over the VDSs and compose the query results (e.g. [10][18]). However, this approach requires maintaining the link list of the VDSs, and may degrade the system performance as it requires distributing the application query over different servers and regrouping the result.

On the other hand, the notion of home and visited domains are also used by mobile telephone networks like GSM. The main idea used in these networks is that users have their "home domains" in which their context is gathered but when they roam to another domain this domain becomes a "visited domain". When a mobile device moves into a different domain, the server of the visited domain inter-links the mobile device and its home server. The home server redirects query statements to the server of the visited domain, which finally dispatches it to the mobile device.

This is achieved by using the Home Location Register (HLR) and Visitor Location Register (VLR) approach of the GSM user profile database [23]. This approach addresses the location-awareness problem by minimizing the invocation of multiple updates in the home node each time a mobile user changed his/her location. However, the effectiveness of this mechanism is questionable for other types of context information, as it requires the application to submit their queries through a web of pointers from the home node to the visited node of the mobile user [24].

In fact, the main problem of context dissemination across domains originates from the observation is that in a distributed system there is an obvious trade-off between costs of updates and costs of requests; i.e. between the communication cost introduced by the fully replicating context data to the home node and the degree of replication that is eventually necessary. This has a direct impact on the achieved system performance and on the provided context precision. For example, when the volume of context data or the rate of change is high, providing high precision context value tends to degrade the performance; on the contrary, optimal performance can only be achieved by sacrificing the precision of the context copy. In the proposed approach, as we will see, the context consumers play a decisive role in the process of context replication as well as the update rate of the relevant context data.

V. THE PROPOSED APPROACH

Basically, when a CS receives a query referring to an entity's context information stored in the local repository the procedure is straightforward. When the required context information is not stored in the local repository it has to be retrieved from a remote CS. An efficient look-up mechanism for finding this context information is essential for the scalability of the whole system. To achieve this mechanism, we choose to synchronize the context information with the HDS only when there is a consumer for this information. This choice is made for the following reasons:

(i) Efficient cross-domain query handling: having all context information related to an entity in one place (HDS) can be exploited during the query resolution phase in order for the applications to retrieve the context information more efficiently. That is, handling a query submitted to the system requires considering the context information in the entity's HDS replicated from different domains instead of sending sub-queries to all VDSs. Thus, the querying response time decreases significantly.

(ii) Privacy ensuring: the alternative to publish the actual data at the HDS would be to only keep references to the relevant visited context server. However, this weakens the privacy support as the context data is stored by the foreign domain that provides the sensor infrastructure. Thus, we choose, as we will see, to design a protocol between CSs which force the context information to be centralized in the HDS. This way, enforcing user's privacy policy will be feasible.

(iii) Cross-domain reasoning: it becomes possible to reason about the context information across different domains (e.g. tracking and understanding user's tendency) and to identify the contextual situations which span different domains (see [6] for example). Moreover, this enforces the idea that each domain should have its own inference mechanism and in the home domain a cross-domain inference mechanism becomes possible.

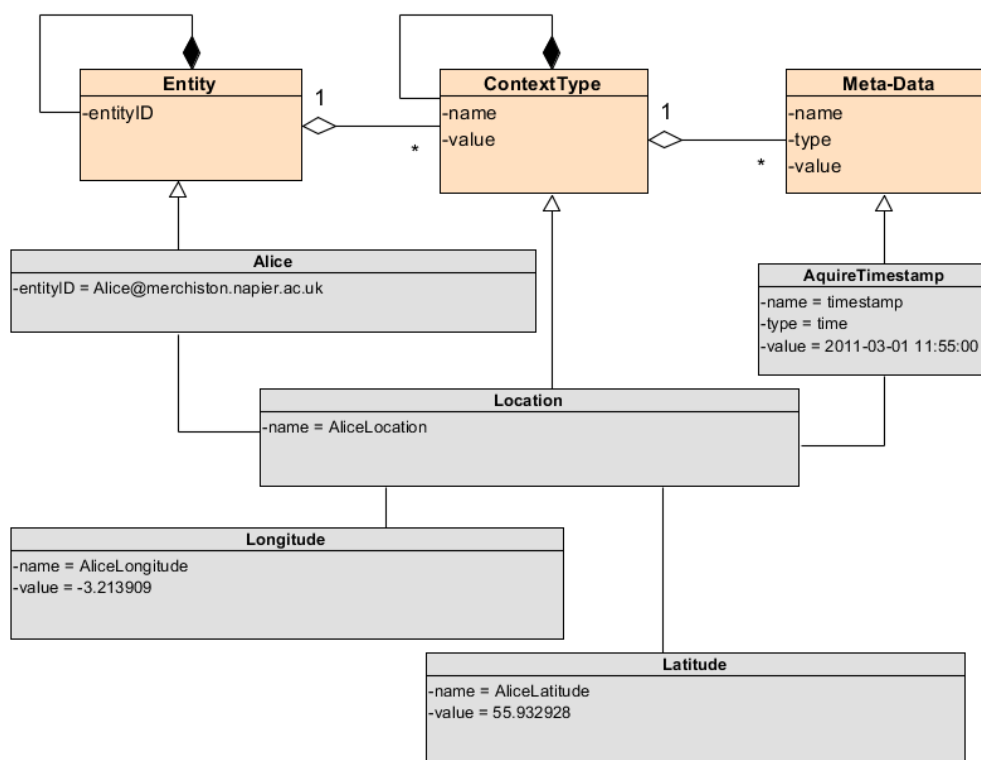


Fig. 1. The proposed context meta-model

(iv) High efficiency: it would be more efficient if we establish context replica on the HDS depending on how often the context change and at the same time on the context consumers needs. In situation of roaming users across domains, additional restrictions may arise (e.g. concerning the limited network connectivity, device power consumption, privacy enforcement, etc.), rendering imperative the need to establish an optimized mechanism in support of optimized context information dissemination among domains taking into account the explicit requirements of consumers.

In the following subsections, we present our designed and implemented framework, *ubique*, which aims at optimizing and controlling the amount of exchanged context information in such a way that context information can efficiently and easily flow from context providers to consumers. *ubique* envisions a highly distributed and loosely coupled solution in order to exchange context information between context providers, CSs, and applications.

Therefore, *ubique* context management framework aims at: (i) enable the discovery of context providers, (ii) standardize context exchange between providers and consumers, (iii) federate contexts among CSs, (iv) standardize and enforce privacy, (v) allow context providers to publish on demand where there is a consumer, (vi) relieve CSs from the burden of replicating frequent updates to the HDS, and (vii) prohibit overloading the context consumers with context information that does not interest them for the time being.

A. *ubique* Context Meta-Model

Context information can be represented in many ways. For *ubique* context modeling, we choose an approach based on XML. As illustrated in Fig. 1, the context information is represented in terms of context elements, which provide information about *context entities*, *context types* and *meta-data*.

The main assumption in the proposed model is the representation of relationships between entity and information: context entities (such as persons, places, events, etc.) are identified and classified by an ID. Each context entity is associated with a set of context types (such as address, location, etc.) which may include other context types. Further, each context type may be characterized by a set of metadata which contain, for example, source of information, timestamps, expiration time, and any Quality-of-Context information such as accuracy and confidence.

B. Context Management Componentss

The *ubique* context management framework is designed for the discovery and exchange of context information across domains. It provides the relevant context information for the service or application, using distributed sensing infrastructure and centralized storing mechanisms. We define *ubique* context management framework as a set of components which are loosely coupled to provide relevant context information both by sensing and interpreting mechanisms. These key components or building blocks are depicted in Fig. 2, and described below.

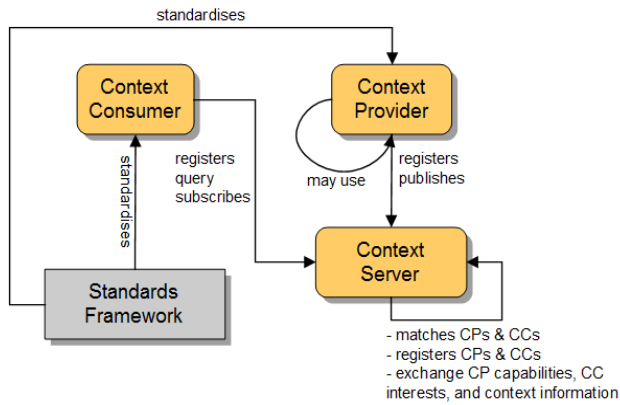


Fig. 2. *ubique* components

Context Consumer (CC) is a software entity that uses the CS interface to register its context interest or query. The CC receives the requested context information asynchronously by submitting context interest and synchronously by submitting context query to the CS. A CC exposes interfaces to start receiving context information from the corresponding CS when they become available. These interfaces adhere to standards defined in the Standards Framework (SF).

Context Provider (CP): is a software entity that uses the CS interface to register its capability of providing context information. A CP exposes interfaces to publish context information to the corresponding CS on-demand. These interfaces adhere to standards defined in the SF. It is registered in the CS so that context consumers can discover and introspect it. Note that any software agent, reasoner, or storage component can be a CP as long as it adheres to the interfaces defined in SF. Usually, CPs wrap context sources such as GPS receiver or temperature sensor to provide their information.

Context Server (CS): provides a registration service for CPs to register/update/unregister their capabilities that uniquely describe their functionalities and for CCs to register/update/unregister their context interests that can be matched against the available CPs, and enables the discovery of various context providers. Additionally, it provides services to exchange the CCs' context interests and CPs capabilities between CSs as we will see later.

Standards Framework (SF): A set of specifications describing the CP capabilities, the CC interests and queries, the interfaces to exchange commands and context information between different components, a format to exchange an atomic context information element, as well as a format for privacy tags.

In *ubique* we rely on the reasonable assumption that a CS is identified by its Internet domain name and that the CS is responsible for managing the context information available in its domain. Additionally, each entity (sensor, user, application, etc.) has a unique ID that should be registered in one of the CSs.

For example Alice ID could be Alice@merchiston.napier.ac.uk as she is a registered

user in the CS of the domain `merchiston.napier.ac.uk` which is Alice's HDS.

C. Context Interfaces and Operations

ubique provides three different interfaces which allows integrating CSs, CCs, and CPs into the eco- system. In the following we describe the main interfaces and the main corresponding operations.

1)Integrating Context Providers: The provided operations allow registering CPs and their information with the CS as well as providing a discovery function with which participating components can check for available CPs.

registerContextProvider: This operation is used by the CP to advertize its capabilities in terms of the types of context information it can provide and the relevant entities playing a role in this information. Additionally, the registration provides a set of available CP meta-data (which mention information about the provider as well as quality of context information it provides). For example, the user's location can be measured with different quality by location sensors like GPS, CellId, WLAN-in-range, etc. Finally, registration provides further information about the registered entities. The CP capabilities XML scheme is depicted in Fig. 3.

Basically, the CP specifies in its capabilities its ID, the domain its information is originated from, and one or more *capability*. Each *capability* specifies its ID, the entities having the context information, and the supported context types. Optionally, it specifies the meta-data about these context types, its different attributes (features), and collection policies.

discoverContextProviders operation is used by CCs to get the list of available CPs and their capabilities for later query.

sendCPCCommand: This operation is used by the CS to command a specific CP to start/stop publishing its information. The command message contains a reference (tuple ID) where the context information should be pushed.

2)Integrating Context Consumers: The provided operations allow registering CCs with the CS, querying (synchronously), as well as subscribing in order to be notified about context information (asynchronous).

queryContextServer: This operation is used by the CC to synchronously request for context information. The CC specify its interest in terms of the needed context types of specific entity(ies), as well as additional constraints on the CPs and context types meta-attributes.

subscribeContextConsumer: This operation enables long-lasting monitoring of the system. Basically, the logic of this operation is similar to the latter operation, but the request context information is returned in the form of an asynchronous "notify" callback operation. Fig. 4 depicts the CC interest XML scheme. The CC can specify one or more *interests*. Each *interest* specifies its ID, the entities the CC is interested to get their context information, and the interested context types. Optionally, it specifies the condition(s) on the context types, the domain(s) this information is originated from, the required feature(s) from the CP, and the ID of a specific CP.

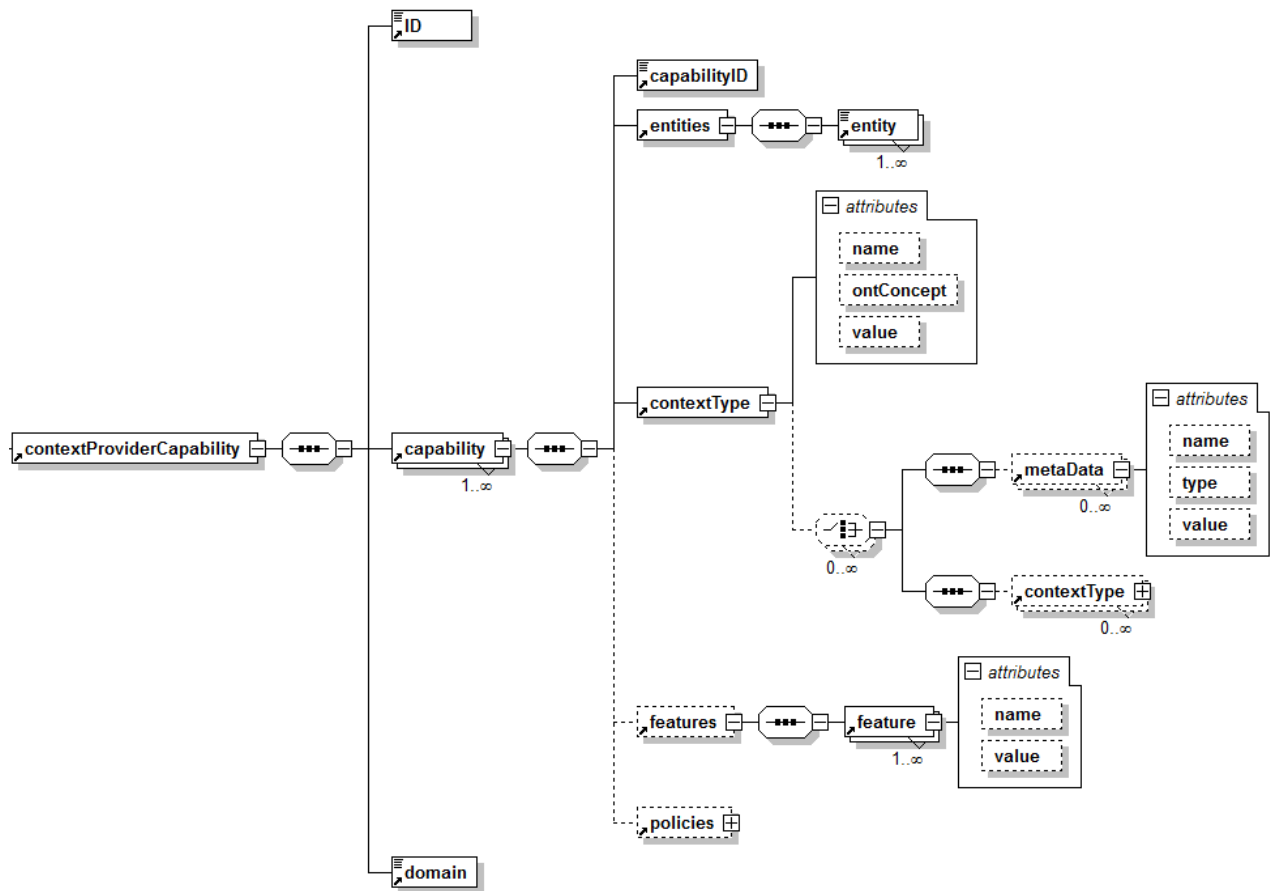


Fig. 3. CP capabilities XML scheme

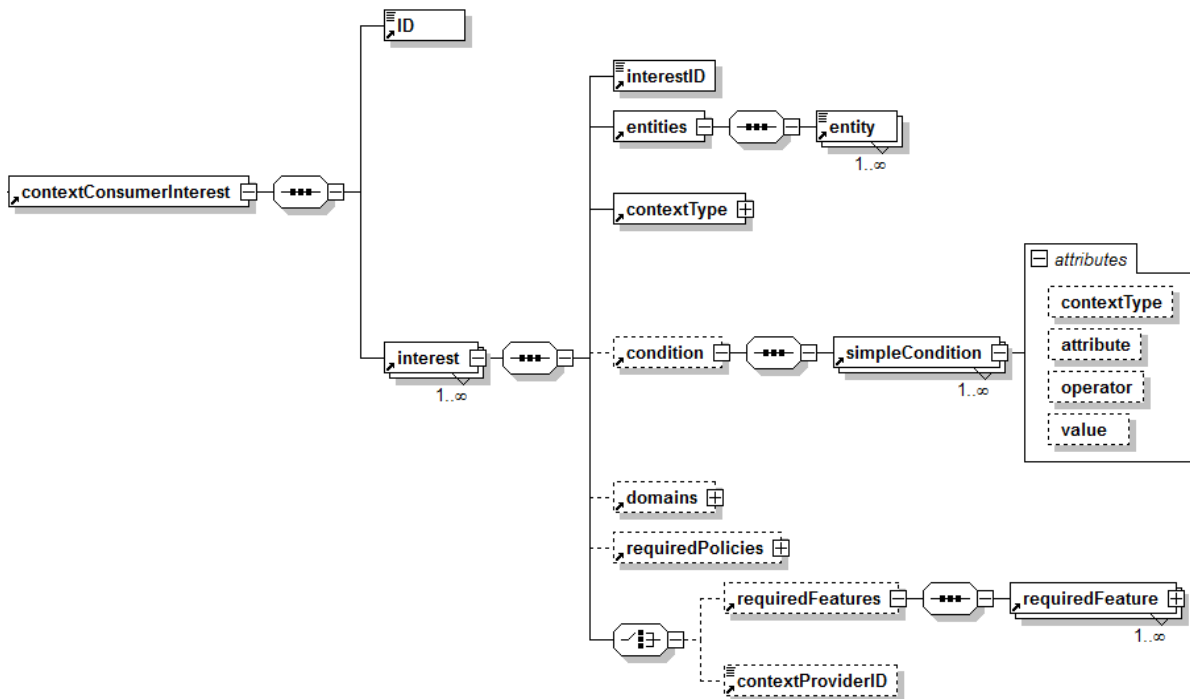


Fig. 4. CC interest XML scheme

sendCCCommand: This operation is used by the CS to command a specific CC to start/stop receiving the information it has subscribed to. The command message contains a reference (tuple ID) where the context information should be popped.

3)Federation between CSs: as already mentioned, every CS is responsible for providing and storing context information related to entities registered in it. Since the sensor infrastructure in each domain may provide context information about roaming entities, a collaboration protocol is needed between CSs in order to federate this information to the entities' HDSs. We can distinguish here between three types of information exchanged between CSs:

- CP Capabilities: CPs may advertise their ability to provide context information about entities not registered in the current domain. For example, a GPS sensor of Alice mobile phone can provide location information about Alice@domain1.com to the CS available in domain1.com (Alice's HDS). However, when Alice move to domain2.com, then this CP advertise its capability to provide Alice location information to CS of the domain2.com. In this case, CS of domain2.com should federate the CP capability to domain1.com (Alice's HDS) which is responsible to handle all queries related to Alice.

- CC Interests: A CS may receive context interest about entities not registered in it. In this case, the CS should federate these interests to the HDS of the corresponding entities.

- Context information: The idea is that each CS has to maintain a repository for all CP capabilities able to provide context information about its registered entities as well as all CC interests related to these entities. Any change in this repository (i.e. addition, updating, or deletion of CP capabilities or CC interests) should trigger a matching function which tries to bind a CP with a CC. When a match is found, a new tuple has to be created; a *startPublishing* command message has to be sent to the CP (via *sendCPCCommand* operation) along with the corresponding CC interest and tuple ID; and a *startReceiving* command has to be sent to the CC (via *sendCCCommand* operation) along with the tuple ID. The CP now has all the information necessary to know what kind of context types, for which entities, and when to publish to the tuple (e.g. regularly or for a context changes greater than a specific threshold, etc.). Note here that when, for example, an application is interested in Alice location in domain2.com, the CS of domain1.com (Alice's HDS) will create a tuple in CS of domain1.com and ask the CP of Alice location to start publishing in this tuple. In other words, all the context information related to Alice, even those emerging from foreign domains, will be kept in her HDS. This way, we have more control about ensuring entities privacy. This mechanism is illustrated in the example usage in the Section 6. Fig. 5 depicts the XML scheme of the published context information which

we call it *contextlet*. Basically, each *contextlet* specifies the CP ID, the interest ID (so that the CC knows that this information is related to which interest he has submitted), the domain from which this information is originated, the entity in question, a list of the requested context types and their values.

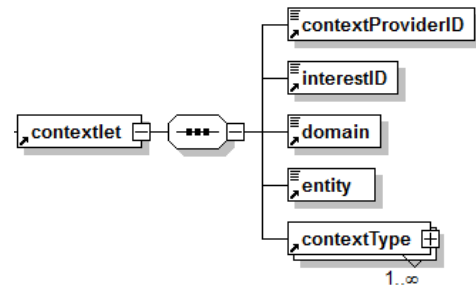


Fig. 5. Contextlet XML scheme

D. Privacy

Privacy is about protecting users' personal information, which may include also context information e.g. location, mood, etc. Obtained context information might be severely misused, e.g., to track users. In context aware environments, the devices belonging to the user communicate with the available CSs all the time, thus revealing privacy sensitive information about the user. In *ubique* approach, to ensure the confidentiality of the privacy-sensitive information, users have the flexibility to define their own privacy policy covering all types of context information that may be distributed in different domains.

Obviously, the sensor infrastructure in each domain may report context information related to entities out of the scope of the current domain which in turn weaken the privacy ensuring mechanism and loosen control over the context originated in different domains. In this case we need a mechanism with which the context information of the foreign entities can be moved to their HDS with the following conditions: (i) there is a corresponding consumer for this information, and (ii) revealing this information does not violate the privacy policy specified by the user. That is, when the CS finds a match between a CP and CC, it retrieves the privacy policy of the entity the CC specifies its interest in getting context information. If this request does not violate the user's privacy then the CP is asked to start publishing the required context information at the entity's HDS; otherwise, an "access denied" response is sent to the CC. Fig. 6 shows the privacy tag schema used in *ubique*. Each user (or each entity in general) has the flexibility to specify its privacy policy for each context type and for each domain. The *privacyTag* specifies for each context type the CCs having the right to get access to the context information and the time intervals during which this context information can be revealed to them.

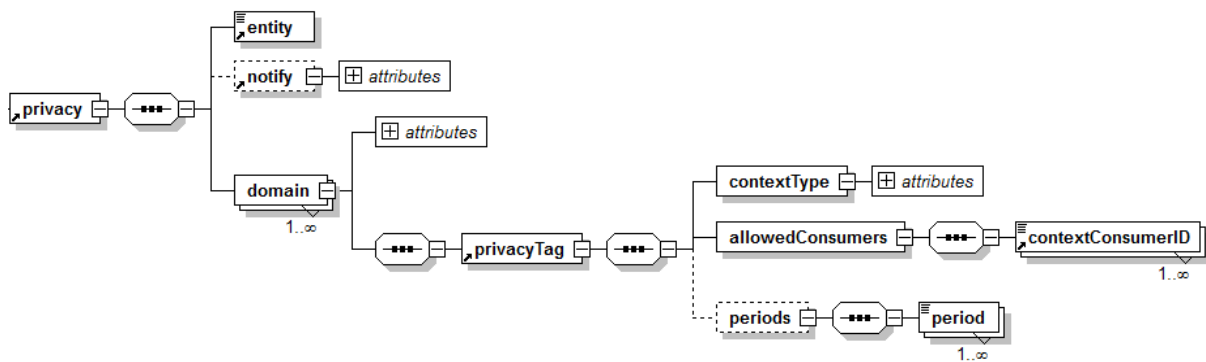


Fig. 6. Privacy XML scheme

Finally, secure storage of context information requires proper authentication and authorization to access it. Therefore, we assume here that each CC is a computational entity registered in one of the CSs which means that it has a unique ID and password, and it must be authenticated by its CS.

E. *ubique* Implementation

Fig. 7 illustrates the proposed domain-based context-aware computing eco-system. In general, the system should integrate distributed hardware and software components and provide naming scheme for those entities. The eco-system starts from a single system with client-server architecture; then multiple systems federate together through server-to-server communication to form the eco-system. A single system usually manages local clients, such as users and devices in a specific domain.

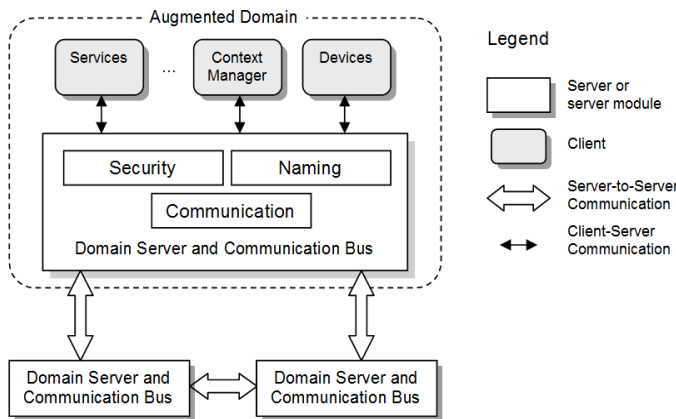


Fig. 7. Domain-based context-aware eco-system

The server is called Domain Server and Communication Bus. As indicated by its name, the server provides core functionalities, such as security and naming, and acts as a communication infrastructure for clients available in its administrative domain. The naming scheme is similar to that of e-mail systems. Each server has a unique domain name; clients have their names concatenated to the server name. Clients from different systems can also communicate with each other with the server-to-server communication. Clients could be devices, such as sensors, and applications that provide services to the user. Clients can be also services that provide functionalities

the server does not provide such as the context manager (see Fig. 7). Clients have to be authenticated by the server to use the system.

Notice that the server does not provide context management service itself, leaving that responsibility to a separate client, the context manager. The context manager can be easily replaced or upgraded without affecting the whole system. The client-server and server-to-server communication interfaces are standardized, which facilitates the system extensibility.

In order to robustly implement the *ubique* approach, relying on a standard or already established protocol is obviously a preferred choice. The eXtensible Messaging and Presence Protocol (XMPP) [25] (also known as a Jabber protocol) is widely adopted open protocol for instant messaging and is designed for near real-time communication. In the following section we describe Jabber technologies by which *ubique* is inspired and based on.

1) *Jabber Overview*

Jabber is an extensible instant messaging (IM) system. More precisely, Jabber is a set of streaming XML protocols and technologies that enable any two entities on the Internet to exchange messages, presence, and any other structured information in near real-time.

The Internet Engineering Task Force (IETF) has standardized the core Jabber protocol as the XMPP protocol [26]. The architecture of the Jabber system is distributed. A Jabber server has a number of registered clients. Clients on the same server interact through that server; clients on different servers interact through server-to-server communication. Jabber enables message transfer not only between people, as in traditional IM systems, but also between any two entities. An entity can be a person, a device, or a software service. Each entity has a unique Jabber ID (JID). A JID is similar to an e-mail address. For example, a JID for Alice is Alice@merchiston.napier.ac.uk. Each entity is allowed to have multiple resources. For example, Alice may have a laptop and a cell phone which could be identified as Alice@merchiston.napier.ac.uk/dell and Alice@merchiston.napier.ac.uk/nokia respectively.

Furthermore, Jabber enriches the communication support beyond chat to many other interaction semantics thanks to the

XMPP extensions. The Jabber Software Foundation develops extensions to XMPP through a standards process centered on XMPP Extension Protocols (XEPs) [27]. Examples of these extensions are the Jabber RPC [XEP-0009], ad-hoc commands [XEP-0050], streaming audio and video [XEP-0166], and so on.

In addition, Jabber has an interesting pubsub facility [XEP-0060], in which both publishers and subscribers are Jabber entities. A publisher publishes a message item to a topic, and then all the topic subscribers will be notified to receive the newly published item. In this communication mechanism, since the publisher does not know who will receive the message, and a subscriber does not know who sent it, the time-coupling and reference-decoupling between publishers and subscribers are assured. This pubsub mechanism is ideal for implementing *ubique*, where context providers and consumers can be associated and disassociated dynamically.

2) Jabber and Domain-based Context Management

As aforementioned, the proposed domain-based context management architecture is based on Jabber technologies. Jabber has been chosen because its design, architecture, and features match our requirements: In the pervasive environment the interaction between different entities should be generic and not in a particular format. Jabber provides a rich set of communication mechanisms (see Section 5.5.1). Moreover, the context management infrastructure should support the interaction between different users, devices, and software components in a universal way. In Jabber systems, any entity that implements the XMPP-Core and its extensions protocols can establish a connection with a Jabber server and interact with other entities on any Jabber server. Thus the open architecture and standardization of the Jabber platform ease its adoption to build *ubique*.

Other than these capabilities, Jabber has other advantages such as its increasing popularity and community support; the availability of a set of servers, clients, and software libraries supporting a low-barrier entry for developers; and its adoption of XML to communicate messages between entities make it possible to leverage existing XML tools and libraries.

3) Jabber and Context Manager

Jabber entities can be implemented either as clients or as external server components. Clients use the protocols defined in “XMPP Core” to connect to the Jabber server; external components use the “Jabber Component Protocol” (JCP) [XEP-0114] for the connection. These two types of entities are functionally similar; thus for a given service, we can implement it as either a client or a component. However, unlike client components whose contact lists and subscription are maintained by the Jabber server, external component has to manage its subscriptions and contact lists by itself. The naming convention for external components is different from client components. For example, the context manager JID might be `context@merchiston.napier.ac.uk` if it is implemented as a client, and `context.merchiston.napier.ac.uk`, if it is implemented as an external component.

In *ubique* the context manager has been implemented as an external Jabber component. The choice of considering the context manager as an extension to the Jabber server functions is more of design decision than a functional one. Fig. 8 shows the architectures of the context manager: *Context*. The pubsub server is also a Jabber component. Context component connects to a Jabber server using JCP. The actual context data (contextlets) is stored in the pubsub so that the pubsub server will notify the subscriber of any context changes.

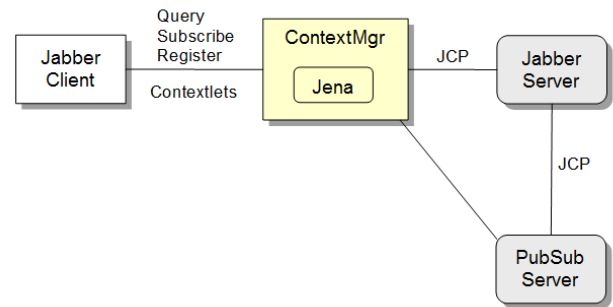


Fig. 8. The context manager external component

In Fig. 9, two Jabber servers are inter-connected; one of them connects to a CP and the other connects to a CC. The context manager, *Context*, connects to the Jabber server as a Jabber external component. The continuous lines represent the transport connections which are the actual routes for transferring data. On the other hand, the dashed lines indicate logical connections which means the communication between two end points does not happen directly, but through physical ones.

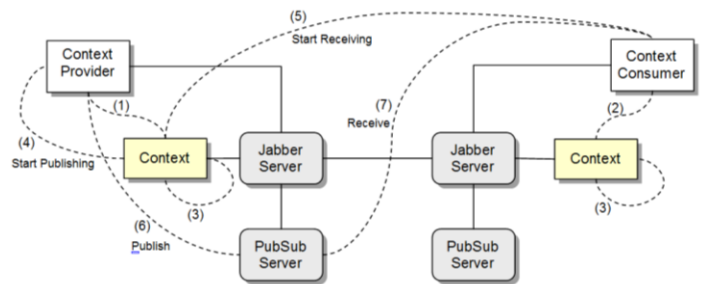


Fig. 9. *ubique* components interactions

When the system starts up, both CP and CC logon to their Jabber servers which may or may not be the same one. Then, the capabilities of each CP and the interests of each CC are registered into the corresponding Jabber server (Step 1 and 2). Thus the context manager can match the published CPs’ capabilities with the CCs’ interests or queries (Step 3). If the context manager decides that the CC interest matches the CP capability and this does not violate any entity’s privacy, then it creates a tuple space in the local PubSub server and sends the *startPublishing* command message to the CP (Step 4) and the *startReceiving* message command to the CC (Step 5) along with the tuple space ID embedded in the message. Once the CP publishes a new contextlet (Step 6), the CC can receive it

asynchronously (Step 7). For the CC query, when the context manager decides which CP can have the requested context information it queries that CP and returns the result to the CC synchronously.

ubique is built on top of a number of technologies, such as Jabber (we use OpenFire [28] as a XMPP server), OWL, Jena [29], and XML. It leverages these enabling technologies to achieve the goal of controlling the context information dissemination between administrative domains in a way that is efficient in terms of saving network bandwidth and devices energy, as well as respecting people privacy in the pervasive environment. The system has a clear architecture and is highly extensible.

VI. CASE STUDY ON *UBIQUE* CONTEXT USAGE

Alice and her husband Bob work as lecturers in Edinburgh Napier University in Merchiston campus. Alice has a daughter, Carol, who studies in the same university in Sighthill campus. Alice would like to keep updated about her husband activities and her daughter location.

Fig. 10 depicts the sequence of exchanging information between different components: CPs, CCs, and CSs.

This is described as follows: The CP ActivityProvider@merchiston.napier.ac.uk registers the following capability in its HDS and wait for confirmation (Step 1).

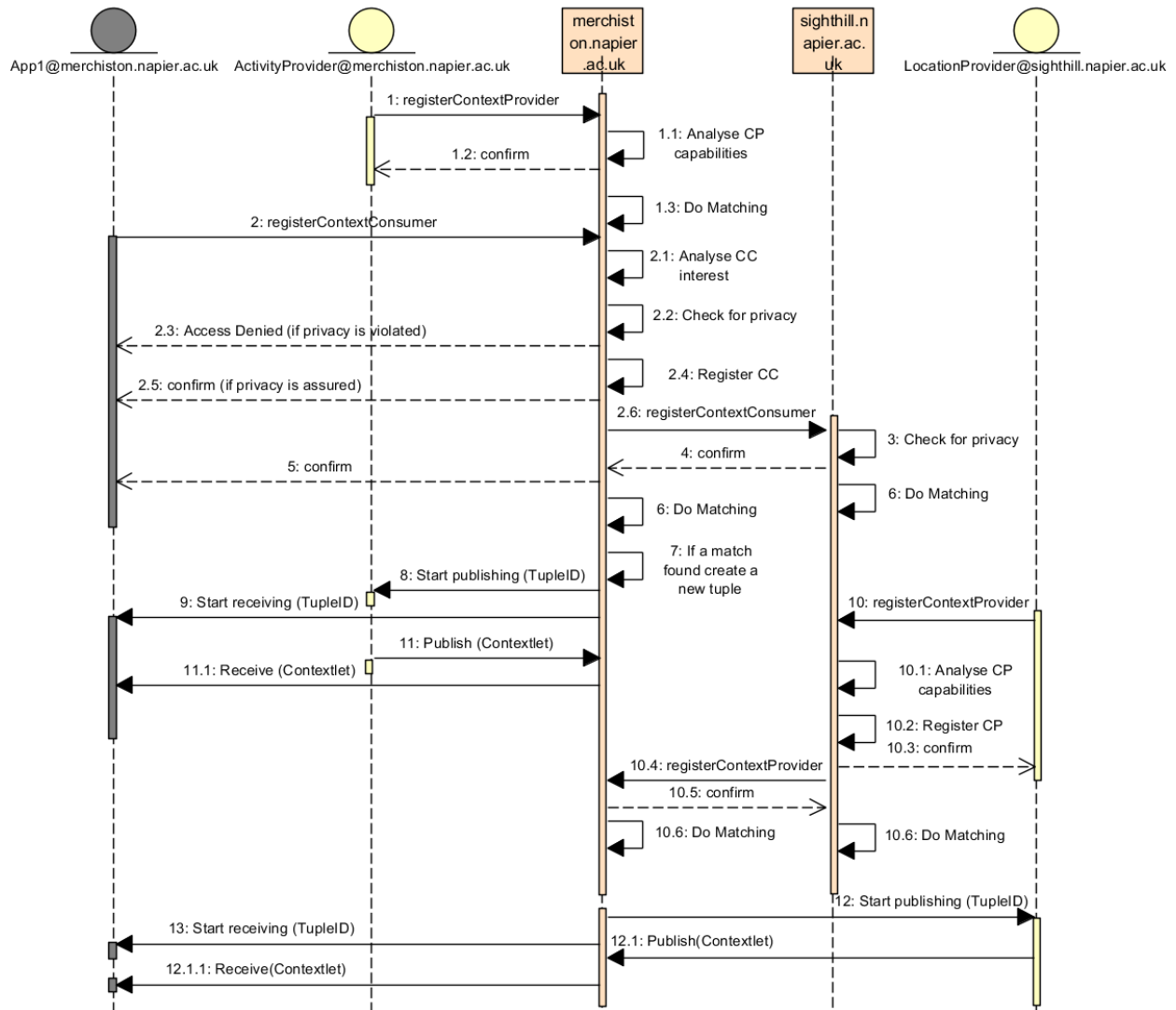


Fig. 10. Interaction between different components

```
<contextProviderCapability>
  <ID>CP1</ID>
  <capability>
    <capabilityID>CPC1</capabilityID>
    <entities>
      <entity>Bob@merchiston.napier.ac.uk</entity>
      <entity>John@merchiston.napier.ac.uk</entity>
    </entities>
    <contextType name="activity" ontConcept="http://www.napier.ac.uk/ontologies/percom.owl#Activity"/>
    <features>
      <feature name="confidence" type="float" value="0.85"/>
    </features>
  </capability>
</contextProviderCapability>
```

Fig. 11. Example of the activity provider advertized capabilities

The CS analyzes the received CP capability to see if any of the supported entities is not registered in it. Because this CP does not provide context information about entities not registered in merchiston.napier.ac.uk no further interaction with other CSs has to be taken. Obviously, any change in the available CPs or CCs triggers the matching function.

For the sake of simplicity and without loss of generality, the example application App1@merchiston.napier.ac.uk is registered in Alice's HDS. It registers the following CC interest (Step 2):

```
<contextConsumerInterest>
  <ID>CC1</ID>
  <interest>
    <interestID>CC1</interestID>
    <entities>
      <entity>Bob@merchiston.napier.ac.uk</entity>
    </entities>
    <contextType name="activity" ontConcept="http://www.napier.ac.uk/ontologies/percom.owl#Activity"/>
    <condition>
      <simpleCondition contextType="Activity" operator="gt" attribute="timestamp" value="2011-03-01 10:30:00"/>
    </condition>
    <domains>
      <domain>merchiston.napier.ac.uk</domain>
    </domains>
    <requiredFeatures>
      <requiredFeature featureName="confidence" operator="gt" value="0.8"/>
    </requiredFeatures>
  </interest>
  <interest>
    <interestID>CC2</interestID>
    <entities>
      <entity>Carol@merchiston.napier.ac.uk</entity>
    </entities>
    <contextType name="location" ontConcept="http://www.napier.ac.uk/ontologies/percom.owl#Location"/>
    <condition>
      <simpleCondition contextType="Latitude" attribute="minAccuracy" operator="lt" value="0.000005"/>
      <simpleCondition contextType="Longitude" attribute="minAccuracy" operator="lt" value="0.000005"/>
    </condition>
  </interest>
</contextConsumerInterest>
```

Fig. 12. Example of an application context interest

This CC interest shows that the application is interested to know the location of Carol in any domain and the activity of Bob in the merchiston.napier.ac.uk domain.

Note here that any CP registered in merchiston.napier.ac.uk domain or in any of its sub-domains is eligible to be matched with the interest CC1. For each context interest, the CS checks for the corresponding entity privacy before registering it. Fig. 13 shows an example of Carol privacy tag.

```
<privacy>
  <entity>Carol@merchiston.napier.ac.uk</entity>
  <notify value="mailto:carol@merchiston.napier.ac.uk"/>
  <domain name="sighthill.napier.ac.uk">
    <privacyTag>
      <contextType name="location" ontConcept="http://www.napier.ac.uk/ontologies/percom.owl#Location"/>
      <allowedConsumers>
        <contextConsumerID>App1@merchiston.napier.ac.uk</contextConsumerID>
        <contextConsumerID>App4@sighthill.napier.ac.uk</contextConsumerID>
      </allowedConsumers>
      <periods>
        <period>...</period>
      </periods>
    </privacyTag>
  </domain>
</privacy>
```

Fig. 13. Example of a privacy policy

If the privacy is violated, an "access denied" message should be sent to the application; otherwise the following context interest will be registered and a confirmation message should be sent to the application.

The CS of merchiston.napier.ac.uk finds out that there is a match between the CP capability whose ID is CPC1 (Fig. 11) and the CC interest whose ID is CC1 (Fig. 12), therefore, it creates a tuple and sends the necessary commands so that ActivityProvider@merchiston.napier.ac.uk starts publishing contextlets in the created tuple and App1@merchiston.napier.ac.uk starts receiving the published contextlets. Fig. 14 shows an example of the contextlet sent by the activity provider. Alice may like to send Bob a congratulations message when he finishes his presentation.

```
<contextlet>
  <contextProviderID>CP1</contextProviderID>
  <interestID>CC1</interestID>
  <domain>merchiston.napier.ac.uk</domain>
  <entity>Bob@merchiston.napier.ac.uk</entity>
  <contextType name="activity" value="FinishPresenting" >
    <metaData name="timestamp" type="time" value="2011-03-01 11:55:00"/>
  </contextType>
</contextlet>
```

Fig. 14. Example of contextlet received from activity provider

In merchiston.napier.ac.uk there is no provider for Carol location. When Carol roams to sighthill.napier.ac.uk the CP LocationProvider@sighthill.napier.ac.uk reports its ability (Fig. 15) to provide Carol as well as other entities locations to CS of sighthill.napier.ac.uk.

```
<contextProviderCapability>
  <ID>CP2</ID>
  <capability>
    <capabilityID>CPC1</capabilityID>
    <entities>
      <entity>Carol@merchiston.napier.ac.uk</entity>
      <entity>Sally@sighthill.napier.ac.uk</entity>
    </entities>
    <contextType name="location" ontConcept="http://www.napier.ac.uk/ontologies/percom.owl#Location">
      <contextType name="latitude" ontConcept="http://www.napier.ac.uk/ontologies/percom.owl#Latitude">
        <metaData name="minAccuracy" type="float" value="0.000002"/>
      </contextType>
      <contextType name="longitude" ontConcept="http://www.napier.ac.uk/ontologies/percom.owl#Longitude">
        <metaData name="minAccuracy" type="float" value="0.000002"/>
      </contextType>
    </contextType>
  </capability>
</contextProviderCapability>
```

Fig. 15. Example of the location provider advertized capabilities

The CS of sighthill.napier.ac.uk finds out that the location provider is able to provide Carol location which is not registered in it; thus, it federates the CP capability depicted in Fig. 16 to Carol HDS: merchiston.napier.ac.uk (Step 10.4 in Fig. 10). Notice that this capability is the same of Fig. 15 except that the entities not registered in merchiston.napier.ac.uk have been removed.

```
<contextProviderCapability>
  <ID>CP2</ID>
  <capability>
    <capabilityID>CPC1</capabilityID>
    <entities>
      <entity>Carol@merchiston.napier.ac.uk</entity>
    </entities>
    <contextType name="location" ontConcept="http://www.napier.ac.uk/ontologies/percom.owl#Location">
      <contextType name="latitude" ontConcept="http://www.napier.ac.uk/ontologies/percom.owl#Latitude">
        <metaData name="accuracy" type="float" value="0.000002"/>
      </contextType>
      <contextType name="longitude" ontConcept="http://www.napier.ac.uk/ontologies/percom.owl#Longitude">
        <metaData name="accuracy" type="float" value="0.000002"/>
      </contextType>
    </contextType>
  </capability>
</domain>sighthill.napier.ac.uk</domain>
</contextProviderCapability>
```

Fig. 16. The location provider capabilities federated to the Carol HCS

After the re-matching process, the CS of merchiston.napier.ac.uk finds out that there is a CP able to provide Carol position. Therefore, as in the previous

case, it creates a tuple and sends the necessary commands to the corresponding entities; however, this time the locally published contextlets are pushed by a CP from other domain. Fig. 17 shows an example of a contextlet published by the location provider indicating Carol location.

```
<contextlet>
  <contextProviderID>CP2</contextProviderID>
  <interestID>CCI2</interestID>
  <domain>sighthill.napier.ac.uk</domain>
  <entity>Carol@merchiston.napier.ac.uk</entity>
  <contextType name="location" >
    <metaData name="timestamp" type="time" value="2011-03-01 11:55:00"/>
    <contextType name="latitude" value="55.923215" >
      <metaData name="timestamp" type="time" value="2011-03-01 14:15:00"/>
      <metaData name="accuracy" type="float" value="0.000002"/>
    </contextType>
    <contextType name="longitude" value="-3.286835">
      <metaData name="timestamp" type="time" value="2011-03-01 14:15:00"/>
      <metaData name="accuracy" type="float" value="0.000002"/>
    </contextType>
  </contextType>
</contextlet>
```

Fig. 17. Example of Carol location contextlet

Fig. 18 depicts screenshots of the example application. The cyan circles represent roughly the domain border of each CS. Each small dot circle represents a contextlet.

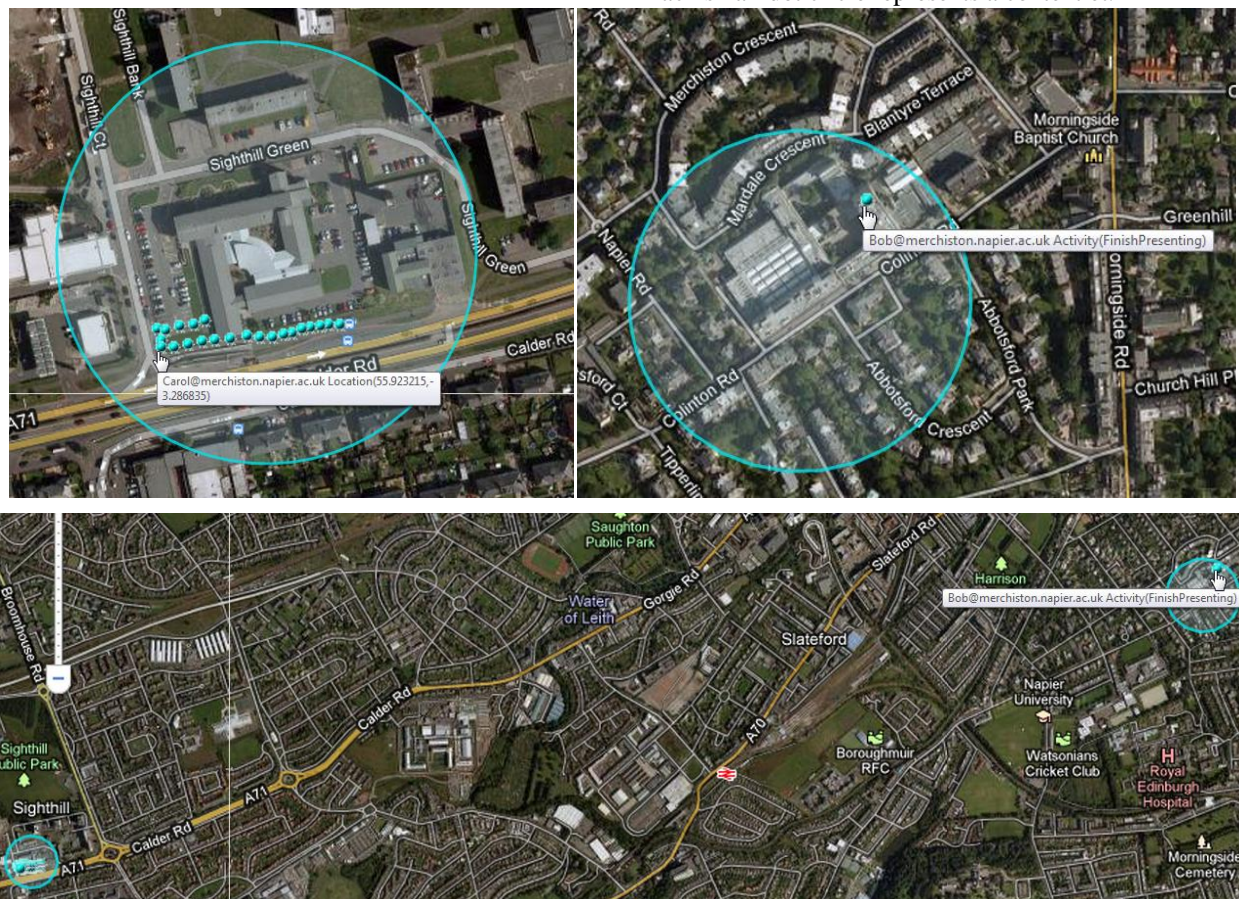


Fig. 18. Screenshots of the example application

VII. EVALUATION

In this section, we first analyze the suitability of the *ubique* approach according to the requirements of context management in pervasive applications as proposed in Section 2. The efficiency of *ubique* is then evaluated via a set of experiments based on the case study (Section 6).

A. Analysis of *ubique* vs. the requirements

Domains of context perception: This requirement, which is compliant with the principle of system boundary of pervasive applications, is achieved by using CS in each domain and the federation between CSs across different domains. Additionally, the notion of home domain CS simplifies application developments as it is the reference point for any context information related to the entities registered in it.

Uniform API interface and protocol: By providing the *ubique*'s set of open and generic APIs, context is made available to third party application developers to build new services without having to define specific mechanisms for context distribution and management between domains. In addition, these APIs and the proposed protocol between different entities enable external providers and consumers to be integrated into the *ubique* system to provide or consume context information.

Efficient context information dissemination: Since the communication resources are limited, and since most context information gathered by a context server will not be necessarily used by any application, *ubique* considers filtering and replicating only the context information that is explicitly required by an application.

Cross-domain reasoning: *ubique* provides an enabling infrastructure to support reasoning about the context information across different domains and to identify the contextual situations which span different domains. Moreover, this enforces the idea that each domain should have its own inference mechanism whereas in the HDS a cross-domain inference becomes possible.

Dynamic matching between context providers and consumers: In *ubique* the matching function of the context manager ensures efficient context information dissemination. In addition, since the CPs specify their capabilities in providing context information that correspond to different domains, an application can specify in its interests or queries the domain(s) from which it is interested in retrieving the context information.

Support for privacy: Since the context information is centralized in one CS (HDS), enforcing user's privacy policy which spans different domains is feasible. In addition, the dissemination protocol between CPs and CSs on one hand, and the between CSs on the other hand, ensures that the context information will not be stored everywhere and that this information will be disseminated only if the receiver has the privilege to get it.

B. Performance evaluation

The efficiency of *ubique* has been evaluated in terms of update latency. As part of the case study, evaluation experiments were done using four CSs distributed in four

university campuses (Merchiston, Craighouse, Sighthill, and Craiglockhart) which store the context information available in their corresponding campuses. All 4 servers have the same hardware capability: Pentium 4, 3.40GHz and 4GB RAM. The aim is to measure the latency average of federating the contextlets from one CS to another. Fig. 19 shows the variation of the latency time (milliseconds) with respect to the number of contextlets simultaneously federated. Obviously the latency increases when the volume of data increases; however, the results show that the increase is not in a linear pace with the amount of contextlets, i.e. the latency is higher when the amount of contextlets is over 150. The latency could reach around 1.5s for sending 200 contextlets simultaneously, which is reasonable and acceptable even for the highly dynamic context information e.g. noise level.

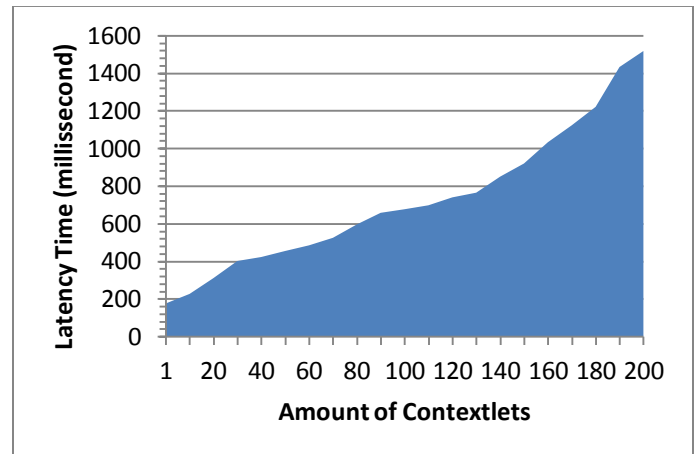


Fig. 19. *ubique* performance evaluation

VIII. CONCLUSION

The essence of context-awareness is to let applications and users take full advantage of the available context information e.g., users' or devices' locations. The requirement for universal context access demands for a middleware solution as an essential requirement for building context-aware systems. In order to address these new challenges, it is essential to establish innovative data storage and dissemination mechanisms. The architecture of *ubique* presented in this paper hides the increasing complexity of context management from applications and incorporates advanced mechanisms that support mobile users. In *ubique*, a Jabber-based context information dissemination protocol has been adopted. The storage and dissemination of the context information is performed by federation between distributed CSs. *ubique* brings several unique features to cross domain context management as discussed in section 7, all of which have been verified by the case studies.

Further research plans involve exploring the use of the middleware in more complex scenarios, extending *ubique* to support the geographic location based access to context information, the extension of the privacy protection scheme to consider not only specified domains but also domain types (e.g. a restaurant or a swimming pool), and *ubique* extension to support context queries on the basis of the entities' and domains' types.

ACKNOWLEDGMENT

The work in this paper has been sponsored by the Lawrence Ho Research Fund (LH-Napier2012).

REFERENCES

- [1] M. Weiser, "The Computer for the 21st Century," *Communications*, vol. 3, no. 3, pp. 3-11, 1991.
- [2] D. Preuveneers, K. Victor, Y. Vanrompay, P. Rigole, M. K. Pinheiro, and Y. Berbers, "Context-Aware Adaptation in an Ecology of Applications," in *Context-Aware Mobile and Ubiquitous Computing for Enhanced Usability: Adaptive Technologies and Applications*, 2009, pp. 1-25.
- [3] R. C. A. da Rocha, "Context Management for Distributed and Dynamic Context-Aware Computing," PhD Thesis, 2009.
- [4] T. Kindberg and A. Fox, "System Software for Ubiquitous Computing," *Pervasive Computing*, IEEE, vol. 1, pp. 70-81, 2002.
- [5] M. Valla et al., "The Context API in the OMA Next Generation Service Interface," in *Proceedings of ICIN 2010*, 2010.
- [6] Z. Jaroucheh, X. Liu, and S. Smith, "Recognize contextual situation in pervasive environments using process mining techniques," *Journal of Ambient Intelligence and Humanized Computing*, vol. 2, no. 1, pp. 53-69, Dec. 2010.
- [7] A. K. Dey, G. D. Abowd, and D. Salber, "A conceptual Framework and a Toolkit for Supporting the Rapid Prototyping of Context-Aware Applications," *Human-Computer Interaction*, vol. 16, no. 2, pp. 97-166, 2001.
- [8] M. Román, C. Hess, R. Cerqueira, and R. H. Campbell, "A Middleware Infrastructure for Active Spaces," *IEEE Pervasive Computing*, vol. 1(4), pp. 74-83, 2002.
- [9] A. Dearle et al., "Architectural Support for Global Smart Spaces," in *Lecture Notes In Computer Science; Vol. 2574. Proceedings of the 4th International Conference on Mobile Data Management*, 2003, pp. 153-164.
- [10] J. I. Hong and J. A. Landay, "Architecture for privacy-sensitive ubiquitous computing," in *2nd International Conference on Mobile Systems, Applications, and Services*, 2004, vol. p, pp. 177-189.
- [11] K. Henriksen, J. Indulska, T. McFadden, and S. Balasubramaniam, "Middleware for Distributed Context-Aware Systems," in *On the Move to Meaningful Internet Systems 2005: CoopIS, DOA. Proceedings of the OTM Confederated International Conferences: CoopIS, DOA and ODBASE 2005, Part 1.*, 2005, vol. 3760, pp. 846-863.
- [12] S. L. Kiani, M. Riaz, S. Lee, and Y.-K. Lee, "Context Awareness in Large Scale Ubiquitous Environments with a Service Oriented Distributed Middleware Approach," in *Fourth Annual ACIS International Conference on Computer and Information Science (ICIS'05)*, 2005, vol. 5, pp. 513-518.
- [13] M. Grossmann, M. Bauer, N. Hönle, U.-P. Käppeler, D. Nicklas, and T. Schwarz, "Efficiently Managing Context Information for Large-Scale Scenarios," in *Third IEEE International Conference on Pervasive Computing and Communications*, 2005, no. PerCom, pp. 331-340.
- [14] P. Floreen et al., "Towards a Context Management Framework for MobiLife," in *In IST Mobile & Wireless Communications Summit*, 2005.
- [15] M. Klemettinen, *Enabling Technologies for Mobile Services: The MobiLife Book*. 2007.
- [16] M. Strohbach, M. Bauer, E. Kovacs, C. Villalonga, and N. Richter, "Context Sessions – A Novel Approach for Scalable Context Management in NGN Networks," in *MNCNA '07 Proceedings of the 2007 Workshop on Middleware for next-generation converged networks and applications*, 2007, pp. 1-6.
- [17] G. Percivall, C. Reed, and J. Davidson, *Open Geospatial Consortium Inc . OGC White Paper OGC ® Sensor Web Enablement : Overview And High Level Architecture .*, no. December. 2007, pp. 1-14.
- [18] G. Chen, M. Li, and D. Kotz, "Data-centric middleware for context-aware pervasive computing," *Pervasive and Mobile Computing*, vol. 4, no. 2, pp. 216-253, 2008.
- [19] D. Lee and R. Meier, "A hybrid approach to context modelling in large-scale pervasive computing environments," *Proceedings of the Fourth International ICST Conference on COMMUNICATION SYSTEM SOFTWARE and middlewARE - COMSWARE '09*, p. 1, 2009.
- [20] J. Zebedee, P. Martin, K. Wilson, and W. Powley, "An Adaptable Context Management Framework for Pervasive Computing," in *Context-Aware Mobile and Ubiquitous Computing for Enhanced Usability*, 2009, pp. 114-146.
- [21] H. K. Pung, T. Gu, W. Xue, et al., "Context-Aware Middleware for Pervasive Elderly Homecare," *IEEE Journal on Selected Areas in Communications (JSAC)*, Special issue on wireless healthcare, vol. 27, no. 4, pp. 510-524, 2009.
- [22] G. Castelli and F. Zambonelli, "Contextual Data Management and Retrieval: a Self-organized Approach," in *2009 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, 2009, pp. 535-538.
- [23] A. Mehrotra, *GSM System Engineering. Mobile Communications Series*, Artech House Publishers., 1997.
- [24] I. Roussaki, M. Strimpakou, C. Pils, N. Kalatzis, and N. Liampotis, "Distributed Context Management in Support of Multiple Remote Users," in *Context-Aware Mobile and Ubiquitous Computing for Enhanced Usability*, 2009, pp. 84-113.
- [25] XMPP, "XMPP Standards Foundation," <http://www.xmpp.org/>, 2004. .
- [26] P. Saint-Andre, "Extensible Messaging and Presence Protocol (XMPP): Core," <http://www.ietf.org/rfc/rfc3920.txt>, 2004. .
- [27] "XMPP Standards Foundation (XSF). XMPP Extensions," <http://xmpp.org/xmpp-protocols/xmpp-extensions/>, 2010. .
- [28] OpenFire, "OpenFire Server," <http://www.igniterealtime.org/projects/openfire/index.jsp>, 2010.
- [29] "Jena2 Semantic Web Toolkit," <http://jena.sourceforge.net>, 2010.

An efficient user scheduling scheme for downlink Multiuser MIMO-OFDM systems with Block Diagonalization

Mounir Esslaoui and Mohamed Essaaidi
Information and Telecommunication Systems Laboratory
Abdelmalek Essaadi University
Tetouan, Morocco

Abstract—The combination of multiuser multiple-input multiple-output (MU-MIMO) technology with orthogonal frequency division multiplexing (OFDM) is an attractive solution for next generation of wireless local area networks (WLANs), currently standardized within IEEE 802.11ac, and the fourth-generation (4G) mobile cellular wireless systems to achieve a very high system throughput while satisfying quality of service (QoS) constraints. In particular, Block Diagonalization (BD) scheme is a low-complexity precoding technique for MU-MIMO downlink channels, which completely pre-cancels the multiuser interference. The major issue of the BD scheme is that the number of users that can be simultaneously supported is limited by the ratio of the number of base station transmit antennas to the number of user receive antennas. When the number of users is large, a subset of users must be selected, and selection algorithms should be designed to maximize the total system throughput. In this paper, the BD technique is extended to MU-MIMO-OFDM systems and a low complexity user scheduling algorithm is proposed to find the optimal subset of users that should transmit simultaneously, in light of the instantaneous channel state information (CSI), such that the total system sum-rate capacity is maximized. Simulation results show that the proposed scheduling algorithm achieves a good trade-off between sum-rate capacity performance and computational complexity.

Keywords—MU-MIMO; OFDM; scheduling; precoding; Block Diagonalization;

I. INTRODUCTION

The forthcoming breed of wireless standards, commonly referred to as fourth generation (4G) systems (LTE-Advanced, WiMAX or IEEE 802.11ac), are expected to satisfy the increasing demand for high data rates while satisfying quality of service (QoS) constraints. In particular, multiple-input multiple-output (MIMO) technology has been also acknowledged as one of the most promising techniques to achieve dramatic improvement in physical-layer (PHY) performance [1], [2]. While orthogonal frequency division multiplexing (OFDM) has long been regarded as an efficient approach to mitigate the effects of inter-symbol interference (ISI) in frequency-selective channels by dividing the entire

channel into many narrow parallel sub-channels [3]. Therefore, the combination of these two technologies, which is called MIMO-OFDM, is an efficient way for providing high data rate reliable communications [4].

It is well known that multiuser MIMO (MU-MIMO) techniques are capable of significantly increasing capacity compared to traditional MIMO wireless systems [5]. Precoding techniques greatly influence the performance of MU-MIMO transmission by increasing the system capacity and/or reducing the complexity of the receiver when channel state information (CSI) is available at the transmitter [6]. It is assumed that the sum-rate capacity of MIMO broadcast channels can be achieved by applying dirty paper coding (DPC) at the transmitter [7]. Unfortunately, implementing this technique in practice is still a challenging task because of the complicated encoding and decoding schemes [8], especially when the number of users is large. To avoid the complexity of DPC, there have been a lot of interests in developing low-complexity precoding methods.

Block Diagonalization (BD) is an alternative and more practical precoding technique for downlink broadcast MU-MIMO channel and considered as an extension of zero forcing beamforming (ZFBF) for multiantenna receivers [9]. In particular, BD transforms the MU-MIMO downlink channel into parallel single-user MIMO (SU-MIMO) channels and completely nulls all interference between users without inverting the channel. However, this nulling operation imposes a constraint that the number of users that can be simultaneously supported with BD is limited by the ratio of the number of base station (BS) transmit antennas to the number of user receive antennas [9], which is not always feasible. Hence, user scheduling becomes necessary. Optimal scheduling method involves searching through all possible combinations of user subsets. However, as the number of users grows, the size of the search space becomes computationally very complex. To reduce this complexity, a large number of existing user selection approaches (e.g., [10], [11], [12]) has been employed to achieve a sum-rate capacity close to the one promised by DPC. Unfortunately, these frameworks solely targets single-carrier architectures.

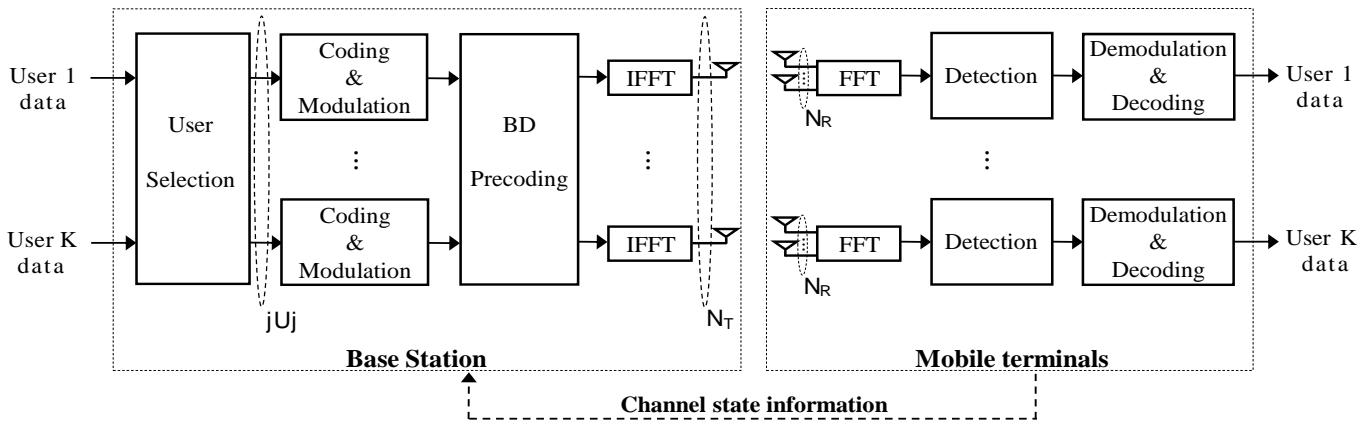


Fig. 1. Block diagram of MU-MIMO-OFDM downlink system with N_T transmit antennas and K users, each with N_R receive antennas

Recently, authors in [13] proposed a low-complexity user selection algorithm for multicarrier networks based on OFDM, where the simultaneously transmitting users must share the utilization of all the subcarriers in the system (i.e., frequency is not used for multiple access as in OFDMA systems). Nevertheless, the work in [13] was limited to single-antenna receivers. In this paper, the BD technique for inter-user interference pre-cancellation is extended to MU-MIMO-OFDM systems and a low-complexity scheduling algorithm, based on [13], is developed considering the availability of multiple-antenna receivers. Numerical results show that the proposed scheme aim to select a subset of users such that the total system throughput is nearly maximized and close to the one obtained by the complex exhaustive search algorithm.

The remainder of this paper is organized as follows: Section II introduces the system model. Block Diagonalization for MU-MIMO-OFDM is presented in Section III. Scheduling algorithms for MU-MIMO-OFDM are described in Section IV and the numerical results are shown in Section V. Finally, Section VI summarizes the main outcomes of the paper and provides hints for further work.

- Notational remark

Boldface letters denote matrix-vector quantities while non-bold letters are used for scalars. The operation $(\cdot)^T$ and $(\cdot)^H$ represent the transpose and the Hermitian transpose of a matrix, respectively, and $\|\mathbf{A}\|_F^2$ denote the Frobenius-norm (F-norm) of a matrix \mathbf{A} . $|\mathcal{U}|$ is the cardinality of subset \mathcal{U} , $\mathbb{E}(\cdot)$ denotes the expectation operator, $\text{Tr}(\cdot)$ is the trace and \mathbb{C} is the set of complex numbers.

II. SYSTEM MODEL

Consider the downlink of a MU-MIMO-OFDM system, as illustrated in Fig. 1, with a single base station (BS) equipped with N_T transmit antennas and $K \geq N_T$ geographically dispersed mobile users, each equipped with N_R receive antennas. The system operates in a total bandwidth W that is exploited by means of N_c OFDM subcarriers. The BS broadcasts to all K users simultaneously over all OFDM subcarriers. Each user k receives from the base station L_k data streams on every subcarrier with $L_k \leq N_R$, resulting in a total of $N_c L_k$ data streams per user. Hence, the BS transmits a

total of $N_c \sum_{k=1}^K L_k$ data streams over all the N_c parallel subcarriers to all the K mobile terminals.

The data streams vector on subcarrier q , $\mathbf{s}_k[q] \in \mathbb{C}^{N_R \times 1}$, contains the data symbols of user k . The $N_T \times 1$ overall data vector of transmitted symbols from the BS antennas on subcarrier q for all K users is

$$\mathbf{x}[q] = [\mathbf{s}_1^T[q] \mathbf{s}_2^T[q] \cdots \mathbf{s}_K^T[q]]^T. \quad (1)$$

The received signal at the k th user on subcarrier q for an arbitrary OFDM symbol is given by

$$\mathbf{y}_k[q] = \mathbf{H}_k[q] \mathbf{x}[q] + \mathbf{n}_k[q], \quad k = 1, \dots, K \quad (2)$$

where $\mathbf{H}_k[q] \in \mathbb{C}^{N_R \times N_T}$ represents the channel gain matrix corresponding to the k th user over the q th subcarrier and $\mathbf{n}_k[q]$ is a zero-mean circularly symmetric additive complex gaussian noise with covariance matrix $\mathbf{R}_\eta = \sigma_\eta^2 \mathbf{I}_{N_R}$. The transmitted signal is subject to an average power constraint P_T , which implies $\text{Tr}(\mathbb{E}(\mathbf{x}[q] \mathbf{x}[q]^H)) \leq P_T$.

Let $\mathcal{U} = \{u_1, \dots, u_{|\mathcal{U}|}\}$ ($|\mathcal{U}| \leq N_T/N_R$) denote a set of users to whom the BS, in light of the available CSI, simultaneously transmits data during a given time slot. Each user's data stream, on subcarrier q , is preprocessed at the transmitter with the precoding matrix $\mathbf{W}_k[q] \in \mathbb{C}^{N_T \times N_R}$, so that the transmitted signal is a linear function that can be written as

$$\mathbf{x}[q] = \sum_{k=1}^{|\mathcal{U}|} \mathbf{W}_k[q] \mathbf{s}_k[q], \quad (3)$$

and thus, the resulting received signal vector for user k on subcarrier q may be rewritten as

$$\mathbf{y}_k[q] = \sum_{j=1}^{|\mathcal{U}|} \mathbf{H}_k[q] \mathbf{W}_j[q] \mathbf{s}_j[q] + \mathbf{n}_k[q] \quad (4)$$

$$= \mathbf{H}_k[q] \mathbf{W}_k[q] \mathbf{s}_k[q] + \sum_{j=1, j \neq k}^{|\mathcal{U}|} \mathbf{H}_k[q] \mathbf{W}_j[q] \mathbf{s}_j[q] + \mathbf{n}_k[q] \quad (5)$$

$$= \mathbf{H}_k[q] \mathbf{W}_k[q] \mathbf{s}_k[q] + \mathbf{c}_k[q] + \mathbf{n}_k[q], \quad (6)$$

where the term $\mathbf{c}_k[q]$ in (6) corresponds to the multi-user interference that represents the major impairment in this scenario.

Transmission strategies for MU-MIMO-OFDM downlink channels are characterized by how we treat the multi-user interference $\mathbf{c}_k[q]$. The availability of channel knowledge at

both ends of the link allows the transmitter to design the precoding matrices to pre-cancel the interference before transmission. Block Diagonalization (BD), proposed in [9] for single-carrier systems, is a precoding technique based on the orthogonalization of the signals that completely eliminate all multiuser interference followed by the waterfilling algorithm [14] to maximize the sum-rate capacity. An extension of the BD approach to multicarrier architecture based on OFDM systems is described in the next section.

III. BLOCK DIAGONALIZATION FOR MU-MIMO-OFDM

The key idea of the Block Diagonalization approach is to find the precoding matrices $\{\mathbf{W}_k[q]\}_k^{|U|}$ for each user k on all subcarriers to pre-eliminate the multiuser interference such that

$$\mathbf{H}_k[q]\mathbf{W}_j[q] = 0, \forall j \neq k. \quad (7)$$

This decomposes the MU-MIMO channel into a set of parallel single-user MIMO (SU-MIMO) channels. Hence the received signal (6) for user k on subcarrier q can be simplified to

$$\mathbf{y}_k[q] = \mathbf{H}_k[q]\mathbf{W}_k[q]\mathbf{s}_k[q] + \mathbf{n}_k[q]. \quad (8)$$

Define the aggregate channel matrix and precoding matrix, on subcarrier q , for all the selected users in the subset \mathcal{U} , respectively, as

$$\mathbf{H}_u[q] = [\mathbf{H}_1^T[q]\mathbf{H}_2^T[q] \cdots \mathbf{H}_{|u|}^T[q]]^T \quad (9)$$

$$\mathbf{W}_u[q] = [\mathbf{W}_1[q]\mathbf{W}_2[q] \cdots \mathbf{W}_{|u|}[q]]. \quad (10)$$

The zero-interference constraint (7) makes the product $\mathbf{H}_u[q]\mathbf{W}_u[q]$ block diagonal and forces $\mathbf{W}_k[q]$ to lie in the null space of the matrix $\bar{\mathbf{H}}_k[q]$ defined as [9]

$$\bar{\mathbf{H}}_k[q] = [\mathbf{H}_{u_1}^T[q] \cdots \mathbf{H}_{u_{k-1}}^T[q]\mathbf{H}_{u_{k+1}}^T[q] \cdots \mathbf{H}_{u_{|u|}}^T[q]]^T. \quad (11)$$

This constraint allows us to define the dimension condition necessary to guarantee that all users can be accommodated. The condition guarantees that data can be transmitted to user k , on the q th subcarrier, if the null space of $\bar{\mathbf{H}}_k[q]$ has a dimension greater than 0. This is satisfied when $\text{rank}(\bar{\mathbf{H}}_k[q]) < N_T$. So for any $\mathcal{H}_u[q]$, block digitalization is possible if

$$\max\{\text{rank}(\bar{\mathbf{H}}_{u_1}[q]) \cdots \bar{\mathbf{H}}_{u_{|u|}}[q]\} < N_T. \quad (12)$$

Assuming the dimension condition is satisfied for all users, let $L_k[q]$ be the rank of $\bar{\mathbf{H}}_k[q]$ and let the singular value decomposition (SVD) of $\bar{\mathbf{H}}_k[q]$ for each user $k \in \mathcal{U}$, on subcarrier q , be defined as

$$\bar{\mathbf{H}}_k[q] = \bar{\mathbf{U}}_k[q]\bar{\Sigma}_k[q][\bar{\mathbf{V}}_k^1[q]\bar{\mathbf{V}}_k^0[q]]^H, \quad (13)$$

where $\bar{\mathbf{U}}_k[q]$ and $\bar{\Sigma}_k[q]$ are the left singular vector matrix and the matrix of singular values of $\bar{\mathbf{H}}_k[q]$, respectively. $\bar{\mathbf{V}}_k^1[q]$ contains the first $L_k[q]$ right singular vectors and $\bar{\mathbf{V}}_k^0[q]$ holds the last $(N_T - L_k[q])$ right singular vectors.

$$\mathcal{R}_{\text{BD}}(\mathcal{U}) = \max_{\mathbf{W}_u[q], \mathbf{H}_k[q]\mathbf{W}_j[q]=0, j \neq k} \frac{1}{N_c} \sum_{q=1}^{N_c} \log_2 \left| \mathbf{I} + \frac{1}{\sigma_n^2} \mathbf{H}_u[q]\mathbf{W}_u[q]\mathbf{W}_u^H[q]\mathbf{H}_u^H[q] \right| \quad (21)$$

Thus, $\bar{\mathbf{V}}_k^0[q]$ constitutes an orthogonal basis for the null space of $\bar{\mathbf{H}}_k[q]$. To satisfy the zero-interference constraint, the precoding matrix can be chosen as

$$\mathbf{W}_k[q] = \bar{\mathbf{V}}_k^0[q]\mathbf{A}_k[q], \quad (14)$$

where $\mathbf{A}_k[q]$ is the $L_k[q] \times N_R$ transmit beamformer matrix for user k on subcarrier q . By doing so, the downlink system reduces to $|\mathcal{U}|$ parallel non-interfering single user MIMO channels, where the equivalent independent channel after precoding for the k th user on the q th subcarrier is expressed as

$$\tilde{\mathbf{H}}_k[q] = \mathbf{H}_k[q]\bar{\mathbf{V}}_k^0[q]. \quad (15)$$

Hence, $\mathbf{A}_k[q]$ can be viewed as a transmit beamformer matrix on $\tilde{\mathbf{H}}_k[q]$ for user k on subcarrier q and can be found through the SVD of the projection of the channel of the k th user on the null space of $\bar{\mathbf{H}}_k[q]$, resulting in the product $\mathbf{H}_k[q]\bar{\mathbf{V}}_k^0[q]$. The SVD of the product is expressed as

$$\mathbf{H}_k[q]\bar{\mathbf{V}}_k^0[q] = \mathbf{U}_k[q] \begin{bmatrix} \Sigma_k[q] & 0 \\ 0 & 0 \end{bmatrix} [\mathbf{V}_k^1[q]\mathbf{V}_k^0[q]]^H, \quad (16)$$

where $\Sigma_k[q]$ is a $\tilde{L}_k \times \tilde{L}_k$ matrix of singular values of $\mathbf{H}_k[q]\bar{\mathbf{V}}_k^0[q]$, with $\tilde{L}_k = \text{rank}(\mathbf{H}_k[q]\bar{\mathbf{V}}_k^0[q])$. $\mathbf{V}_k^1[q]$ is the matrix that holds the first \tilde{L}_k right singular vectors of $\mathbf{H}_k[q]\bar{\mathbf{V}}_k^0[q]$. The product of $\bar{\mathbf{V}}_k^0[q]$ and $\mathbf{V}_k^1[q]$ now produces an orthogonal basis of dimension \tilde{L}_k and represents the transmission vectors that maximize the information rate for user k on subcarrier q subject to producing zero interference. The beamformer matrix $\mathbf{A}_k[q]$ can hence be written as

$$\mathbf{A}_k[q] = \mathbf{V}_k^1[q]\mathbf{\Lambda}^{1/2}[q], \quad (17)$$

where $\mathbf{\Lambda}[q]$ is a diagonal power loading matrix, on the q th subcarrier, whose entries are the allocated powers whose optimal values are obtained by the waterfilling algorithm [14] applied to the diagonal elements of the matrix $\Sigma[q]$ that is expressed as

$$\Sigma[q] = \begin{bmatrix} \Sigma_1[q] & & \\ & \ddots & \\ & & \Sigma_{|u|}[q] \end{bmatrix}. \quad (18)$$

For each user $k \in \mathcal{U}$ on subcarrier q , the precoding matrix is then

$$\mathbf{W}_k[q] = \bar{\mathbf{V}}_k^0[q]\mathbf{V}_k^1[q]. \quad (19)$$

Thus, the aggregate precoding matrix on each subcarrier q becomes

$$\mathbf{W}_u[q] = [\bar{\mathbf{V}}_{u_1}^0[q]\mathbf{V}_{u_1}^1[q]\bar{\mathbf{V}}_{u_2}^0[q]\mathbf{V}_{u_2}^1[q] \cdots \bar{\mathbf{V}}_{u_{|u|}}^0[q]\mathbf{V}_{u_{|u|}}^1[q]]\mathbf{\Lambda}^{1/2}[q]. \quad (20)$$

The sum-rate capacity of the system resulting from BD over all subcarriers is expressed as (21) and (22), shown at the bottom of the page.

$$= \max_{\mathbf{H}_k[q] \mathbf{W}_j[q]=0, j \neq k} \frac{1}{N_c} \sum_{q=1}^{N_c} \sum_{k=1}^{|\mathcal{U}|} \log_2 \left| \mathbf{I} + \frac{1}{\sigma_n^2} \mathbf{H}_k[q] \mathbf{W}_k[q] \mathbf{W}_k^H[q] \mathbf{H}_k^H[q] \right|. \quad (22)$$

The waterfilling algorithm maximizes the system's sum-rate capacity. With $\mathbf{W}_U[q]$ chosen as in (14), the sum-rate capacity of the BD method in (22) becomes

$$\mathcal{R}_{\text{BD}}(\mathcal{U}) = \max_{\Lambda[q]} \frac{1}{N_c} \sum_{q=1}^{N_c} \log_2 \left| \mathbf{I} + \frac{\Sigma^2[q] \Lambda[q]}{\sigma_n^2} \right|, \quad (23)$$

where this maximization is subject to the power constraint P_T . The block diagonalization technique applied is summarized in Algorithm 1.

Algorithm 1 : Block Diagonalization algorithm.

1: For $q = 1 \dots N_c$ and $k = 1 \dots |\mathcal{U}|$, compute SVD of $\bar{\mathbf{H}}_k[q]$.

$$\bar{\mathbf{H}}_k[q] = \bar{\mathbf{U}}_k[q] \bar{\Sigma}_k[q] [\bar{\mathbf{V}}_k^1[q] \bar{\mathbf{V}}_k^0[q]]^H.$$

2: Compute the SVD of the projection of $\mathbf{H}_k[q]$ on the right null space of $\bar{\mathbf{H}}_k[q]$

$$\mathbf{H}_k[q] \bar{\mathbf{V}}_k^0[q] = \mathbf{U}_k[q] \begin{bmatrix} \Sigma_k[q] & 0 \\ 0 & 0 \end{bmatrix} [\mathbf{V}_k^1[q] \mathbf{V}_k^0[q]]^H.$$

3: Use the waterfilling algorithm on the diagonal elements of $\Sigma[q]$ to determine the optimal power loading matrix $\Lambda_k[q]$ under the power constraint P_T .

4: Set the aggregate precoding matrix on each subcarrier q as

$$\mathbf{W}_U[q] = [\bar{\mathbf{V}}_1^0[q] \mathbf{V}_1^1[q] \bar{\mathbf{V}}_2^0[q] \mathbf{V}_2^1[q] \dots \bar{\mathbf{V}}_{|\mathcal{U}|}^0[q] \mathbf{V}_{|\mathcal{U}|}^1[q]] \Lambda^{1/2}[q].$$

IV. SCHEDULING ALGORITHMS FOR MU-MIMO-OFDM

A. Optimal scheduling algorithm

Let $K_{\max} = \lceil N_T/N_R \rceil$ be the maximum number of simultaneously selected users at a given time slot, where $\lceil \cdot \rceil$ is the ceiling function. When K is large, there are possibly many supportable subsets \mathcal{U} that satisfy the dimensionality constraint, i.e., $|\mathcal{U}| \leq K_{\max}$. Consider the sum-rate problem; the multiuser diversity gain is achieved by selecting an optimal subset of users that maximizes the sum-rate capacity in (23). Mathematically, the optimal sum-rate capacity of the BD scheme with multiuser diversity is given by

$$\mathcal{R}_{\text{BD}}^{\text{opt}}(\mathcal{U}) = \max_{\substack{\mathcal{U} \subset \{1,2,\dots,K\} \\ 1 \leq |\mathcal{U}| \leq K}} \mathcal{R}_{\text{BD}}. \quad (24)$$

Therefore, the optimal scheduling requires an exhaustive search through all $\sum_{i=1}^{K_{\max}} \binom{K}{i}$ possible combinations of subsets of simultaneously supported users and is computationally very complex especially when the number of users is large. Thus, significant research efforts have been made to find suboptimal low-complexity algorithms that can achieve a significant fraction of the best performance provided by a full search method. In this paper, we propose a low-complexity multicarrier user selection algorithm for BD.

B. Multicarrier user selection algorithm for BD

The multicarrier user selection algorithm for BD (MUS-BD), summarized in Algorithm 2, selects a subset \mathcal{U} of users, up to K_{\max} , from a total user pool $\Omega_1 = \{1, 2, \dots, K\}$ such that the sum-rate capacity (23) is maximized for a given

time slot. The base station will then simultaneously transmit to all users in \mathcal{U} using all the OFDM subcarriers of the system. The algorithm works as follows: The step (I) initializes counters and the user pool Ω_1 with the total K active users in the cell. In step (II.a), the matrix $\mathbf{G}_k[q]$ defined as

$$\mathbf{G}_k[q] = \mathbf{H}_k[q] - \sum_{j=1}^{i-1} \mathbf{H}_k[q] \hat{\mathbf{V}}_{u_j}^{1H}[q] \hat{\mathbf{V}}_{u_j}^1[q], \quad (25)$$

is found as the component of the channel matrix $\mathbf{H}_k[q]$, for each user k on the q th subcarrier, orthogonal to the subspace spanned by $\{\mathbf{G}_{u_1}[q] \dots \mathbf{G}_{u_{i-1}}[q]\}$ where $\{u_1 \dots u_{i-1}\}$ denote the indexes of the previously selected users. Step (II.b) selects the best user u_i , at each iteration i , the one with the maximum squared F-norm of its user channel component \mathbf{G}_k averaged over all subcarriers, that as,

$$u_i = \operatorname{argmax}_{k \in \Omega_i} \frac{1}{N_c} \sum_{q=1}^{N_c} \|\mathbf{G}_k[q]\|_{\text{F}}^2. \quad (26)$$

Algorithm 2 : Multicarrier user selection algorithm for BD

(I) Initialization:

$$\Omega_1 = \{1, 2, \dots, K\}; K_{\max} = \left\lfloor \frac{N_T}{N_R} \right\rfloor; \mathcal{U} = \emptyset; i = 1$$

(II) Main Loop:

while $(|\Omega_i| \neq 0) \ \& \ (|\mathcal{U}| < K_{\max})$

(a) Orthogonality measure computation:

for Each user $k \in \Omega_i$ **do**

for Each subcarrier q **do**

$$\mathbf{G}_k[q] = \mathbf{H}_k[q] - \sum_{j=1}^{i-1} \mathbf{H}_k[q] \hat{\mathbf{V}}_{u_j}^{1H}[q] \hat{\mathbf{V}}_{u_j}^1[q]$$

end for

end for

(b) Orthogonal user selection

$$u_i = \operatorname{argmax}_{k \in \Omega_i} \frac{1}{N_c} \sum_{q=1}^{N_c} \|\mathbf{G}_k[q]\|_{\text{F}}^2$$

$$\mathcal{U} = \mathcal{U} \cup \{u_i\}; \Omega_i = \Omega_i \setminus \{u_i\}$$

(c) Singular value decomposition:

for Each subcarrier q **do**

$$\mathbf{G}_{u_i}[q] = \hat{\mathbf{U}}_{u_i}[q] \hat{\Sigma}_{u_i}[q] [\hat{\mathbf{V}}_{u_i}^1[q] \hat{\mathbf{V}}_{u_i}^0[q]]^H$$

end for

(d) Intermediate user grouping:

$$\Omega_{i+1} = \left\{ k \in \Omega_i, k \notin \mathcal{U}: \frac{1}{N_c} \sum_{q=1}^{N_c} \frac{\|\mathbf{H}_k[q] \hat{\mathbf{V}}_{u_i}^{1H}[q]\|_{\text{F}}^2}{\|\mathbf{H}_k[q]\|_{\text{F}}^2 \cdot \|\hat{\mathbf{V}}_{u_i}^1[q]\|_{\text{F}}^2} < \alpha \right\}$$

$$i = i + 1$$

(III) Output: The selected user subset \mathcal{U}

Then the user u_i is included in the subset \mathcal{U} of selected users and discarded from the total user pool Ω_i . Singular value decomposition of the matrix $\mathbf{G}_{u_i}[q]$ is computed in step (II.c) for the selected user u_i at the i th iteration over all subcarriers as,

$$\mathbf{G}_{u_i}[q] = \widehat{\mathbf{U}}_{u_i}[q] \widehat{\Sigma}_{u_i}[q] [\widehat{\mathbf{V}}_{u_i}^1[q] \widehat{\mathbf{V}}_{u_i}^0[q]]^H, \quad (27)$$

in order to obtain the first $\widehat{L}_i[q]$ right singular vectors $\widehat{\mathbf{V}}_{u_i}^1[q]$, where $\widehat{L}_i[q] = \text{rank}(\mathbf{G}_i[q])$, used to compute the orthogonalization process in step (II.a) for each iteration $i > 1$. Finally, in step (II.d), in order to reduce the computational complexity of the algorithm, the users whose orthogonality coefficient α_k , given by,

$$\alpha_k = \frac{1}{N_c} \sum_{q=1}^{N_c} \frac{\|\mathbf{H}_k[q] \widehat{\mathbf{V}}_{u_i}^1[q]\|_F^2}{\|\mathbf{H}_k[q]\|_F^2 \|\widehat{\mathbf{V}}_{u_i}^1[q]\|_F^2} \quad (28)$$

exceeds a certain threshold α , are discarded from the total user pool. Note that α is a small positive value used to update the set of orthogonal users and to avoid loss in the sum-rate capacity. Hence, the selection of optimal value of α becomes crucial. Its selection will be discussed using simulation in Section V. The user selection process stops when the number of users in the active subset \mathcal{U} equals the maximum number K_{\max} or the list of remaining users is empty. Once the users have been chosen, the BD algorithm described in Section III is used to compute the precoding matrices for the selected users.

V. NUMERICAL RESULTS

The simulations consider the use of parameters currently found in the latest WLAN standard IEEE 802.11n and that will surely also form part of the forthcoming IEEE 802.11ac norm. The system has been configured to operate at 5.25 GHz carrier frequency on a bandwidth of $W = 20$ MHz with $N_c = 64$ subcarriers. The channel profile used to generate the frequency-selective channel responses correspond to profiles B (residential) from channel models developed within the IEEE 802.11n standard [15]. The BS is assumed to operate with either $N_T = 4$ and $N_T = 8$ transmit antennas while the number of receive antennas at each mobile station has been fixed to $N_R = 2$. Unless otherwise specified, the BS is assumed to transmit with total power P_T to ensure that every subcarrier, on average, operates with an $\text{SNR}[q] = 10$ dB.

In Fig. 2 we plot the sum-rate capacity, obtained using (23), under the proposed MUS-BD algorithm as a function of the orthogonality parameter α . The plots are obtained by averaging over 1000 independent channel realizations. The total number of users in the cell has been fixed respectively to 10, 50 and 100. It is observed that, the orthogonality parameter α is mainly driven by the number of transmit antennas and the number of active users in the cell. Note that, smaller thresholds result in small user subset at each step of the algorithm, which in turn lowers the search complexity. However, too small threshold incurs throughput loss due to the reduced multiuser diversity gain. Nevertheless, if it is too large, non-orthogonal users are selected and the channel gains are reduced. Hence, it is very important to find the optimal value of α that yield sum-rate capacity close to that with exhaustive search method. In this figure, we can see that optimal values decrease as the

number of users grows. For $N_T = 4$, (see Fig. 1.a), optimal

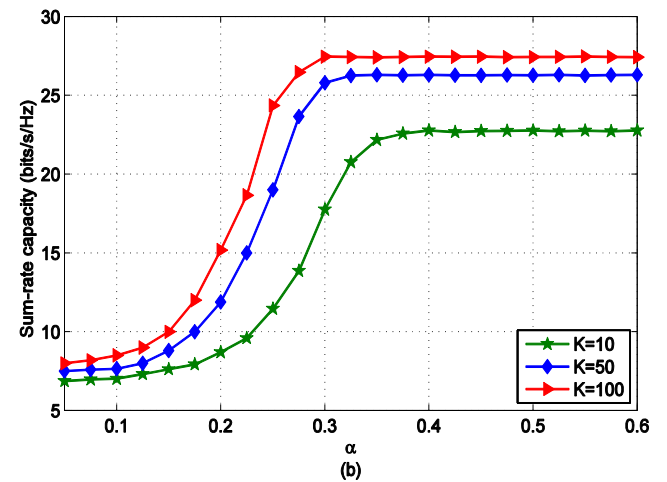
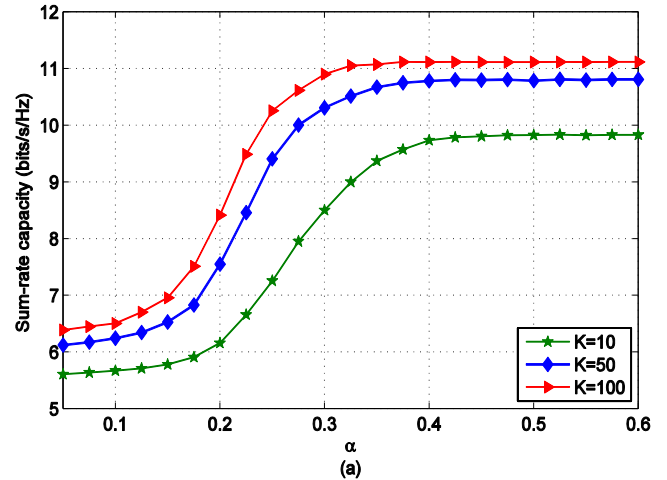


Fig. 2. Sum-rate capacity vs. α . $\text{SNR}[q] = 10$ dB. (a) $N_T = 4$, $N_R = 2$, $K_{\max} = 2$. (b) $N_T = 8$, $N_R = 2$, $K_{\max} = 4$.

threshold which maximizes the sum-rate capacity are in the range $[0.35 - 0.45]$. However, for large N_T (e.g. $N_T = 8$ in Fig. 1.b), this trend changes slightly and thus the optimal threshold that results in maximum throughput are in the range $[0.3 - 0.4]$. For the rest of simulations optimal value α , which maximizes the sum-rate capacity, has been considered.

Fig. 3 plots the sum-rate capacity of MUS-BD, optimal scheduling and conventional opportunistic multicarrier TDMA (opp-MC-TDMA) [16] versus the number of users in the cell. In conventional opp-MC-TDMA scheme, the BS selects the single user experiencing the best channel realization at each time slot that will be allocated all the spectrum and power resources. It can be observed that, as the number of users and transmit antennas become larger, more simultaneously selected users will be found by the proposed algorithm and thus the system sum-rate capacity will be improved. Both MUS-BD and optimal scheduling algorithms manage to achieve more than 65% and 200% increase in sum-rate capacity for $N_T = 4$ (see Fig. 1.a) and $N_T = 8$ (see Fig. 1.b), respectively, compared to conventional opp-MC-TDMA technique. Moreover, the

proposed MUS-BD algorithm performs very close to the optimal scheduling algorithm with much lower computational

than 90% of the sum-rate capacity offered by the optimal scheduling method with much lower computational complexity.

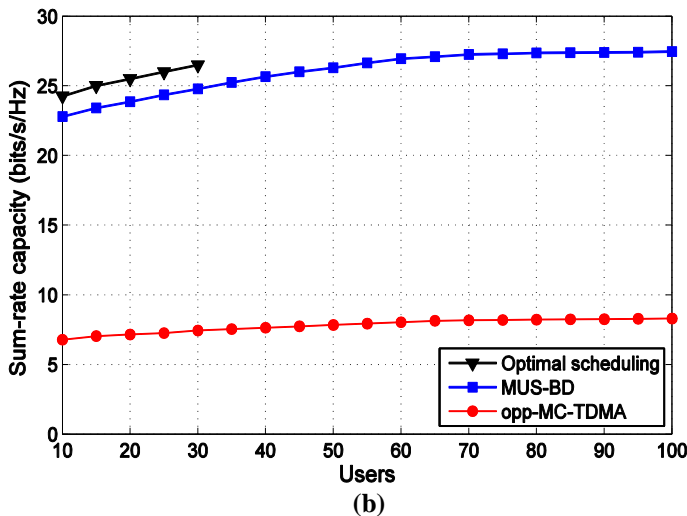
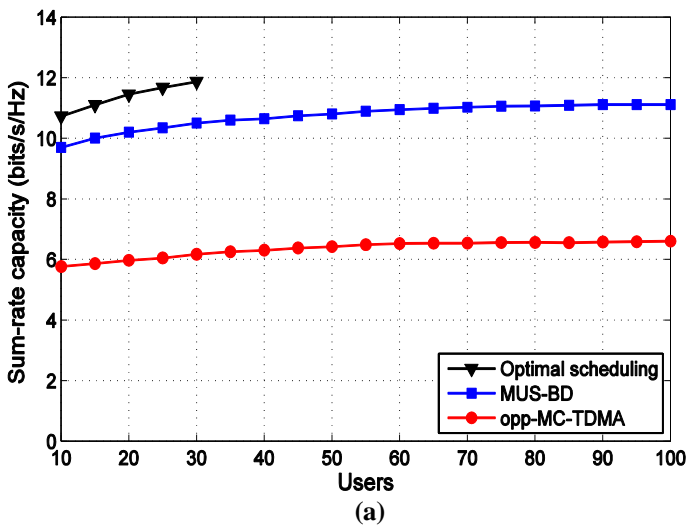


Fig.3. Sum-rate capacity vs. number of users K . $SNR[q] = 10$ dB.
(a) $N_T = 4, N_R = 2, K_{max} = 2$. (b) $N_T = 8, N_R = 2, K_{max} = 4$.

complexity. Note that, for optimal user selection scheme, it was only feasible to compute values for up to $K = 30$ users in the cell due to the high computational complexity requirements.

Finally, Fig. 4 compares the sum-rate capacity of MUS-BD, optimal scheduling and opp-MC-TDMA as a function of the SNR. In this scenario the total number of users in the cell has been fixed to $K = 10$ and the number of transmit antennas at BS to $N_T = 4$. It can be seen that in the low SNR regime, e.g. $SNR[q] = 0$ dB, the performance of all schemes are very close to each other because fewer users are scheduled to obtain the larger transmit diversity. However, at high SNR regime, both MUS-BD and optimal scheduling schemes tend to select the maximum number of users that can be supported simultaneously by BS and thus the gain in sum-rate capacity gap is noticeable compared to the opp-MC-TDMA technique. Moreover, the proposed MUS-BD algorithm achieves more

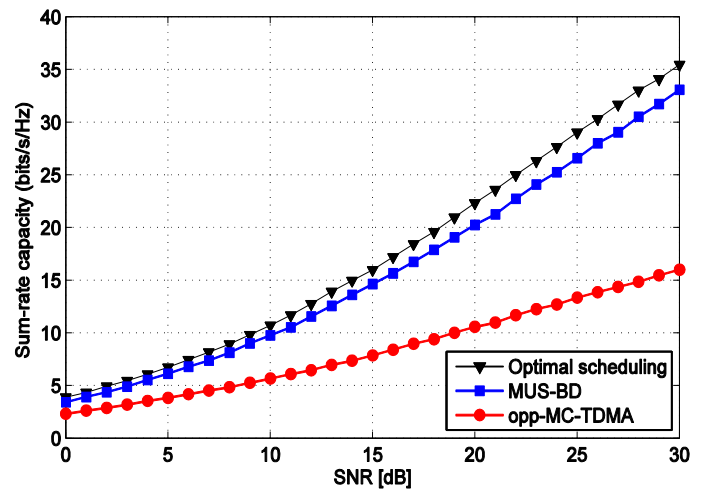


Fig.4. Sum-rate capacity vs. SNR. $K = 10, N_T = 4, N_R = 2, K_{max} = 2$.

VI. CONCLUSIONS

This paper has introduced a suboptimal user selection algorithm for multiuser MIMO-OFDM systems employing Block Diagonalization. Its goal is to select a subset of users to maximize the total sum-rate capacity while keeping the complexity low. The proposed scheme is composed of a user selection step that selects the users experiencing good channel realizations, and a subcarrier-specific precoder based on BD designed to pre-eliminate the multiuser interference. Simulation results have shown that the proposed scheme achieves a very significant increase in sum-rate capacity compared to conventional opportunistic MC-TDMA technique and performs close to the optimal scheduling method with much lower computational complexity.

The user selection algorithm proposed in this paper requires complete channel knowledge at the transmitter and, furthermore, it did not consider any of the fairness issues MU-MIMO brings along. Further work will address both issues by considering partial channel knowledge at the BS and introducing mechanisms to improve fairness among users.

ACKNOWLEDGEMENT

The authors would like to thank the editor and the anonymous reviewers, whose careful consideration of the manuscript improved the presentation of the material.

REFERENCES

- [1] E. Telatar, "Capacity of multi-antenna Gaussian channels," Eur. Trans. Telecommun., vol. 10, no. 6, pp. 585–595, Nov. 1999.
- [2] G. Foschini and M. Gans, "On limits of wireless communications in a fading environment when using multiple antennas," Wireless Personal Communications, vol. 6, pp. 311–335, Mar 1998.
- [3] R. van Nee and R. Prasad, OFDM for wireless multimedia communications Artech House, 2000.
- [4] A. J. Paulraj, D. A. Gore, R. U. Nabar, and H. Bolcskei, "An overview of MIMO communications a key to gigabit wireless," Proceedings of the IEEE, vol. 92, no. 2, pp. 198–218, 2004.

AUTHORS PROFILE

- [5] D. Gesbert, M. Kountouris, R. Heath, C.-B. Chae, and T. Salzer, "Shifting the MIMO paradigm," *IEEE Sig. Proces. Mag.*, vol. 24, no. 5, Sep. 2007.
- [6] A. Scaglione, P. Stoica, S. Barbarossa, G. Giannakis, and H. Sampath, "Optimal designs for space-time linear precoders and decoders," *IEEE Trans. on Signal Processing*, vol. 6, pp. 311–335, Mar. 1998.
- [7] M. Costa, "Writing on dirty paper," *IEEE Trans. Info. Theory*, vol. 29, pp. 439–441, May. 1983.
- [8] N. Jindal and A. Goldsmith, "Dirty-paper coding versus tdma for mimo broadcast channels," *IEEE Trans. on Information Theory*, vol. 51, no. 5, pp. 1783–1794, May 2005.
- [9] Q. Spencer, A. Swindlehurst, and M. Haardt, "Zero-forcing methods for downlink spatial multiplexing in multiuser MIMO channels," *IEEE Trans. on Signal Processing*, vol. 52, no. 2, Feb. 2004.
- [10] Zukang Shen; Runhua Chen; Andrews, J.G.; Heath, R.W.; Evans, B.L., "Low complexity user selection algorithms for multiuser mimo systems with block diagonalization," *Signal Processing, IEEE Transactions on*, vol. 54, no. 9, p. 3658,3663, Sep 2006.
- [11] M. Eslami and W. A. Krzymien, "Efficient transmission technique for mimo-ofdm broadcast channels with limited feedback," in *Spread Spectrum Techniques and Applications, 2008. ISSSTA '08. IEEE 10th International Symposium on*, aug. 2008, pp. 237 –241.
- [12] T. Yoo and A. Goldsmith, "On the optimality of multiantenna broadcast scheduling using zero-forcing beamforming," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 3, pp. 528–541, Mar. 2006.
- [13] M. Esslaoui, F. Riera-Palou, and G. Femenias, "Opportunistic multiuser MIMO for OFDM networks." in *Proc. IEEE 8th Multi-Carrier Systems & Solutions*, May 2011.
- [14] N. Jindal, W. Rhee, S. Vishwanath, S. A. Jafar, and A. Goldsmith, "Sum power iterative water-filling for multi-antenna gaussian broadcast channels," *IEEE Trans. on Information Theory*, vol. 51, no. 4, pp. 1570–1580, April 2005.
- [15] J. Kermaol, L. Schumacher, K. Pedersen, P. Mogensen, and F. Frederiksen, "A stochastic MIMO radio channel model with experimental validation," *IEEE JSAC*, vol. 20, no. 6, pp. 1211–1226, Aug 2002.
- [16] P. Viswanath, D. Tse, and R. Laroia, "Opportunistic beamforming using dumb antennas," *IEEE Trans. on Inf. Theory*, vol. 48, no. 6, Jun. 2002.

Mounir Esslaoui received the Master degree in Electronics and Telecommunications from Abdelmalek Essaadi University, Tetouan, Morocco in 2008. He is currently working towards the Ph.D. degree in information and Communication sciences at Abdelmalek Essaadi University. He was an exchanged Ph.D. student at Mobile Communications Group laboratory, University of the Balearic Islands, Spain, from 2009 till 2012. His research is currently focused on the forthcoming generation of wireless communication networks with particular emphasis on multiuser MIMO and multicarrier systems.

Mohamed Essaaidi received the Ph.D. degree in Electrical Engineering in 1997 from Abdelmalek Essaadi University, Tetouan, Morocco. He is the current Director of National College of IT (ENSIAS) of Mohammed 5th Souissi University, Rabat, Morocco and he was a Professor of Electrical and Computer Engineering at Abdelmalek Essaadi University, Morocco from 1993 till 2011. He is an IEEE Senior Member, the founder and Chairman of the IEEE Morocco Section, founder of IEEE Computer & Communication Societies Joint Morocco Chapter, Founder and Chair of IEEE Antennas and Propagation Society and Microwave Theory and Techniques Society Morocco Joint Chapter and founder of IEEE Education Society Morocco Chapter. He has been also the founding Director of the Morocco Office of Arab Science and Technology Foundation, ASTF (2006-2009) and the Coordinator of ASTF RD&I Network of Electro-Technology since 2006. He has also founded several IEEE Student Branches in different Moroccan universities and engineering schools.

He has authored and co-authored 5 books and more than 120 papers in international refereed journals and conferences in the field of Electrical, Information and Communication Technologies. He has been the Editor-in-Chief of International Journal on Information and Communication Technologies, Serial Publications, India since 2007. He is also an active member of the editorial boards of several IEEE and other indexed international journals in the field of information and communication technologies.

Dr. Essaaidi holds four patents on antennas for very high data rate UWB and multi-band wireless communication systems and high resolution medical imaging systems. Furthermore, he has co-organized / been involved in the juries of several national and international competitions aiming at fostering research, development and innovation such as Moroccan Engineers Week 2006, 2007, "Made in Morocco", Arab Science and Technology Foundation (ASTF) "Made in Arabia" Competitions in 2007 and 2009, Qatar Foundation Stars of Science 2010 and Intel Science Competition 2011.

QOS, Comparison of BNP Scheduling Algorithms with Expanded Fuzzy System

Amita Sharma
M.Tech Scholar (CSE)
SBBSIET, Jalandhar

Harpreet Kaur
AP, CSE
SBBSIET, Jalandhar

Abstract—Parallel processing is a field in which different systems run together to save the time of the processing and to increase the performance of the system. It has been also seen that it works somewhat up to the load balancing concept. Previous algorithms like HLFET, MCP, DLS, ETF, have shown that they can reduce the burden of the processor by working simultaneous working system. In our research work, we have combined HLFET, MCP, DLS, ETF with FUZZY logic to check out what effect it makes to the parameters which has been taken from the previous work done like Makespan, SLR, Speedup, Process Utilization. It has been found that the fuzzy logic system works better than the single algorithm.

Keywords—Parallel Processing; DAG, BNP; Fuzzy logic; Multiprocessor.

I. INTRODUCTION

Parallel processing is one of the emerging concepts that is used to execute number of tasks on different number of processors at the same [5] time. With the help of parallel processing one is able to solve complex and computation intensive problems in an effective way. Depending upon nature of nodes the parallel processing system can be divided into two categories known as homogenous or heterogeneous parallel system. In homogenous environment the number of processor used for executing the different tasks are similar in capacity and in case of heterogeneous environment the tasks are allocated on various processors of different capacity and speed.

Independent of the environment the objective of parallel processing is to improve the execution speed and to minimize the makespan of task execution. This is done by using the different precious and competent task scheduling algorithm. The objective of task scheduling algorithm is to allocate the different tasks to different processor so that execution speeds of the task increases and the overall execution time of the task decreases. Task scheduling algorithms take care of all the above said factors and can be represented by Directed Acyclic Graphs (DAG)[15].

II. MULTIPROCESSOR SCHEDULING ALGORITHMS [3]

A. *Bounded Number of Processors (BNP) scheduling algorithms:* These algorithms schedule the DAG to a bounded number of processors directly. The processors are assumed to be fully-connected.

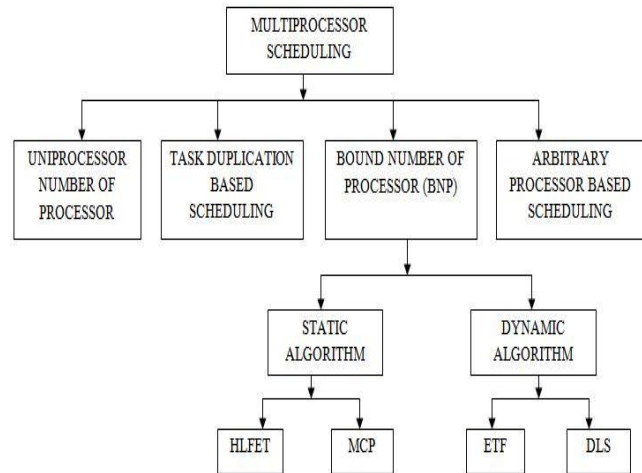


Fig. 1. Multiprocessor scheduling algorithms

- B. *Unbounded Number of Clusters (UNC) scheduling algorithms:* These algorithms schedule the DAG to an unbounded number of clusters. The processors [1][26] are assumed to be fully-connected. The technique employed by these algorithms is also called clustering.
- C. *Task Duplication Based (TDB) scheduling algorithms:* These algorithms also schedule the DAG to an unbounded number of clusters but employ task duplication technique to further reduce the completion time.
- D. *Arbitrary Processor Network (APN) scheduling algorithms:* These algorithms perform scheduling and mapping on the target architectures in which the processors are connected via a network of arbitrary topology. In this paper, BNP scheduling algorithms are discussed and their performance is analyzed.

BNP class of algorithms is categorized into two categories:

Static Algorithms[2] : These algorithms use list scheduling approach. Therefore in static algorithms once the task prioritization phase is finished then and only then the processor selection phase begins.

1. HLFET Algorithm

It is one of the simplest algorithms. Here the HLFET stands for Highest Level First with Estimated Time.

Algorithm Steps:

- 1) Calculate the static Level of the nodes in the DAG.
- 2) Insert all the nodes into a list and sort the list according to descending order of Static Level of the nodes.
- 3) While not the end of the list do
 - Remove the node n_i from the list.
 - Compute the earliest start execution time of n_i for all the processor present in the system.
 - Map the node n_i to the processor that has the least early start execution time.

2. MCP Algorithm

MCP stands for Modified Critical Path. It uses the Latest Start Time attribute for mapping the nodes to processors.

Algorithm Steps:

- 1) Calculate the Latest Start Time (LST) of all the nodes in the DAG.
- 2) Insert all the nodes into a list and sort the list according order of Latest Start Time.
- 3) While not the end of the list do
 - Remove the node from the list
 - Compute the earliest start execution time of n_i for all the processors present in the system.
 - Map the node n_i to the processor that has the least early start execution time.

Dynamic Algorithms: These algorithms also use list scheduling approach. In Dynamic algorithms both the task prioritization phase and processor selection phase goes on side by side.

1. ETF Algorithm

ETF stands for Earliest Task First. This algorithm computes the earliest execution start time for all nodes and selects one with lowest value for scheduling. In this algorithm the ready node stands for that node which has all its parents scheduled.

Algorithm Steps:

- 1) Calculate the Static Level of each node in the DAG.
- 2) In the beginning the ready node list contains only the entry node.
- 3) While the ready node list is not empty do
 - Compute the earliest start time of all the nodes in the ready node list on each processor.
 - Select the node with earliest start time. If two or more nodes have same earliest execution start time values then the node with highest Static Level is selected.
 - Map the selected node to the processor.
 - Add new ready nodes to the ready node list.

2. DLS Algorithm

DLS stands for Dynamic Level Scheduling. It uses the Dynamic Level attribute for scheduling the nodes.

Algorithm Steps:

- 1) Calculate the Static Level of node in DAG.
- 2) In the beginning the ready node list contains only the entry node.
- 3) While the ready node list is not empty do
 - Compute the earliest start time of every node in the ready node list on each processor.
 - Calculate the Dynamic Level of every node in the list
 - Select the node with the largest Dynamic Level.
 - Schedule the node onto the processor.
 - Add new ready nodes to the ready node list.

III. DAG MODEL:

The DAG is generic model of a parallel program consisting of a set of processes among which there are dependencies. Each process is an indivisible unit of execution, expressed by node. A node has one or more inputs and can have one or more output to various nodes. When all inputs are available, the node is triggered to execute. After its execution, it generates its output. In this model, a set of node ($n_1, n_2, n_3 \dots \dots n_n$) are connected by a set of a directed edges, which are represented by (n_i, n_j) where n_i is called the Parent node and n_j is called the child node. A node without parent is called an Entry node and a without child node called an Exit node. The weight of a node, denoted by $w(n_i)$, represents the process execution time of a process. Since [3] each edge corresponds to a message transfer from one process to another, the weight of an edge, denoted by $c(n_i, n_j)$, is equal to the message transmission time from node n_i to n_j . Thus, $c(n_i, n_j)$ becomes zero when n_i and n_j are scheduled to the same processor because intraprocessor communication time is negligible compared with the inter processor communication time. The node and edge weights are usually obtained by estimations. Some [15] variations in the generic DAG model are described below:

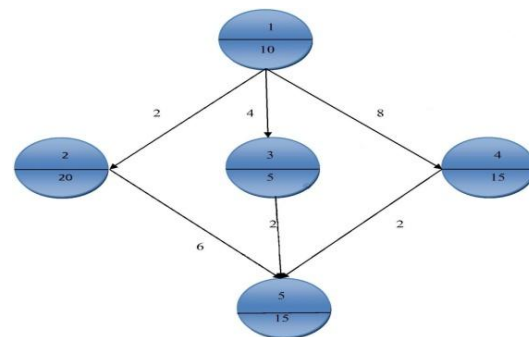


Fig. 2. DAG Model

IV. FUZZY LOGIC

Fuzzy logic is a rigorous mathematical field, and it provides an effective vehicle for modeling the uncertainty [29] in human reasoning. Fuzzy set is uniquely determined by its membership function (MF), and it is also associated with a linguistically meaningful term. Fuzzy logic provides a systematic tool to incorporate human experience. It is based on three core concepts, namely, fuzzy sets, linguistic variables,

and possibility distributions. The importance of fuzzy logic derives from the fact that most modes of [32] human thinking and especially common sense reasoning are approximate in nature. The essential features of fuzzy logic are as follows:

- In fuzzy logic everything is a matter of degree.
- Any logical system can be fuzzified.
- In fuzzy logic, knowledge is interpreted as a collection of elastic or, equivalently, fuzzy constraint on a collection of variables.
- Inference is viewed as a process of propagation of elastic constraints.

V. EXPERIMENTAL RESULTS AND DISCUSSIONS

The analytical results of the entire algorithms under all three case scenarios are determined. Following parameters are implemented for given scheduling algorithms and results are shown graphically:

Makespan: It is defined as the completion time of the algorithm. Lesser the Makespan less time to execute the algorithm more efficient is the algorithm. Makespan is calculated by measuring the finishing time of the exit task by the algorithm.

Processor Utilization: In multiprocessor system, processor work in parallel. There can be some scenario when large amount of work is done by one processor and lesser by others, which the work distribution is not proportionate

Processor Utilization (%) = (total time taken of Scheduled tasks/Makespan)*100

Speed Up: It is defined as the ratio of time taken by serial algorithm work to the time taken by the algorithm to perform the same work.

Speed Up = Time taken by serial algorithm/Time taken by parallel Algorithm

Scheduled length Ratio (SLR): It is defined as the ratio of Makespan of the algorithm to Critical Path values of the DAG. The lesser the values of SLR the more efficient is the algorithm, but The SLR cannot be less than the Critical path values.

Scheduled length Ratio = Makespan / Critical Path

5.2.1 Scenario 1: 5 Task Nodes : From the graphs shown below ,we can say that results from all the algorithms using 5 Task Nodes the parameters Makespan, SLR, SpeedUp are throughout the same.

Fig. 3. Values of different parameters for 5 nodes.

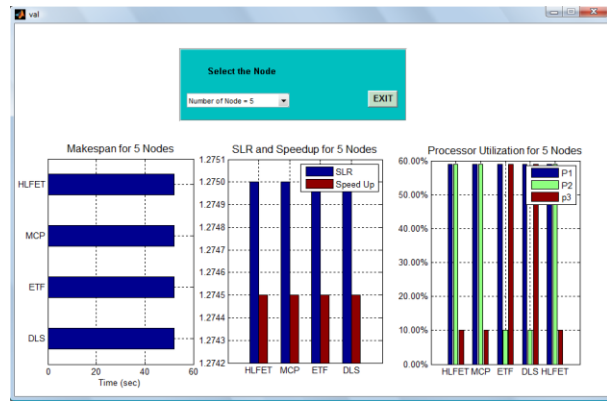


Fig.3.1 Makespan for 5 nodes

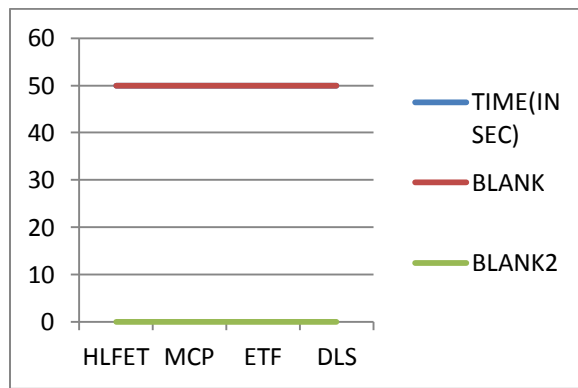


Fig.3.2: SLR & Speed Up for 5 Nodes

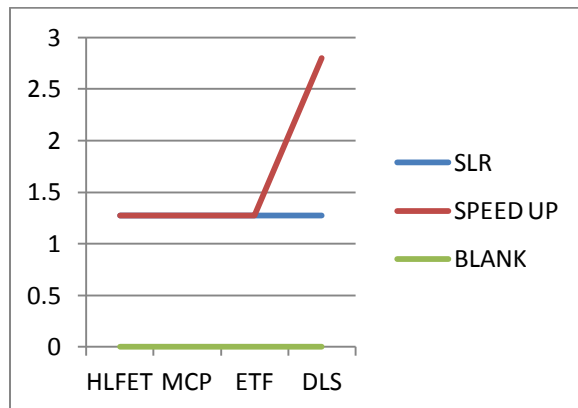
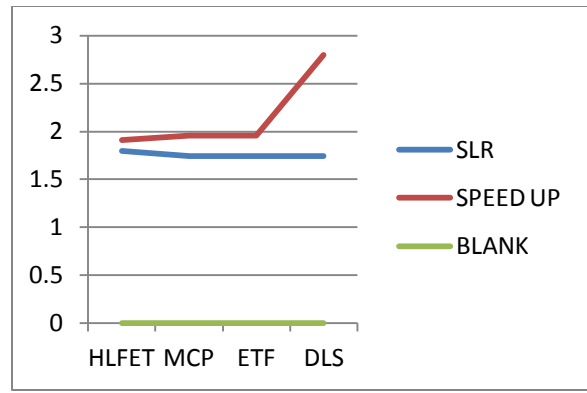
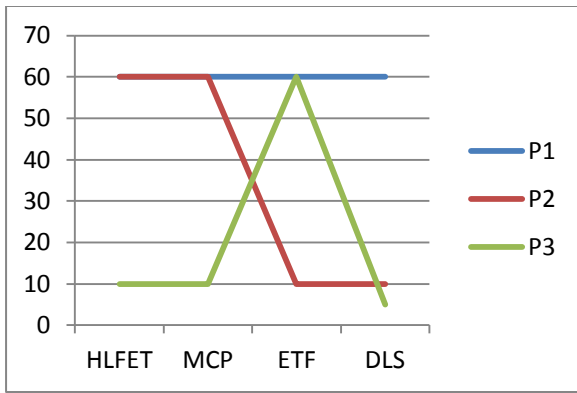


Fig.3.3: Process Utilization for 5 Nodes



5.2.2 Scenario 2: 10 Task Nodes: Makespan of HLFET,MCP,ETF is equal but Makespan of DLS is higher than the others. Similarly, SLR and Speedup is same for MCP,ETF, DLS but different for HLFET. The same applies for Process Utilization.

Fig.4.2:SLR & Speed up for 10 nodes

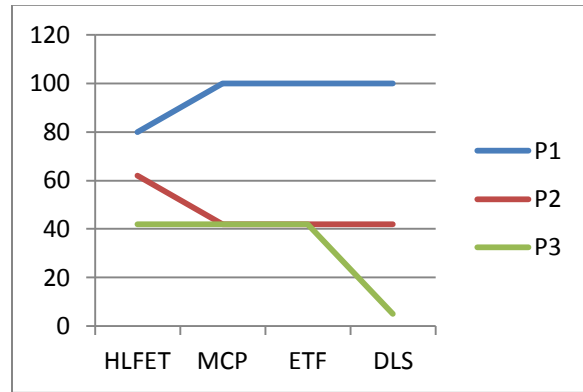
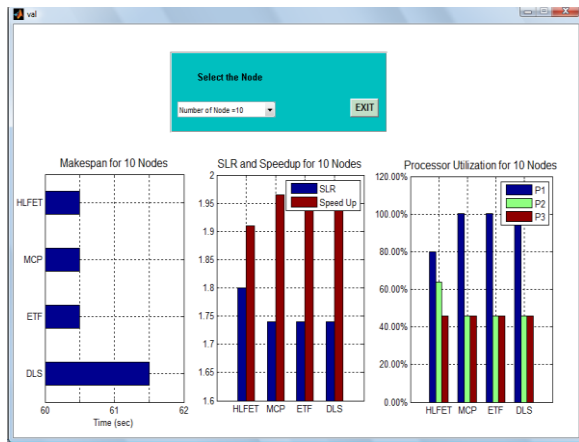


Fig.4.3: Process Utilization for 10 Nodes

5.2.3 Scenario 3: 15 Task Node: MCP shows the highest Makespan and Process Utilization values than others whereas ETF shows the lowest Speedup and SLR values.

Fig. 4. Values of different parameters for 10 nodes.

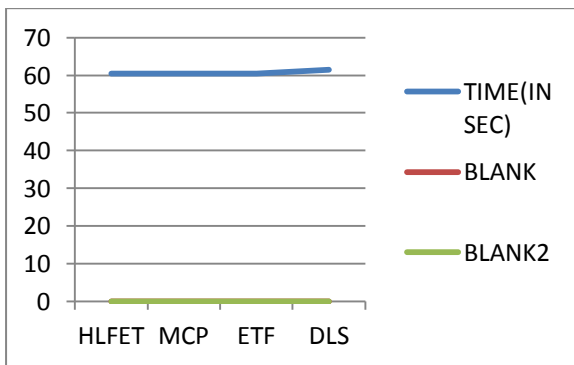
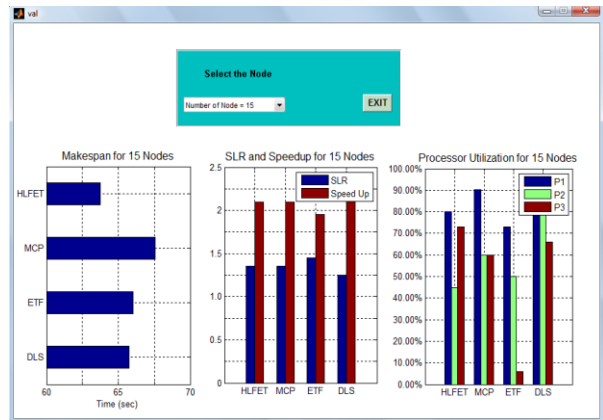


Fig.4.1:Makespan for 10 nodes

Fig. 5. Values of different parameters for 15 nodes.

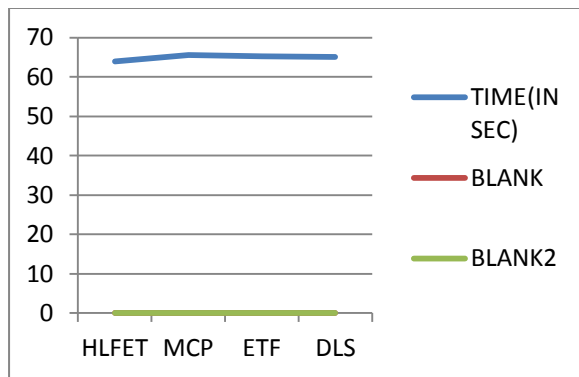


Fig.5.1: Makespan for 15 nodes

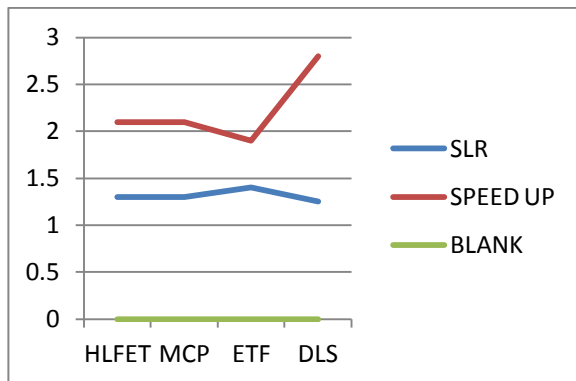


Fig.5.2: SLR & Speed up for 10 nodes

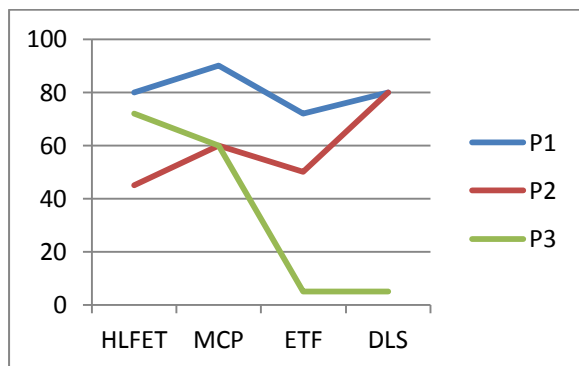


Fig.5.3: Process Utilization for 15 nodes

VI. CONCLUSION

After comparative analysis , following results were obtained:

- Makespan of DLS showed large increase in amount with 10 nodes and MCP for 15nodes.
- Process Utilization remained same for 5 nodes.MCP,ETF,DLS utilized process efficiently than HLFET for 10 nodes with 15 nodes.MCP proved to be better than other algorithms whereas ETF showed large drop in utilization rate.
- SLR remained the same for 5 nodes while HLFET for 10 nodes and ETF for 15 nodes showed lesser SLR.
- ETF was the algorithm with higher Speedup.

VII. FUTURE SCOPE

Although fuzzy logic is very interesting to be implemented but it has a loop hole into it. There is a rule base for each and every type of architecture into it. If we would be introducing a new random node into the system, the fuzzy logic won't have any information about it unless and until we don't mention it into the program by creating a new rule base. Hence in future, if somebody can find a method to optimize the rule set on run time so that the system takes a slight less time to create those rule sets which would make a system little smoother. This task can be achieved with the help of some of the latest methods formally known as Neural Classifications. There are some other methods also which can be tried for genetic algorithm such as PS,PSO,BFO.

REFERENCES

- [1] T. Hagraş; J. Janeček (2003) "Static vs. Dynamic list-scheduling performance comparison" Acta Polytechnica vol. 43 no.6.
- [2] Parneet Kaur; Dheerendra Singh; Gurbinder Singh ; Navneet Singh (2011) "Analysis, comparison and performance evaluation of BNP scheduling algorithms in parallel processing International journal of information technology and knowledge management" volume 4, no. 1, pp. 279-284.
- [3] Ranjit Rajak (2012) "Comparison of bounded number of processors (BNP) class of scheduling algorithms based on matrices" no.2(34) issn 1512-1232 3 5.
- [4] Hui Jin; Xian-he sun; Ziming zheng; Zhiling lan ; Bing xie (2009) "International symposium on cluster computing and the grid performance under failures of Dag-based parallel computing".
- [5] Er. Navneet Singh; Er. Gagandeep Kaur; Er. Parneet Kaur ;Dr. Gurdev Singh(2012) "Analytical performance comparison of bnp scheduling Algorithms".
- [6] Samriti, Sandeep Gill , Ankur Bharadwaj , Navpreet singh, Harsimran Singh, Jashwinder Singh (2012) "Analysis of hlfet and mcp task scheduling algorithms", International journal of modern engineering research (ijmer) vol.2, issue.3, may-june 2012 pp-1176-1180 issn: 2249-6645.
- [7] Rinkle Aggarwal, Lakhwinder Kaur, Himanshu Aggarwal(2009) "Design and Reliability analysis of a new fault-tolerant multistage Interconnection network" , Icgst-cnir journal, volume 8, issue 2.
- [8] Albert p. C. Chan; Daniel w. M. Chan, m.asce; and John f. Y. Yeung (2009) "Overview of the application of "fuzzy techniques" In construction management research journal of construction engineering and management" asce / november / 1241
- [9] Mutasim nour, Shireen y.m (2006) "Adaptive fuzzy logic speed controller with torque adapted gains function for pmsm drive", Journal of engineering science and technology vol. 1, no. 1 59-75 © school of engineering, taylor's college .
- [10] Amrita Sarkar ;G.Sahoo; U.C.Sahoo(2012) "Application of fuzzy logic in transport Planning International journal on soft computing" (ijsc) vol.3, no.2, may 2012
- [11] José m. Merigó, anna m. Gil-lafuente(2009) "Some basic results of fuzzy research in the isi web of knowledge", ninth international conference on intelligent systems design and applications
- [12] T. Trigo de la vega, p. Lopez-garcía, s. Muñoz-hernandez "Towards fuzzy granularity control in Parallel/distributed computing".
- [13] Manik Sharma, Dr. Gurdev Singh, Harsimran Kaur (2012) "A study of BNP parallel Task scheduling algorithms Metric's for Distributed database system", International journal of distributed and parallel systems (ijdps) vol.3, no.1.
- [14] M.kaladevi ; Dr.s.sathiyabama (2010) "A comparative study of scheduling algorithms for real time task" ,International journal of advances in science and technology, vol. 1, no. 4.
- [15] Nidhi Arora (2012) " Comparative study of task duplication based scheduling algorithms for parallel systems department of computer science and engineering" ,International journal of computer applications (0975 – 8887) volume 58– no.19.
- [16] Ali m. Alakeel (2012) "A fuzzy dynamic load balancing algorithm For homogenous distributed systems".

- [17] Ravi Rastogi; Nitin Durg Singh Chauhan ; Mahesh Chandra Govil (2011) “Disjoint paths multi-stage interconnection networks stability problem”, Ijcsi international journal of computer science issues, vol. 8, issue 4, no 1.
- [18] Ravi rastogi ; Nitin ; Durg singh chauhan ; Mahesh Chandra Govil (2011) “ On stability problems of omega and 3-disjoint paths omega multi-stage interconnection networks”, Ijcsi international journal of computer science issues, vol. 8, issue 4, no 2
- [19] Rinkle Rani Aggarwal (2012) “Design and performance evaluation of a new irregular fault-Tolerant multistage interconnection network”, Ijcsi international journal of computer science issues, vol. 9, issue 2, no 3.
- [20] Ms. Rita Mahajan, Dr. Renu Vig , Ms. Preeti Abrol (2012) “ Ftelmin network: a new min with Fault- tolerance characteristics”, vol 5.
- [21] Erick Cantú-Paz (2008) “A survey of parallel genetic algorithms”, Vol 5.
- [22] Yudan Liu¹, Raja Nassar, Chokchai (box) leangsuksun , Nichamon naksinehaboon, Mihaela Paun, Stephen L. Scott (2008) “ An optimal checkpoint/restart model for a Large scale high performance computing System”
- [23] Michael Isard Mihai Budiu, Yuan Yu Microsoft Research, Silico (2004) , “Dryad: Distributed Data-Parallel Programs from Sequential Building Blocks” Vol 5.
- [24] Mamta Ghai ;Karamjit Kaur Cheema (2010) “ Design and reliability analysis of new fault-tolerant Irregular multistage interconnection network International journal of computer applications” (0975 – 8887) Volume 10– no.1.
- [25] Jorge R. Ramos; Vernon Rego (2004) “Efficient implementation of Multiprocessor scheduling Algorithms on a simulation Testbed” .
- [26] Ishfaq Ahmad, Yu-Kwong Kwok; Min-You Wu (1996) “Analysis, Evaluation, and Comparison of Algorithms for Scheduling Task Graphs on Parallel Processors”.
- [27] S. N. Sivanandam, S. Sumathi and S. N. Deepa (2007) “Introduction to Fuzzy Logic using MATLAB”, Springer-Verlag Berlin Heidelberg.
- [28] Li Xin Wang (1997) “A course in Fuzzy system & control”.
- [29] Yue Wu; Biaobiao Zhang; Jiabin Lu; K. -L. Du(2011) “Fuzzy Logic and Neuro-fuzzy Systems: A Systematic Introduction”, International Journal of Artificial Intelligence and Expert Systems (IJAE), Volume (2) : Issue (2).
- [30] Claudio Moraga(2005) “Introduction to Fuzzy Logic” FACTA UNIVERSITATIS (NI `S) SER.: ELEC. ENERG. vol. 18, No. 2, 319-328.
- [31] Z.Zalila;A.Cuquemella;C.Penet;A.Chika;B.Lorentz;D.Deschamps;C.Assemat (2007)“ Fuzzy Logic & Fuzzy Interference System: Introduction & properties” .
- [32] Srinivasa Rao D; Seetha M; Krishna Prasad MHM(2012) “Comparison of Fuzzy and Neuro Fuzzy Image Fusion Techniques and its Applications” International Journal of Computer Applications (0975 – 8887) Volume

BIOGRAPHIES:



She is Convener of CSE Dept. at Sant Baba Bhag Singh Institute of Engineering & Technology ,Padhiana. She has completed her B.Tech in 2003 and M.Tech in 2006 and pursuing Ph. D from Punjab Technical University, Jalandhar. She has published more than 15 papers in National and International Conferences and 7 papers in International Journals. She has guided more than 100 students in B.Tech Major Projects and 30 in M.Tech Major Projects.



Amita Sharma, A dynamic lecturer, who just not has a flair for teaching but a passion to learn. She has worked at K.C. Group of Institutes, Nawanshahr. She has Completed her B.tech (Computer Science & Engineering) in 2008 and M.tech (Computer Science & Engineering) in 2012. She has published more than 5 national and international papers. She is an active researcher and her areas of interest are Parallel Computing, Software testing and MANET (Moblile Aadhoc Network). She has guided more than 60 students in B. tech Major Projects

LASyM: A Learning Analytics System for MOOCs

Yassine Tabaa

Information and Communication Systems Laboratory
College of Sciences, Abdelmalek Essaadi University
Tetouan, Morocco

Abdellatif Medouri

Information and Communication Systems Laboratory
College of Sciences, Abdelmalek Essaadi University
Tetouan, Morocco

Abstract—Nowadays, the Web has revolutionized our vision as to how deliver courses in a radically transformed and enhanced way. Boosted by Cloud computing, the use of the Web in education has revealed new challenges and looks forward to new aspirations such as MOOCs (Massive Open Online Courses) as a technology-led revolution ushering in a new generation of learning environments. Expected to deliver effective education strategies, pedagogies and practices, which lead to student success, the massive open online courses, considered as the “linux of education”, are increasingly developed by elite US institutions such MIT, Harvard and Stanford by supplying open/distance learning for large online community without paying any fees, MOOCs have the potential to enable free university-level education on an enormous scale. Nevertheless, a concern often is raised about MOOCs is that a very small proportion of learners complete the course while thousands enrol for courses. In this paper, we present LASyM, a learning analytics system for massive open online courses. The system is a Hadoop based one whose main objective is to assure Learning Analytics for MOOCs’ communities as a mean to help them investigate massive raw data, generated by MOOC platforms around learning outcomes and assessments, and reveal any useful information to be used in designing learning-optimized MOOCs. To evaluate the effectiveness of the proposed system we developed a method to identify, with low latency, online learners more likely to drop out.

Keywords—Cloud Computing; MOOCs; Hadoop; Learning Analytics.

I. INTRODUCTION

Nowadays, Cloud Computing [1] has laid the ground for a new generation of educational systems, by providing scalable anytime/anywhere services simply accessed through the Web from multiple devices without worrying how/where those services are installed, maintained or located. The Web [2] ushered in a new era of possibilities and expectations for transforming education as it was stated by many studies and reports. With its promise of virtually “infinite resources”, Cloud Computing has consolidated the ubiquity of the Web in several learning aspects [3] and made feasible widened access to quality educational materials and courses. Thus, any educational institution can exploit and share its teaching expertise and learning resources through a global online presence. In 2012, new endeavors such as edX [4], Coursera[5] and Udacity [6] introduced more than 200 online costless college courses made accessible to any person connected to the Internet. These courses are called MOOCs, Massive Open Online Courses [7], and they exploit web technologies [2] to offer free online education to as many persons as possible. In May 2012, Harvard and MIT inaugurated the non-profit edX

and, since then, the University of Texas and the University of California Berkeley have joined them. The for-profit MOOC platform, Courseara, was initiated after the joint of 33 colleges and it exposes contributions from Princeton, Stanford, Penn, Duke, Ohio State, the University of Virginia and other colleges. Another for-profit MOOC platform, Udacity, was co-created by Stanford professor Sebastian Thrun, David Stavens, and Mike Sokolsky. Although actually most MOOCs do not offer credit, students can learn at their own pace and receive electronic certificate of accomplishment.

Leading US massive open online course providers have each almost increased the number of universities offering courses; for instance, Coursera delivers some 332 courses to its 3.1 million students since its launch in 2012, thereby generating big data as Web logs of activities and learning operations. However, course completion rates have gotten a lot of attention: as reported in Katy synthesis [8] which compares the ratio of students completing a course to total number of students registered on a variety of courses provided by several MOOC providers, for some courses the rate does not reach 2%. Accordingly, two interesting aspects may well be enhanced in future designed MOOCs; namely, decreasing the high MOOCs’ dropout rates, and optimizing learning operations through MOOC platforms.

In this paper, we propose LASyM, a Learning Analytics System for MOOCs, whose core aim is to mine MOOCs’ big data, essentially generated by user through learning operations on MOOC platforms, using a Cloud based Hadoop [9] to ultimately analyse students’ behaviour with the intent of increasing the impact of analytics on teaching and learning in such environments.

The rest of the paper is structured as follows. Section 2 presents briefly MOOCs and their types. Section 3 discusses benefits of learning analytics coupled with big data. Section 4 presents a background overview and discusses related work in the context of MOOCs, Learning Analytics and Hadoop based platforms. Section 5 introduces the proposed system and describes a small-scale environment. Finally, section 6 concludes the paper and describes the future research directions.

II. MOOCs : AN OVERVIEW

The progress of both information technologies and the education context run in a parallel course. In particular, educational exchange means knew an exponential growth around the end of the 20th century. By the 21st century, these means became more sophisticated and innovative [10]. Basically, Internet-based learning stood in lieu of any

educational transfer means used antecedently. Lately, mobile technologies joined the learning environment virtualizing classrooms and education sources. In this virtual numeric learning environment, the responsibility of the instructor has become an administrative one, and the educative material is simply advocated based on a general interesting context. Moreover, the number of students that an instructor can successfully manage, the main instructor's capacity indicator, is no more an issue of significance. The use of sophisticated information technologies for information transfer and student activities evaluation destroys the obstacles of human competences' limitedness, and makes the concept of unlimited class sizes achievable [10].

There is no commonly accepted definition of MOOCs, even during the period that this paper was being written the Wikipedia definition of MOOCs evolved. On September 2012, Wikipedia defined a MOOC as "a course where the participants are distributed and course materials are also dispersed across the web", adding that "this is possible only if the course is open, and works significantly better if the course is large. The course is not a gathering, but rather a way of connecting distributed instructors and learners across a common topic or field of discourse" [7]. By January 2013, the definition had become: "a type of online course aimed at large-scale participation and open access via the web. MOOCs are a recent development in the area of distance education, and a progression of the kind of open education ideals suggested by open educational resources" [7].

Massive Open Online Courses were perceived by Stephen Downes, and George Siemens, as an approach to address information excess, react to students' inquiries for pertinent knowledge, integrate IT progress, and decrease education's fee [11]. The intended objectives of this suggested online educational model was to gather unlimited number of learners, course materials, and information transfer means. The proposed model would not be subject to any limitations except for technological capabilities and their related costs. While MOOCs are considered a relatively new initiative, the concept was first discussed in 2008, but wasn't really taken up to any great extent until the last couple of years. The term MOOC (Massive Open Online Course) was coined by Dave Cormier [12] and created a buzz in 2012 which has already been described as the "Year of the MOOC".

There seems to be many definitions of MOOCs; however, two key features characterize this new educational technology: open accessibility and scalability. Thus, MOOC participants do not need to be registered in a school or a university nor paying fees in order to take part of a MOOC course. Indeed, there are two types of courses offered through the MOOCs platforms: cMOOCs and xMOOCs [13]. The first type [12], described as the good MOOCs by George Siemens, who, with Stephen Downes, early put forward these courses in Canada, is essentially based on a philosophy of connectedness and sustains the social dimension of learning and active practices; thereby, this type of course encourages knowledge production rather than knowledge consumption. While xMOOCs, the most adopted by higher education worldwide [4-6], consider the instructor-guided lesson as the centre of the course and offer to

large numbers of students the opportunity to study high quality courses with prestigious universities.

A MOOC system is consisted of five main elements [14]: Instructors, learners, topic, material, and context.

Instructors – Simplify the learning process via making available appropriate material, initiate communication between learners, and manage evaluations with regards to intended learning outcomes.

Learners – Anyone who wants to learn about the topic. Learners could be pursuing a formal degree or not. Learners who are simply interested with no precise objective are as well authorized to enrol.

Topic – The topic is discovered through the learner, instructor, material, and context. It is introduced all over the learning system and not just residing in a warehouse. It is adequately limited to allow emphasis but adequately wide to provide extensive coverage.

Material – Resides in diverse sites and is of multiple types and is accessed via various technological solutions.

Context – Represents the different actors forming a learning environment. This can incorporate online social networks, IT solutions, conventional information origins, diverse kinds of information transfer schemes, communication systems, intended learning outcomes, and the group constituting every course offering.

In MOOC platforms, information provided to learners is considered starting points from which they can jump off and pursue an information trajectory in accordance with their concerns. Accordingly, learners are able to communicate with one another through forums set up to help them discover common fields, find help and extra materials, and constitute particular groups so as to investigate shared topics more thoroughly. Indeed, the objective is to conceive a community of learners whereby everyone contributes by information and perspectives besides those provided by the instructor, and to get in an exploration ride. A course offered through a MOOC platform can be subject to a predefined time schedule or not, and can incorporate videos of different sources, links to websites and other online resources, some extra study materials, support forums, and all this can be accessed through multiple devices connected to the internet over wired, wireless, or cellular connections [11]. The learner chooses through which mean information is transferred may it be class forums, online social networks, or any other virtual domain. The strongest feature of a MOOC platform is elasticity [14].

III. BACKGROUND AND RELATED WORK

A. Learning Analytics

Learning Analytics was defined in the 1st international Conference on Learning Analytics and Knowledge (LAK 2011) as "The measurement, collection, analysis and reporting of data about learners... for purposes of understanding and optimising learning and the environments in which it occurs". The main goal of LA (Learning analytics) in distance learning is primarily improving learning efficiency and learning operations effectiveness, as well as providing educators,

learners, and decision makers with actionable insight to online course level activities. Specifically, learning analytics centres on the learning process through online platforms, including the analysis of the relationship between learners, contents, and eventually instructors.

B. Big data

Big Data is defined as huge amount of unstructured information and content that can be gleaned from “infinite” activity on the internet, generally non-traditional sources, such as web logs, click streams, social media, emails, sensors, images, and videos. The ability to analyze and exploit big data offers massive opportunities for real-time intelligence about responses to products, services and even political decisions.

Thus, several business activities can benefit from opportunities that big data can engender. Common use cases include, but are not limited to: sentiment analysis, marketing campaign analysis, fraud detection, and research and development. Nowadays, big data analytics is considered as a top IT priority for most organizations. Certainly, learning analytics and big data will have a significant role to play in the future of higher education.

C. Apache Hadoop

Hadoop [9] is an open source project sponsored by the Apache Software Foundation. Inspired by Google's MapReduce [15] paradigm, it's a Java-based programming framework that supports the processing of large data sets in a distributed computing environment. Cloud based Hadoop offers the possibility of running scalable applications on systems with thousands of nodes dealing with thousands of terabytes.

Hadoop framework is actually used by major players including Facebook, Twitter, Yahoo and IBM, largely for applications involving analytics, search engines and advertising. Hadoop MapReduce is a system for parallel processing of large data sets, and a number of related projects such as Apache HBase, Hive and Zookeeper.

D. related work

Andrew NG and Daphne Koller, two Stanford University computer professors, founded Coursera [5] MOOC in 2012. Since then, it has attracted an international student community of some 3.1 million students to its 332 courses. With the array of international partners, Coursera has already begun to offer courses in Spanish, Chinese, French and Italian. MITx [16] is the first prototype designed by MIT to support MOOC courses. “Circuits and Electronics”, also known as 6.002x, is the first course available on MITx, debuted in May 2012. EdX [4], successor of MITx, is a nonprofit organization put forward by Harvard and M.I.T., that allows a large community of academic institutions to take advantage of the MITx infrastructure and offers MOOC courses.

Most of the work in the Learning Analytics has focused on the LMS (Learning Management Systems), such as Morris et Al. [17] who concluded that frequency of participation and time spent on tasks are important for successful online learning through LMS. Macfadyen and Dawson [18] advocate for early-warning reporting tools that can identify and flag “at-risk”

students on LMS based platforms and allow instructors to develop early intervention strategies. Yanyan et Al. [19] proposed an improved mix framework for opinion leader identification in online learning communities, the study rank opinion leaders based on four distinguishing features: expertise, novelty, influence, and activity.

In the Cloud, Amazon ElasticMapReduce [20] offers Apache Hadoop as a hosted service on the Amazon AWS (Amazon Web Services) cloud environment that provides resizable compute capacity. Tabaa and Medouri [21] proposed a novel implementation of cloud computing platform SPC (Scientific Private Cloud) which offers a highly scalable data intensive distributed computing to perform complex tasks on massive amounts of data such as clustering, matrix computation, data mining, information extraction, etc.

IV. LEARNING ANALYTICS FOR MOOCs

Before introducing the proposed system, we would like, first, to analyze the lifecycle of the Big Data generated by MOOCs. Then, we classify student types or MOOC profiles that we shall identify while analyzing data using a simple method explained below based on “MOOC”ers behaviour.

A. Lifecycle of MOOCs' big data

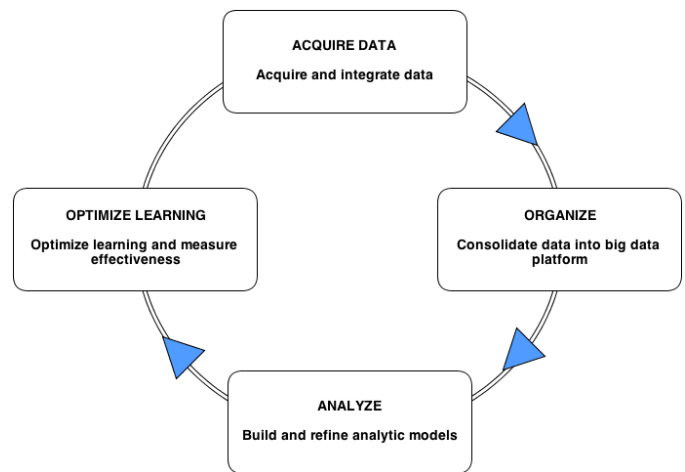


Fig. 1. Lifecycle of MOOCs' big data

As depicted in figure 1, the life cycle of big data generated by MOOCs can be described as follows:

a) *Acquire:* Generated data are captured periodically at source, typically as part of learning operations such as viewing materials, posting, surveys, user profiles, social media...etc.

b) *Organize:* Data is transferred from various sources and consolidated into a big data platform in order to prepare it for processing.

c) *Analyze:* Data stored in the big data platform is processed using various analysis modules, either in batches or a real-time processing.

d) *Optimize Learning:* The results of the “Analyze” phase are presented to MOOCs' stakeholders, enabling actions and automated interventions to be taken to provide early assistance to “at-risk” learners.

B. MOOC student patterns

Based on the Phil classification [22] of student types in a coursera-MOOC style, we redefine selected groups in the following modified classification list, that presents learners' profiles usually found in MOOC environments:

“Ghosts” – As long as a MOOC course is activated, this category of students registers to the course but at no time signs in. This category is usually the largest in terms of the number of enrolled students.

“Observers” – This category of students actually registers for the course, signs in, and might as well explore course materials. However, they do not carry out any kind of evaluations apart from basic quizzes found on lecture videos.

“Non-completers” – The majority of students fall into this category; they have recourse to MOOCs course materials to assist them study for and succeed in other courses. Essentially, those students attempt to use different course resources but do not accomplish the whole course.

“Passive Participants” – These students might consume each course material: watch lectures, complete quizzes, and interact with other learners and lecturers. Nevertheless, they do not participate in the course homework and projects.

“Active Participants” – Active participants are students who actually planned to take part of a MOOC course; they attend the lectures, accomplish the homework, interact with other participants, and complete all evaluations forms.

Following, we consider a learner as a more likely profile to dropout or “at-risk” student if he belongs to one of the following categories: “Ghosts”, “Observers”, “Non-completers” or “Passive Participants”.

C. A method to identify “at-risk” students in MOOC environments

As depicted in Fig. 2, in order to identify “at-risk” students, we suggest a simple method based on two principal characteristics: interaction and persistence. These indicators can be measured by essentially analyzing learners' behaviour and activities, such as the number of viewed videos, downloaded lectures, and replayed quizzes and surveys.

Persistence indicates user's concentration stability on a course in temporal terms. Although, it is a complex phenomenon that results in student completion of an online course, we can measure students' persistence through an important indicator, namely the number of viewed course materials. Thus we define the persistence of a student as:

$$P_c(s) = \frac{|VM_c(t)|}{(1 - \alpha)|NM_c(t)|}$$

$|VM_c(t)|$ denotes the number of viewed materials (namely slides, lecture sequences, tutorials...etc), of a course (c) in an instant (t) and $|NM_c(t)|$ is the total number of materials of a course (c) released in an instant (t). α is an adjustment factor with a value range of [0, 1]. On the other hand, interaction indicator is used to measure students' interaction on a specific course. Scoring of course-specific student is performed based

on two assumptions: (1) the more assessments and surveys a user has participated in, the more interest he /she has on this course; and (2) the more correct submitted assessments, the more comforting to continue learning operations.

$$I_c(s) = \frac{(1 - \beta) \cdot (AC_c(s) + SR_c(s))}{(AP_c(s) + SP_c(s))}$$

Where β is an adjustment factor with a value range of [0, 1].

Student's interaction in a course (c) is calculated based on the number of students' participation in assessments and surveys, which is designated respectively as $AP_c(s)$ and $SP_c(s)$. $AC_c(s)$ and $SR_c(s)$ denotes respectively the number of correct submitted assessments and replayed surveys.

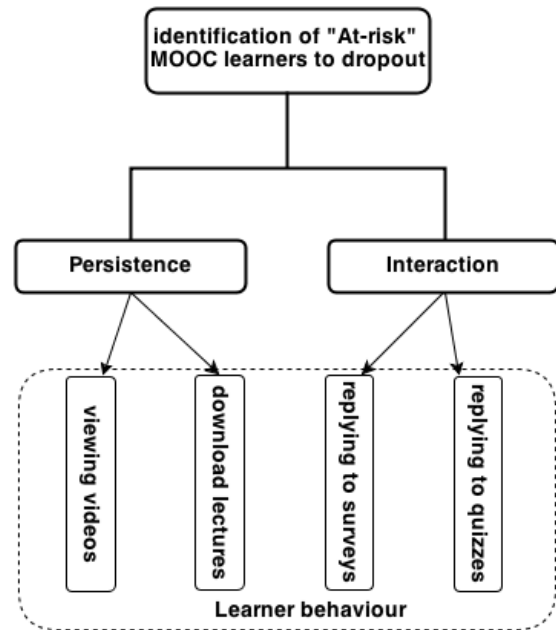


Fig. 2. Method to identify “at-risk” learners

We define $ED_c(s)$, the Engagement Degree of a learner in a MOOC course (c) as:

$$ED_c(s) = P_c(s) \times I_c(s)$$

This will result in a $ED_c(s)$ value range of [0, 1] that can be evaluated according to an adjustable threshold to identify if a learner is a potential “at-risk” profile.

V. LASyM ARCHITECTURE AND DESIGN

Fig. 3 depicts the general organization of LASyM, which envisions a learning analytics system, that enables MOOC stakeholders and providers to adjust content and provide support to their users by leveraging a private cloud based Hadoop [16], capable of processing the huge amount of captured learner-produced and learner-related data from MOOC platforms in order to minimize the time delay between the capture and use of data.

The core component of the proposed system is the analytics engine. This module of LASyM acts as a processing engine by

supplying building and deploying distributed learning analytics applications based on multiple frameworks such as MapReduce. Indeed, the main role of this component is first classifying and then processing the huge amount of hidden information in MOOC users' data.

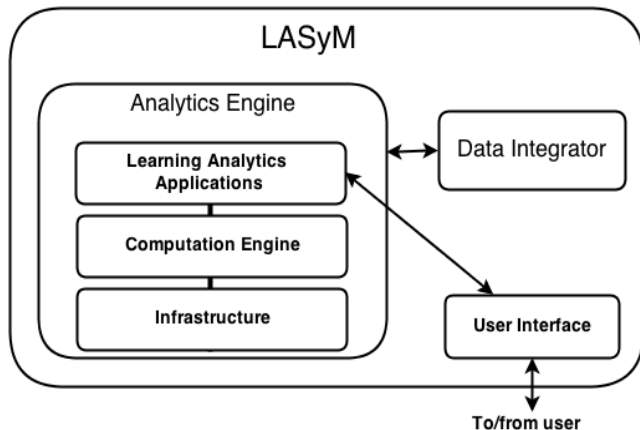


Fig. 3. LASyM architecture

The Data Integrator component is responsible of capturing data at sources before being transferred to the analytics platform. Data sources vary from students' engagement and behaviour to students' interests and preferences. Using various analysis modules, massive amounts of data will be provided and that can be mined to better optimize online learning experiences.

Finally, LASyM implements a user interface for accessing the corresponding learning analytics applications through a Web interface which also enables users to submit learning analytics jobs and explore results.

VI. LASyM : IMPLEMENTATION AND EVALUATION

In this section, we present the evaluation of our LASyM prototype. The section starts with the description of the experimental environment and infrastructure where the LASyMs' components were deployed. Then, in order to show the effectiveness of the proposed system, a small-scale scenario which implements a MapReduce-based application to identify "at-risk" learners is set up. Subsequently, evaluation of results will be presented.

A. Experimental setup

The experiments were conducted on a small scale implementation through operating a private cloud based Hadoop already deployed [21], composed of one master acting as the Resource Manager and 12 nodes, each node is a virtual machine with 2.4 GHz, 2 GB of RAM memory and 20 GB disk space allocated for HDFS (Hadoop Distributed File System), and embedding additional component, namely the data integrator and a MapReduce-based application to identify "at-risk" learners. Thus, we established a prototype of the LASyM system as shown in the Fig. 4. To execute the experiment, we used a sample of amplified dataset gathered from a typical MOOC deployed on jamiaati.org platform based on Stanford Class2go open source, itself deployed in the same private

cloud. In our experiment, the parameters α and β were set at 0.2 and 0.1 respectively.

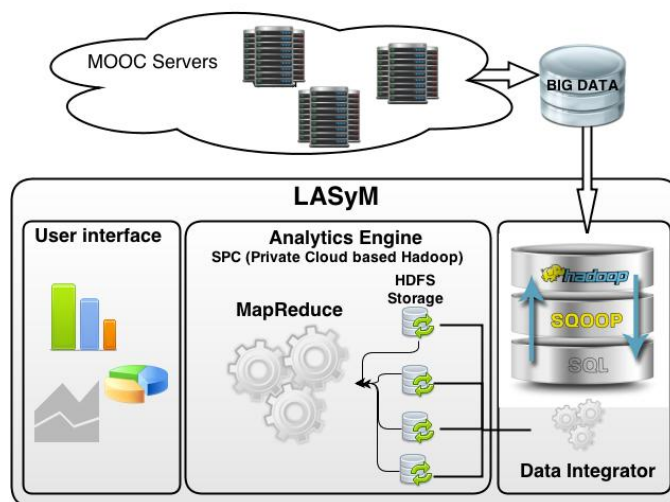


Fig. 4. LASyM architecture

The MapReduce-based application in LASyM which implements the previously developed method to identify "at-risk" learners is reduced to a single MapReduce job, meaning there is no overhead from executing multiple jobs in sequence.

B. Analytics engine

As can be seen above, the implemented LASyM analytics engine is a Hadoop-based component deployed on a private Cloud. It's a batch-oriented delivery system of analytics based on the SPC (Scientific Private Cloud) [16] previously deployed to provide researchers with a next generation of scientific platform based on the new generation of Hadoop, commonly named Hadoop 2.0. It's the main component of the LASyM, which will process Map/Reduce jobs to analyze all data acquired and organized by the data integrator component.

Indeed, there are several benefits of using such infrastructure which include the following:

- Rapid provisioning: Deploying analytics engine in the Private Cloud based Hadoop in few minutes.
- Multi-tenant frameworks: Using Cloud based Hadoop for other ends than MapReduce as a processing framework.
- High Availability: High availability with no single point of failure.
- Multi-purpose cloud infrastructure: Sharing infrastructure between Hadoop and non-Hadoop learning analytics applications.

C. Data Integrator

The Data Integrator captures data at sources, generally from relational databases, before being transferred to the analytics platform. In this experiment, the Data Integrator is responsible of extracting information from user profiles and orders stored within a relational database. Then, data is consolidated and

transferred into the analytics engine respecting the HDFS (Hadoop Distributed File System) file system. The Data Integrator implements the Apache sqoop [23] open source originally developed by cloudera [24] and currently an Apache top line project. It's a tool designed for efficiently transferring massive data between Hadoop and structured data stores such as relational databases.

Therefore, the data integrator component will collect and then import data from the MySQL database into the Hadoop Distributed File System (HDFS) [9], extracts results from processing Hadoop jobs and then exports data to MySQL tables to be able to easily explore analytics results. This component also provides the ability to schedule and automate import/export tasks.

D. Evaluation

After organizing and gathering data from the MySQL database using the Data Integrator component, an experiment was conducted in order to evaluate the performance of our LASyM implementation.

TABLE I. RUN TIMES FOR "AT-RISK" IDENTIFIER APPLICATION USING LASyM UNDER VARYING NODES NUMBER

LASyM nodes	Learners enrolled in course (c) ^a				
	10 K	100 K	1M	2M	4M
1 node	132	232	495	1578	4649
2 nodes	101	155	266	957	2760
4 nodes	88	113	234	668	1017
8 nodes	79	96	198	361	489
12 nodes	64	79	163	278	365

^a The experiment uses a dataset collected in the 5th week of a MOOC course

Knowing that the average duration of a MOOC course is 5 weeks, we executed the developed MapReduce-based application into LASyM in different number of parallel nodes, as shown on Table 1, to be able to calculate learning analytics speedup. Learning analytics speedup measures how many times processing learning analytics through LASyM is faster than running the same MapReduce jobs on a single node. If its value is greater than 1, it means there is at least some gain from doing the work using LASyM system. A speedup equal to the number of nodes is considered ideal and means that the system has a perfect scalability.

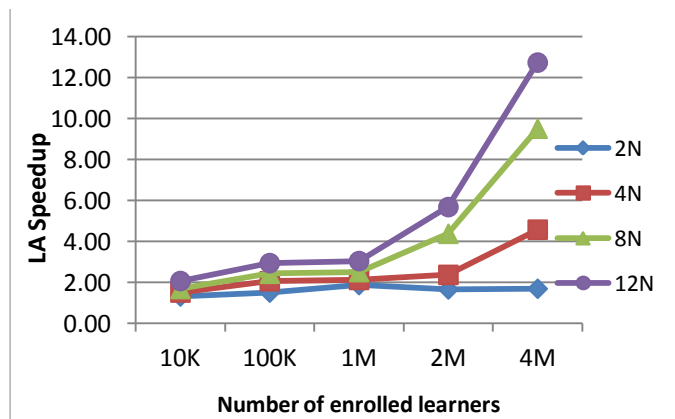


Fig. 5. Learning analytics speedup using LASyM

Fig. 5 depicts the calculated speedup. Thus, we can see that when the number of enrolled learners is small, there is only a small advantage in using LASyM: the speedup is slightly above 1 for 2 LASyM nodes and only reaches 2.06 in 12 nodes. However, when dealing with an important number of enrolled MOOC learners, the speedup grows significantly. With a number of 4 millions enrolled learners, the Learning analytics speedup using eight nodes into LASyM reaches 9.51, that can be interpreted as an ideal gain from using the proposed system to identify "at-risk" learners even if the number of enrolled learners is very high. With such system, composed of a larger number of nodes, the speedup reach the number of nodes, indicating that learning analytics process got the full benefit from using 12 nodes. The calculated results of learning analytics speedup for all cases suggest that this system has a good scalability to deal with such operations.

VII. CONCLUSION AND PERSPECTIVES

Most of the previous work related to learning analytics has focused on the learning analytics for LMS (learning Management Systems) or simply draw attention to the high dropout rate on MOOCs. However, few studies have addressed the issue of "at-risk" students' identification. As LA programs grow to include analysis of unstructured data, universities will need to develop skill and capacity to offer Hadoop based platforms and retrieval services. Indeed, the purpose of this work was, first, to design a learning analytics system capable to deal with the huge amount of unstructured data generated by a MOOC platform, as well as to develop a Hadoop MapReduce based application for automatic identification of "at-risk" students in MOOC environments.

To evaluate the performance of the proposed system, we conducted an experiment on an amplified typical MOOC dataset, where each learner is observed in terms of two features, namely interaction and persistence. We conclude that by using LASyM and our "at-risk" identification method implemented into this proposed system, we greatly reduced the latency time to analyze the huge amount of MOOCs' generated data, allowing us to identify "at-risk" learners at different stages of learning operations through the MOOC platform in a reasonable time.

In our case, we experimented LASyM for MOOC environments, which use MySQL as their DBMS, nevertheless, the system can be implemented for alternative SQL and/or NoSQL based MOOCs.

In future work we will take in consideration more indicators capable of identifying in a more precise manner "at-risk" profiles. Also, we will develop a component that enables MOOC providers to implement early intervention strategies in order to minimize the high rates of dropout in MOOC platforms.

REFERENCES

- [1] R. Buyya, C. S. Yeo, S. Venugopal, J. Broberg, and I. Brandic, "Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility," *Future Generation Computer Systems*, vol. 25, no. 6, pp. 599-616, 6//, 2009.
- [2] C. Allison, A. Miller, I. Oliver, R. Michaelson, and T. Tiropanis, "The Web in education," *Computer Networks*, vol. 56, no. 18, pp. 3811-3824, 12/17, 2012.

- [3] N. Sultan, "Cloud computing for education: A new dawn?," International Journal of Information Management, vol. 30, no. 2, pp. 109-116, Apr, 2010.
- [4] M. Harvard. "edX," <https://www.edx.org/>.
- [5] "Coursera," <https://www.coursera.org/>.
- [6] "Udacity," <https://www.udacity.com/>.
- [7] Wikipedia. "Massive open online course," 2012-10 and 2013-01; http://en.wikipedia.org/wiki/Massive_open_online_course.
- [8] K. Jordan. "MOOC completion rates," <http://www.katyjordan.com/MOOCproject.html>.
- [9] "Apache Hadoop web page," <http://hadoop.apache.org/>.
- [10] C. Long. "A new higher education online business model: Open and non-profit," <http://blogs.reuters.com/muniland/2012/09/15/a-new-higher-education-online-business-model-open-and-non-profit/>.
- [11] M. Jenny, J. M. Sui Fai, and W. Roy, "Networked Learning Conference 2010."
- [12] S. D. George Siemens, and Dave Cormier. "CMOOCs initiators."
- [13] J. Daniel, Making Sense of MOOCs: Musings in a Maze of Myth, Paradox and Possibility.
- [14] R. Kop, "The challenges to connectivist learning on open online networks: learning experiences during a massive open online course," The International Review of Research in Open and Distance Learning, vol. 12, no. 3, 2011.
- [15] J. Dean, and S. Ghemawat, "Mapreduce: Simplified data processing on large clusters," Communications of the Acm, vol. 51, no. 1, pp. 107-113, Jan, 2008.
- [16] "MITX Home," <http://www.mitx.org/>.
- [17] L. V. Morris, C. Finnegan, and S.-S. Wu, "Tracking student behavior, persistence, and achievement in online courses," The Internet and Higher Education, vol. 8, no. 3, pp. 221-231, 0/3rd/, 2005.
- [18] L. P. Macfadyen, and S. Dawson, "Mining LMS data to develop an "early warning system" for educators: A proof of concept," Computers & Education, vol. 54, no. 2, pp. 588-599, 2//, 2010.
- [19] Y. Li, S. Ma, Y. Zhang, R. Huang, and Kinshuk, "An improved mix framework for opinion leader identification in online learning communities," Knowledge-Based Systems, vol. 43, no. 0, pp. 43-51, 5//, 2013.
- [20] "Amazon Elastic MapReduce," <http://aws.amazon.com/fr/elasticmapreduce/>.
- [21] Y. Tabaa, and A. Medouri, "Towards a next generation of scientific computing in the Cloud," IJCSI International Journal of Computer Science Issues, vol. 9, no. 3, 2012.
- [22] P. Hill. "Emerging Student Patterns in MOOCs," <http://mfeldstein.com/emerging-student-patterns-in-moocs-a-revised-graphical-view/>.
- [23] "Apache Sqoop," <http://sqoop.apache.org/>.
- [24] "Cloudera platform," <http://www.cloudera.com/content/cloudera/en/home.html>.

Research on Chinese University Students' Media Images

--Based on Content Analysis of "China Youth Daily" and "Qilu Evening News"

Chengliang Zhang

School of Literature and Art
Ludong University
Yantai, China

Haifei Yu

School of Marxism
Ludong University
Yantai, China

Abstract—At present, university students, as the "after 90" and a new generation of young intellectuals, are being paid generally attentions by mass media. Nevertheless, university students' public images are on a decline as they have negative news appeared ceaselessly. Contemporary university students are becoming a group of people who are gazed at fixedly by the media. Moreover, the media keeps gazing at them and help them to build university students' media images. However, this kind of media behavior affects public judgments on university students' images. Furthermore, in the eye of the public, university students' images become serious distortion.

Keywords—University students' media image; Content analytic method; the public opinion; Synergistic effect

I. INTRODUCTION

Communication was seen as a magic bullet that transferred ideas or feelings or knowledge or motivations almost automatically from one mind to another...In the early days of communication study, the audience was considered relatively passive and defenseless, and communication could shoot something into them [1]. Although it is over exaggerated functions of communication by the "magic bullet theory", nevertheless, researchers have found out that audiences' judgments towards images of some certain groups of people were indeed affected by some media factors in the real activities of communication [2]. The public understandings and judgments towards some certain things in their minds are truly affected by traditional media and propagandistic advertisements, popular culture, and suggestions of friends as well as actual impressions [3]. In the contemporary social structure, the mess media productions (audiences of movies, books and magazines) are main resources to support the public to make judgments on images of certain groups of people. Furthermore, the related research methods are usually utilized from experimentalism to rational critique, and other methods include "utilization and satisfaction" method[4], recipient analysis method[5] and cultural study method[6], etc. Furthermore, the mass media play an intermediary role on transmitting information and strengthening information of social culture and social value judgment[7]. Hence, media's social cultural function supplies audiences with a theoretical framework which plays the corbelled function between social pressures and building images of groups [8].

In western society, university students' images which media pay attentions on is mainly about body image which is a psychological concept [9]. And this kind of attention is to discuss university students' features and temperamental formation. In the view of construction on media images, the mass media even provide a standard value of measurement for university students' feature and weight by their discussions on university students' images [10]. However, this kind of discussion on body images is a general discussion which cannot help the public to treat certain groups of people separately. On the contrary, in China, the word image has more abundant meanings. University students' images include lots of evaluation system such as their temperamental evaluation system, evaluation system on their reputation and morality, etc [11].

On the research issue of certain groups of people's media images, western researchers tend to particularly emphasis on cultural ethnic[12], religious union[13] and sexual discrimination[14] and they do not separate other certain groups of people according to people's occupation and the nature of their job additionally. There were no researches or reports that showed prejudice towards certain groups of people's media images.

Comparing with a more stable media evaluation system in western democratic society, China is in a period of social transformation currently. Its group evaluations are somehow uncertain [15]. In recent years, Chinese media have labeled "group construction" for group evaluation such as label "the peasant laborer", "the youth", "the women", "the female doctor", "the after 90", "the teacher" and many other groups and people from various of fields in the need of media communication and even in the need of prejudiced requirements of parts of audiences[16]. As a special social group, the university students have been paid particular attention by the media. Most of university students' images usually come from media reports, in some extent, it is the media that build university students' images[17]. Furthermore, a tendency of media images' construction also affects the public evaluative judgments on university students[18]. In this paper, the author analyses construction and evolution of university students' media images in China by using content analysis method. The author tends to open the images building process of certain groups of people by the media and unclose the effects of media evaluation on social evaluation.

II. RESEARCH QUESTIONS

In order to show the constructive and evolutionary condition of university students' media images in China, we make following research questions.

Q1: What topics do the mass media care about university students?

Q2: What kinds of report topics do the mass media use to build the framework of university students' media images?

Q3: Whether negative evaluations of university students by the public are connected with university students' media images built by the mass media?

III. METHODS

In this paper, the author uses data collecting method which includes media content analysis and deep survey and interview method. In the method of media content analysis, the author takes some newspapers that have considerable social influence as referenced examples, while the interview method is used to confirm participants' potential hypothesis [19] and its correlation with media content analysis. Finally, complete content and organizational editing before formatting. Please take note of the following items when proofreading spelling and grammar:

A. Media Content Analysis Method

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, sc, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

- **Research Subject.** In this paper, the author uses content analysis method to ensure university students' images in news reports. For the sake of ensure the unity of science on data connection, and then we set up a research scope of news media on print media. And we just do research on these national and local newspapers that are daily published, have wider transmitted scope and more audiences. We take "China Youth Daily" and "Qilu Evening News" as research samples. In these two newspapers, "China Youth Daily" is a major newspaper with important influence in contemporary China, and it's a comprehensive daily newspaper with distinct youthful characteristics. While the later is the largest circulation newspaper with the greatest social influence in Shandong province. It has above 1.35 million daily circulations. In the statistical results published by World Association of Newspapers in 2012, the circulation of "Qilu Evening News" is No.37 among world daily newspapers [20]. Since the media pay attention on university students are monthly changed according to job seeking period and holiday time, in order to ensure realism and effectiveness of statistics, the author general investigates university students-related reports of these two newspapers in the whole year of 2012 to acquire research data.
- **Unit of Analysis.** In this paper, the author sets unit of analysis as reports that are related to university students. And the concept of unit is that it is a standard

classification of content analysis in terms of content quantitative content. In this paper, the author takes pieces as the basic classified unit in the process of doing surveys and collecting data, and takes pieces of news on a main topic of "university students" into account.

- **Statistical Categories.** The establishment of these categories needs to conforming five principles. A: categories should inflect research purposes. B: these categories should be exhaustion. C: they should be mutually exclusive. D: they should be independent. E: they should be in single categories [21]. Then there are some statistical categories in this paper.

a) *Report quantity:* it means the number of statistical reports and the proportion of each report in the newspaper (unit: square centimeter).

b) *Report topics:* make some classifications for collected reports which are related to university students' images according to their main idea. These classifications are further study and postgraduate examination, campus entertainment and sports life, moral honesty condition, job seeking and working condition, consumption and financing condition, scientific research and social practical life, employees' life and starting a business, awareness of asserting rights and interests, physical and mental health, public-spirited activity, criminal offences, social assistance and so on.

c) *Standpoint:* the standpoint is expressed in three aspects. First, a somehow advantageous standpoint to university students, second, a somehow disadvantageous stand point to university students and third, a neutral standpoint or unable to judge.

d) *Presentation of images:* the Youth Development Department of China Youth University for political Sciences did a "survey on university students' public images" in Beijing from July to August, 2005. In this survey, it related to a category of "university students' images". Then we put this category as reference. In addition, we also consider the description of university students' images in the sample reports topics that are mentioned above as reference, too. Therefore, we make classifications for university students' images present in the media. There are several categories. A: what are university students' study attitudes? B: whether they are actively take part in campus entertainment and sports life or not? C: what are university students' consumption and financing condition? D: how are their moral honesty conditions? E: what performances do they have in job seeking as well as working? F: how about their scientific research and social practical life? G: are there any problems in their employees' life and starting a business? H: are there any problems in their physical and mental health? I: how are their awareness of asserting rights and interests?

A. Deep survey and Interview Method

We did questionnaire survey in agencies, enterprises and institutions in Jinan and some parts of Yantai. Then we gave questionnaires to service personals, administrative staffs, university students, middle school and high school teachers, retirees and other social groups of people. After getting there questionnaires back, we analyzed them by a statistical software

called SPSS to acquire effective data of university students' public images.

In this survey, we sent 214 questionnaires and 197 effective questionnaires were sent back. And 17 questionnaires are ineffective, and then effective percentage is 92.1%. Since the main topic of this survey is whether the framework of university students' images which are built by the media has influenced the public's judgment or not. Therefore, in the questionnaire, questions are mainly designed about contact conditions of interviewees with the university students and the media, and their judgments on moral condition of the university students, etc. (shown in table 1)

IV. ATA PROCESSING AND ANALYZING

A. Media Content Analysis

- Statistic on reports of Newspapers. Among the 357 sample reports that got in 1997, 202 reports came from "China Youth daily" which was 56.6% of the whole samples, while 155 reports came from "QILU Evening News" which was 43.4% of the whole. Moreover, among the 1198 sample reports that got in 2012, 515 reports came from "China Youth daily" which was 43.0% of the whole samples, while 683 reports came from "Qilu Evening News" which was 57.0% of the whole. (Shown in table 2)

- Statistic on Report Pages of Newspapers and Months When Reports Are Published. In 1997, there were 53 reports about "university students" on the front page of "China Youth Daily", was 26.2% of whole reports throughout the year. That number of "Qilu Evening News" was 21, with 13.6%. In 2012, there were 44 reports about "university students" on the front page of "China Youth Daily", was 8.5% of whole reports throughout the year. While that number of "Qilu Evening News" was 96, with 14.1%.
- In the view of statistic results of reports in the year 1997 and 2012, reports were not equally distributed in every month. There were more than 140 reports in July or August and there were 149 reports which in August were the much more one compared with that in July. And there were 53 reports in February. Then the number of reports in August was triple than that in February. Besides, the number of reports in November, December and January was more than that in other months since these three months were the end of terms or holidays. Hence, all these statistics shown that the media pay more attentions on university students' life out of campus, such as their social practical life, their employees' life and starting a business, postgraduate examinations, etc..

TABLE I. STATISTIC INFORMATION OF INTERVIEWEES

Name of company/institution/ organization/agency	The number of people	Department	city
<i>Yantai transportation group, Co., Ltd.</i>	25	Operating department Business department	Yantai
<i>Yantai Lianmin property group Co., Ltd</i>	28	Zhongcheng digital products mart	Yantai
<i>Yantai supply and marketing oil company</i>	23	Sales department	Yantai
<i>Yantai municipal office, Sat</i>	24	Agency	Yantai
<i>Ludong University</i>	31	Literature school	Yantai
<i>Shandong University</i>	41	Literature& journalism school	Jinan
<i>Jinan motive power Co., Ltd.</i>	13	Planning department	Jinan
<i>Shandong network radio-television station</i>	12	"Life help" program	Jinan

TABLE II. COMPARATIVE ANALYSIS ON THE NUMBER OF REPORTS ON NEWSPAPER

Category Percentage	China Youth Daily		Qilu Evening News		In total	
	1997	2012	1997	2012	1997	2012
The number of reports	202	515	155	683	357	1198
Percentage	56.6%	43.0%	43.4%	57.0%	100%	100%

- **Statistic on Tendency to Certain Groups of People.** In recent years, many people identified as certain social groups, for instance, “the peasant laborer”, “the after 80” and so on. In this paper, the author tries to find out if the media have identified University students as a “certain social group” by searching appearing frequencies of the word “university student” in the titles of news reports. It was pointed out by statistic results that there were 44 reports with the word “university student” in the titles of “China Youth Daily” in 1997, and it was 21.8% of the whole. And there were 35 reports with the word “university student” in the titles of “Qilu Evening News” in 1997, and it was 22.6% of the whole. Therefore, that means the media did not have special tendency on university students.

In 2012, there were 637 news reports with the word “university student” in the titles. Among these reports, “China Youth Daily” had 228 reports, was 44.3% of the whole, while “Qilu Evening News” had 409 reports, was 59.9% of the whole. Therefore, that means “university students” have been seen as a certain group in this society in the process of building framework of university students’ images by the media.

A sociologist Gordon W. Allport thought that tendentious attitude towards some group lied on accumulation of individual episode description. “People would abominate someone belonged to a certain group and even hostile to them” [22]. People had this attitude only because these ones were someone belonged to this certain group, and person who were belonged to this certain group must have all unpleasant characteristics of the group.

So it is shown that the university students, as a special social group, do exist by comparing “individual episode description” and “entirety background description”. An American famous journalist Iyengar [23] and a politician Ansolabehere[24] made a conclusion of writing news reports, one was “episodic” while the other was “thematic”. The former was based on individual cases and the later was based on an entire view of cases’ background and environment. The entirety background description more cared about a whole view angle. According to the statistics of “China Youth Daily” and “Qilu Evening News” in 2012, the ratio of news reports with individual episode description to news reports with entirety background description was 358: 276 (the other 564 reports did not have these two ways of description). Hence, it is obviously that individual cases can not represent the whole image of social groups, nevertheless, effects of individual cases’ accumulation will lead university students’ entire images to negation.

- **Statistic on Report Topics.** The author divides all the collecting news reports that talk about “university students” into several parts according to their main topics. These parts are students’ study and further study attitudes, postgraduate examination, campus entertainment and sports life, moral honesty condition, job seeking and work, consumption and financing condition, scientific research and social practical life, their employees’ life and starting a business, awareness of asserting rights and interests, physical and mental

health, participation on public-spirited activity, criminal offences, getting social assistance, etc. According to statistic results, the media mainly report university students’ campus entertainment and sports life, their job seeking and working condition, their employees’ life and starting a business and their social assistance. The percentage of these four parts is all above 10%, and they are 52.8% of the whole quantity of reports. (shown in table 3)

- **Statistic on Standpoint of Reports.** In the macroscopic level, advantageous reports on university students were in the leading role in 1997, while neutral reports or uncertain reports were in the leading role on university students in 2012, it was about 41.0%. Disadvantageous reports were a bit more than advantageous reports with 30.0% and 29.0%. (shown in table 4)

In the microcosmic level, statistic on using prejudiced words can make the media’s standpoint more clearly.

In addition, using prejudiced words is a common expressing way for ensuring the media’s reporting standpoints. An American journalist Meerill[25] put forward a famous method called “prejudiced types of words” to explore whether a journalist intended to use prejudiced (over advantageous or over disadvantageous) nouns, adjectives, verbs and adverbs to describe reported subjects[26]. In the statistic of prejudiced words used in 1997 and 2012, there were no such words used in reports in 1997.

However, such words appeared many times in the newspaper “China Youth Daily” and “Qilu Evening News”, especially in “Qilu Evening News”. They include “female university students”, “play truant”, “being cheated”, “cohabitation”, “being indulgent”, “unhealthy mental condition”, etc. The utilization of such words makes university students’ images more negative and also affects the public judgments to university students’ images.

- **University Students’ Media Images.** Based on a thought to help to build a framework of university students’ media images, the mass media design some topics with clear value tendency which constitute this framework. In this framework, there are many parts of contents which are study attitude, campus entertainment and sports life, moral honesty condition, job seeking and work, consumption and financing condition, scientific research and social practical life, employees’ life and starting a related values while ignore some reports with neutral judgments and uncertain opinions.

Then the numerical value=the percentage of positive images on each item-the percentage of negative images on each item. In 1997, university students’ images shown on “China Youth Daily” and “Qilu Evening News” were mainly positive images (figure omitted), except for negative images of moral honesty condition (only -0.28) and consumption and financing condition (only -0.56). However, descriptive figures of university students’ images made by the media had obviously changed in 2012. (Shown in fig.1)

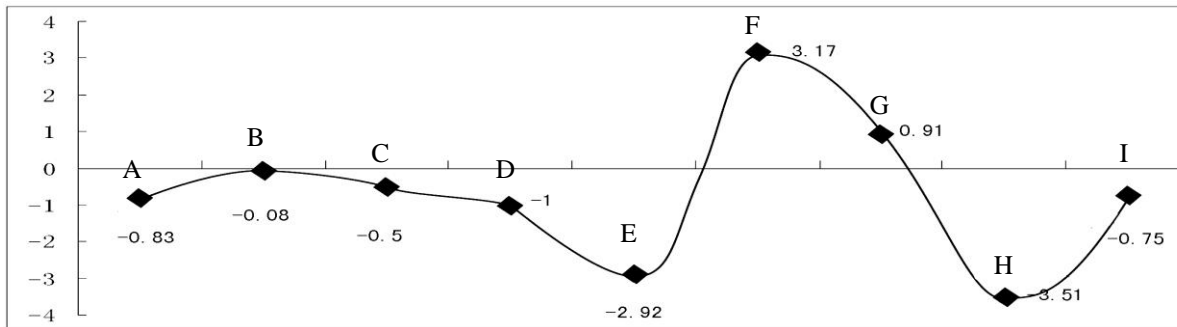


Fig. 1. University Students' Media Images

TABLE III. DISTRIBUTIVE CONDITION OF TOPICS ON "CHINA YOUTH DAILY" AND "QILU EVENING NEWS"

Types of report topics	The number of reports		Percentage	
	1997	2012	1997	2012
Study and further study attitudes, postgraduate examination	63	107	17.7%	8.9%
Campus entertainment and sports life	55	155	15.4%	12.9%
Moral honesty condition	20	85	5.6%	7.1%
Job seeking and work	66	228	18.5%	19.0%
Consumption and financing condition	14	73	3.9%	6.1%
Scientific research and social practical life	44	92	12.3%	7.7%
Employees' life and starting a business	8	128	2.2%	10.7%
Awareness of asserting rights and interests	8	66	2.2%	5.5%
Physical and mental health	4	19	1.1%	1.7%
Participation on public-spirited activity	33	78	9.2%	6.5%
Criminal offences	3	16	0.9%	1.3%
Getting social assistance	33	122	9.2%	10.2%
Others	6	29	1.8%	2.4%
In total	357	1198	100%	100%

TABLE IV. DISTRIBUTIVE CONDITION OF STANDPOINT OF REPORTS

Standpoint of reports	The number of reports		percentage	
	1997	2012	1997	2012
Advantageous	204	348	57.1%	29.0%
Neutral or uncertain	118	491	33.1%	41.0%
Disadvantageous	35	359	9.8%	30.0%
In total	357	1198	100%	100%

Business, awareness of asserting rights and interests and physical and mental health. In the need of directly expressing demands, we design a figure based on related values while ignore some reports with neutral judgments and uncertain opinions.

- 1) A. Study attitude; B. Campus entertainment and sports life; C. Moral honesty condition
- 2) D. Job seeking and working condition; E. Consumption and financing condition
- 3) F. Scientific research and social practical life; G. Employees' life and starting a business
- 4) H. Awareness of asserting rights and interests; I. Physical and mental health

In figure 1, it points out that most values of images are negative in the reports except for two values of positive images, the scientific research and social practical life and the employees' life and starting a business. Besides, the absolute d-value is generally small. According to data in figure one, we make a statistic of several items whose absolute values are more than or equal to one in proportion. And we take these items as university students' images. These items are actively do scientific research and take social practice (3.17), bad performance in job seeking and work (-1.00), bad condition on consumption and finance (-2.92) and weak awareness of asserting rights and interests (-3.51) in order.

B. Interview survey analysis

If we say interviewees who are not university students have special visual angles on university students, then it is more

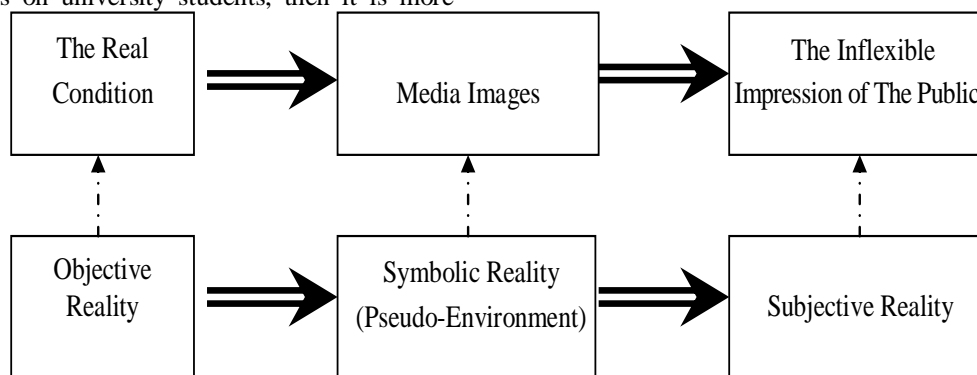


Fig. 2. A Formative Process of University Students' Media Image

Elisabeth Noelle-Neumann points out that the media have three influencing factors that influence human's environmental cognitive activities. Initially, most reports from most medias have highly similarity (Resonance effect). Besides, transmitting activities for similar information has continuity and repetition in time (Cumulative effect). Finally, the media information can be reached to a scope which is in the unprecedented universality (Ubiquitous effect) [28]. This paper shows that reports' contents of the surveyed media changed in these 15 years and this kind of change is similar. Moreover, this kind of change means that university students' media images are no longer positive as they were 15 years ago, while they are turned to neutral images and somehow are turned to be little negative

valuable to analyzing and thinking about attitudes of university students themselves. We found that university students who had been interviewed had close understanding on their images. In another word, they saw themselves not as independent audiences who kept long distance from images, on the contrary, they integrated themselves (or individual or groups) into the university students' images [27]. Specifically speaking, the media image is just an intermediary point when they thought about their existence. There are many particular concepts about appraisalment of media image by interviewees who are university students. A concept that is initially used as well as mostly used is deformation. It means they think that university students' images are deformed and other groups' appraisements come from the media or their own feelings.

According to opinions from Walter Lippmann, in this modern society with highly developed mass media, people's behaviors are closely related with three sorts of "realities". The first one is an "objective reality" in this real world. The second one is a "symbolic reality" (pseudo-environment) which selective hinted by the media. The last one is a "picture about

the outside world" which people describe in their minds, thus is the "subjective reality". Moreover, the real condition in the group of university students is the "objective reality" said Walter Lippmann while university students' images made by the media is the "symbolic reality", and the inflexible impression of the public towards university students is "subjective reality" (Dai Yuanguang, 2007). We would like to show thus in a figure. (shown in fig.2)

images. This kind of change has also been put into news consciousness of many media. Since lots of media especially those mainly work on social news selective build and transmit university students' negative images, then audiences certainly will have multiple resonances in psychology with these media. Furthermore, this continuous and repetitive transmission of university students' images will gradually pass audiences an illusion that degraded university students are everywhere[29].

C. Correlation between Media Images and Interview Survey Analysis

All kinds of social evaluations about university students may not match with their real living conditions. Maybe some of

these evaluations have large deviations. Then how this deviation formed? Some scholars think that social evaluation is an evaluated activity with basic characteristics of meaning understanding and value judgment. Its rationality is totally different with real and fake logical characteristics pursued by the fact-based judgment. Hence, there are diversified evaluated

systems exist in elements of complex social evaluation. Among these systems, the evaluated tendency made by construction of media's framework cannot be neglected.

In the acquired data, statistic results of communicating conditions with university students by interviewees in daily life are shown in table 5.

TABLE V. COMMUNICATING CONDITIONS WITH UNIVERSITY STUDENTS BY INTERVIEWEES IN DAILY LIFE AND INTERVIEWEES' EVALUATED CONDITIONS

Category		Frequently communication	Occasionally talk	Only meet	Never know	In total
Tendency of Evaluation	<i>Positive evaluation</i>	57	16	1	2	76
	<i>Neutral evaluation</i>	18	13	4	7	42
	<i>Negative evaluation</i>	6	55	7	11	79
In total (persons)		81	84	20	12	197
Percentage (%)		41.1	42.6	10.2	6.1	100

According to statistics in table 5, interviewees who frequently communicate with university students are usually relatives of them, such as parents and children or siblings. The percentage of these interviewees is 41.1. The percentage of those interviewees who not frequently communicate with university students is 58.9. It is also presented in table 5 that positive evaluation by interviewees is a bit lower than negative evaluation in 2012. But there is no obvious distinction between these two kinds of evaluations. However, if we do a further statistic on communicating degrees with university students, we find out that there is a big different in evaluation towards university students' entire images by those interviewees who have opposite communicating degrees with university students. The higher communicating degrees with them, the higher identification on their morality interviewees have. And those interviewees who frequently communicate with university students think that the percentage of university students' good honesty (very honest+ comparatively honest) is 88.9. While those interviewees who occasionally talk to university students think the percentage is 19. And 88.9% is higher than 19%. Furthermore, those interviewees who only meet university students think that the percentage of university students' good honesty is 5%. Obviously, 19% is higher than 5%.

Apparently, those interviewees who frequently communicate with university students are usually their relatives, and they make judgments through communication. They have more direct and visualize judgments, however, their evaluations may be mixed with "consanguinity" effects, etc. Therefore, their evaluation may not reliable. Meanwhile, in order to be suitable with the main topic that whether university students' public images are related to media images, the author removes those interviewees who frequently communicate with university students when do relevance analysis on samples of university students' public images. The author just sorts out statistics data of those interviewees who do not have frequently communication with university students.

Most of interviewees are in Shandong province. Among interviewees who often read newspapers, most of them read "QILU Evening News". Since the author uses content analysis method to analyze "China Youth Daily" and "QILU Evening News", therefore, statistics data of interviewees can be used as an effective supplement to media content analysis. For the convenience of research, the author makes a "two dimensional figure" for statistic of interviewees' communicating degree with the media and their evaluations to university students' images. (shown in table 6)

TABLE VI. STATISTIC ON INTERVIEWEES' EVALUATIONS OF UNIVERSITY STUDENTS' PUBLIC IMAGES AND INTERVIEWEES' COMMUNICATING DEGREE WITH THE MEDIA

Category		Evaluations on university students' public images (persons)		In total (person)
		<i>Prefer to negative evaluation</i>	<i>Prefer to positive evaluation</i>	
Communicating degree with the surveyed media	<i>Frequently communication</i>	68	17	85
	<i>Not frequently communication</i>	15	16	31
In total		83	33	116

In the process of statistic integration with relative analysis, people always use Chi-square test. Purpose of Chi-square test is

$$\chi_p^2 = \sum_{i=1}^k \frac{(A_i - T_i)^2}{T_i}$$

In this formula, A_i is a count. T_i is the theoretical value (expected value) when H_0 is true.

The Chi-square value is 11.15 though calculating. According to Chi-square theory, if Chi-square value is more than 6.63 in the two dimensional figure above, these two indexes have 99% possibilities of correlation[30]. And the real measuring value in this figure is more than 6.63. Therefore, it means there is a connection between communicating degrees with surveyed media by interviewees and interviewees' preference for negative evaluations. Thus is, more contacts with the media, more negative evaluations the interviewees have.

V. RESULTS

A. Topic Scopes That The Mass Media Pay Attention to on University Students Is Expanded.

It can be seen in the statistics that media reports' tendency had been quietly changed in the past 15 years. Their main topics have obviously and gradually inclined from further study and postgraduate examination, scientific research and social practical life to getting social assistance and employees' life and starting a business except for topics on job seeking and working condition and campus entertainment and sports life. Besides, the topic on university students' images is become increasingly diversified. These changes have formed a fragmented tendency towards topics correspond to the language environment in this Internet era.

B. The Mass Media Build Framework of University Students' Media Images by "Certain Groups of People"

1) The Appearance of A Concept "Image".

In the research of public relations, the concept "image" appears frequently in recent years. This terminology covers to a more complex descriptive system. And as a terminology with various meanings, the concept "image" means quality, reputation and morality in Chinese[31]. The formation of "image" is a two-way interactive process, one way is from inside to outside while another is from outside to inside. It is a practical product of the development of the concept "image" in public relationship that the appearance of university students' media images as a certain group. In addition, this concept is also an epitome of label "certain groups of people" by the media.

2) The Process of Constructing "Certain Groups of People" by The Media.

With diversified reports of the media, prejudiced reports about university students are gradually produced in the brewing process. The framework of university students as the "certain group", made by the mass media by means of emphasizing individual episode description has caused apparently negative transfer. By expressing macroscopic standpoints and utilizing

to find out difference between observed value and expected value. A formula of Chi-square test is,

microcosmic expressions, the mass media make university students' group characteristics which have been labeled become more and more apparently. All these steps have pushed to build the framework of university students' media images.

In the view of the media, perhaps each report is real, however, a macroscopic framework made up by repeated adding lots of real microcosmic realities on it may not be real. A Chinese young scholar Sun Wei puts forward a flatness theory for the mass media's contents. This theory was expounded in four aspects. In the theory, Sunwei thinks that in the view of relations between real and unreal, "reality of the media" has to be inevitably constructed based on a single reality which is supposed to be separated from entirety. Therefore, the mass media are good at manifesting a "spot" of reality and not good at manifesting a "flatness" of reality. That means what the media point out is a kind of flatness reality[32]. On the macroscopic level which is continuously constructed (gaze fixedly) by the mass media, the "symbolic reality" (university students' images in the media) is departed from the "objective reality" (real condition of university students). And this phenomenon has caused "macroscopic inconsistent with the facts" (university students' images have serious distorted).

C. The Public's Negative Evaluations to University Students Are Highly Related with The Mass Media Images.

Results of correlated calculation for the media images and interview survey show that there is a close correlation between negative evaluations by the public towards university students and report topics from the media. In the statistic results, although interviewees who are university students think "the media do distorted reconstruction for university students' images", nevertheless, this is not a truly reflection of university students. In fact, these interviewees also have negative evaluations on university students' entire images by themselves when they have answered questions in the questionnaires. Their evaluate results are certainly correlated with university students' media images which we have got. According to the social comparison theory [33] proposed by an American social psychologist Leon Festinger, individual does self-evaluation under the comparative dimension of others (organizations) when individual lacks of objective facts. University students' self-recognition and evaluation are influenced by the social comparison theory. Meanwhile, they tend to reference framework of images constructed by the media to make self-evaluation [34]. Finally, they make similar evaluations with university students' media images.

VI. DISCUSSION

In this paper, the author discusses three fields of construction and evolution of university students' media images in China. Firstly, it is the construction and change of university students' media images. Secondly, it is the formation and identification of the concept "certain groups of people". Thirdly, it is correlation between social evaluation and the media evaluation. Research shows that there are different understandings of media images in Chinese and western society[35]. Western media believe that individual images are

related to appearance and internal feelings, while Chinese media emphasis on individual reputation as well as morality.

A. *The Construction and Change of University Students' Media Images.*

Images are carries of culture. All the prejudiced information included in university students' images is not manifested in form of destruction and inhibition in modern media. This kind of information is expressed by constructing a new standard (lack of pursuits, be addicted to reality, not being elites in certain fields) to give pressures to those university students who are looking for diversified development [36]. Then this construction of the new standard is covert and it influences feelings of a society as well as individuals with its invisible existence. And this construction and changes are gradually formed by time passes. Furthermore, it is usually hard to thoroughly eliminate its potential effects in a short time [37].

B. *The Formation and Identification of the Concept "Certain Groups of People".*

The process of forming the concept of certain groups of people is a process formed from microcosmic view to macroscopic view, and is a process formed from fragmentation to concretion. In 1970s, a theoretical physical scientist, Prof. H. Haken, from University of Stuttgart, Germany created synergy theory [38]. Main principles of synergy theory are synergistic effects and order parameter. Synergy effects is that all subsystems' synergy behaviors cause some common influences that are much more effective than influences caused by each subsystem inside a complex large system. And these synergy behaviors cause unified and associated influence. The construction of media images plays an order parameter role in the process of forming the concept of certain groups of people. It controls individual opinions, forces people to have a generally similar public opinions to maintaining its existence[39]. The order parameter role played by the media images can be seen as a process, and this process is exactly the one that forms and identifies the concept of certain groups of people. In addition, the collected statistics on negative reports of university students' study attitude, campus entertainment and sports life, moral honesty condition, job seeking and working condition, consumption and financing condition, scientific research and social practical life, employees' life and starting a business, awareness of asserting rights and interests, physical and mental health, etc that have been mentioned above, are numbers of the order parameter. And the larger the number, the stronger orderly conclusion will be produced, then the deeper influence to the ordinary public, then the harder to overcome and change audiences in each subsystem.

C. *Correlation between Social Evaluation and the Media Evaluation.*

Production of values of the social public opinion is a process of social evaluation, and the media' s impetus is the great reference value. University students' images constructed by the media evaluation turn to be a stubborn and unchangeable image through interpreting and recreating of individual images and even human communication. It follows that media transmission have essential social guidance responsibility. Of course, the influence of the media should not

be limitless exaggerated. Some suggestions can be applied by negotiation or even by "opposite" effects (the second and the third hypothesis of Hall) to counteract negative factors of the media evaluation.

VII. CONCLUSIONS

If a sort of news framework has been repeatedly and exaggeratedly used, it will father upon audiences a kind of certain opinion. And audiences once accept this opinion, then they will utilize this news framework unconsciously when they think about images of this certain groups of people. Hence, they will come up with improperly judgments. Audience is like a mirror of the media. We expect that the media may check themselves by understanding audiences and change their images.

Therefore, under professionalized tendency of university students, the media must do self-criticism for their prejudiced construction of university students' images framework. They need to comprehensively and objectively construct the news framework of university students' images with rationally attitudes. They need to effectively guide the public opinions. Since the financial crisis in 2008 all over the world, university students' employments have been stroked as well. Then some media increase their reports that are adapted to current affairs on university students who start their own business. These reports increase the proportions of positive reports on university students and have some certain effects on correcting university students' media images. However, this kind of adjustments of reports is also a stress reaction, and it lacks of conditions and guarantees to normally operate. It is a question worth to be considerate that how to build a long-term testing mechanism for news framework to adjust news reports' tendency in time and to entirely and correctly pass objective realities to audiences in pseudo-environment by the media.

ACKNOWLEDGMENTS

This paper is sponsored by humanities and social sciences fund of ministry of education of China. Project's name is "Research on New Models of The Convergence of Communication Based on Entropy Theory" (12YJC860054). This paper is a staged research achievement.

REFERENCES

- [1] Ai Zhanga & Yi Luob & Hua Jiange, An inside-out exploration of contemporary Chinese public relations education, *Public Relations Review*, December 2011, Pages 513–521
- [2] Ang, 1985I. Ang, *Watching Dallas*, Methuen, London (1985)
- [3] Chen Xinyan, Survey on The Macroscopic Unreality of News in View of Distortion of University Students' Images by The Media [EB/OL]. <http://media.people.com.cn/GB/4975805.html>
- [4] Dai Yuanguang, *Research Theories And Methods of The Science of Communication*[M]. Fudan University Press, 2003.(11).P108
- [5] Dittmar, 2005H. Dittmar, Vulnerability factors and processes linking sociocultural pressures and body dissatisfaction, *Journal of Social and Clinical Psychology*, 24 (2005), pp. 1081–1087
- [6] Gubrium and Holstein, 2001J.F. Gubrium, J.A. Holstein, *Handbook of interview research: Context & method*, Sage, Thousand Oaks, CA (2001)
- [7] Guo Qingguang, *Comments on Journalism & Communication*[M]. China Renmin Press, 1999, P.221.
- [8] Gunn, 1988C. Gunn, *Vacationscape: Designing Tourist Regions*, (2nd ed.)Van Nostrand Reinhold, New York (1988)
- [9] [38] [39] H. Haken, Synergetics and computers, *Journal of Computational and Applied Mathematics*, June 1988, Pages 197–202

- [10] Han Henan, Being Gazed at And Expelling—The Investigation on Female Media Images by University Students. *Modern Communication*, 2004(3) P. 99-100
- [11] Herrman Haken, Synergetic - The Mystery of Natural Constitutions, Translated by Ling Fuhua, Shanghai: Shanghai Translation Press, 2005, P. 2
- [12] Hou Yingzhong, College Students Image in Mainstream Media: A Study of Reports in “China Youth Daily”, “Guangzhou Daily” and “Yangcheng Evening News”, *Journal of Guangdong University of Foreign Studies*, 2010(02) P.54
- [13] Huang Ruigang, The Synergetic Effect of the Government-Media Interaction in Crisis Control, *Journal of International Communication*, 2010, (05), P.36
- [14] Jerald Greenberga & Claire E. Ashton-Jamesb & Neal M.Ashkanasy, Social comparison processes in organizations, *Organizational Behavior and Human Decision Processes*, January 2007, Pages 22–41
- [15] Jie Zhang & Xiangguang Zhang, Chinese college students’ SCL-90 scores and their relations to the college performance, *Asian Journal of Psychiatry*, Available online 22 November 2012
- [16] Julien Mercille, 2005A. Julien Mercille, Media effects on image: The Case of Tibet, *Annals of Tourism Research*, 10(2005), pp. 1039–1055
- [17] Justin J. Lehmiller, Alvin T. Lawb, Teceta Thomas Tormalac, The effect of self-affirmation on sexual prejudice, *Journal of Experimental Social Psychology*, 3(2010), pp.276–285
- [18] Kellner, 1995D. Kellner, *Media Culture: Between the Modern and the Postmodern*, Routledge, New York (1995)
- [19] Leon Festinger & Mark R. Allyn & Charles W. White, The perception of color with achromatic stimulation, *Vision Research*, June 1971, Pages 591–612
- [20] Li Peng, Jiajia Zhang, Min Li, Peipei Li, Yu Zhang, Xin Zuo, Yi Miao, Ying Xu, Negative life events and mental health of Chinese medical students: The effect of resilience, personality and social support, *Psychiatry Research*, 3(2012), pp.138–141
- [21] McQuail, 1984D. McQuail, *Mass Communication Theory: An Introduction*, (3rd ed.)Sage, London (1984)
- [22] Phelps, 1986 A. Phelps, Holiday Destination Image: The Problem of Assessment—An Example Developed in Menorca, *Tourism Management*, 7 (1986), pp. 168–180
- [23] Qi Guijie & Wang Ruijin, *The Theory and Method and Technological Application on Information System*, Jinan, Shandong University Press, 2005, P.60
- [24] R.F. Rodgers,P. Salès, H. Chabrol, Psychological functioning, media pressure and body dissatisfaction among college women, *European Review of Applied Psychology*, April 2010, pp. 89–95
- [25] Rosengren et al, 1985E. Rosengren, L. Wenner, P. Palmgreen, *Media Gratifications Research: Current Perspectives*, Sage, Beverly Hills (1985)
- [26] Schramm, 1971W. Schramm, *The Nature of Communication Between Humans*, W. Schramm, D. Roberts (Eds.), *The Process and Effects of Mass Communication*, Harper and Row, New York (1971)
- [27] Shi Qingsheng, *The Communicative Principles Analysis*[M]. Anhui University Press, 2001.(12), P149
- [28] Siobhan S. O’Riordan,Byron L. Zamboanga, Aspects of the media and their relevance to bulimic attitudes and tendencies among female college students, *Eating Behaviors*, Volume 9, Issue 2, 4 (2008) , pp. 247–250
- [29] Spettigue and Henderson, 2004W. Spettigue, K.A. Henderson, Eating disorders and the role of the media, *The Canadian Child and Adolescent Psychiatry Review*, 13 (1) (2004), pp. 16–19
- [30] Tian Ping, On Media Image of “Female College Students”, *Journal of Shaoguan University*, 2008 (04) P.32
- [31] Vassilis Saroglou & Bahija Lamkaddem & Matthieu Van Pachterbeke & Coralie Buxant, Host society’s dislike of the Islamic veil: The role of subtle prejudice, values, and religion, *International Journal of Intercultural Relations*, 9(2009), Pages 419–428
- [32] Xu Zhansheng, On Brand Strategies of “QILU Evening News”[J]. *Youth Journalist*, 2012, (10), P.11
- [33] Yan Xiong, Comments on Research of The Media Attentions on Contemporary University Students’ Images, *News World*, 2012(11) P.18-21
- [34] Yuning Wu, College Students’ Evaluation of Police Performance: A Comparison of Chinese and Americans, *Journal of Criminal Justice*, July–August 2010, Pages 773–780
- [35] Wolton, 1993Wolton, D. 1993 A la recherche du public. Special edited issue of *Hermès* (11–12)
- [36] Zhou Chi & Lv Xiaohu & Wang Guangyin, Vertically And Horizontally Comments on Evolution of University Students’ Images of New China, *China Youth Study*[J]. 1991, (5), P.16-17
- [37] Zhou Ning, Cold War Thinking and Double Standard, *Journal of International Communication*, 2006, (9), P.18

Secure Medical Images Sharing over Cloud Computing environment

Fatma E.-Z. A. Elgamal

Information Technology dept.
Faculty of Computers and
Information Sciences, Mansoura
University
Mansoura, Egypt

Noha A. Hikal

Information Technology dept.
Faculty of Computers and
Information Sciences, Mansoura
University
Mansoura, Egypt

F.E.Z. Abou-Chadi

Electronics and Communications
Engineering dept.
Faculty of Engineering, Mansoura
University,
Mansoura, Egypt

Abstract—Nowadays, many applications have been appeared due to the rapid development in the term of telecommunication. One of these applications is the telemedicine where the patients' digital data can transfer between the doctors for farther diagnosis. Therefore, the protection of the exchanged medical data is essential especially when transferring these data in an insecure medium such as the cloud computing environment, where the security is considered a major issue. In this paper, two security approaches were presented to guarantee a secure sharing of medical images over the cloud computing environment by providing the mean of trust management between the authorized parties of these data and also allows the privacy sharing of the Electronic Patients' Records string data between those parties while preserving the shared medical image from the distortion. The first approach apply spatial watermarking technique while the second approach implements a hybrid spatial and transform techniques in order to achieve the needed goal. The experimental results show the efficiency of the proposed approaches and the robustness against various types of attacks.

Keywords—Cloud computing; Electronic Patients' Records; Cloud drops; encryption; spatial synchronization; authentication; Hybrid image watermarking; spatial watermarking; Discrete cosine Transform

I. INTRODUCTION

In recent years and as a result of the fast development in the technology and telecommunications, a lot of digital applications such as the telemedicine start to emerge. This application facilitates the transmission and sharing of the patient's medical data by the healthcare professionals for further diagnosis works [1].

Cloud computing, the environment that offers resources encapsulation on the Internet in the form of dynamic, scalable, and virtualized services [2], presents a variety of on demand services to the public such as the telemedicine services. Over this environment, the user can enjoy a lot of benefits offered by this computing paradigm like transmission, storage, and further processing needs on the user data. In spite of the cloud computing advantages, it has a number of disadvantages such as the data security which considered a major problem that face the users of this technology since they outsource their data to distributed storage systems and not a local ones [3]. Therefore, when transferring user's data over the cloud environment, especially the medical data, this kind of data which contains crucial information about the patients, a high level of protection

of the integrity and confidentiality [4] of these data have to be guaranteed to overcome any attacking attempts that may face these transmitted data.

One of the solutions to achieve the required trust management between the cloud computing parties is to use any watermarking technique which in turn classifies into two main domains, spatial domain and transformed domain. In the spatial domain which is the most straightforward embedding method, the watermarks are embedded directly in the cover image pixels values [5]. While in the transform domain, the transform coefficients of the cover image are used to embed the watermarks in [1]. Despite the simplicity and the shorter required execution time benefits of the spatial domain, the main drawback of the implemented schemes in this domain is that they divide the cover image into fixed-size blocks of pixels so the hidden data are inserted in the LSB's of each pixel in every block and this can decrease the visibility of the resulted watermarked image which is not acceptable especially when dealing with medical images [6]. On the other hand, the transform domain methods can guarantee more robustness against attacks but needs more processing powers and computation times [7].

In recent years and in order to overcome this problem, medical image exchanging over cloud environments has gained a great interest. The medical images present in the cloud can provide the necessary details to the doctors and the patient can seek the treatment in different branch hospital, reduce the information and computational resource maintenance in the hospital. Furthermore, existing medical equipments can be rebuilt to be more efficient and low-cost as medical terminal units. Different proposals were introduced in [8, 9] to deal the exchanging, storing and sharing on medical images in the way that verifying data integrity, availability, and confidently.

This paper introduces two approaches aims to provide the mean of trust management between data parties over the cloud computing environment. The two methods achieve the required goal through providing three levels of authentication, from data owner to the destinations, from the data owner to the cloud service provider and finally from the destination to data owner. For the first approach, the idea of the spatial watermarking techniques has been exploited. While in the second approach, a hybrid model based on the idea of the spatial and the transform techniques were implemented.

In addition to offering the trust management, the proposed approaches allow secure sharing of the Electronic Patients' Record (EPR), which is string data helps to speed up the clinical communication, reduce the diagnostic errors by providing more accurate and timely clinical information and also the EPR assist doctors in diagnosis and treatment [10]. So and since it is considered a sensitive data, the proposed approaches guarantee the protection of them while they are transferred.

The reminder of this paper is organized as follows; Section 2 describes the spatial embedding based approach. Section 3 combine the first approach with discrete cosine Transform to provide hybrid spatial and transform embedding approach. Section 4 presents the experimental results while section 5 shows the paper conclusion.

II. SPATIAL SYNCHRONIZATION AND DYNAMIC EMBEDDING APPROACH

The three main stages of this approach are shown in Fig. 1. The first stage dynamically embeds the EPR data into the original medical image. Then, the cloud model is applied to the medical image to extract the approximated version. Finally, the encryption process is done using a symmetric negotiated private key between the authorized parties of the data.

A. Spatial domain dynamic embedding/extraction algorithm

The purpose here is to hide the EPR data into the original shared medical image in an effective way that does not affect the visual quality of the medical image using Dynamic Embedding algorithm [6]. The main task is to exploit the overall capacity of the cover image in order to guarantee a high visibility which is a necessity especially when dealing with medical images. Moreover, this method provides a flexibility of cover images' size rather than restricting its size to be more than or equal the fourfold size of the embedded data as in static embedding techniques.

In other words, this step gets the benefits of the shorter execution time associated with spatial watermarking

algorithms. But in the same time and due to the usage of the dynamic embedding algorithm, it can overcome the drawback of inserting the hiding data in the LSB of the cover image block pixels that decrease the visibility of the resulting image.

In addition to the dynamic embedding algorithm, symmetric secret key (K_1) was applied to perform a spatial synchronization embedding/extraction processes through using this key as a seed in a pseudo random number generator (PRNG) in order to generate random arrangement of the used pixels for the embedding/extraction processes within the medical image. To accomplish this, the Mersenne Twister algorithm [11] was applied which is a pseudo random number generator (PRNG) that in turn uses some kind of mathematical formulas or pre-calculated tables to generate a sequence of numbers that appear random but it is not truly random. It is completely determined by an arbitrary initial state called seed state that can be represented by K_1 in this work. The reason for using Mersenne Twister algorithm is because it has a huge period length of $2^{19937} - 1$, very fast, has good equidistributional properties and passing most statistical tests [12].

Spatial synchronization dynamic embedding phase: The cover medical image (CMI) and the EPR string data (D_i) are the inputs of this step where their sizes, $|CMI|$ and l respectively, are used to dynamically determine the size of each block of which the cover image is divided in and for the LBS used for the embedding process. In addition, to the added security, changing the known static embedding ways and using secret key (K_1) for rearranging the pixels used in the embedding process. Dynamic embedding process also improves the visibility by regulating the embedding steps according to the used inputs. This is required especially for the medical images where high quality is a major aspect that has to be guaranteed. Fig. 2 illustrates the steps of this phase and how the dynamic idea is applied for the required embedding process.

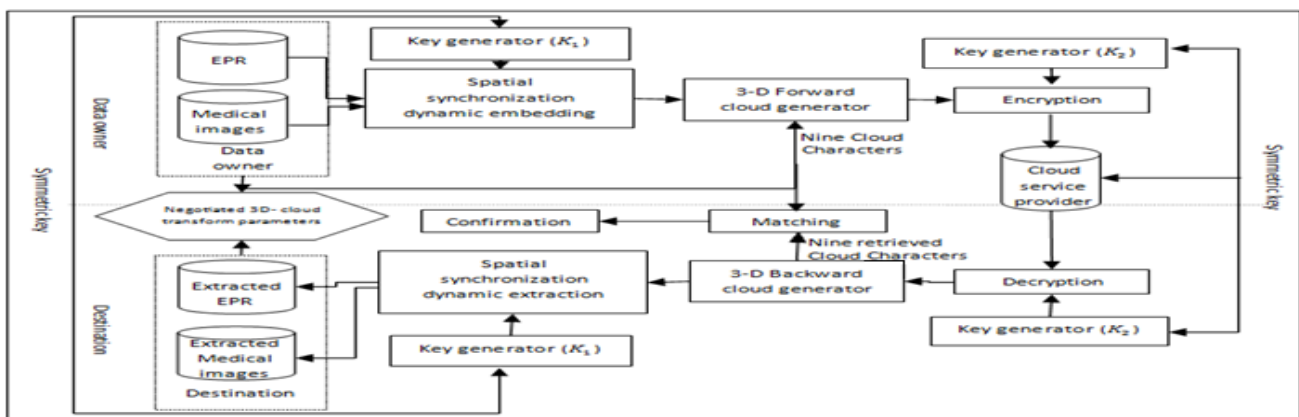


Fig. 1. The proposed scheme framework

1) *Spatial synchronization dynamic extraction phase:* Fig. 3 show the required steps for the extraction phase. It illustrates how the EPR data can be extracted in a dynamic manner using the same key used in the embedding phase.

B. The 3D cloud generation

The aim of this step is to generate approximated shared medical image after embedding the EPR data to form both the required image to be shared and at the same time represents one level of authentication that is used by the destination of the data in order to confirm the identity of data owner. To accomplish this, three dimension cloud model is applied, with six cloud characters $E_x, E_y, E_z, E_{nx}, E_{ny}, E_{nz}, H_{ex}, H_{ey}, H_{ez}$ used for the required confirmation step. This is an expansion form of one dimension cloud model [13] where the expected value (E_x) is the point that is most representative of the qualitative concept, the entropy (E_n) is The uncertainty measurement of the qualitative concept which is determined by both the randomness and the fuzziness of the concept to represent the measurement of randomness and the value region in which the drop is acceptable by the concept, and the hyper-entropy (H_e) is the second-order entropy of the entropy.

These values are the general concepts that are applicable in one-dimensional and can be extended to higher dimensional situations. According to these cloud characteristics, the next step is to perform a "forward cloud generator" that aims to generate cloud drops to express the concept quantitatively. Then, to extract the cloud characteristics, the "Backward cloud generators" is applied to the previously generated cloud drops.

Therefore, by expanding the one-dimensional cloud model into a three-dimensional model, where E_x, E_y and E_z are refers here to the components of the RGB colour format of the original image. Algorithm 3 and algorithm 4 represent the forward and backward cloud generators in the three-dimensional model as shown in Fig. 4 and Fig. 5.

Algorithm 1: Spatial synchronization dynamic embedding algorithm
 Input: the cover medical image (CMI) and D_i .
 Output: The watermarked medical image WMI
 Step 1) Divide CMI into blocks (B) with sizes (BS) changes according to the size of the CMI and l . So, BS will be:

$$BS = \left\lfloor \frac{|CMI|}{l} \right\rfloor$$

Where: $|CMI|$ is the size of the CMI.
 Step 2) Determine the number of LSB where the hidden data will be replaced in each block pixel (B_i), $1 < i <= BS$ through:

$$Nb = \frac{|D_i|}{BS}$$

Step 3) Since Nb may not be integer, the number of used bits in each pixel B_i of B is obtained as:

$$Ub_i = \begin{cases} \lfloor Nb \rfloor, & \text{if } i = 1, \dots, BS * \lfloor Nb \rfloor - |D_i| \\ \lfloor Nb \rfloor, & \text{otherwise} \end{cases}$$

Step 4) Use a pseudorandom generator with K_1 to embed the D_i bits into the corresponding rearranged pixels bits inside B_i 's according to Ub_i until finally construct the WMI.

Fig. 2. Spatial synchronization dynamic embedding algorithm [6]

Algorithm 2: Spatial synchronization dynamic extraction algorithm
 Input: The watermarked medical image WMI, l .
 Output: EPR data
 Step 1) Calculate BS, Nb and Ub_i values respectively through Fig. 2.
 Step 2) Apply Ub_i in each block pixel determined by K_1 , which generates spatial schedule of the right sequence of the embedded pixels, to retrieve the embedded EPRs' bits.
 Step 3) Use the retrieved bits to finally reconstruct the required EPR data.

Fig. 3. Spatial synchronization dynamic extraction algorithm [6]

Algorithm 3: Three dimensional Forward Cloud Generator
 Input: ($E_x, E_y, E_z, E_{nx}, E_{ny}, E_{nz}, H_{ex}, H_{ey}, H_{ez}$), WMI
 Output: the Approximated Shared Image (ASI)
 Step 1) Generates three-dimensional normally distributed random vector ($E_{nx}'_i, E_{ny}'_i, E_{nz}'_i$) where:

$$E_{nx}'_i = NORM(E_{nx}, H_{ex}^2)$$

$$E_{ny}'_i = NORM(E_{ny}, H_{ey}^2)$$

$$E_{nz}'_i = NORM(E_{nz}, H_{ez}^2)$$

Step 2) Generates three-dimensional normally distributed random vector (x_i, y_i, z_i) where x_i, y_i and z_i are cloud drops in each of the images' dimensions.:

$$x_i = NORM(E_x, E_{nx}'_i^2)$$

$$y_i = NORM(E_y, E_{ny}'_i^2)$$

$$z_i = NORM(E_z, E_{nz}'_i^2)$$

Step 3) Repeat Steps 1 to 3, in the entire WMI pixels to generate the required approximated shared image (ASI).

Fig. 4. Three dimensional Forward Cloud Generators

Algorithm 4: Three dimensional Backward Cloud Generator
 Input: Approximated Shared Image (ASI)
 Output: ($E_x', E_y', E_z', E_{nx}', E_{ny}', E_{nz}', H_{ex}', H_{ey}', H_{ez}'$).
 Step 1) Calculate E_x', E_y' and E_z' :

$$E_x' = \bar{X} = \frac{1}{n} \sum_{i=1}^n x_i,$$

$$E_y' = \bar{Y} = \frac{1}{n} \sum_{i=1}^n y_i,$$

$$E_z' = \bar{Z} = \frac{1}{n} \sum_{i=1}^n z_i$$

Step 2) Calculate E_{nx}', E_{ny}' and E_{nz}' :

$$E_{nx}' = \sqrt{\frac{\pi}{2}} \left(\frac{1}{n} \sum_{i=1}^n |x_i - E_x'| \right),$$

$$E_{ny}' = \sqrt{\frac{\pi}{2}} \left(\frac{1}{n} \sum_{i=1}^n |y_i - E_y'| \right),$$

$$E_{nz}' = \sqrt{\frac{\pi}{2}} \left(\frac{1}{n} \sum_{i=1}^n |z_i - E_z'| \right)$$

Step 3) Calculate H_{ex}', H_{ey}' and H_{ez}' using the variances S_x^2, S_y^2 and S_z^2 :

$$S_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{X})^2, \text{ then } H_{ex}' = \sqrt{S_x^2 - E_{nx}'^2}$$

$$S_y^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{Y})^2, \text{ then } H_{ey}' = \sqrt{S_y^2 - E_{ny}'^2}$$

$$S_z^2 = \frac{1}{n-1} \sum_{i=1}^n (z_i - \bar{Z})^2, \text{ then } H_{ez}' = \sqrt{S_z^2 - E_{nz}'^2}$$

Fig. 5. Three dimensional Backward cloud Generator

C. Encryption/Decryption Technique

In this step, cryptographic algorithm [14] with pseudorandom number generator was applied for the encryption/decryption. The idea is that, the owner of the data uses a private key K_2 to generate a spatial schedule, which is used to encrypt the required approximated shared image. The goal of the detector is to use K_2 for the decryption process. For simplicity, symmetric technique is assumed, where the encrypting and corresponding detection key is identical [15].

K_2 is used as a seed to a pseudo random number generator (PRNG) using the Mersenne Twister algorithm, for the reasons illustrated in subsection II.A, to provide random arrangement of the pixels for the encryption/decryption processes on the shared image.

The resulting schedule rearranges the image pixels randomly, in spatial domain, to encrypt the approximated watermarked image. The used key is substantial to desynchronize the encrypted image at the destination. In other words, it helps the owner to be ensured about the identity of the data recipient. In other words the usage key provides a mean of authentication between the data owner and the service provider and also between the data owner and the destination of the shared data since they are the legal recipients of the data.

III. SPATIAL SYNCHRONIZATION, DYNAMIC AND TRANSFORM EMBEDDING APPROACH

In this approach, the same stages of Fig. 1 are performed but rather than using the dynamic embedding algorithm only to perform the embedding process, the second approach uses both dynamic embedding along with the Discrete Cosine Transform DCT algorithm and Inverse Discrete Cosine Transform IDCT that are widely used transform algorithms [16] to perform the embedding process.

The purpose of the extra step here is to get the benefit of the DCT algorithm that provides more robustness against attacks than the spatial embedding algorithms so it can help to preserve the hidden data much better than the first approach. While the first approach guarantee more fast computations than this approach.

The inserted steps were in the embedding/extraction stages. Therefore the new algorithms are as shown in Fig. 6 and Fig. 7.

IV. EXPERIMENTAL RESULTS

The results of the proposed schemes have been carried out inside MATLAB environment with a set of $350 \times 350 \times 3$ MR images obtained from standard web portal for MRI images [17]. Moreover, MRI images from standard web portal [18] were used for further investigation of the proposed scheme efficiency. Then, in order to test the quality of the both schemes, numbers of quality metrics were applied. These metrics include Mean square Error (MSE) and peak signal to noise ratio that were calculated using (1) and (2) respectively. Structural similarity (SSIM) index address was also applied to measure the local images similarities and it was measured through (3). The number of changing pixel rate (NPCP) and the unified averaged changed intensity (UACI) metrics to test the number of changed pixels and the number of averaged changed intensity respectively between encrypted/decrypted images [19]

Algorithm 5: Spatial synchronization dynamic and transform embedding algorithm
 Input: the cover medical image (CMI) and D_i .
 Output: The watermarked medical image WMI
 Step 5) Divide CMI into blocks (B) with sizes (BS) changes according to the size of the CMI and l . So, BS will be:

$$BS = \left\lfloor \frac{|CMI|}{l} \right\rfloor$$

Where: $|CMI|$ is the size of the CMI.
 Step 6) Determine the number of LSB where the hidden data will be replaced in each block pixel (B_i), $1 \leq i \leq BS$ through:

$$Nb = \frac{|D_i|}{BS}$$

Step 7) Since Nb may not be integer, the number of used bits in each pixel B_i of B is obtained as:

$$Ub_i = \begin{cases} \lfloor Nb \rfloor, & \text{if } i = 1, \dots, BS * \lfloor Nb \rfloor - |D_i| \\ \lfloor Nb \rfloor, & \text{otherwise} \end{cases}$$

Step 8) Use a pseudorandom generator with K_1 to rearranged pixels bits inside B_i 's.
 Step 9) Transform the rearranged pixels using DCT.
 Step 10) Embed the D_i bits into the rearranged transformed pixels according to Ub_i .
 Step 11) Retransform the resulted pixels after the embedding process using IDCT.

Fig. 6. Spatial synchronization dynamic and transform embedding algorithm

Algorithm 6: Spatial synchronization dynamic extraction algorithm
 Input: The watermarked medical image WMI, l .
 Output: EPR data
 Steps:

Step 4) Calculate BS, Nb and Ub_i values respectively through Fig. 2.
 Step 5) Apply Ub_i in each block pixel determined by K_1 , which generates spatial schedule of the right sequence of the embedded pixels.
 Step 6) Transform the rearranged pixels using DCT.
 Step 7) Retrieve the embedded EPRs' bits from the transformed pixels values.
 Step 8) Use the retrieved bits to finally reconstruct the required EPR data.

Fig. 7. Spatial synchronization dynamic and transform extraction algorithm

were also calculated using (4), (5) and (6) respectively. Finally, to measure the rate of the bits error (BER), (7) was used in order to check the performance of the proposed schemes in the presence of the attacking attempts.

$$MSE = \frac{1}{MP} \sum_{i=0}^{M-1} \sum_{j=0}^{P-1} [OMI(i,j) - RMI(i,j)]^2 \quad (1)$$

$$PSNR = 10 \log_{10} \left(\frac{R^2}{MSE} \right) \quad (2)$$

Where R is the maximum fluctuation in the input image data type, M, P are the sizes of the original medical image (OMI) and the retrieved medical images (RMI) respectively [20]

$$SSIM(OMI, RMI) = LC(OMI, RMI)^\alpha \times CC(OMI, RMI)^\beta \times SC(OMI, RMI)^\lambda \quad (3)$$

Where: OMI, RMI are the original and the reconstructed medical images respectively? *LC* is the luminance, *CC* is the contrast and *SC* is the structure of OMI and RMI. α, β and λ are ≥ 1 and are used to weight the importance of each of the three components. [20]

$$D(i, j) = \begin{cases} 0, & \text{if } OMI(i, j) = RMI(i, j) \\ 1, & \text{if } OMI(i, j) \neq RMI(i, j) \end{cases} \quad (4)$$

$$NPCR: N(OMI, RMI) = \sum_{i,j} \frac{D(i,j)}{T} \times 100\% \quad (5)$$

$$UACI: U(OMI, RMI) = \sum_{i,j} \frac{|OMI(i,j) - RMI(i,j)|}{F.T} \times 100\% \quad (6)$$

Where *F* denotes the largest supported pixel value of the image format and *T* represents the size of the OMI and RMI [19].

$$BER = \frac{100}{l} \sum_{i=1}^l \begin{cases} 1, & D'_i = D_i \\ 0, & D'_i \neq D_i \end{cases} \quad (7)$$

Where D_i and D'_i are the i^{th} bit of the embedded and the recovered EPR data respectively and *l* is the length of the EPR data [21].

Before going through these measurements, the processing time of the both approaches were illustrated in Tables 1 and 2. These results were computed on a personal computer worked with Intel (R) Core (TM) i3 CPU, 2.53 GHz and installed memory (RAM) of 2.00GB (1.86 GB usable).

The results in the tables shows that the first approach that apply only the dynamic embedding algorithm has less processing time since it performs the embedding/extraction processes in the spatial domain and dealing with the pixels bits directly. While in the second and because of converting the pixels to their DCT coefficients this consume some additional time to perform the embedding/extraction processes.

TABLE I. PERFORMANCE EVALUATION OF THE FIRST APPROACH

	Processing time (sec)	Addition / subtraction	Multiplication / Division	Special functions
EPR spatial synchronization dynamic embedding step:	0.23	4631	6	2298
3D Forward Cloud Generator:	2.37	735000	735001	735000
Encryption step:	0.84	367500	1	2
Decryption step:	0.66	367500	1	2
3D Backward Cloud Generator:	1.35	10637	6	54
EPR spatial synchronization dynamic extraction step:	0.19	1249	6	2296

TABLE II. PERFORMANCE EVALUATION OF THE SECOND APPROACH

	Processing time (sec)	Addition / subtraction	Multiplication / Division	Special functions
EPR Spatial synchronization, dynamic and transform embedding step:	0.4	4631	6	5146
3D Forward Cloud Generator:	2.28	735000	735001	735000
Encryption step:	0.87	367500	1	2
Decryption step:	0.66	367500	1	2
3D Backward Cloud Generator:	2.26	10637	6	54
EPR Spatial synchronization, dynamic and transform extraction step:	0.28	1249	6	3720

Now, to evaluate the images quality, Table 3 and 4 were constructed. Tables 3 show that the first approach guarantees lossless reconstruction of the transferred medical images in the absence of the attacks. Table 4 shows some degree of distortion in the second approach due to the quantization operations performed during the embedding stage. But it still provides acceptable quality results as shown especially for the *SSIM* that considered as an ideal metric for testing similarities in medical images due to focusing on the local rather than global image similarity and placing more emphasis on the Human Visual System (*HVS*) than *PSNR* [20]. In general, the high results of the both approaches have been achieved due to the usage of the dynamic embedding algorithm that exploits the overall capacity of the cover image for the embedding process. In addition to that, applying *Enx*, *Eny*, *Enz* values less than or equal to 0.1 and *Hex*, *Hey*, *Hez* values equal to zero helps to generate an approximated images that most represents the original images.

Fig. 8 and Fig. 9 shows the resulting images after each step of the both approaches respectively. Start with Fig. 8 (a) that represents the original images, then after adding the EPR data shown in Fig. 10(a) through using the dynamic embedding algorithm the results will be as shown in Fig. 8 (b). The 3D-CT approximation images are shown in Fig. 8 (c) where E_x, E_y and E_z equals to colour channels pixels of the medical images, E_{nx}, E_{ny} and E_{nz} values equal to 0.1, 0.01 and 0.02, H_{ex}, H_{ey} and H_{ez} equals to zero. The encrypted versions are then shown in Fig. 8 (d). These images can reside in the cloud service provider CSP where the first level of authentication can takes place between the data owner and the CSP through K_2 . Then, after the arrival of these encrypted images to the destination, the destination uses K_2 to decrypt the images which in turn provide the second level of authentication and obtain the results shown in Fig. 8 (e). Then the destination performs 3D backward cloud generator to retrieve the cloud characters ($E_x', E_y', E_z', E_{nx}', E_{ny}', E_{nz}', H_{ex}', H_{ey}'$ and H_{ez}') to accomplish the confirmation of the data owner identity that provides the third level of authentication. Finally the destination applies K_1 to

finally get the hidden EPR data as shown in Fig. 10(b) and Fig. 10(c) from the two approaches respectively.

In Fig. 9 that refers to the second approach results, the same procedures were applied except for the embedding step which accomplished in the second approach through using dynamic embedding algorithm along with the DCT technique to provide more robustness against attacks.

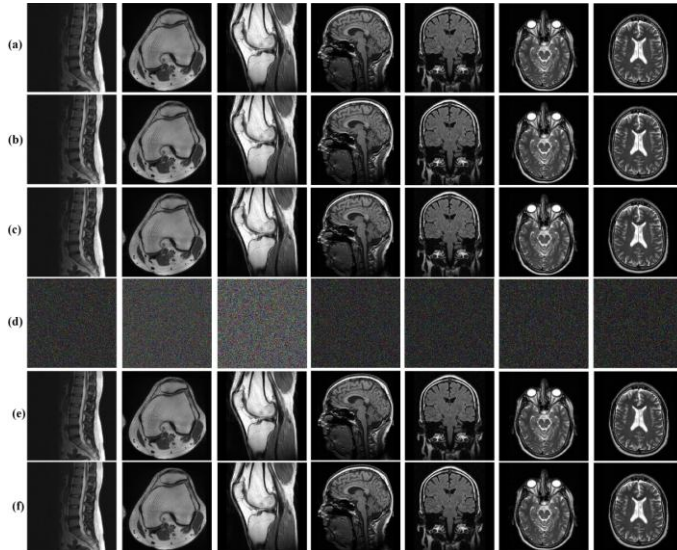


Fig. 8. The results of the first approach (a) The original medical images, (b) The images after embedding process, (c) The approximated images (d) The encrypted images, (e) The decrypted images, (f) The reconstructed lossless medical images

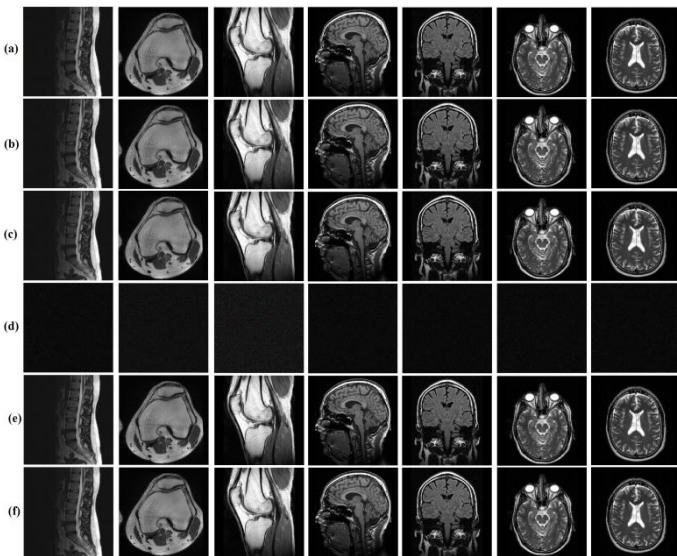


Fig. 9. The results obtained from the second approach (a) The original medical images, (b) The images after embedding process, (c) The approximated images (d) The encrypted images, (e) The decrypted images, (f) The reconstructed medical images

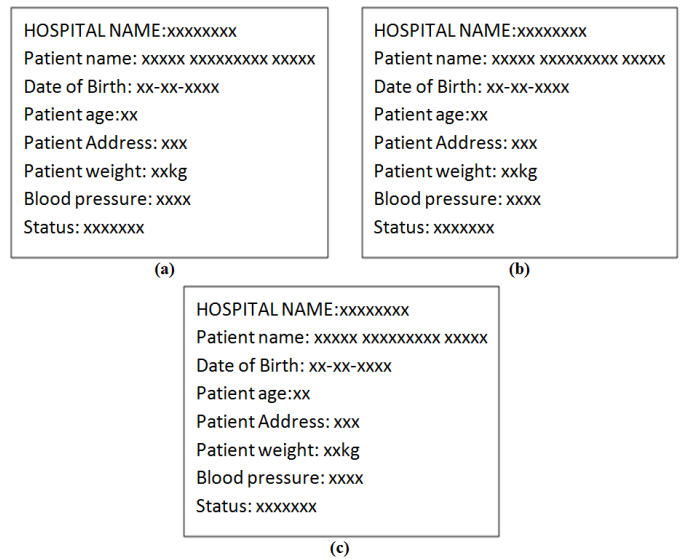


Fig. 10. The used EPR data, (a) Original EPR data, (b) Retrieved EPR data using first approach, (c) Retrieved EPR data using second approach

The scheme presented in [22] also provides a lossless retrieval of the shared image while preserving the resulting images from pixels expansion. But it offers only one level of authentication. The presented approaches here preserve the images from pixels expansion; guarantee high level of visibility of the retrieved images and at the same time offers three levels of authentication between the all authorized parties of the shared data, the owner of the data, the cloud service provider and finally the destination of the shared data which consists of the shared image and the EPR data.

Finally, the robustness of the proposed approaches has been evaluated through calculating the MSE and the PSNR for the attacked images as presenting in Tables 5 and 6 and calculating the BER for the hidden EPR data as shown in Tables 7 and 8.

Tables 5 and 6 shows that the proposed schemes provide higher degree of robustness with respect to [23, 24, 25, 26]. This means that the both approaches help to deliver the shared data to the other side of the communication with an acceptable level of quality.

For the BER results, the illustrated results in Tables 7 and 8 shows that the proposed schemes provide higher degrees of robustness under salt and pepper noise than the other attacks types. This is because salt and pepper noise affects random pixels and so not the whole embedded data were be altered. For the other attacks that affect the entire image pixels, the effect on the embedded data would be larger and hence higher BER values. The tables also shows that the second approach provides higher robustness results than first approach in most types of attacks and this is due to applying discrete cosine transform in the embedding step that helps to achieve higher robustness results against attacks as compared with performing the embedding process using dynamic embedding that is a spatial embedding algorithm.

For the salt and pepper noise, the second approach shows less robustness results than the first approach since the salt and pepper noise in the first approach affects the pixels themselves and do not propagate the distortion leading to minimum BER values. While in the second approach the distortion propagates and causes modifications in the coefficients values that increase the effects on the BER results as compared with the first approach. Moreover, the second approach, that is a hybrid between the spatial and the transform embedding processes, shows acceptable results with respect to [27].

TABLE III. QUALITY EVALUATION OF THE FIRST APPROACH

Image	MSE	PSNR (db)	SSIM	NPCR	UACI
Image 1	0	Infinity	1.000	0	0
Image 2	0	Infinity	1.000	0	0
Image 3	0	Infinity	1.000	0	0
Image 4	0	Infinity	1.000	0	0
Image 5	0	Infinity	1.000	0	0
Image 6	0	Infinity	1.000	0	0
Image 7	0	Infinity	1.000	0	0

TABLE IV. QUALITY EVALUATION OF THE SECOND APPROACH

Image	MSE	PSNR (db)	SSIM	NPCR	UACI
Image 1	0.0940	58.3988	0.9997	0.3461	0.0031
Image 2	0.0239	64.3420	1.000	0.3306	0.0014
Image 3	0.4962	51.1744	0.9987	0.6604	0.0140
Image 4	0.1408	56.6460	0.9999	0.3412	0.0036
Image 5	0.1337	56.8688	0.9999	0.3861	0.0035
Image 6	0.1017	58.0587	0.9997	0.2612	0.0029
Image 7	0.1350	56.8263	0.9998	0.2718	0.0034

TABLE V. PSNR VALUES UNDER DIFFERENT ATTACKS IN THE FIRST APPROACH

Attack type	MSE	PSNR (db)
Non attacked image	0	Infinte
Salt and pepper noise (0.001)	11.2154	37.6326
Salt and pepper noise (0.01)	115.8620	27.4914
Salt and pepper noise (0.1)	1183.3	17.3999
Speckle noise (0.001)	2.1763	44.7536
Speckle noise (0.01)	19.9383	35.1339
Speckle noise (0.1)	180.5496	25.5648
Average Filter 3x3	2166.9	14.7725
Motion (10,45)	2223.5	14.6603
Rotation (25°)	3777.7	12.3585
Rotation (45°)	3776	12.3605
Blurring	2518.6	14.1192

TABLE VI. PSNR VALUES UNDER DIFFERENT ATTACKS IN THE SECOND APPROACH

Attack type	MSE	PSNR (db)
Non attacked image	0.0940	58.3988
Salt and pepper noise (0.001)	11.3072	37.5973
Salt and pepper noise (0.01)	114.0611	27.5594
Salt and pepper noise (0.1)	1183.4	17.3994
Speckle noise (0.001)	2.2749	44.5612
Speckle noise (0.01)	20.1879	35.0799
Speckle noise (0.1)	180.7737	25.5595
Average Filter 3x3	2167	14.7722
Motion (10,45)	2223.7	14.6601
Rotation (25°)	3777.1	12.3593
Rotation (45°)	3775.3	12.3613
Blurring	2518.7	14.1191

TABLE VII. BER RESULTS FOR THE EPR DATA UNDER DIFFERENT ATTACKS IN THE FIRST APPROACH

Attack type	BER
Non attacked image	0
Salt and pepper noise (0.001)	0
Salt and pepper noise (0.01)	0.2809
Salt and pepper noise (0.1)	5.1966
Speckle noise (0.001)	53.6517
Speckle noise (0.01)	53.3708
Speckle noise (0.1)	51.6854
Average Filter 3x3	48.4551
Motion (10,45)	53.6517
Rotation (25°)	50.5618
Rotation (45°)	49.6489
Blurring	49.7893

TABLE VIII. BER RESULTS FOR THE EPR DATA UNDER DIFFERENT ATTACKS IN THE SECOND APPROACH

Attack type	BER
Non attacked image	0
Salt and pepper noise (0.001)	0
Salt and pepper noise (0.01)	1.3343
Salt and pepper noise (0.1)	8.3567
Speckle noise (0.001)	45.1545
Speckle noise (0.01)	48.5253
Speckle noise (0.1)	48.8764
Average Filter 3x3	48.7360
Motion (10,45)	47.5421
Rotation (25°)	48.8062
Rotation (45°)	49.4382
Blurring	48.7360

CONCLUSION

The presented paper introduces two approaches with aim of providing the mean of the trust management between the parties of the cloud computing environment that considered as unsecure environment to deal with. Both approaches provide three levels of authentication that are from the owner to the destination of the data. The second one is between the owner of the data and the cloud service provider. The third level is from the destination of the data to its owner. The first approach exploits the advantage of the spatial embedding techniques that considered fast and require less processing time. This approach uses dynamic embedding algorithm to increase the visibility of the shared images. The second approaches implements discrete cosine transform along with the dynamic embedding algorithm to get the advantage of the DCT in providing more robustness against attacks while preserving the fastness and the highest visibility results obtained from dealing with dynamic embedding algorithm. The future work has an aim of implementing the hybrid model using other transform domain embedding techniques and evaluates the results in order to maximize the robustness against the attacking attempts. Also it has an aim of applying the proposed approaches in other sorts of medical data and test the consequent performance.

REFERENCES

[1] Sonika C. Rathi, Vandana S. Inamdar, "Analysis of watermarking techniques for medical images preserving ROI", Computer Science & Information Technology (CS & IT 05) - open access-Computer Science Conference Proceedings (CSCP) , pp. 297-308 , 2012.

[2] Borko Furht, Armondo Escalante, HandBook of Cloud computing, Springer Science + business Media, LLC 2010.

- [3] Danwei Chen and Yanjun He, "A Study on Secure Data Storage Strategy in Cloud Computing", *Journal of Convergence Information Technology*, Volume 5, Number 7, September 2010.
- [4] Mustafa Ulutas, Güzin Ulutas, Vasif V. Nabiyev." Medical image security and EPR hiding using Shamir's secret sharing scheme". *The Journal of Systems and Software* 84 (2011), 341–353.
- [5] Jasni Mohamad Zain and Malcolm Clarke, "Reversible Region of Non-Interest (RONI) watermarking for authentication of DICOM Images", *IJCSNS International Journal of Computer Science and Network Security*, vol. 7, No.9, pp. 19-28, September 2007.
- [6] Z. Eslami and J. Zarepour Ahmadabadi, "Secret image sharing with authentication-chaining and dynamic embedding", *The Journal of Systems and Software*, vol. 84, pp. 803–809, May 2011.
- [7] Siau-Chuin Liew, Siau-Way Liew and Jasni Mohd Zain, "Reversible Medical Image Watermarking For Tamper Detection and Recovery with Run Length Encoding Compression", *World Academy of Science, Engineering and Technology*, vol. 48, pp.799-803, December 2010.
- [8] Chao-Tung Yang; Lung-Teng Chen; Wei-Li Chou; Kuan-Chieh Wang, "Implementation of a Medical Image File Accessing System on Cloud Computing". 2010 IEEE 13th International Conference on Computational Science and Engineering (CSE), DOI:10.1109/CSE.2010.48, pp.:321-326
- [9] G.Kanagaraj,A.C.Sumathi. "Proposal of an open-source Cloud computing system for exchanging medical images of a Hospital Information System". 2011 3rd International Conference on Trendz in Information Sciences and Computing (TISC), DOI:10.1109/TISC.2011.6169102 .Page(s): 144 – 149.
- [10] House of Commons, Health Committee. *The Electronic Patient Record. Sixth Report of Session 2006–07, Volume I, Ordered by The House of Commons to be printed 25 July 2007.*
- [11] MATLAB version 7.6.0.324 (R2008a), 2008, computer software, The MathWorks Inc., Natick.
- [12] D.P. Kroese, T. Taimre, Z.I. Botev, *Handbook of Monte Carlo Methods*. Wiley Series in Probability and Statistics, John Wiley & Sons, New York, 2011, ch. 1, pp. 7.
- [13] Deyi Li, Yi Du, *Artificial Intelligence with Uncertainty*, Chapman and Hall/CRC, pp. 107–151, September 27, 2007.
- [14] A.Menezes, P.van Oorschot, and S.Vanstone, *Handbook of Applied Cryptography*, CRC Press, 1996.
- [15] Teddy Furon, *Watermarking for alternative requirements*, INRIA, Université de Rennes 1, 2005.
- [16] Mr.Navnath S. Narawade and Dr.Rajendra D.Kanphade," DCT Based Robust Reversible Watermarking For Geometric Attack". *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)*, Volume 1, Issue 2, July – August 2012.
- [17] RadLink centre (2000) RadLink Diagnostic Imaging, <http://radlink.com.sg/> [Accessed 17/1/2013].
- [18] SoftWays' Medical Imaging Group (2003) *Magnetic Resonance - Technology Information Portal*, <http://www.mr-tip.com/> [Accessed 17/1/2013].
- [19] Yue Wu, Joseph P. Noonan, Sos Aгаian, "NPCR and UACI Randomness Tests for Image Encryption", *Journal of Selected Areas in Telecommunications (JSAT)*, April Edition, 2011.
- [20] Farhad Rahimi and Hossein Rabbani, "A dual adaptive watermarking scheme in contourlet domain for DICOM images", *BioMedical Engineering OnLine*, 2011.
- [21] Chun-Shien Lu, *Multimedia security: steganography and digital watermarking techniques for protection of intellectual property*. Idea Group Inc (IGI), 2005, pp. 100.
- [22] Hao-Kuan Tso and Der-Chyuan Lou, "Medical image protection using secret sharing scheme". In *Proceedings of the 6th International Conference on Ubiquitous Information Management and Communication (ICUIMC '12)*. ACM, New York, NY, USA, Article 93, 4 pages, (2012).
- [23] Surya Pratap Singh, Paresh Rawat and Sudhir Agrawal, "A Robust Watermarking Approach using DCT-DWT". *International Journal of Emerging Technology and Advanced Engineering (ISSN 2250-2459, Volume 2, Issue 8, August 2012)*.
- [24] U. M. Gokhale and Y. V. Joshi, "A New Watermarking Algorithm Based on Image Scrambling and SVD in the Wavelet Domain". *ACEEE Int. J. on Network Security*, Vol. 02, No. 03, July 2011.
- [25] Khaled Loukhaoukha, Jean-Yves Chouinard, and Abdellah Berdai, "A Secure Image Encryption Algorithm Based on Rubik's Cube Principle". *Hindawi Publishing Corporation Journal of Electrical and Computer Engineering* Volume 2012, Article ID 173931, 13 pages.
- [26] Patil Ramana Reddy, Munaga.V.N.K.Prasad and D.Srinivasa Rao, "Digital Image Watermarking Using SPIHT". *International Journal of Recent Trends in Engineering*, Vol 2, No. 3, November 2009.
- [27] Ghazali Bin Sulong, Harith Hasan, Ali Selamat, Mohammed Ibrahim and Saparudin, " A New Color Image Watermarking Technique Using Hybrid Domain". *IJCSI International Journal of Computer Science Issues*, Vol. 9, Issue 6, No 1, November 2012.

Revisit of Logistic Regression: Efficient Optimization and Kernel Extensions

Takumi Kobayashi

National Institute of Advanced Industrial Science and Technology
Umezono 1-1-1, Tsukuba, 305-8568, Japan
Email: takumi.kobayashi@aist.go.jp

Kenji Watanabe

Wakayama University
Sakaedani 930, Wakayama, 640-8510, Japan
Email: k-watanabe@vrl.sys.wakayama-u.ac.jp

Nobuyuki Otsu

National Institute of Advanced Industrial Science and Technology
Umezono 1-1-1, Tsukuba, 305-8568, Japan
Email: otsu.n@aist.go.jp

Abstract—Logistic regression (LR) is widely applied as a powerful classification method in various fields, and a variety of optimization methods have been developed. To cope with large-scale problems, an efficient optimization method for LR is required in terms of computational cost and memory usage. In this paper, we propose an efficient optimization method using non-linear conjugate gradient (CG) descent. In each CG iteration, the proposed method employs the optimized step size without exhaustive line search, which significantly reduces the number of iterations, making the whole optimization process fast. In addition, on the basis of such CG-based optimization scheme, a novel optimization method for kernel logistic regression (KLR) is proposed. Unlike the ordinary KLR methods, the proposed method optimizes the kernel-based classifier, which is naturally formulated as the linear combination of sample kernel functions, directly in the reproducing kernel Hilbert space (RKHS), not the linear coefficients. Subsequently, we also propose the multiple-kernel logistic regression (MKLR) along with the optimization of KLR. The MKLR effectively combines the multiple types of kernels with optimizing the weights for the kernels in the framework of the logistic regression. These proposed methods are all based on CG-based optimization and matrix-matrix computation which is easily parallelized such as by using multi-thread programming. In the experimental results on multi-class classifications using various datasets, the proposed methods exhibit favorable performances in terms of classification accuracies and computation times.

I. INTRODUCTION

A classification problem is an intensive research topic in the pattern recognition field. Especially, classifying the feature vectors extracted from input data plays an important role; e.g., for image (object) recognition [1] and detection [2], motion recognition [3], natural language processing [4]. Nowadays, we can collect a large amount of data such as via internet, and thus large-scale problems have been frequently addressed in those fields to improve classification performances.

In the last decade, the classification problems have been often addressed in the large margin framework [5] as represented by support vector machine (SVM) [6]. While those methods are basically formulated for linear classification, they are also extended to kernel-based methods by employing kernel functions and produce promising performances. However, they are mainly intended for binary (two) class problems and it

is generally difficult to extend the method toward the multi-class problems without heuristics such as a one-versus-rest approach. Several methods, however, are proposed to cope with the multi-class problems, e.g., in [7]. Another drawback is that the optimization in those methods has difficulty in parallelization. The SVM-based methods are formulated in quadratic programming (QP). Some successful optimization methods to solve the QP, such as sequential minimal optimization (SMO) [8], are based on a sequential optimization approach which can not be easily parallelized. Parallel computing currently developed such as by using GPGPU would be a key tool to effectively treat large-scale data.

On the other hand, logistic regression has also been successfully applied in various classification tasks. Apart from the margin-based criterion for the classifiers, the logistic regression is formulated in the probabilistic framework. Therefore, it is advantageous in that 1) the classifier outputs (class) posterior probabilities and 2) the method is naturally generalized to the multi-class classifiers by employing a multi-nominal logistic function which takes into account the correlations among classes. While the optimization problem, i.e., objective cost function, for logistic regression is well-defined, there is still room to argue about its optimization method in terms of computational cost and memory usage, especially to cope with large-scale problems. A popular method, *iterative reweighted least squares*, is based on the Newton-Raphson method [9] requiring significant computation cost due to the Hessian.

In this paper, we propose an efficient optimization method for the logistic regression. The proposed method is based on non-linear conjugate gradient (CG) descent [10] which is directly applied to minimize the objective cost. The non-linear CG is widely applied to unconstrained optimization problems, though requiring an exhaustive line search to determine a step size in each iteration. In the proposed method, we employ the optimum step size without the line search, which makes the whole optimization process more efficient by significantly reducing the number of iterations. In addition, we propose a novel optimization method for kernel logistic regression (KLR) on the basis of the CG-based optimization scheme. Unlike the ordinary KLR methods, the proposed method optimizes the kernel-based classifier, which is naturally formulated as the linear combination of sample kernel functions as in SVM, di-

TABLE I. NOTATIONS

N	Number of samples
C	Number of classes
\mathbf{x}_i	Feature vector of the i -th sample ($\in \mathbb{R}^L$)
\mathbf{y}_i	Class indicator vector of the i -th sample ($\in \{0, 1\}^C$) in which only the assigned class component is 1 and the others 0
\mathbf{X}	Matrix containing feature vectors \mathbf{x}_i in its columns ($\in \mathbb{R}^{L \times N}$)
\mathbf{Y}	Matrix containing class vectors \mathbf{y}_i in its columns ($\in \{0, 1\}^{C \times N}$)
\mathcal{H}	Reproducing kernel Hilbert space (RKHS)
$k(\cdot, \cdot)$	Kernel function in RKHS \mathcal{H}
\mathbf{K}	Kernel Gram matrix of $[k(\mathbf{x}_i, \mathbf{x}_j)]_{i,j=1,\dots,N}^{j=1,\dots,N}$ ($\in \mathbb{R}^{N \times N}$)
$[\cdot]_{1:C-1}$	Operator extracting the 1~ $C-1$ -th rows of a matrix/vector
$[\cdot]_{i=1,\dots,W}^{j=1,\dots,W}$	Operator constructing the matrix of the size $\mathbb{R}^{H \times W}$ where the lower/upper index is for the row/column.
\cdot^\top	Transpose of a matrix/vector
$\langle \cdot, \cdot \rangle$	Frobenius inner product of matrices, i.e., $\langle \mathbf{A}, \mathbf{B} \rangle = \text{trace}(\mathbf{A}^\top \mathbf{B})$

rectly in the reproducing kernel Hilbert space (RKHS), not the linear coefficients of samples. Subsequently, multiple-kernel logistic regression (MKLR) is also proposed as multiple-kernel learning (MKL). The MKL combines the multiple types of kernels with optimizing the weights for the kernels, and it has been addressed mainly in the large margin framework [11]. The proposed MKLR is formulated as a convex form in the framework of logistic regression. In the proposed formulation, by resorting to the optimization method in the KLR, we optimize the kernel-based classifier in sum of multiple RKHSs and consequently the linear weights for the multiple kernels. In summary, the contributions of this paper are as follows;

- Non-linear CG in combination with the optimum step size for optimizing logistic regression.
- Novel method for kernel logistic regression to directly optimize the classifier in RKHS.
- Novel method of multiple-kernel logistic regression.

Note that all the proposed methods are based on the CG-based optimization and the computation cost is dominated by matrix-matrix computation which is easily parallelized.

The rest of this paper is organized as follows: the next section briefly reviews the related works of optimization for logistic regression. In Section III, we describe the details of the proposed method using non-linear CG. And then in Section IV and Section V we propose the novel optimization methods for kernel logistic regression and for multiple-kernel logistic regression. In Section VI, we mention the parallel computing in the proposed methods. The experimental results on various types of multi-class classification are shown in Section VII. Finally, Section VIII contains our concluding remarks.

This paper contains substantial improvements over the preliminary version [12] in that we develop the kernel-based methods including MKL and give new experimental results.

A. Notations

We use the notations shown in Table I. Basically, the big bold letter, e.g., \mathbf{X} , indicates a matrix, its small bold letter with the index, e.g., \mathbf{x}_i , denotes the i -th column vector, and the small letter with two indexes, e.g., x_{ic} , indicates the c -th component of the i -th column vector \mathbf{x}_i , corresponding to the c -th row and i -th column element of \mathbf{X} .

To cope with multi-class problems, we apply the following multi-nominal logistic function for the input $\mathbf{z} \in \mathbb{R}^{C-1}$:

$$\sigma_c(\mathbf{z}) = \begin{cases} \frac{\exp(z_c)}{1 + \sum_{k=1}^{C-1} \exp(z_k)} & (c < C) \\ \frac{1}{1 + \sum_{k=1}^{C-1} \exp(z_k)} & (c = C) \end{cases}, \quad \boldsymbol{\sigma}(\mathbf{z}) = \begin{bmatrix} \sigma_1(\mathbf{z}) \\ \vdots \\ \sigma_C(\mathbf{z}) \end{bmatrix} \in \mathbb{R}^C,$$

where $\sigma_c(\mathbf{z})$ outputs a posterior probability on the c -th class and $\boldsymbol{\sigma}(\mathbf{z})$ produces the probabilities over the whole C classes.

II. RELATED WORKS

The (multi-class) logistic regression is also mentioned in the context of the maximum entropy model [13] and the conditional random field [14]. We first describe the formulation of linear logistic regression. The linear logistic regression estimates the class posterior probabilities $\hat{\mathbf{y}}$ from the input feature vector $\mathbf{x} \in \mathbb{R}^L$ by using the above logistic function:

$$\hat{\mathbf{y}} = \boldsymbol{\sigma}(\mathbf{W}^\top \mathbf{x}) \in \mathbb{R}^C,$$

where $\mathbf{W} \in \mathbb{R}^{L \times C-1}$ is the classifier weight matrix. To optimize \mathbf{W} , the following objective cost is minimized:

$$J(\mathbf{W}) = - \sum_i^N \sum_c^C y_{ic} \log \sigma_c(\mathbf{W}^\top \mathbf{x}_i) \rightarrow \min_{\mathbf{W}}. \quad (1)$$

There exists various methods for optimizing the logistic regression, as described below. Comparative studies on those optimization methods are shown in [13], [15].

A. Newton-Raphson method

For simplicity, we unfold the weight matrix \mathbf{W} into the long vector $\mathbf{w} = [\mathbf{w}_1^\top, \dots, \mathbf{w}_{C-1}^\top]^\top \in \mathbb{R}^{L(C-1)}$. The derivatives of the cost function (1) is given by

$$\nabla_{\mathbf{w}_c} J = \sum_i^N \mathbf{x}_i (\hat{y}_{ic} - y_{ic}) \in \mathbb{R}^L, \quad \nabla_{\mathbf{w}} J = \begin{bmatrix} \nabla_{\mathbf{w}_1} J \\ \vdots \\ \nabla_{\mathbf{w}_{C-1}} J \end{bmatrix} \in \mathbb{R}^{L(C-1)},$$

where $\hat{y}_{ic} = \sigma_c(\mathbf{W}^\top \mathbf{x}_i)$, and the Hessian of J is obtained as

$$\mathbf{H}_{c,k} = \nabla_{\mathbf{w}_c} \nabla_{\mathbf{w}_k}^\top J = \sum_i^N y_{ic} (\delta_{ck} - y_{ik}) \mathbf{x}_i \mathbf{x}_i^\top \in \mathbb{R}^{L \times L},$$

$$\mathbf{H} = \begin{pmatrix} \mathbf{H}_{1,1} & \cdots & \mathbf{H}_{1,C-1} \\ \vdots & \ddots & \vdots \\ \mathbf{H}_{C-1,1} & \cdots & \mathbf{H}_{C-1,C-1} \end{pmatrix} = [\mathbf{H}_{ck}]_{c=1,\dots,C-1}^{k=1,\dots,C-1} \in \mathbb{R}^{L(C-1) \times L(C-1)},$$

where δ_{ck} is the Kronecker delta. This Hessian matrix is positive definite, and thus the optimization problem in (1) is convex. For the optimization, the Newton-Raphson update is described by

$$\mathbf{w}^{new} = \mathbf{w}^{old} - \mathbf{H}^{-1} \nabla_{\mathbf{w}} J = \mathbf{H}^{-1} (\mathbf{H} \mathbf{w}^{old} - \nabla_{\mathbf{w}} J) = \mathbf{H}^{-1} \mathbf{z}, \quad (2)$$

where $\mathbf{z} \triangleq \mathbf{H} \mathbf{w}^{old} - \nabla_{\mathbf{w}} J$. This update procedure, which can be regarded as reweighted least squares, is repeated until convergence. Such a method based on Newton-Raphson, called *iterative reweighted least squares* (IRLS) [16], is one of the commonly used optimization methods.

This updating of \mathbf{w} in (2) requires the inverse matrix computation for the Hessian. In the case of large-dimensional feature vectors and large number of classes, it requires much computational cost to compute the inverse of the large Hessian matrix. To cope with such difficulty in large-scale data, various optimization methods have been proposed by making the update (2) efficient. Komarek and Moore [17] regarded (2) as the solution of the following linear equations, $\mathbf{H} \mathbf{w}^{new} = \mathbf{z}$, and they apply Cholesky decomposition to efficiently solve it. On the other hand, Komarek and Moore [18] applied linear conjugate-gradient (CG) descent to solve these linear equations [19]. The CG method is applicable even to large-dimensional Hessian \mathbf{H} . Recently, Lin et al. [20] employed

trust-region method [9] to increase the efficiency of the Newton-Raphson update using the linear CG. Note that the method in [20] deals with multi-class problems in a slightly different way from ordinary multi-class LR by considering one-against-rest approach.

B. Quasi Newton method

As described above, it is inefficient to explicitly compute the Hessian for multi-class large dimensional data. To remedy it, Malouf [13] and Daumé III [21] presented the optimization method using limited memory BFGS [22]. In the limited memory BFGS, the Hessian is approximately estimated in a computationally efficient manner and the weight \mathbf{W} is updated by using the approximated Hessian $\hat{\mathbf{H}}$.

C. Other optimization methods

Besides those Newton-based methods, the other optimization methods are also applied. For example, Pietra et al. [23] proposed the method of *improved iterative scaling*, and Minka [15] and Daumé III [21] presented the method using non-linear CG [10] with an exhaustive line search.

In this study, we focus on the non-linear CG based optimization due to its favorable performance reported in [15] and its simple formulation which facilitates the extensions to the kernel-based methods.

D. Kernel logistic regression

Kernel logistic regression [24], [25], [26] is an extension of the linear logistic regression by using kernel function. By considering the classifier function $f_c(\cdot)$, $c \in \{1, \dots, C-1\}$, the class posterior probabilities are estimated from \mathbf{x} by

$$\hat{\mathbf{y}} = \sigma([f_c(\mathbf{x})]_{c=1, \dots, C-1}) \in \mathbb{R}^C.$$

As in the other kernel-based methods [27], $f_c(\cdot)$ is represented by the linear combinations of sample kernel functions $k(\mathbf{x}_i, \cdot)$ in the reproducing kernel Hilbert space (RKHS) \mathcal{H} :

$$f_c(\cdot) = \sum_i^N w_{ci} k(\mathbf{x}_i, \cdot) \Rightarrow \sigma([f_c(\mathbf{x})]_{c=1, \dots, C-1}) = \sigma(\mathbf{W}^\top \mathbf{k}(\mathbf{x})),$$

where $\mathbf{W} = [w_{ci}]_{i=1, \dots, N}^{c=1, \dots, C-1} \in \mathbb{R}^{N \times C-1}$ indicates the (linear) coefficients of the samples for the classifier and $\mathbf{k}(\mathbf{x}) = [k(\mathbf{x}_i, \mathbf{x})]_{i=1, \dots, N} \in \mathbb{R}^N$ is a kernel feature vector. Ordinary kernel logistic regression is formulated in the following optimization problem;

$$J(\mathbf{W}) = - \sum_i^N \sum_c^C y_{ic} \log \{ \sigma_c(\mathbf{W}^\top \mathbf{k}(\mathbf{x}_i)) \} \rightarrow \min_{\mathbf{W}}.$$

This corresponds to the linear logistic regression in (1) except that the feature vectors are replaced by the kernel feature vectors $\mathbf{x}_i \mapsto \mathbf{k}(\mathbf{x}_i)$ and the classifier weights $\mathbf{W} \in \mathbb{R}^{N \times C-1}$ are formulated as the coefficients for the samples.

III. EFFICIENT OPTIMIZATION FOR LINEAR LOGISTIC REGRESSION

In this section, we propose the optimization method for linear logistic regression which efficiently minimizes the cost even for the large-scale data. The proposed method is based on non-linear CG method [10] directly applicable to the optimization as in [15], [21]. Our contribution is that the step

size required in CG updates is optimized without an exhaustive line search employed in an ordinary non-linear CG method, in order to significantly reduce the number of iterations and speed-up the optimization. The non-linear CG can also save memory usage without relying on the Hessian matrix. The proposed method described in this section serves as a basis for kernel-based extensions in Section IV and V.

A. Non-linear CG optimization for linear logistic regression

We minimize the following objective cost with the regularization term, L_2 -norm of the classifier weights $\mathbf{W} \in \mathbb{R}^{L \times C-1}$:

$$J(\mathbf{W}) = \frac{\lambda}{2} \|\mathbf{W}\|_F^2 - \sum_i^N \sum_c^C y_{ic} \log \{ \sigma_c(\mathbf{W}^\top \mathbf{x}_i) \} \rightarrow \min_{\mathbf{W}}, \quad (3)$$

where $\|\mathbf{W}\|_F^2 = \langle \mathbf{W}, \mathbf{W} \rangle$ and λ is a regularization parameter. The gradient of J with respect to \mathbf{W} is given by

$$\nabla_{\mathbf{W}} J = \lambda \mathbf{W} + \mathbf{X} [\hat{\mathbf{Y}} - \mathbf{Y}]_{1:C-1}^\top \in \mathbb{R}^{L \times C-1},$$

where $\hat{\mathbf{Y}} = [\hat{y}_i = \sigma(\mathbf{W}^\top \mathbf{x}_i)]_{i=1, \dots, N} \in \mathbb{R}^{C \times N}$.

The non-linear CG method utilizes the gradient $\nabla_{\mathbf{W}} J$ to construct the conjugate gradient, and the cost (3) is minimized iteratively. At the l -th iteration, letting $\mathbf{G}^{(l)} \triangleq \nabla_{\mathbf{W}} J(\mathbf{W}^{(l)})$, the conjugate gradient $\mathbf{D}^{(l)} \in \mathbb{R}^{L \times C-1}$ is provided by

$$\mathbf{D}^{(l)} = -\mathbf{G}^{(l)} + \beta \mathbf{D}^{(l-1)}, \quad \mathbf{D}^{(0)} = -\mathbf{G}^{(0)},$$

where β is a CG update parameter. There are various choices for β [10]; we employ the update parameter in [28]:

$$\beta = \max \left\{ \frac{\langle \mathbf{G}^{(l)}, \mathbf{G}^{(l)} - \mathbf{G}^{(l-1)} \rangle}{\langle \mathbf{D}^{(l-1)}, \mathbf{G}^{(l)} - \mathbf{G}^{(l-1)} \rangle}, 0 \right\} - \theta \frac{\langle \mathbf{G}^{(l)}, \mathbf{W}^{(l)} - \mathbf{W}^{(l-1)} \rangle}{\langle \mathbf{D}^{(l-1)}, \mathbf{G}^{(l)} - \mathbf{G}^{(l-1)} \rangle}, \quad (4)$$

where we set $\theta = 0.5$ in this study. Then, the classifier weight \mathbf{W} is updated by using the conjugate gradient:

$$\mathbf{W}^{(l+1)} = \mathbf{W}^{(l)} + \alpha \mathbf{D}^{(l)}, \quad (5)$$

where α is a step size, the determination of which is described in the next section. These non-linear CG iterations are repeated until convergence.

B. Optimum step size α

The step size α in (5) is critical for efficiency in the optimization, and it is usually determined by an exhaustive line search satisfying Wolfe condition in an ordinary non-linear CG [10]. We optimize the step size α so as to minimize the cost function:

$$\alpha = \arg \min_{\alpha} J(\mathbf{W} + \alpha \mathbf{D}), \quad (6)$$

$$J(\mathbf{W} + \alpha \mathbf{D}) = \frac{\lambda}{2} \|\mathbf{W} + \alpha \mathbf{D}\|_F^2 - \sum_i^N \sum_c^C y_{ic} \log \{ \sigma_c(\mathbf{W}^\top \mathbf{x}_i + \alpha \mathbf{D}^\top \mathbf{x}_i) \}.$$

Here, we introduce auxiliary variables, $\mathbf{P} = \mathbf{W}^\top \mathbf{X} \in \mathbb{R}^{C-1 \times N}$, $\mathbf{Q} = \mathbf{D}^\top \mathbf{X} \in \mathbb{R}^{C-1 \times N}$ and $\hat{\mathbf{Y}} = [\hat{y}_i = \sigma((\mathbf{W} + \alpha \mathbf{D})^\top \mathbf{x}_i) = \sigma(\mathbf{p}_i + \alpha \mathbf{q}_i)]_{i=1, \dots, N}$, and thereby the gradient and Hessian of J with respect to α are written by

$$\frac{dJ}{d\alpha} = \lambda \{ \alpha \langle \mathbf{D}, \mathbf{D} \rangle + \langle \mathbf{W}, \mathbf{D} \rangle \} + \sum_i^N \mathbf{q}_i^\top [\hat{\mathbf{y}}_i - \mathbf{y}_i]_{1:C-1} \triangleq g(\alpha),$$

$$\frac{d^2 J}{d\alpha^2} = \lambda \langle \mathbf{D}, \mathbf{D} \rangle + \sum_i^N \sum_c^{C-1} \hat{y}_{ic} q_{ic} \left(q_{ic} - \sum_k^{C-1} \hat{y}_{ik} q_{ik} \right) \triangleq h(\alpha).$$

Algorithm 1 : Logistic Regression by non-linear CG

Input: $\mathbf{X} = [\mathbf{x}_i]_{i=1,\dots,N} \in \mathbb{R}^{L \times N}$, $\mathbf{Y} = [\mathbf{y}_i]_{i=1,\dots,N} \in \{0,1\}^{C \times N}$

1: **Initialize** $\mathbf{W}^{(0)} = \mathbf{0} \in \mathbb{R}^{L \times C-1}$, $\hat{\mathbf{Y}} = [\frac{1}{C}] \in \mathbb{R}^{C \times N}$
 $\mathbf{G}^{(0)} = \mathbf{X}[\hat{\mathbf{Y}} - \mathbf{Y}]_{1:C-1}^\top \in \mathbb{R}^{L \times C-1}$,
 $\mathbf{D}^{(0)} = -\mathbf{G}^{(0)} \in \mathbb{R}^{L \times C-1}$
 $\mathbf{P} = \mathbf{W}^{(0)\top} \mathbf{X} = \mathbf{0} \in \mathbb{R}^{C-1 \times N}$, $l = 1$

2: **repeat**

3: $\mathbf{Q} = \mathbf{D}^{(l-1)\top} \mathbf{X} \in \mathbb{R}^{C-1 \times N}$

4: $\alpha = \arg \min_{\alpha} J(\mathbf{W}^{(l-1)} + \alpha \mathbf{D}^{(l-1)})$: see Section III-B

5: $\mathbf{W}^{(l)} = \mathbf{W}^{(l-1)} + \alpha \mathbf{D}^{(l-1)}$, $\mathbf{P} \leftarrow \mathbf{P} + \alpha \mathbf{Q}$

6: $\hat{\mathbf{Y}} = [\hat{y}_i = \sigma(\mathbf{p}_i)]_{i=1,\dots,N}$
 $J^{(l)} = J(\mathbf{W}^{(l)}) = \frac{\lambda}{2} \|\mathbf{W}^{(l)}\|_F^2 - \sum_i^N \sum_c^C y_{ic} \log \hat{y}_{ic}$

7: $\mathbf{G}^{(l)} = \mathbf{X}[\hat{\mathbf{Y}} - \mathbf{Y}]_{1:C-1}^\top$

8: $\beta = \max \left\{ \frac{\langle \mathbf{G}^{(l)}, \mathbf{G}^{(l)} - \mathbf{G}^{(l-1)} \rangle}{\langle \mathbf{D}^{(l-1)}, \mathbf{G}^{(l)} - \mathbf{G}^{(l-1)} \rangle}, 0 \right\} - \theta \frac{\langle \mathbf{G}^{(l)}, \mathbf{W}^{(l)} - \mathbf{W}^{(l-1)} \rangle}{\langle \mathbf{D}^{(l-1)}, \mathbf{G}^{(l)} - \mathbf{G}^{(l-1)} \rangle}$

9: $\mathbf{D}^{(l)} = -\mathbf{G}^{(l)} + \beta \mathbf{D}^{(l-1)}$, $l \leftarrow l + 1$

10: **until** convergence

Output: $\mathbf{W} = \mathbf{W}^{(l)}$

Since this Hessian is non-negative, the optimization problem in (6) is convex. Based on these quantities, we apply Newton-Raphson method to (6),

$$\alpha^{new} = \alpha^{old} - \frac{g(\alpha^{old})}{h(\alpha^{old})}. \quad (7)$$

This is a one-dimensional optimization and it terminates in only a few iterations in most cases. By employing so optimized step size α , the number of CG iterations is significantly reduced compared to the ordinary non-linear CG method using a line search [15], [21].

The overall algorithm is shown in Algorithm 1. In this algorithm, the number of matrix multiplication which requires large computation time is reduced via updating the quantities \mathbf{P}, \mathbf{Q} ; as a result, the matrix multiplication is required only two times (line 3 and 7 in Algorithm 1) per iteration.

IV. NOVEL OPTIMIZATION FOR KERNEL LOGISTIC REGRESSION

As reviewed in Section II-D, the kernel logistic regression has been formulated in the optimization problem with respect to the coefficients $\mathbf{W} \in \mathbb{R}^{N \times C-1}$ over the samples by simply substituting the kernel features $\mathbf{k}(\mathbf{x}_i)$ for the feature vectors \mathbf{x}_i . It optimizes the classifier in the subspace spanned by the kernel functions of samples, which tends to cause numerically unfavorable issues such as plateau. We will discuss these issues in the experiments. In contrast to the ordinary method, we propose a novel method for kernel logistic regression that directly optimizes the classifier f_c in RKHS \mathcal{H} , not the coefficients of the samples, by employing the scheme of the non-linear CG-based optimization described in Section III.

By introducing regularization on the classifier f_c , $c \in \{1, \dots, C-1\}$, the kernel logistic regression is optimized by $J(\{f_c\}_{c=1,\dots,C-1})$

$$= \frac{\lambda}{2} \sum_c^{C-1} \|f_c\|_{\mathcal{H}}^2 - \sum_i^N \sum_c^C y_{ic} \log \left\{ \sigma_c([\mathbf{f}_c(\mathbf{x}_i)]_{c=1,\dots,C-1}) \right\} \rightarrow \min_{\{f_c\}}$$

and the gradient of J with respect to f_c is given by

$$\mathbf{g}_c(\cdot) = \lambda f_c(\cdot) + \sum_i^N (\hat{y}_{ic} - y_{ic}) \mathbf{k}(\mathbf{x}_i, \cdot), \quad (8)$$

where $\hat{y}_{ic} = \sigma_c([\mathbf{f}_c(\mathbf{x}_i)]_{c=1,\dots,C-1})$ and we use $f_c(\mathbf{x}) = \langle \mathbf{f}_c(\cdot), \mathbf{k}(\mathbf{x}, \cdot) \rangle_{\mathcal{H}}$. The conjugate gradient is obtained as

$$\begin{aligned} \mathbf{d}_c^{(l)}(\cdot) &= -\mathbf{g}_c^{(l)}(\cdot) + \beta \mathbf{d}_c^{(l-1)}(\cdot) \\ &= -\lambda f_c^{(l)}(\cdot) - \sum_i^N (\hat{y}_{ic}^{(l)} - y_{ic}) \mathbf{k}(\mathbf{x}_i, \cdot) + \beta \mathbf{d}_c^{(l-1)}(\cdot), \quad (9) \end{aligned}$$

$$\mathbf{d}_c^{(0)}(\cdot) = -\mathbf{g}_c^{(0)}(\cdot) = -\lambda f_c^{(0)}(\cdot) - \sum_i^N (\hat{y}_{ic}^{(0)} - y_{ic}) \mathbf{k}(\mathbf{x}_i, \cdot),$$

and the classifier f_c is updated by

$$f_c^{(l)}(\cdot) = f_c^{(l-1)}(\cdot) + \alpha \mathbf{d}_c^{(l-1)}(\cdot). \quad (10)$$

Based on these update formula, if the initial classifier $f_c^{(0)}(\cdot)$ is a linear combination of the sample kernel functions $\mathbf{k}(\mathbf{x}_i, \cdot)$, it is recursively ensured that all of the functions $f_c^{(l)}(\cdot)$, $\mathbf{g}_c^{(l)}(\cdot)$ and $\mathbf{d}_c^{(l)}(\cdot)$ can also be represented by such linear combinations as well. In addition, at the optimum, the classifier function eventually takes the following form,

$$\lambda f_c(\cdot) + \sum_i^N (\hat{y}_{ic} - y_{ic}) \mathbf{k}(\mathbf{x}_i, \cdot) = 0, \therefore f_c(\cdot) = \frac{1}{\lambda} \sum_i^N (y_{ic} - \hat{y}_{ic}) \mathbf{k}(\mathbf{x}_i, \cdot).$$

Thus, the above-mentioned linear combination is actually essential to represent $f_c^{(l)}$. In this study, by initializing the classifier $f_c^{(0)} = \mathbf{0}$, such representations are realized; we denote $f_c(\cdot) = \sum_i^N w_{ci} \mathbf{k}(\mathbf{x}_i, \cdot)$, $\mathbf{g}_c(\cdot) = \sum_i^N g_{ci} \mathbf{k}(\mathbf{x}_i, \cdot)$ and $\mathbf{d}_c(\cdot) = \sum_i^N d_{ci} \mathbf{k}(\mathbf{x}_i, \cdot)$. Consequently, the updates (8), (9) and (10) are applied only to those coefficients:

$$\mathbf{G}^{(l)} = \lambda \mathbf{W}^{(l)} + [\hat{\mathbf{Y}} - \mathbf{Y}]_{1:C-1}^\top \in \mathbb{R}^{N \times C-1}, \quad (11)$$

$$\mathbf{D}^{(l+1)} = -\mathbf{G}^{(l)} + \beta \mathbf{D}^{(l)}, \mathbf{D}^{(0)} = -\mathbf{G}^{(0)} \in \mathbb{R}^{N \times C-1}, \quad (12)$$

$$\mathbf{W}^{(l+1)} = \mathbf{W}^{(l)} + \alpha \mathbf{D}^{(l)} \in \mathbb{R}^{N \times C-1}, \quad (13)$$

where $\hat{\mathbf{Y}} = [\hat{y}_i = \sigma(\mathbf{W}^{(l)\top} \mathbf{k}_i)]_{i=1,\dots,N} \in \mathbb{R}^{C \times N}$, α is a step size and the CG update parameter β is given in a manner similar to (4) by

$$\beta = \max \left\{ \frac{\langle \mathbf{K} \mathbf{G}^{(l)}, \mathbf{G}^{(l)} - \mathbf{G}^{(l-1)} \rangle}{\langle \mathbf{K} \mathbf{D}^{(l-1)}, \mathbf{G}^{(l)} - \mathbf{G}^{(l-1)} \rangle}, 0 \right\} - \theta \frac{\langle \mathbf{K} \mathbf{G}^{(l)}, \mathbf{W}^{(l)} - \mathbf{W}^{(l-1)} \rangle}{\langle \mathbf{K} \mathbf{D}^{(l-1)}, \mathbf{G}^{(l)} - \mathbf{G}^{(l-1)} \rangle}.$$

A. Optimum step size α

As in Section III-B, the step size α is determined so as to minimize the cost:

$$\alpha = \arg \min_{\alpha} J(\{f_c + \alpha d_c\}_{c=1,\dots,C-1}).$$

Let $\mathbf{P} = [f_c(\mathbf{x}_i)]_{c=1,\dots,C-1}^{i=1,\dots,N} = \mathbf{W}^\top \mathbf{K} \in \mathbb{R}^{C-1 \times N}$, $\mathbf{Q} = [d_c(\mathbf{x}_i)]_{c=1,\dots,C-1}^{i=1,\dots,N} = \mathbf{D}^\top \mathbf{K} \in \mathbb{R}^{C-1 \times N}$ and $\hat{\mathbf{Y}} = [\hat{y}_i = \sigma(\mathbf{p}_i + \alpha \mathbf{q}_i)]_{i=1,\dots,N} \in \mathbb{R}^{C \times N}$, and the gradient and Hessian of J with respect to α are written by

$$\begin{aligned} J(\{f_c + \alpha d_c\}_{c=1,\dots,C-1}) &= \frac{\lambda}{2} (\alpha^2 \langle \mathbf{Q}^\top, \mathbf{D} \rangle + 2\alpha \langle \mathbf{Q}^\top, \mathbf{W} \rangle + \langle \mathbf{P}^\top, \mathbf{W} \rangle) - \sum_i^N \sum_c^C y_{ic} \log \hat{y}_{ic}, \\ \frac{dJ}{d\alpha} &= \lambda \{ \alpha \langle \mathbf{Q}^\top, \mathbf{W} \rangle + \langle \mathbf{Q}^\top, \mathbf{D} \rangle \} + \langle \mathbf{Q}, [\hat{\mathbf{Y}} - \mathbf{Y}]_{1:C-1} \rangle \triangleq g(\alpha), \\ \frac{d^2J}{d\alpha^2} &= \lambda \langle \mathbf{Q}^\top, \mathbf{D} \rangle + \sum_i^N \sum_c^{C-1} \hat{y}_{ic} q_{ic} \left(q_{ic} - \sum_k^{C-1} \hat{y}_{ik} q_{ik} \right) \triangleq h(\alpha). \end{aligned}$$

The step size α is optimized by Newton-Raphson in (7).

The overall algorithm is shown in Algorithm 2. Although

Algorithm 2 : Kernel Logistic Regression by non-linear CG

Input: $\mathbf{K} \in \mathbb{R}^{N \times N}$, $\mathbf{Y} = [\mathbf{y}_i]_{i=1, \dots, N} \in \{0, 1\}^{C \times N}$
1: **Initialize** $\mathbf{W}^{(0)} = \mathbf{0} \in \mathbb{R}^{N \times C-1}$, $\hat{\mathbf{Y}} = [\frac{1}{C}] \in \mathbb{R}^{C \times N}$
 $\mathbf{G}^{(0)} = [\hat{\mathbf{Y}} - \mathbf{Y}]_{1:C-1}^T \in \mathbb{R}^{N \times C-1}$,
 $\mathbf{D}^{(0)} = -\mathbf{G}^{(0)} \in \mathbb{R}^{N \times C-1}$
 $\mathbf{P} = \mathbf{W}^{(0)T} \mathbf{K} = \mathbf{0} \in \mathbb{R}^{C-1 \times N}$,
 $\mathbf{Q} = \mathbf{D}^{(0)T} \mathbf{K} \in \mathbb{R}^{C-1 \times N}$, $l = 1$,
2: **repeat**
3: $\alpha = \arg \min_{\alpha} J(\{f_c^{(l-1)} + \alpha d_c^{(l-1)}\}_{c=1, \dots, C-1})$: see Section IV-A
4: $\mathbf{W}^{(l)} = \mathbf{W}^{(l-1)} + \alpha \mathbf{D}^{(l-1)}$, $\mathbf{P} \leftarrow \mathbf{P} + \alpha \mathbf{Q}$
5: $\hat{\mathbf{Y}} = [\hat{y}_i = \sigma(\mathbf{p}_i)]_{i=1, \dots, N}$,
 $J^{(l)} = J(\{f_c^{(l)}\}_{c=1, \dots, C-1}) = \frac{\lambda}{2} \langle \mathbf{P}, \mathbf{W}^{(l)} \rangle - \sum_i^N \sum_c^C y_{ic} \log \hat{y}_{ic}$
6: $\mathbf{G}^{(l)} = [\hat{\mathbf{Y}}^{(l)} - \mathbf{Y}]_{1:C-1}^T$
7: $\mathbf{R} = \mathbf{G}^{(l)T} \mathbf{K}$
8: $\beta = \max \left\{ \frac{\langle \mathbf{R}^T, \mathbf{G}^{(l)} - \mathbf{G}^{(l-1)} \rangle}{\langle \mathbf{Q}^T, \mathbf{G}^{(l)} - \mathbf{G}^{(l-1)} \rangle}, 0 \right\} - \theta \frac{\langle \mathbf{R}^T, \mathbf{W}^{(l)} - \mathbf{W}^{(l-1)} \rangle}{\langle \mathbf{Q}^T, \mathbf{G}^{(l)} - \mathbf{G}^{(l-1)} \rangle}$
9: $\mathbf{D}^{(l)} = -\mathbf{G}^{(l)} + \beta \mathbf{D}^{(l-1)}$
10: $\mathbf{Q} \leftarrow -\mathbf{R} + \beta \mathbf{Q}$, $l \leftarrow l + 1$
11: **until** convergence
Output: $f_c = \sum_i^N w_{ci}^{(l)} k(\mathbf{x}_i, \cdot)$

only the coefficients are updated, the proposed method directly optimizes the classifier f_c itself by minimizing J with respect to f_c . In that point, the method differs from the ordinary optimization in kernel logistic regression.

V. MULTIPLE KERNEL LOGISTIC REGRESSION

In recent years, such a method that integrates different kernel functions with the optimized weights for a novel kernel has attracted keen attentions, which is called multiple kernel learning (MKL). By combining multiple types of kernels, the heterogeneous information, which is complementary to each other, can be effectively incorporated to improve the performance. The MKL has been mainly addressed in the framework of large margin classifiers [11]. In this section, we formulate MKL in the proposed scheme of kernel logistic regression described in Section IV.

For MKL, we first consider combined RKHS as in [29]. Suppose we have M types of kernel functions, k_1, \dots, k_M , and corresponding RKHS's $\mathcal{H}_1, \dots, \mathcal{H}_M$ each of which is endowed with an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_m}$. We further introduce the slightly modified Hilbert space \mathcal{H}'_m in which the following inner product with a scalar value $v_m \geq 0$ is embedded:

$$\mathcal{H}'_m = \left\{ f | f \in \mathcal{H}_m, \frac{\|f\|_{\mathcal{H}_m}}{v_m} < \infty \right\}, \langle f, g \rangle_{\mathcal{H}'_m} = \frac{\langle f, g \rangle_{\mathcal{H}_m}}{v_m}.$$

This Hilbert space \mathcal{H}'_m is a RKHS with the kernel $k'_m(\mathbf{x}, \cdot) = v_m k_m(\mathbf{x}, \cdot)$ since

$$f(\mathbf{x}) = \frac{\langle f(\cdot), v_m k_m(\mathbf{x}, \cdot) \rangle_m}{v_m} = \langle f(\cdot), v_m k_m(\mathbf{x}, \cdot) \rangle_{\mathcal{H}'_m}.$$

Finally, we define the RKHS $\bar{\mathcal{H}}$ as direct sum of \mathcal{H}'_m : $\bar{\mathcal{H}} = \bigoplus_m^M \mathcal{H}'_m$, in which the associated kernel function is given by

$$\bar{k}(\mathbf{x}, \cdot) = \sum_m^M k'_m(\mathbf{x}, \cdot) = \sum_m^M v_m k_m(\mathbf{x}, \cdot).$$

Based on the $\bar{\mathcal{H}}$, we estimate the class posterior probabilities as

$$\hat{\mathbf{y}} = \sigma([\bar{f}_c(\mathbf{x})]_{c=1, \dots, C-1}),$$

where $\bar{f}_c \in \bar{\mathcal{H}}$ is the classifier function in the combined RKHS. We formulate the multiple-kernel logistic regression (MKLR)

in

$$J(\{\bar{f}_c \in \bar{\mathcal{H}}\}_{c=1, \dots, C-1}, \mathbf{v}) \quad (14)$$

$$= \frac{\lambda}{2} \sum_c^{C-1} \|\bar{f}_c\|_{\bar{\mathcal{H}}}^2 - \sum_i^N \sum_c^C y_{ic} \log [\sigma_c([\bar{f}_c(\mathbf{x}_i)]_{c=1, \dots, C-1})] \rightarrow \min_{\{\bar{f}_c\}, \mathbf{v}}$$

$$\Leftrightarrow J(\{f_{mc} \in \mathcal{H}_m\}_{m=1, \dots, M, c=1, \dots, C-1}, \mathbf{v}) \quad (15)$$

$$= \frac{\lambda}{2} \sum_m^M \frac{1}{v_m} \sum_c^{C-1} \|f_{mc}\|_{\mathcal{H}_m}^2 - \sum_i^N \sum_c^C y_{ic} \log \left[\sigma_c \left(\left[\sum_m^M f_{mc}(\mathbf{x}_i) \right]_{c=1, \dots, C-1} \right) \right] \rightarrow \min_{\{f_{mc}\}, \mathbf{v}}$$

$$s.t., \sum_m^M v_m = 1, v_m \geq 0, \forall m$$

where $\bar{f}_c(\mathbf{x}) = \sum_m^M f_{mc}(\mathbf{x})$ and f_{mc} belongs to each RKHS \mathcal{H}_m . The derivative of the cost J in (15) with respect to f_{mc} is given by

$$\frac{\partial J}{\partial f_{mc}} = \frac{\lambda}{v_m} f_{mc} + \sum_i^N (\hat{y}_{ic} - y_{ic}) k_m(\mathbf{x}_i, \cdot),$$

where $\hat{y}_i = \sigma \left(\left[\sum_m^M f_{mc}(\mathbf{x}_i) \right]_{c=1, \dots, C-1} \right)$ and we use $f_{mc}(\mathbf{x}) = \langle f_{mc}(\cdot), k_m(\mathbf{x}, \cdot) \rangle_{\mathcal{H}_m}$. At the optimum $\frac{\partial J}{\partial f_{mc}} = 0$, the classifier eventually takes the following form;

$$\bar{f}_c = \sum_m^M f_{mc} = \frac{1}{\lambda} \sum_i^N (y_{ic} - \hat{y}_{ic}) \sum_m^M v_m k_m(\mathbf{x}_i, \cdot).$$

Multiple kernels are linearly combined with the weight v_m . Thus, the above-defined MKLR enables us to effectively combine multiple kernels, which can be regarded as multiple kernel learning (MKL).

The regularization term in the costs (14) and (15) is an upper bound of the mixed norm as follows.

$$\frac{\lambda}{2} \sum_c^{C-1} \|\bar{f}_c\|_{\bar{\mathcal{H}}}^2 = \frac{\lambda}{2} \sum_m^M \frac{1}{v_m} \sum_c^C \|f_{mc}\|_{\mathcal{H}_m}^2 \geq \frac{\lambda}{2} \left(\sum_m^M \sqrt{\sum_c^C \|f_{mc}\|_{\mathcal{H}_m}^2} \right)^2$$

where the equality holds for $v_m = \frac{\sqrt{\sum_c^C \|f_{mc}\|_{\mathcal{H}_m}^2}}{\sum_m^M \sqrt{\sum_c^C \|f_{mc}\|_{\mathcal{H}_m}^2}}$. The right-hand-side is similar to group LASSO, and such regularization induces sparseness on the multiple kernels [30]; namely, we can obtain the sparse kernel weights in MKLR. It is noteworthy that the proposed MKLR in (14) and (15) is a convex optimization problem since the regularization term as well as the second term are convex (ref. Appendix in [29]).

We alternately minimize the objective cost (14) with respect to two variables $\{\bar{f}_c\}_{c=1, \dots, C-1}$ and $\mathbf{v} = [v_m]_{m=1, \dots, M}$.

A. Optimization for \bar{f}

The gradients of the cost J in (14) with respect to \bar{f}_c in the RKHS $\bar{\mathcal{H}}$ is given by

$$\frac{\partial J}{\partial \bar{f}_c} = \lambda \bar{f}_c(\cdot) + \sum_i^N \{\hat{y}_{ic} - y_{ic}\} \bar{k}(\mathbf{x}_i, \cdot),$$

where we use $\bar{f}_c(\mathbf{x}) = \langle \bar{f}_c(\cdot), \bar{k}(\mathbf{x}, \cdot) \rangle_{\bar{\mathcal{H}}}$. This is the same form as in the kernel logistic regression in (8) by replacing kernel function $k(\mathbf{x}, \cdot) \mapsto \bar{k}(\mathbf{x}, \cdot)$ and $\mathbf{K} \mapsto \bar{\mathbf{K}} = \sum_m^M v_m \mathbf{K}^{[m]}$ where $\mathbf{K}^{[m]} \in \mathbb{R}^{N \times N}$ is the Gram matrix of the m -th type of kernel k_m . Therefore, the optimization procedure described

in Section IV is also applicable to this optimization. The classifiers \bar{f}_c and the conjugate gradients d_c are represented by linear combinations of kernel functions $k(\mathbf{x}_i, \cdot)$;

$$\bar{f}_c^{(l)} = \sum_i^N w_{ci}^{(l)} \bar{k}(\mathbf{x}_i, \cdot) = \sum_i^N w_{ci}^{(l)} \sum_m^M v_m k_m(\mathbf{x}_i, \cdot), \quad (16)$$

$$d_c^{(l)} = \sum_i^N d_{ci}^{(l)} \bar{k}(\mathbf{x}_i, \cdot),$$

and the update for \bar{f} is performed by

$$\bar{f}_c^{(l)}(\cdot) = \bar{f}_c^{(l-1)}(\cdot) + \alpha d_c^{(l-1)}(\cdot).$$

Note that only the coefficients $\mathbf{W}^{(l)} = [w_{ci}^{(l)}]_{i=1, \dots, N}^{c=1, \dots, C-1}$, $\mathbf{D}^{(l)} = [d_{ci}^{(l)}]_{i=1, \dots, N}^{c=1, \dots, C-1}$ are updated by (11)~(13).

B. Optimization for \mathbf{v}

To update the kernel weights \mathbf{v} , the following cost using the updated \bar{f}_c in (16) is minimized with respect to \mathbf{v} :

$$J(\mathbf{v}) = \frac{\lambda}{2} \sum_m^M v_m \langle \mathbf{P}^{[m]\top}, \mathbf{W}^{(l)} \rangle - \sum_i^N \sum_c^C y_{ic} \log \left\{ \sigma_c \left(\sum_m^M v_m \mathbf{p}_i^{[m]} \right) \right\},$$

where $\mathbf{P}^{[m]} = \mathbf{W}^{(l)\top} \mathbf{K}^{[m]} \in \mathbb{R}^{C-1 \times N}$. The derivative of this cost function with respect to \mathbf{v} is

$$\frac{\partial J}{\partial v_m} = \frac{\lambda}{2} \langle \mathbf{P}^{[m]\top}, \mathbf{W}^{(l)} \rangle + \langle \mathbf{P}^{[m]}, [\hat{\mathbf{Y}} - \mathbf{Y}]_{1:C-1} \rangle,$$

$$s.t., \sum_m^M v_m = 1, v_m \geq 0, \quad (17)$$

where $\hat{\mathbf{Y}} = [\hat{y}_i = \sigma(\sum_m^M v_m \mathbf{W}^{(l)\top} \mathbf{k}_i^{[m]})]_{i=1, \dots, N}$. We apply reduced gradient descent method [31] to minimize the cost while ensuring the constraints (17). The descent direction denoted by \mathbf{e} is computed in a manner similar to [29] as follows.

$$\mu = \arg \max_m \left\{ \sum_c^{C-1} \|f_{mc}\|_{\mathcal{H}_m}^2 = v_m^2 \langle \mathbf{P}^{[m]\top}, \mathbf{W} \rangle \right\},$$

$$e_m = \begin{cases} 0 & (v_m = 0 \wedge \frac{\partial J}{\partial v_m} - \frac{\partial J}{\partial v_\mu} > 0) \\ -\frac{\partial J}{\partial v_m} + \frac{\partial J}{\partial v_\mu} & (v_m > 0 \wedge m \neq \mu) \\ \sum_{v \neq \mu, v_\nu > 0} \frac{\partial J}{\partial v_\nu} - \frac{\partial J}{\partial v_\mu} (m = \mu) \end{cases}.$$

After computing the descent direction \mathbf{e} first, we then check whether the maximal admissible step size (to set a certain component, say v_ν , to 0 in that direction) decreases the objective cost value. In that case, v_ν is updated by setting $v_\nu = 0$ and \mathbf{e} is normalized to meet the equality constraint. By repeating this procedure until the objective cost stops decreasing, we obtain both the modified \mathbf{v}' and the final descent direction \mathbf{e} . Then, the kernel weights are updated by $\mathbf{v}^{new} = \mathbf{v}' + \alpha \mathbf{e}$, where α is the step size.

The optimal step size is also computed in a manner similar to the method in linear logistic regression (Section III-B). Let $\mathbf{P}^{[m]} = \mathbf{W}^{(l)\top} \mathbf{K}^{[m]} \in \mathbb{R}^{C-1 \times N}$, $\mathbf{P} = \sum_m^M v_m' \mathbf{W}^{(l)\top} \mathbf{K}^{[m]} = \sum_m^M v_m' \mathbf{P}^{[m]}$, $\mathbf{Q} = \sum_m^M e_m \mathbf{P}^{[m]}$ and $\hat{\mathbf{Y}} = [\hat{y}_i = \sigma(\mathbf{p}_i + \alpha \mathbf{q}_i)]_{i=1, \dots, N}$, and the step size α is optimized by (7) using the following derivatives,

$$\frac{dJ}{d\alpha} = \frac{\lambda}{2} \langle \mathbf{P}^\top, \mathbf{W}^{(l)} \rangle + \langle \mathbf{Q}, [\hat{\mathbf{Y}} - \mathbf{Y}]_{1:C-1} \rangle \triangleq g(\alpha),$$

$$\frac{d^2 J}{d\alpha^2} = \sum_i^N \sum_c^{C-1} \hat{y}_{ic} q_{ic} \left(q_{ic} - \sum_k^{C-1} \hat{y}_{ik} q_{ik} \right) \triangleq h(\alpha).$$

Algorithm 3 : Multiple Kernel Logistic Regression by non-linear CG

Input: $\mathbf{K}^{[m]} \in \mathbb{R}^{N \times N}$, $m \in \{1, \dots, M\}$,
 $\mathbf{Y} = [y_i]_{i=1, \dots, N} \in \{0, 1\}^{C \times N}$

- 1: **Initialize** $\mathbf{v} = [\frac{1}{M}] \in \mathbb{R}^M$, $\bar{\mathbf{K}} = \sum_m^M v_m \mathbf{K}^{[m]}$,
 $\mathbf{W}^{(0)} = \mathbf{0} \in \mathbb{R}^{N \times C-1}$, $\hat{\mathbf{Y}} = [\frac{1}{C}] \in \mathbb{R}^{C \times N}$,
 $\mathbf{G}^{(0)} = [\hat{\mathbf{Y}} - \mathbf{Y}]_{1:C-1}^\top \in \mathbb{R}^{N \times C-1}$,
 $\mathbf{D}^{(0)} = -\mathbf{G}^{(0)} \in \mathbb{R}^{N \times C-1}$,
 $\mathbf{P} = \mathbf{W}^{(0)\top} \bar{\mathbf{K}} = \mathbf{0} \in \mathbb{R}^{C-1 \times N}$,
 $\mathbf{Q} = \mathbf{D}^{(0)\top} \bar{\mathbf{K}} \in \mathbb{R}^{C-1 \times N}$, $l = 1$
- 2: **repeat**
- 3: $\alpha = \arg \min_{\alpha} J(\{\bar{f}_c^{(l-1)} + \alpha d_c^{(l-1)}\}_{c=1, \dots, C-1})$: see Section V-A
- 4: $\mathbf{W}^{(l)} = \mathbf{W}^{(l-1)} + \alpha \mathbf{D}^{(l-1)}$, $\mathbf{P} \leftarrow \mathbf{P} + \alpha \mathbf{Q}$
- 5: **if** $l \bmod \tau = 0$ **then**
- 6: /* Optimization for \mathbf{v} */
- 7: $\mathbf{P}^{[m]} = \mathbf{W}^{(l)\top} \mathbf{K}^{[m]}$, $\forall m$
- 8: Calculate reduced gradient \mathbf{e} and \mathbf{v}' : see Section V-B
- 9: $\alpha = \arg \min_{\alpha} J(\{\mathbf{f}_{mc} \in \mathcal{H}_m\}_{m=1, \dots, M}^{c=1, \dots, C-1}, \mathbf{v}' + \alpha \mathbf{e})$
- 10: $\mathbf{v} = \mathbf{v}' + \alpha \mathbf{e}$
- 11: $\bar{\mathbf{K}} = \sum_m^M v_m \mathbf{K}^{[m]}$, $\mathbf{P} = \sum_m^M v_m \mathbf{P}^{[m]}$
- 12: $\hat{\mathbf{Y}} = [\hat{y}_i = \sigma(\mathbf{p}_i)]_{i=1, \dots, N}$,
 $J^{(l)} = J(\{\bar{f}_c^{(l)}\}_{c=1, \dots, C-1}, \mathbf{v}) = \frac{\lambda}{2} \langle \mathbf{P}, \mathbf{W}^{(l)} \rangle - \sum_i^N \sum_c^C y_{ic} \log \hat{y}_{ic}$
- 13: $\mathbf{G}^{(l)} = [\hat{\mathbf{Y}} - \mathbf{Y}]_{1:C-1}^\top$, $\mathbf{D}^{(l)} = -\mathbf{G}^{(l)}$
- 14: $\mathbf{Q} = \mathbf{D}^{(l)\top} \bar{\mathbf{K}}$
- 15: **else**
- 16: /* Optimization for \bar{f} */
- 17: $\hat{\mathbf{Y}} = [\hat{y}_i = \sigma(\mathbf{p}_i)]_{i=1, \dots, N}$,
 $J^{(l)} = J(\{\bar{f}_c^{(l)}\}_{c=1, \dots, C-1}, \mathbf{v}) = \frac{\lambda}{2} \langle \mathbf{P}, \mathbf{W}^{(l)} \rangle - \sum_i^N \sum_c^C y_{ic} \log \hat{y}_{ic}$
- 18: $\mathbf{G}^{(l)} = [\hat{\mathbf{Y}} - \mathbf{Y}]_{1:C-1}^\top$
- 19: $\mathbf{R} = \mathbf{G}^{(l)\top} \bar{\mathbf{K}}$
- 20: $\beta = \max \left\{ \frac{\langle \mathbf{R}^\top, \mathbf{G}^{(l)} - \mathbf{G}^{(l-1)} \rangle}{\langle \mathbf{Q}^\top, \mathbf{G}^{(l)} - \mathbf{G}^{(l-1)} \rangle}, 0 \right\} - \theta \frac{\langle \mathbf{R}^\top, \mathbf{W}^{(l)} - \mathbf{W}^{(l-1)} \rangle}{\langle \mathbf{Q}^\top, \mathbf{G}^{(l)} - \mathbf{G}^{(l-1)} \rangle}$
- 21: $\mathbf{D}^{(l)} = -\mathbf{G}^{(l)} + \beta \mathbf{D}^{(l-1)}$
- 22: $\mathbf{Q} \leftarrow -\mathbf{R} + \beta \mathbf{Q}$
- 23: **end if**
- 24: $l \leftarrow l + 1$
- 25: **until** convergence

Output: $\bar{f}_c = \sum_i^N w_{ci}^{(l)} \sum_m^M v_m k_m(\mathbf{x}_i, \cdot)$

The overall algorithm is shown in Algorithm 3. Since the dimensionality of \bar{f} is larger than that of \mathbf{v} , the optimization for \mathbf{v} is performed every τ iterations; we set $\tau = 5$ in this study. It should be noted that the optimizations both of \bar{f} and \mathbf{v} are ensured to monotonically decrease the objective cost J via iterations.

VI. PARALLEL COMPUTING

Although the non-linear CG sequentially minimizes the objective cost in an iterative manner, each step of iteration can be easily parallelized. The computational cost per iteration is dominated by the (large) matrix multiplications: lines 3 and 7 in Algorithm 1, line 7 in Algorithm 2, and lines 7, 14 and 19 in Algorithm 3. Those multiplications are parallelized such as by multi-thread programming especially in GPGPU, which effectively scales up the whole optimization procedure.

VII. EXPERIMENTAL RESULTS

We conducted various experiments on multi-class classification by using linear logistic regression (LR), kernel LR (KLR) and multiple-kernel LR (MKLR). The proposed methods were compared to the other related methods in terms of the classification accuracy and computation time.

TABLE II. DATASETS OF DENSE FEATURES. WE APPLY FIVE-FOLD CROSS VALIDATION ON THE DATASETS MARKED BY *, WHILE USING GIVEN TRAINING/TEST SPLITS ON THE OTHER DATASETS.

Dataset	#class	#feature	#training sample	#test sample
SENSIT-VEHICLE	3	100	78,823	19,705
SEMEION*	10	256	1,275	318
ISOLET	26	617	6,238	1,559
MNIST	10	784	60,000	10,000
P53*	2	5,408	13,274	3,318

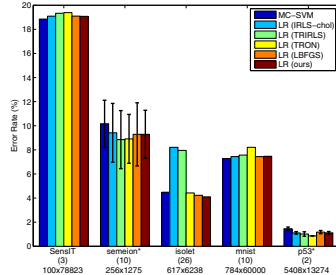


Fig. 1. Error rates on linear classification for dense features. The numbers of classes are indicated in parentheses and the sizes of X (#feature \times #sample) are shown in the bottom.

A. Linear classification

As a preliminary experiment to the subsequent kernel-based methods, we applied linear classification methods.

For comparison, we applied multi-class support vector machine (MC-SVM) [7] and for LR, four types of optimization methods other than the proposed method in Section III:

- IRLS with Cholesky decomposition (IRLS-chol) [17]
- IRLS with CG (TRIRLS) [18]
- IRLS with trust region newton method (TRON) [20]
- limited memory BFGS method (LBFGS) [13] and [21].

All of these methods introduce regularization with respect to classifier norm in a similar form to (3), of which the regularization parameter is determined by three-fold cross validation on training samples ($\lambda \in \{1, 10^{-2}, 10^{-4}\}$). We implemented all the methods by using MATLAB with C-mex on Xeon 3GHz (12 threading) PC; we used LIBLINEAR [32] for MC-SVM and TRON, and the code¹ provided by Liu and Nocedal [22] for LBFGS.

We first used the datasets² of the dense feature vectors, the details of which are shown in Table II. For evaluation, we used the given training/test splits on some datasets and applied five-fold cross validation on the others. The classification performances (error rates) and the computation times for training the classifier are shown in Fig. 1 and Fig. 2, respectively. The computation times are measured in two ways; Fig. 2(a) shows the computation time only for learning the final classifier and Fig. 2(b) is for ‘whole’ training process including both the final learning and three-fold cross validations to determine the regularization parameter. The proposed method is favorably compared to the other methods in terms of error rates and computation time; the method of LR with IRLS-chol which is quite close to the ordinary IRLS requires more training time.

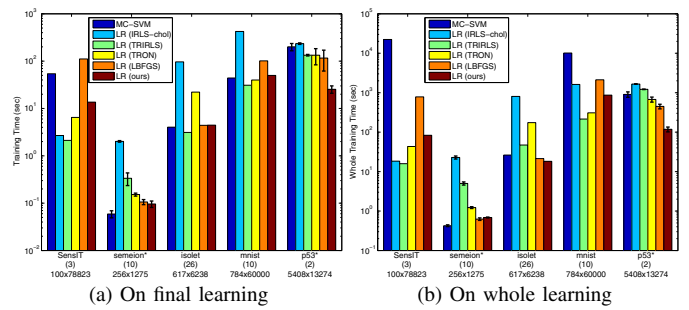


Fig. 2. Computation times (\log -scale) on linear classification for dense features. The computation time for learning final classifier is shown in (a), while that for whole training including 3-CV to determine λ is in (b).

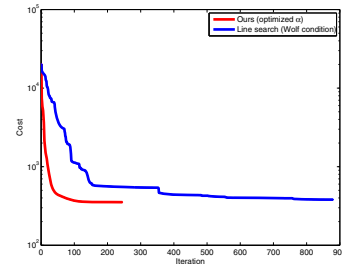


Fig. 3. Comparison to the method using an exhaustive line search. The plot shows the objective cost values through iterations on ISOLET.

We then investigated the effectiveness of the optimized step size α (Section III-B) which is one of our contributions in this paper. Fig. 3 shows how the proposed optimization method works, compared to that using an exhaustive line search. By employing the optimized step size, the objective cost drastically decreases in the first few steps and reaches convergence in a smaller number of iterations.

In the same experimental protocol, we also applied the methods to datasets which contain sparse feature vectors. The details of the datasets³ are shown in Table III. Note that the method of LR with IRLS-chol can not deal with such a huge feature vectors since the Hessian matrix is quite large, making it difficult to solve linear equations by Cholesky decomposition in a realistic time. As shown in Fig. 4 and Fig. 5, the computation times of the methods are all comparable (around 10 seconds) with similar classification accuracies.

Though the performances of the proposed method are favorably compared to the others as a whole, they are different from those of IRLS-based methods (TRIRLS and TRON). The reason is as follows. The objective costs of those methods⁴ are shown in Table IV. The proposed method produces lower objective costs than those by TRIRLS, and thus we can say that the IRLS-based method does not fully converge to global minimum. Although the objective cost function is convex, there would exist plateau [38] which stop the optimization in the IRLS-based methods before converging to the global minimum. Thus, from the viewpoint of optimization, the proposed method produces favorable results.

³REUTERS21578 (UCI KDD Archive) and TDT2 (Nist Topic Detection and Tracking corpus) are downloaded from <http://www.zjucadcg.cn/dengcai/Data/TextData.html>, and RCv1 [35], SECTOR [36] and NEWS20 [37] are from <http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/>.

⁴We do not show the cost of TRON [20] whose formulation is slightly different as described in Section II-A.

¹The code is available at <http://www.ece.northwestern.edu/~nocedal>.
²SEMEION, ISOLET and P53 are downloaded from UCI-repository <http://archive.ics.uci.edu/ml/datasets.html>, and SENSIT-VEHICLE [33] and MNIST [34] are from <http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/>.

TABLE III. DATASETS OF SPARSE FEATURES. WE APPLY FIVE-FOLD CROSS VALIDATION ON THE DATASETS MARKED BY *, WHILE USING GIVEN TRAINING/TEST SPLITS ON THE OTHER DATASETS.

Dataset	#class	#feature	#training sample	#non zeros	#test sample
REUTERS21578	51	18,933	5,926	283,531	2,334
TDT2*	77	36,771	8,140	1,056,166	2,035
RCV1	51	47,236	15,564	1,028,284	518,571
SECTOR	105	55,197	6,412	1,045,412	3,207
NEWS20	20	62,060	15,935	1,272,568	3,993

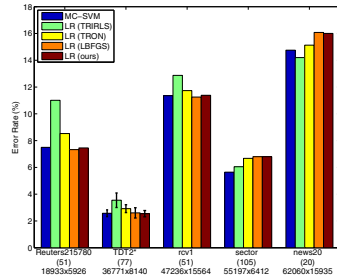


Fig. 4. Error rates on linear classification for sparse features.

B. Kernel-based classification

Next, we conducted the experiments on kernel-based classifications. We applied the proposed kernel logistic regression (KLR) in Section IV and the kernelized methods of the above-mentioned linear classifiers;

- multi-class kernel support vector machine (MC-KSVM) [7]
- KLR using IRLS with CG (TRIRLS) [18]
- KLR using IRLS with trust region newton method (TRON) by [20]
- KLR using limited memory BFGS method (LBFGS) [13], [21].

Note that the KLR methods of TRIRLS, TRON and LBFGS are kernelized in the way described in Section II-D. Table V shows the details of the datasets⁵ that we use, and in this experiment, we employed RBF kernel $k(x, \xi) = \exp(-\frac{\|x-\xi\|^2}{2\sigma^2})$ where σ^2 is determined as the sample variance. The experimental protocol is the same as in Section VII-A.

As shown in Fig. 6, the classification performances of the proposed method are superior to the other KLR methods and are comparable to MC-KSVM, while the computation times of the proposed method are faster than that of MC-KSVM on most datasets (Fig. 7). As discussed in Section VI, we can employ GPGPU (NVIDIA Tesla C2050) to efficiently compute the matrix multiplications in our method on the datasets except for the huge dataset of SHUTTLE, and the computation time is significantly reduced as shown in Fig. 7.

While the proposed method optimizes the classifier in RKHS, the optimization in the other KLR methods is performed in the subspace spanned by the sample kernel functions (Section IV), possibly causing numerically unfavorable issues such as plateau [38], and the optimizations would terminate before fully converging to the global minimum. The objective costs shown in Table VI illustrates it; the proposed method provides lower costs than those of the other KLR methods. In addition, the obtained classifiers, i.e., coefficients W for

⁵USPS [39], LETTER (Statlog), PROTEIN [40] and SHUTTLE (Statlog) are downloaded from <http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/>, and POKER is from UCI repository <http://archive.ics.uci.edu/ml/datasets/>.

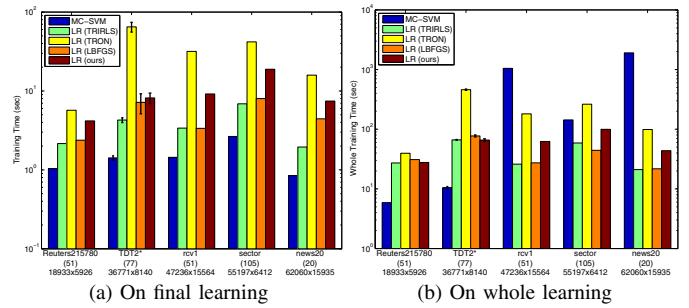


Fig. 5. Computation times on linear classification for sparse features.

TABLE IV. OBJECTIVE COST VALUES OF LR METHODS WITH $\lambda = 10^{-2}$ ON SPARSE DATASETS.

Dataset	Ours	TRIRLS	LBFGS
REUTERS21578	9.98	389.26	10.32
TDT2	12.13	387.58	13.65
RCV1	906.49	15687.17	969.07
SECTOR	1102.08	29841.19	1167.35
NEWS20	1949.60	7139.07	2000.64

samples, are shown in Fig. 8. The proposed method produces near sparse weights compared to those of the other methods and contribute to improve the performance similarly to MC-KSVM, even though any constraints to enhance sparseness are not imposed in the proposed method.

C. Multiple-kernel learning

Finally, we conducted the experiment on multiple-kernel learning. We applied the proposed multiple-kernel logistic regression (MKLR) described in Section V and simpleMKL [29] for comparison. For simpleMKL, we used the code⁶ provided by the author with LIBSVM [41]. The details of the datasets⁷ are shown in Table VII; for multi-class classification, in the dataset of PASCAL-VOC2007, we removed the samples to which multiple labels are assigned. In the datasets of PSORT-, NONPLANT and PASCAL-VOC2007, we used the precomputed kernel matrices provided in the authors' web sites. The dataset of PEN-DIGITS contains four types of feature vectors and correspondingly we constructed four types of RBF kernel in the same way as in Section VII-B.

The classification performances are shown in Fig. 9. As a reference, we also show the performances of KLR with the (single) averaged kernel matrix and the (single) best kernel matrix which produces the best performance among the multiple kernel matrices. The MKL methods produce superior performances compared to those of KLR with single kernel, and the proposed method is comparable to simpleMKL. The obtained kernel weights are also shown in Fig. 10. The weights by the proposed method are sparse similarly to those by simpleMKL, due to the formulation based on the combined RKHS \mathcal{H} in (14) and its efficient optimization using non-linear CG.

As shown in Fig. 11, the computation time of the proposed method is significantly ($10^2 \sim 10^4$ times) faster than

⁶The code is available at <http://asi.insa-rouen.fr/enseignants/~arakotom/code/mkindex.html>

⁷PASCAL-VOC2007 [42] is downloaded from <http://lear.inrialpes.fr/people/guillaumin/data.php>, PEN-DIGITS [43] is from <http://mkl.ucsd.edu/dataset/pendigits>, and PSORT-, NONPLANT [44] are from <http://www.fml.tuebingen.mpg.de/raetsch/suppl/protsubloc>.

TABLE V. DATASETS FOR KERNEL-BASED CLASSIFICATION.

Dataset	#class	#feature	#training sample	#test sample
USPS	10	256	7,291	2,007
LETTER	26	16	15,000	5,000
PROTEIN	3	357	17,766	6,621
POKER	10	10	25,010	1,000,000
SHUTTLE	7	9	43,500	14,500

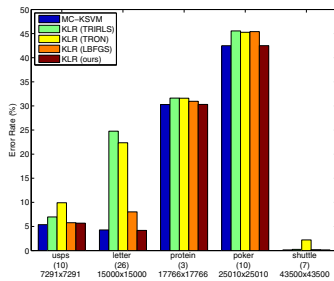


Fig. 6. Error rates on kernel-based classification.

that of simpleMKL. Thus, as is the case with kernel-based classification (Section VII-B), we can say that the proposed method produces comparable performances to simpleMKL with a significantly faster training time.

VIII. CONCLUDING REMARKS

In this paper, we have proposed an efficient optimization method using non-linear conjugate gradient (CG) descent for logistic regression. The proposed method efficiently minimizes the cost through CG iterations by using the optimized step size without an exhaustive line search. On the basis of the non-linear CG based optimization scheme, a novel optimization method for kernel logistic regression (KLR) is also proposed. Unlike the ordinary KLR methods, the proposed method naturally formulates the classifier as the linear combination of sample kernel functions and directly optimizes the kernel-based classifier in the reproducing kernel Hilbert space, not the linear coefficients for the samples. Thus, the optimization effectively performs while possibly avoiding the numerical issues such as plateau. We have further developed the KLR using single kernel to multiple-kernel LR (MKLR). The proposed MKLR, which is also optimized in the scheme of non-linear CG, produces the kernel-based classifier with optimized weights for multiple kernels. In the experiments on various multi-class classification tasks, the proposed methods produced favorable results in terms of classification performance and computation time, compared to the other methods.

REFERENCES

- [1] G. Wang, D. Hoiem, and D. Forsyth, "Building text features for object image classification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 1367–1374.
- [2] S. K. Divvala, D. Hoiem, J. H. Hays, A. A. Efros, and M. Hebert, "An empirical study of context in object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 1271–1278.
- [3] I. Laptev, M. Marszaek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.
- [4] A. Genkin, D. D. Lewis, and D. Madigan, "Large-scale bayesian logistic regression for text categorization," *Technometrics*, vol. 49, no. 3, pp. 291–304, 2007.
- [5] P. J. Bartlett, B. Schölkopf, D. Schuurmans, and A. J. Smola, Eds., *Advances in Large-Margin Classifiers*. MIT Press, 2000.

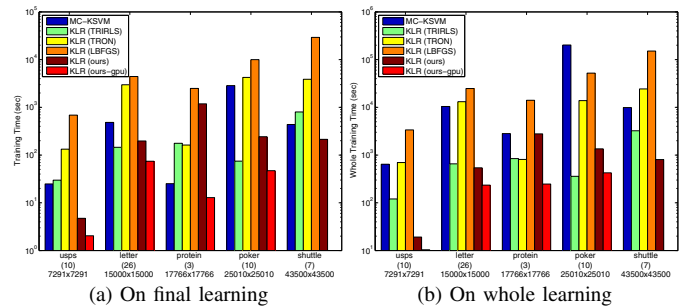


Fig. 7. Computation times on kernel-based classification.

TABLE VI. OBJECTIVE COST VALUES OF KLR METHODS WITH $\lambda = 10^{-2}$ ON KERNEL DATASETS.

Dataset	Ours	TRIRLS	LBFGS
USPS	446.37	914.88	501.15
LETTER	4746.13	12476.41	5789.30
PROTEIN	5866.16	12576.97	10650.96
POKER	22186.19	30168.74	23345.94
SHUTTLE	759.99	1100.07	811.91

- [6] V. Vapnik, *Statistical Learning Theory*. Wiley, 1998.
- [7] K. Crammer and Y. Singer, "On the algorithmic implementation of multiclass kernel-based vector machines," *Journal of Machine Learning Research*, vol. 2, pp. 265–292, 2001.
- [8] J. Platt, "Fast training of support vector machines using sequential minimal optimization," in *Advances in Kernel Methods - Support Vector Learning*, B. Schölkopf, C. Burges, and A. Smola, Eds. Cambridge, MA, USA: MIT Press, 1999, pp. 185–208.
- [9] J. Nocedal and S. Wright, *Numerical Optimization*. Springer, 1999.
- [10] W. W. Hager and H. Zhang, "A survey of nonlinear conjugate gradient methods," *Pacific Journal of Optimization*, vol. 2, pp. 35–58, 2006.
- [11] G. Lanckriet, N. Cristianini, P. Bartlett, L. E. Ghaoui, and M. I. Jordan, "Learning the kernel matrix with semidefinite programming," *Journal of Machine Learning Research*, vol. 5, pp. 27–72, 2004.
- [12] K. Watanabe, T. Kobayashi, and N. Otsu, "Efficient optimization of logistic regression by direct cg method," in *International Conference on Machine Learning and Applications*, 2011.
- [13] R. Malouf, "A comparison of algorithms for maximum entropy parameter estimation," in *The Sixth Conference on Natural Language Learning*, 2002, pp. 49–55.
- [14] C. Sutton and A. McCallum, *An introduction to conditional random fields for relational learning*, L. Getoor and B. Taskar, Eds. MIT Press, 2006.
- [15] T. Minka, "A comparison of numerical optimizers for logistic regression," Carnegie Mellon University, Technical report, 2003.
- [16] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2007.
- [17] P. Komarek and A. Moore, "Fast robust logistic regression for large sparse datasets with binary outputs," in *The 9th International Workshop on Artificial Intelligence and Statistics*, 2003, pp. 3–6.
- [18] —, "Making logistic regression a core data mining tool," in *International Conference on Data Mining*, 2005, pp. 685–688.
- [19] M. R. Hestenes and E. L. Stiefel, "Methods of conjugate gradients for solving linear systems," *Journal of Research of the National Bureau of Standards*, vol. 49, no. 6, pp. 409–436, 1952.
- [20] C.-J. Lin, R. Weng, and S. Keerthi, "Trust region newton methods for large-scale logistic regression," in *International Conference on Machine Learning*, 2007, pp. 561–568.
- [21] H. Daumé III, "Notes on CG and LM-BFGS optimization of logistic regression," Technical report, 2004. [Online]. Available: <http://www.umiacs.umd.edu/~hal/docs/daume04cg-bfgs.pdf>
- [22] D. C. Liu and J. Nocedal, "On the limited memory bfgs method for large scale optimization," *Mathematical Programming*, vol. 45, pp. 503–528, 1989.
- [23] S. D. Pietra, V. D. Pietra, and J. Lafferty, "Inducing features of

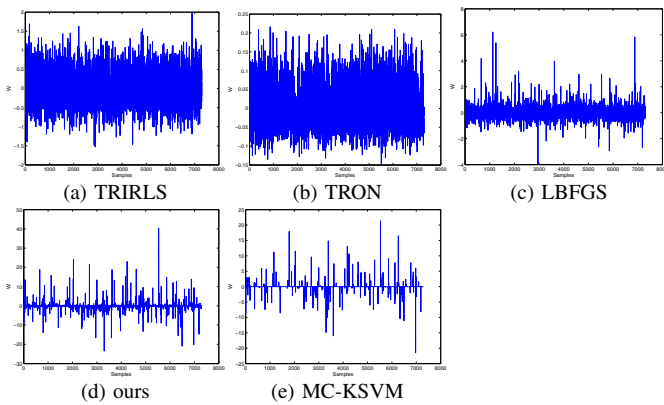


Fig. 8. Classifiers (coefficients w_1 across samples) of class 1 on USPS.

TABLE VII. DATASETS FOR MULTIPLE-KERNEL LEARNING. WE APPLY FIVE-FOLD CROSS VALIDATION ON THE DATASETS MARKED BY *, WHILE USING GIVEN TRAINING/TEST SPLITS ON THE OTHER DATASETS.

Dataset	#class	#kernel	#training sample	#test sample
PSORT*	5	69	1,155	289
NONPLANT*	3	69	2,186	546
PASCAL-VOC2007	20	15	2,954	3,192
PEN-DIGITS	10	4	7,494	3,498

random fields,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 4, pp. 380–393, 1997.

[24] J. Zhu and T. Hastie, “Kernel logistic regression and the import vector machine,” *Journal of Computational and Graphical Statistics*, vol. 14, no. 1, pp. 185–205, 2005.

[25] G. Wahba, C. Gu, Y. Wang, and R. Chappell, “Soft classification, a.k.a. risk estimation, via penalized log likelihood and smoothing spline analysis of variance,” in *The Mathematics of Generalization*, D. Wolpert, Ed. Reading, MA, USA: Addison-Wesley, 1995, pp. 329–360.

[26] T. Hastie and R. Tibshirani, *Generalized Additive Models*. Chapman and Hall, 1990.

[27] B. Schölkopf and A. Smola, *Learning with Kernels*. MIT Press, 2001.

[28] Y. Dai and L. Liao, “New conjugacy conditions and related nonlinear conjugate gradient methods,” *Applied Mathematics and Optimization*, vol. 43, pp. 87–101, 2001.

[29] A. Rakotomamonjy, F. R. Bach, S. Canu, and Y. Grandvalet, “Simplemkl,” *Journal of Machine Learning Research*, vol. 9, pp. 2491–2521, 2008.

[30] F. Bach, “Consistency of the group lasso and multiple kernel learning,” *Journal of Machine Learning Research*, vol. 9, pp. 1179–1225, 2008.

[31] J. Bonnans, J. Gilbert, C. Lemaréchal, and C. Sagastizábal, *Numerical Optimization: Theoretical and Practical Aspects*. Springer, 2006.

[32] R. Fan, K. Chang, C. Hsieh, X. Wang, and C. Lin, “Liblinear: A library for large linear classification,” *Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008, Software available at <http://www.csie.ntu.edu.tw/~cjlin/liblinear>.

[33] M. Duarte and Y. H. Hu, “Vehicle classification in distributed sensor networks,” *Journal of Parallel and Distributed Computing*, vol. 64, no. 7, pp. 826–838, 2004.

[34] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[35] D. D. Lewis, Y. Yang, T. G. Rose, and F. Li, “Rcv1: A new benchmark collection for text categorization research,” *Journal of Machine Learning Research*, vol. 5, pp. 361–397, 2004.

[36] A. McCallum and K. Nigam, “A comparison of event models for naive bayes text classification,” in *AAAI’98 Workshop on Learning for Text categorization*, 1998.

[37] K. Lang, “Newsweeder: Learning to filter netnews,” in *International Conference on Machine Learning*, 1995, pp. 331–339.

[38] K. Fukumizu and S. Amari, “Local minima and plateaus in hierarchical

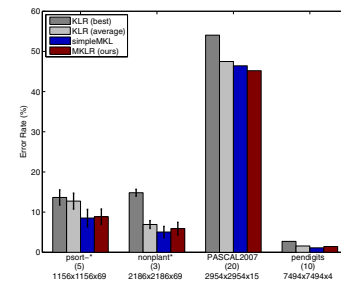


Fig. 9. Error rates and computation times on multiple-kernel learning.

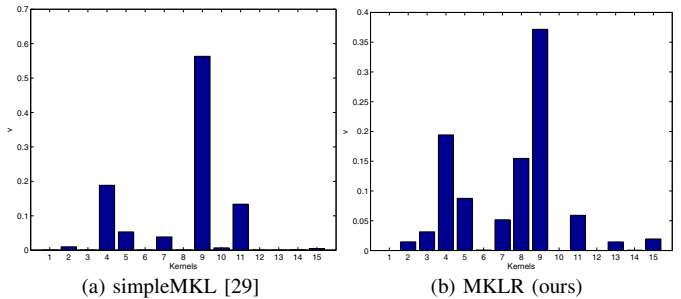


Fig. 10. The obtained kernel weights v on PASCAL-VOC2007.

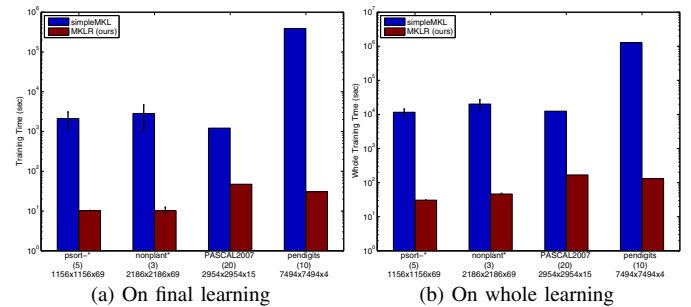


Fig. 11. Computation times on multiple-kernel learning.

structures of multilayer perceptrons,” *Neural Networks*, vol. 13, no. 3, pp. 317–327, 2000.

[39] J. Hull, “A database for handwritten text recognition research,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no. 5, pp. 550–554, 1994.

[40] J.-Y. Wang, “Application of support vector machines in bioinformatics,” Master’s thesis, Department of Computer Science and Information Engineering, National Taiwan University, 2002.

[41] C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machines*, 2001, software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.

[42] M. Guillaumin, J. Verbeek, and C. Schmid, “Multimodal semi-supervised learning for image classification,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 902–909.

[43] F. Alimoglu and E. Alpaydin, “Combining multiple representations and classifiers for pen-based handwritten digit recognition,” in *International Conference on Document Analysis and Recognition*, 1997, pp. 637–640.

[44] A. Zien and C. S. Ong, “An automated combination of kernels for predicting protein subcellular localization,” in *Proceedings of the 8th Workshop on Algorithms in Bioinformatics*, 2008, pp. 179–186.

Study of Current Femto-Satellite Approches and Services

Nizar Tahri, Chafaa Hamrouni, Adel M. Alimi
REGIM-Lab: Research Groups on Intelligent Machines
Department of Electrical Engineering, National Engineering School of Sfax (ENIS)
BP1173, Sfax 3038, Tunisia
nizar.isi@laposte.net, {chafaa.hamrouni, adel.alimi@ieeee.org}

Abstract— The success of space technology evolves according to the technological progression in terms of density of CMOS integration (Complementary on - Silicon Metal) and MEMS (Micro-Electro-Mechanical System) [4]. The need of spatial services has been a significant growth due to several factors such as population increases, telecommunication applications, climate changes, earth control and observation military goals, and so on. To cover this, spatial vehicle generations, specific calculators and Femto-cell systems have been developed. More recently, Ultra - Small Satellites (USS) have been proposed and different approaches, concerning developing of these kind of spatial systems, have been presented in literature. This miniature satellite is capable to fly in the space and to provide different services such as imagery, measures and communications [4, 9, 10]. This paper deals with the study of two different USS Femto-satellite architectures that exist in literature in order to propose a future architecture that can provide an optimization of power supply consumption and ameliorate service communication quality.

Keywords: Femtosatellite, Communication, Spacecraft

I. INTRODUCTION

In the last few decades, a new generation of space mission architecture design is emerging in USS. It will collectively perform missions; both earth-orbiting and inter-satellite communication, in a distributed fashion. Solutions have been already proposed for optimization of complex distributed space mission architectures. However, to support such architectures, a novel approach, with a high volume production of Femto-Satellite less than 100g satellites-on-a-chip or satellite on board at low cost, is required.

In this paper we present the migration for the USS and we detail a probe of two FemtoSatellite approaches. The first one is named Femto-Satellite-On-Chip, the second one is named Femto-Satellite-On-Board or Femto-Satellite based on Commercial-Of-The-Shelf (COTS). Moreover, a comparison study of these approaches is done to show the difference between them and then to propose a future architecture.

The paper is structured as following. The next section expands migration towards the USS and especially toward the Femto-Satellite. Section 2 presents an evaluation

of ten year-research on USS. Section 3 expands migration towards Femto-Sat-COTS. Section 4 shows comparison between PCBSat and Wiki-Sat from two points of view. Section 5 lists the FemtoSatellite theme researches. Section 6 is devoted to discuss the two architectures comparison results. Last but not least, the paper concludes with roadmap of future researches required to realize a specific FemtoSatellite from Tunisia with COTS approach.

II. MIGRATION TOWARD THE ULTRA-SMALL-SATELLITE GENERATION

Classification of satellites frequently depends on its masses. Actually, we talk about satellites which are inferior to 1 kg and sometimes to some grams, as shown on table 1. PicoSat is, eventually, larger than FemtoSatellite, AttoSatellite and ZeptoSatellite [11]. This passage is justified with the technological evolution in terms of integration density. In the last decades, the world of technology manifested mixing between the conception of micro-electromechanic MEMS approach and the conception of electronic CMOS approach [13], which allows the development of capacitors, processors and systems that are completed on a miniature surface even granular. These miniature systems offer the advantage of reducing the cost of production as well as the cost by prototype [11].

Table 1: Ultra-Small Satellite

	Weight	Price
PicoSatellite	~ 1 kg	~ 10000 \$
FemtoSatellite	~ 0.1Kg	~ 100 \$
AttoSatellite	~ 0.01Kg	Few \$
ZeptoSatellite	~ 0.001 Kg	Few \$

Researches on Pico-Sat are very advanced, but only some of these satellites are functioning. Many of Pico-Sat are missed after their launching and almost 30% shows good results [16]. This result, as it is shown on figure 1, represents the major problem of PicoSat when connecting the number of satellites non-functioning with the cost of development.

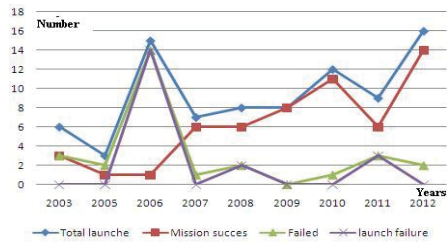


Fig.1: Histogram of cube satellite in the world [16]

Requirements to minimize development cost and time push researches towards satellite generations. In order to offer more chance to the entire world to design their own satellites responding to particular requirements, these generations are developed to be cheaper, more rapid to be on service, simpler to function and with a commercialized technology [1]. Figure 2 displays a comparison according to five parameters including PicoSats and FemtoSats. These parameters offer the advantage of migration from the PicoSat generation to the Femto Sat. We can note that this position of FemtoSat focus on its architecture and its power in terms of services and lifetime.

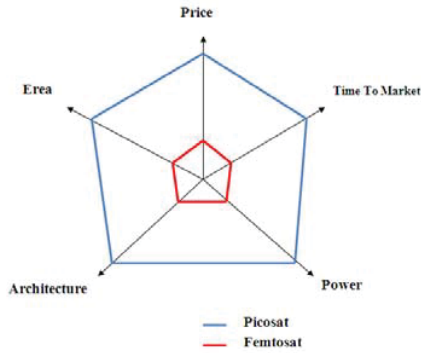


Fig. 2: Comparison between PicoSats and FemtoSats

III. EVALUATION OF TEN YEAR-RESEARCH OF USS

Since the birth of integration technology, with mixing of electric and mechanic systems known as MEMS approach, engineers and researchers of technology space opted a conceive; a satellite according to the miniaturization norms, for instance N-Prize [6]. This perspective began with conception of embedded systems that include all classical satellite subsystems. Meanwhile, the miniaturization has confronted an obstacle which has changed, until now, the axes of the research; it is the insufficiency of energy [1]. In the beginning, researchers migrated toward a satellite on chip by integrating solar energy as a hybrid source. Winsat and Smart Dusts are examples that show this philosophy in Silicone and on chip smaller than fraction of centimeters as shown on figure 3 [4].

The use of Femto solar cells caused problems of integration since the two approaches, CMOS and Solar Cells, aren't compatible [4]. But many studies have shown the feasibility without really having a complete launched system.

	Picture	Reference
PCBSat		[11]
WikiSat		[6]

Fig.3: Two types of FemtoSatellite

This Femto-satellite evolution equally with the technological development of captureurs and commercialized microscopic actuators conception, raised questions on the possibility of designing granular Femto-satellites with discrete commercialized components. Indeed, in 2009, D. Barnhart published the first FemtoSatellite on board [11]. He named this satellite PCBSat according to the didactic norms inherited of his EsaySat ancestor of the Picosat family. This satellite has been viewed as a map on board of some centimeters which integrates the based subsystems of a classic satellite as shown on table 2.

In 2011, J. Tristancho published the first WikiSat generation fruit of a spatial-aerospace competition called N-Prize [6]. This competition has presented particular specifications for the FemtoSat future that are summarized by size inferior to 20 gr, use of COTS components and low conception cost. WikiSat is, actually, referenced as a landmark not only to study the exploration of FemtoSat in certain applications but also to testify certain updated components.

IV. USS-ON CHIP MIGRATION TOWARD FEMTO SATELLITE-ON-COST

In [12], D. Barnhart studied FemtoSatellites. He started his career with Wisnet, a result which is published in 2007 and shown in table 2. This satellite on chip is sized to navigate in low orbits after capturing earthly images. It was equipped with Femto-solar cells planned to produce a low power less than mW [6]. Meanwhile, miniaturization minimized consumed energy, but the sources remain insufficient and limited. The hybrid solution has been studied as a classic solution, however, many constraints have been established. They, also, limited the commercialization of this satellite generation. Among these constraints, we cite the expensive cost of not only the CS solar cell manufacturing but all the embedded Femto-cells, since we need two different technologies such as CMOS for the electronic components, as well as the CS (table 2) [11]. The cost remained hugely high compared to many solutions which were more popular and less complex. Satellites on miniature board are proposed as a concurrent solution. Moreover, it is until 2009 that D.

Bernhat has published his second FemtoSatellite generation designed on board [11].

Table 2: Historic of 10 years FemtoSat generation

	Author	Sat-name	Year	Reference
Femto Sat-On Ship	A. Brett et al	Dust Smart	2002	[14]
	D. Barnhart et al	WISNET	2007	[12]
FemtoSat-On Board	D. Barnhart	PCBSat	2009	[11]
	J. Tristancho et al	WIKISat	2011	[6]

V. COMPARISON BETWEEN TWO USS ON BROAD: PCBSAT AND WIKISAT

Comparison between two COST FemtoSatellites aims, as an objective, at familiarizing with the spatial satellite technology as well as finding a concrete amelioration or exploration based on architectures. Eventually, levels of comparison are classified from general to specific terms based on both D. Bernhart and J. Tina works [6,8,9].

1) High level comparison: functional, structural

Satellite in general, and FemtoSatellite in particular, is composed of six large subsystems that can also be subdivided elementarily. These subsystems assure the navigation, the communication, the management, and the captures by using a dedicated structure and one or more sources of energy. Often, we find other classifications of subsystems such as the navigation which is responsible for the control and the determination of the attitude as well as the position.

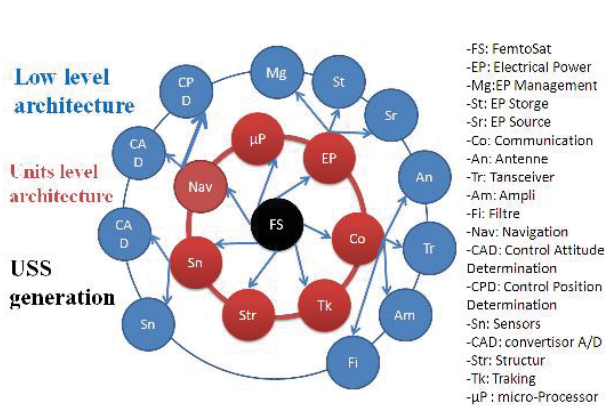


Fig.4: Basic architecture of FemtoSatellite

In this section, we focus on comparing, without technical specifications, two FemtoSatellites as shown on figure 4. In fact, PCBSat is advantaged relatively to WikiSat by the integration of 7 solar cells, a GPS and a Radio Frequency module ZigBee [3]. These modules, respectively, offer more energetic autonomy and more

flexibility in communication, despite the fact that this WikiSat is advantaged by miniaturization. However, as we have said, we haven't found other criteria of general order to compare, for instance; the applications, the services etc... In the section 5, we will study, in details, all the parameters available in the two prototypes.

2) Low level comparison: subsystems

This part highlights the different architectural parameters of two FemtoSatellites. Table 3 summarizes this comparison that displays the possibility of miniaturizing the satellite. Thereby, this trend encapsulated multitasking and multi-discipline systems that still suffer from energetic limitation which influences the service quality and the lifetime of satellite [6, 11].

Table 3: Specification of PCBSat and WikiSat

Subsystems	PCBSAT	WIKISAT
Masse/system mass (gr)	70	19,7
dimensions	9x9x1 cm (PC104)	30x25x7 mm
Cost per prototype(\$)	300	100
Payload	640x480 CMOS imager	1280x1024 CMOS imager
Electrical power subsystem	Solar cells	6 Solar cell 689 mW
	Battery	645 mA Li-ion battery
	Regulator and controller	-3,3 v regulated system bus - Peak power tracking - Battery charge regulation -6-chanel telemetry
Data Handling Subsystem	Mega 128L microcontroller 3.6864 Mhz system clock	ATmega328P MCU 8 MHz (v<3,3)
Communication Subsystem	2,4 GHz 60 mW RF ZigBee protocol Signal strength telemetry	2,4 GHz Wireless radio
Attitude and Orbit Control/Determination Subsystem	Passive aerodynamic Tow sun sensors GPS receiver	ADS: 4 optic sensors PDS : 3D accelerometer 3D gyrometer ACS : magnetorquer
Thermal Control Subsystem	Solar cell and battery Temperature monitors	No planned
Structural Subsystem	TBD	Fiberglass
Propulsion	None planned	None planned

The major specific architectural differences are situated in power and communication subsystems, particularly the antenna module. In fact, this power modification not only increases the weight but also integrates the additional components for control, regulation and distribution of power [11]. Indeed, a multi-source power system has, often, a controller that monitors the charge level, a distributor that directs the flow of current and a

regulator that adapts the tension as well as rewarding the lack of energy from secondary source. These sub-modules occupy more space and consume more energy, the fact that influences the masse and the lifetime. Similarly, the integration of a specific antenna raises the question about the supplement components: amplifier, filter, converter and the transceiver. These components should be targeted after guaranteeing better operation, taking into consideration the size and the weight. For this reason, the actual spacecraft researches still work on the modeling, the conception and the integration of secondary renewable energy sources and the antenna, chip for high gain with low power.

VI. FIELDS OF ACTUAL RESEARCHES ON FEMTOATS ON BOARD

The previous comparison showed that the general architecture kept the same specifications except some in the sensitive modules. Power is among the extremely sensitive modules. In fact, it influences other modules like communication and captures.

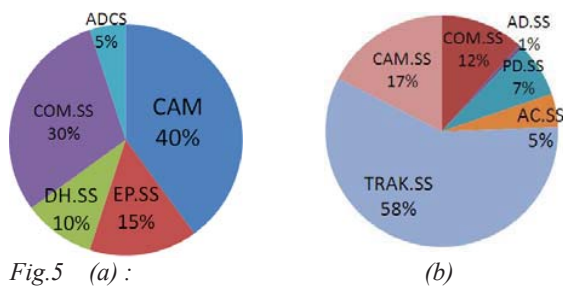


Fig.5 (a) :

(b)

(a): Power distribution PCBSat
(b): Power distribution WikiSat

Power distribution, displayed on figure 5 (part a and b), presents all greedy subsystems in terms of power. Communication and imaging sensors are the most demanding power modules, where the necessity of searching other more optimistic antennas and cameras exists. Similarly, WikiSat has three critical modules; communication, sensor imaging and tracking. All the modules directly influence the lifetime in space which has been already limited by the resource insufficiency. Actually, researches are around some minutes [3]. These limitations are, nowadays, the large perspectives of research. Works are distributed between the updates of commercialized components, the study of certain modules such as communication and the exploration of FemtoSat services.

1) Modeling of subsystems

Modeling, first of all, builds a mathematic model of the phenomena, which is called the digital modeling. Then, this model will be transformed in an observable system where we can change the parameters [7]. Thereby, these digital and analogical modelings are treated in several works such as:

- In 2011, Chang-Chan designed the power system of USS by specifying the solar cells and the battery [11]. In fact, he calculated the provided solar power, masses, and the lifetime of the satellite functioning by his battery.
- In 2012, Sunday studied and analyzed the architecture of the new adaptive miniature satellite generation [2]. He oriented the power efforts of spatial system toward the conception of system on board with commercialized modules keeping a service quality and a masse/power compromise. He designed the weight, the used power and the lifetime.

These modelings, later on, facilitate the conception and even the research of optimistic components. However, they are not followed by concrete realizations, the fact that raises the question of their feasibility. Despite this, certain works, that we will detail later on, have responded at this question as well as finishing the prototypes.

2) Study and conception of FemtoSatellite subsystems: Antenna

Power and communication are the most delicate modules since they allow re-assuring the link between the satellite and its world during life stage in space. Indeed, this link has several levels regarding the type of communication. For instance, Iverson Bell studied this problem in a way to profit the spatial waves in order to get an electromagnetic charge [4].

Besides, the space has ever attached researchers as a renewable power source. This power is found in the form of waves, luminous photos and particular charged gaseous. Exploration of such sources requires a powerful technology and a modeling knowledge in order to design sensor costs. Solar cells are among the studied sensors, but their yields aren't, until now, effective to 100%. Harvested current is trying low a point that the cells' surface should be trying large so that we can get an autonomic power. Chang-Chun estimated this power by 0.025 Kg/W and with a yield of 15% [7]. These results don't support the FemtoSat development, in a way, to insist guarding the CS, without forgetting that explorations of such cells, opposite to the sun, aren't positioned and this time aren't also to exceed 20 minutes at 2 hours cycle per orbital rotation.

The insufficiencies of FemtoSatellite power impose the optimistic communication use, strictly to guarantee that the satellite won't be lost in space or won't be incapable of assuring such a service. Often, we find 3 communication types:

- Extra-Communication: with an earthly control chamber that commands and controls the FemtoSat and the sensor measurements.
- Intra-Communication: it's communication into FemtoSatellites with constellation. This communication assures sharing task or data since FemtoSat is incapable of assuring such a service lonely. This incurability is due to its coverage and its limited power.


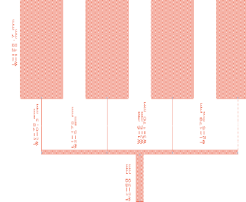
- Inter-Communication: it's the communication with ground station generally situated at high orbits.

These three modules of communication impose the use of a high gain antenna with low power consumption. In 2011, Enric.F proposed a specific antenna for WikiSat [8] regarding the following constraints:

- Weights : 7,6 gr
- Frequency band : UHF 2,4 à 2,5 GHz
- Resistance : 50 ohms
- Coverage : --
- Power: --
- Gain : 16 dB

This antenna displayed in table.4 is testified by altitudes inferior to 50 Km after launching. Tests showed a stability of operation. But this operation wasn't final because WikiSat didn't reach orbits of 200 Km in total functioning.

Table 4: Tow FemtoSat Antenna

	WIKI-Sat antenna	FEMTO-Sat linear antenna
Picture		
Authors	E. Fernandez-Murcia	C.Hamrouni et al
Ref	[8]	[5]

C.Hamrouni et E.Gill studied and designed antennas (table.4) that respond to FemtoSatellite requirement except that the latter isn't testified on a real FemtoSat. The antenna conception isn't to be the first between communication module conceptions. Indeed, noise management and synchronization of communication are also of high complicity order [1]. These reasons are actually behind the delay of functional launching of FemtoSat. However, in this year 2013, it's proved that the European Community will launch its Swarm, composed of three FemtoSatellites. It's a real test of such a kind of satellite.

3) Update of COT Subsystems: Imagery sensor

The basic motivation of FemtoSatellite on board conception is to reduce the production cost and use the COTS components. The objective remains the focus of many studies of certain researchers that opt to find an adequate material configuration with mission, power and weight constraints. Indeed, in 2011, Luis Izquierdo published a new study of FemtoSatellite generation with high resolution- observation -earth system (OES) by 14 Mpix with an optimum power [9]. This study isn't standardized since the OES technology evaluates rapidly. Moreover, these sensors remain insufficient because the FemtoSatellite specifications find that integrating high quality sensor with

low power consumption and more miniaturizing. This energy/quality paradox is generally difficult to manage. The case of solar cell integration is the classic example of this relation. In fact, we look for getting autonomous satellites with renewable power sources but the CS yield remains insufficient if we use a small area/wide cell. Besides, the use of several CS provokes a large mass and temperature that can disrupt the regular operation. For this reason, PCBSAT uses many thermal sensors [11] while it is not the case for WikiSat.

Update of sensors provokes the architecture change. Amelioration of integration and MEMS technology evolution can produce future FemtoSats on board that are more miniature and effective as well as with simpler architecture to manipulate. This trend cannot be reached without getting specific module conception since the real optimization is, by definition, linked to a specific approach.

4) Feasibility application study

Actual space technology becomes accessible for all the world, the fact that proves many countries as Spain which studied WIKISAT in 2011 and Tunisia which studied l'EREP SAT1 in 2009 [15]. This orientation offers many countries, societies and even individuals to build their personal satellites.

Chile makes an example of constellation study with FemtoSatellite for controlling the climatic change from the space in order to guess the security [3]. The countries that are situated on the cyclone trajectory or that have nuclear constructions are the focus of observers in order to reform dynamic data base on climate, atmospheric temperature and gaze concentration.

The FemtoSatellite is the future solution if it guards the following principles [10]:

- Simpler and more functional : KISS (keep it Simple and Safe) [6]
- Low and reliable costs
- Miniature autonomous power

Feasibility study gives the opportunity to valorize and to prove the FemtoSatellite efficacy in order to be the first generation of personal satellite. Indeed, low cost by prototype and launching encourages countries and researchers to study and to design prototypes until their use in the following services:

- Observation system
- Help decision system
- Security and military control system
- Didactic research system
- Commercial imagery system

These systems can be effective in securing populations, the atmosphere security and the earth monitoring. In fact, these are among the advantages of digital space technology. Consequently, the picture of nowadays is viewed not only as a color collection but also as a source of metrological data.

VII. DISCUSSION

We notice that the two FemtoSatellites WIKISat and PCBSat share the same principal of COTS conception except the fact that they differ in the use of some specific modules. The two prototypes focused on the mass and the form guarding a basic modular architecture. PCBSat is purely commercialized and presents more advantages by the use of solar power. However, WIKISat uses a specific antenna limiting itself to a unique source of power in the form of a battery. The advantage of WIKISat in terms of the integration of a specific antenna gives much more flexibility and power in communication and offers a more adaptable structure. E. Fernandez and C. Hamrouni worked on this module in order to produce an antenna generation which is more miniature and adequate instead of LEO communication [5, 8]. Nevertheless, there is no unique conception, the question of communication performance remains raised since it is the most critical element in the FemtoSatellite functionality that has already suffered from power limitation and mass. It also should optimize the consumption, the management and the communication synchronization. Until now, the test of a FemtoSatellite constellation hasn't already been validated and problems of synchronization between FemtoSat with same constellation haven't already been tested too.

The first real event that will test this satellite is expected in the European competition Swarm in March 2013 [10]. P.Sundaramoorth, studied the FemtoSat ability, made this mission the fact that valorizes all optimization on power and communication. Indeed, the triple functionality-power-communication relationship remains a restraint to study for coming generations. The antenna is, consequently, the element-engine in this bilon. It remains a motivating of research field for a FemtoSat generation that looks for maximizing the spatial lifetime by the optimization of power consumption.

VIII. CONCLUSIONS

In this paper we have presented a comparison between two approaches of USS FemtoSatellites. We have showing that engineering space is evolving rapidly in parallel with the evolution of mixed integration technology CMOS / MEMS. This Migration from one generation to another satellite has reached the stage of satellite-on-chip. These types of miniature satellites have specific missions in low orbits. Among others, the miniaturization of this system is coupled with a multi-energy deficiency which limits the lifetime of such a satellite. Communication also suffers from several constraints related to this low energy but also to the nature of the space which is noisy and loaded with thermal and magnetic fields that can affect the operation or the quality of FemtoSatellite services. During this last decade, the research has detailed architecture and has modeled some subsystems without actually launching a prototype FemtoSat in a real application. This lack of real parameters allows the opportunity for researchers to build multiple architectures and explore some specific subsystems to optimize energy and mass keeping a limited quality service. The antenna

presents one of the most principal elements that can provide a real contribution in the future generation of FemtoSatellite on Board. Further works will be done on the novel architecture conception based on a specific patch antenna mixed to a communication module.

IX. REFERENCES

- [1] P.P. Sundaramoorthy, E.Gill, C.J.M.Verhoeven, "Enhancing ground communication of distributed space systems", Acta Astronautica Volume 84, pp 15-23, March-April 2013.
- [2] I. Bell and B. Ilchrist, "Investigating the use of Miniaturization electrodynamic tethers to enhance the capabilities of femtosatellite and other ultra-small-satellite", 26th annual AIAA/USU conference on Small satellite, LOGAN Utah USA, 13-16/08/2012,
- [3] Sunday C. Ekpo, "A System Engineering Consideration for Future-Generations Small Satellites Design", Satellite Telecommunications (ESTEL), 2012 IEEE First AESS European Conference, Centro Congressi Roma Eventi Fontana di Trevi Rome, Italy, pp: 1 – 6, 02 Oct - 05 Oct 2012.
- [4] A. Becerra all, "Feasibility Study of using a Small Satellite Constellation to Forecast, Monitor and Mitigate Natural and Man-made Disasters in Chile and Similar Developing countries", 26th annual AIAA/USU conference on Small satellite 13-16/08/2012, LOGAN Utah USA..
- [5] Ch. Hamrouni, A. Abraham, and A.M. Alimi, "Both Sides Linked Antenna Array for Ultra Small Satellite Communication Subsystem", 2012 International Conference on Innovation, Management and Technology Research (ICIMTR2012), Malacca, Malaysia :21-22 May, 2012
- [6] J. Tristancho and J. Gutierrez-Cabello, "A Probe of Concept for FEMTO-SATELLITES based on Commercial-Of-The-Shelf", Digital Avionics Systems Conference (DASC), 2011 IEEE/AIAA 30th, Seattle, Washington, USA, pp: 8A2-1 - 8A2-9, 16-20 October 2011
- [7] Ch. Chen, "The Satellite Optimization Design Using Normal Cloud Model Method", 2011 International Conference on Network Computing and Information Security
- [8] E. Fernandez-Murcia, Luis Izquierdo, et Joshua Tristancho, "A Synthetic Aperture Antenna for FEMTO-SATELLITES Based On Commercial-Of-The-Shelf", Digital Avionics Systems Conference (DASC), 2011 IEEE/AIAA 30th, Seattle, Washington, USA, pp: 8A3-1 - 8A3-12., 16-20 October 2011 .
- [9] L. Izquierdo, J. Tristancho, "Next Generation of Sensors for FEMTO-SATELLITES Based On Commercial-Of-The-Shelf", Digital Avionics Systems Conference (DASC), 2011 IEEE/AIAA 30th, pp: 8A4-1 - 8A4-7, Seattle Washington USA, 16-20 October 2011.
- [10] P.P. Sundaramoorthy, E. Gill and C.J.M. Verhoeven, "Systematic Identification of Applications for A Cluster of FEMTO-SATELLITES", 61st International Astronautical Congress, Prague,27/09/ 2010
- [11] D.Barnhart, "A Low-Cost FEMTO-SATELLITES To Enable Distributed Space Missions", Acta Astronautica 64 Science direct, 9/03/2009
- [12] D. Barnhart, T. Vladimirova, and Martin N. Sweeting, "System-On-A-Chip Design of Self-Powered Wireless Sensor Nodes for Hostile Environments". Aerospace Conference IEEE, Big Sky Montana, pp: 1- 12, 3-10- 2007
- [13] D. Barnhart and T. Vladimirova, "Design of Self-Powered Wireless System On Chip Sensor Nodes for Hostile Environment", Aerospace Conference IEEE, Seattle Washington, pp: 824 – 827, 18-21 May 2008
- [14] A. Brett, D. Michaeland all, "An Autonomous 16 mm3 Solar-Powered Node for Distributed Wireless Sensor Networks", ICSENS. Hyatt Orlando, Florida, USA, pp : 1510 - 1515 vol.2, 12-14 June 2002 .
- [15] B.Neji, C.Hamrouni, and A.M Alimi, "EREPsSat-1 Scientific Pico Satellite Development", Systems conference, 4th Annual IEEE, San Diego California, pp: 255 – 260, 5-8 april 2010.
- [16] Darren D. Garber, Ph.D.' The CubeSat Challenge: A Review of CubeSat Dependencies and Requirements for Space Tracking Systems". 19 July 2012

Neural Network based Mobility aware Prefetch Caching and Replacement Strategies in Mobile Environment

Hariram Chavan
Information Technology,TEC,
Mumbai University, India
chavan.hari@gmail.com

Suneeta Sane
Computer Technology,
V. J. T. I., Mumbai, India
sssane@vjti.org.in

H. B. Kekre
MPS of Tech. Mgmt & Engg
NMIMS University, India
hb_kekre@nmims.edu

Abstract—The Location Based Services (LBS) have ushered the way mobile applications access and manage Mobile Database System (MDS). Caching frequently accessed data into the mobile database environment, is an effective technique to improve the MDS performance. The cache size limitation enforces an optimized cache replacement algorithm to find a suitable subset of items for eviction from the cache. In wireless environment mobile clients move freely from one location to another and their access pattern exhibits temporal-spatial locality. To ensure efficient cache utilization, it is important to consider the movement direction, current and future location, cache invalidation and optimized prefetching for mobile clients when performing cache replacement. This paper proposes a Neural Network based Mobility aware Prefetch Caching and Replacement policy (NNMPCR) in Mobile Environment to manage LBS data. The NNMPCR policy employs a neural network prediction system that is able to capture some of the spatial patterns exhibited by users moving in a wireless environment. It is used to predict the future behavior of the mobile client. A cache-miss-initiated prefetch is used to reduce future misses and valid scope invalidation technique for cache invalidation. This makes the policy adaptive to clients movement behavior and optimizes the performance compared to earlier policies.

Keywords—Location Based Services, Caching, backpropagation, cache-miss-initiated prefetch, cache replacement policy.

I. INTRODUCTION

The fast development of wireless communication system and advancement in computer hardware technology has led to the seamlessly converged research area called mobile computing. The mobile computing environment enables unrestricted mobility of the mobile user. The notion of anything, anytime, anywhere has been brought by mobile devices such as laptops, PDAs and cell phones. Mobility and portability have created an entire new class of applications in mobile database [1][2].

In last few years, the new fertile area for researchers in mobile computing is LBS. The LBS have ushered the way mobile applications access and manage Mobile Database System. Location based services are based on the location of mobile clients. The different value-added applications which specifically target the mobile clients location as context information are traffic condition, tourist information system, weather information system, emergency system, location-dependent advertising and disaster management systems [3][4][5][7][11][14].

The location dependent advertisement messages can be delivered to users who are within a specific area such as Departmental Store. The complete weather forecasts and weather alert notifications can be delivered to users who are in the area where weather conditions will change. If a user is under circumstances of an emergency, and the exact location of the user can be obtained then other people or organizations can offer help to the user more quickly, easily and efficiently [6][14][26].

LBS, being wireless in nature are plagued by mobility constraints like limited bandwidth, client power and intermittent connectivity. It is observed that prefetching and data caching at mobile client helps to address some of these challenges and acts as an effective antidote for the listed limitations [1][2][3][4][7][11]. In general, there are three important issues involved in the client cache management [13][19]:

- 1) A cache replacement policy finds a suitable subset of items for eviction from the cache when there is no adequate space to accommodate a new data item.
- 2) A cache prefetching policy finds suitable subsets of data items in which user maybe interested in future; and
- 3) A cache invalidation scheme maintains data consistency between the client cache and the server.

Prefetching improves the system performance in mobile environments but consumes system resources, such as bandwidth and power. A simple consequence of a cache miss is a user based reactive information query and thus slower data access [27]. So when a cache miss happens, instead of sending an uplink request only for the cache-missed data item, the client requests several associated data items to reduce future cache misses. The most important aspect is, the client can prefetch more than one associated data items by an uplink request with a little additional cost. Thus, prefetching several data items in one uplink request can save additional uplink requests [14][15][16]. In this paper, we have optimized the issue of prefetching by implementing cache-miss-initiated prefetching.

A cache invalidation scheme maintains data consistency between the client cache and the server. There are two kinds of cache invalidation methods for mobile databases: temporal-dependent invalidation caused by data updates and location-dependent invalidation caused by client mobility. The

temporal-dependent invalidation is handled by Invalidation Report (IR) schemes which are variations of the basic IR approach. In location-dependent services, a previously cached data value may become invalid when the client moves to a new location. The valid scope of an item is defined in the region within which the item is valid (i.e scope invalidation scheme) [17][18][19].

In this paper, we propose a neural network based mobility aware cache replacement policy that takes into account both the temporal and spatial locality of clients access pattern. The proposed policy is called Neural Network based Mobility aware Prefetch Caching and Replacement Policy (NNMPCR). We validate with simulation (NS2) results that mobile clients using NNMPCR are able to achieve significant improvement in cache hit ratio compared to clients using other existing cache replacement policies.

This paper is organized as follows. Section II provides a survey of existing cache replacement policies. Section III describes the Neural Network as a Location Predictor: preliminaries. In Section IV, the data prefetching concept is described. In section V, we propose a new cache replacement strategy, NNMPCR for mobile clients. Simulation model of NNMPCR is presented in section VI. The results are presented in Section VII. Section VIII concludes the paper.

II. RELATED WORK

Through extensive literature survey, it is realized that the issue of data caching in the mobile environment was first addressed in [1][2] by D. Barbara et.al. Since then caching in the mobile environment has been addressed by many researchers. Most of the existing cache replacement policies use cost functions to incorporate different factors including access frequency, update rate, size of objects, temporal score, spatial score etc. [3][4][5][6][10][18].

The traditional temporal-dependent Least Recently Used (LRU) policy has been studied widely in the literature. LRU drop a page from buffer based on last reference. This policy suffers from decision making with very limited data and has been modified by ONeil et.al [8]. The LRU-K [8] discriminates between frequent and infrequent pages based on the built-in notion of aging and is able to cope-up with evolving access patterns. The LRU-K policy has been designed by considering the history of the last k references (where $k \geq 2$). However, in LBS, the replacement policy must consider temporal as well as spatial dependency of data

Furthest Away Replacement (FAR) [9] is a location-aware cache replacement policy. FAR selects replacement victim based on the current location and movement direction of mobile client. The data items which are furthest from the mobile client will be evicted first. The assumption is that they wont be visited in the near future. FAR does not take into account access probability, valid scope area and random mobility of mobile client.

Mobility-Aware Replacement Scheme (MARS) [6] is a gain-based cache replacement policy for the mobile environment. The cost function of MARS considers temporal and spatial score together while making cache replacement decisions. Access probabilities, update and query rate results into

temporal score. The spatial score consists of client location, the location of data objects and client movement direction. For LBS, spatial score dominates and must consider the impact of clients anticipated location while deciding cache replacement which still remains unexplored.

Probability Area Inverse Distance (PAID) proposed by Baihua Zheng et. al [18] is based on Cache-Efficiency Based scheme (CEB). CEB is used for balancing the overhead and the precision of performance criterion. PAID, undertakes the valid scope area, data distance and access probability for cache replacement. The valid scope of a data value is defined as the geographical area within which the data value is valid and has been used for cache invalidation. PAID neither takes into account the data size nor data updates while cache replacement. The CEB concept is based on inscribe circle within polygon to represent the valid scope area. If the polygon is thin, then the inscribed circle covers less valid scope area. This problem is analyzed, and a modified CEB with a median circle is proposed by Kahkashan Tabassum et. al [10]. The median circle radius is in between the inscribe circles radius and the outer circle radius.

Ajey Kumar, Manoj Misra and A. K. Sarje [3][4][5] proposed Predicted Region based Cache Replacement Policy (PPRRP) and Weighted Predicted Region based Cache Replacement Policy (WPPRRP) policies. The cost function of these policies considers access probability, weighted data distance from predicted region, valid scope area and data size in cache. Predicting future user influence region and assigning priority to the data items in the current vicinity of mobile client helps to increase the cache hit ratio.

III. NEURAL NETWORK AS A LOCATION PREDICTOR: PRELIMINARIES

The mathematical model of an artificial neural network (ANN) emulates the functioning of a biological neural system. The architecture of feed forward network is shown in Fig.1. It does not have feedback connections, but errors are back-propagated during supervised training as shown in Fig.2.

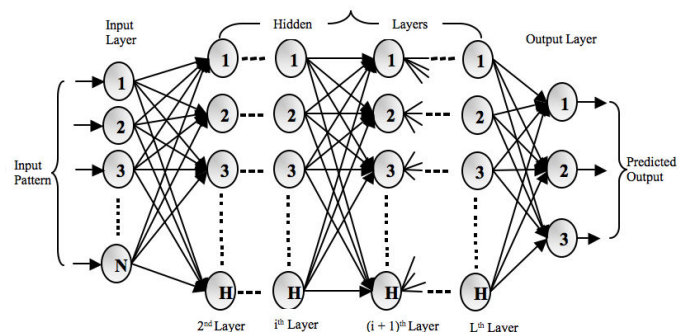


Fig. 1. Multilayer feedforward network architecture

The input layer neurons are linear, whereas neurons in the hidden and output layers have sigmoidal activation functions. For input neurons

$$f(x) = x \quad (1)$$

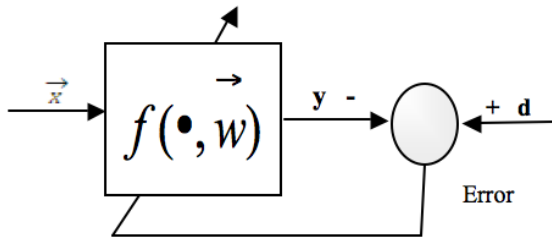


Fig. 2. Neural network training model with error backpropagation

For sigmoidal neurons in the hidden and output layers,

$$f(x) = (1 + e^{-x})^{-1} \quad (2)$$

Assuming we have N training vector pairs (input, desired)

$$T_p = \{X_k, D_k\}_{k=1}^N \quad (3)$$

Where X_k is the k^{th} input pattern and D_k is the k^{th} desired vector response when X_k^{th} input is applied to the network. The gradient of the pattern error is used to reduce the global error over the entire training set. For k^{th} training pair (X_k, D_k) the error is

$$E_k = D_k - f(y_k) \quad (4)$$

For a neuron j with activation function $f(x)$, the delta learning rule for j 's i^{th} weight w_{ji} is given by:

$$\Delta w_{ji} = \eta(d_j - y_j)f'(v_j)x_i \quad (5)$$

Where,

η	Is a small constant called learning rate
$f(x)$	Is the neuron activation function (sigmoid)
d_j	Is the desired output
y_j	Is the actual output
v_j	Is the weighted sum of the neurons inputs
x_i	Is the i^{th} input

The v_j is given by

$$v_j = \sum x_i w_{ji} \text{ and } y_j = f(v_j) \quad (6)$$

The error gradient for each pattern is computed and used to change the weights between layers of network. Adjusting the weights between the pairs of layers and recalculating the outputs is an iterative process and carried out until the errors fall below a tolerance level. Learning rate parameters scale the adjustments to weights. Massive parallelism, learning, adaptivity and fault tolerance are the desirable characteristics of ANN. Researchers from many scientific disciplines are designing ANNs to solve a variety of problems and location prediction is one of them [20].

Intuitively location (mobility) prediction uses the historical movement patterns of mobile client to determine his/her possible future locations [21][22][23]. Knowing in advance where a mobile client is heading allows us to take proactive measures. The motion of mobile client within the mobile network and

The back-propagation algorithm [24][29]:

1. Initialize the weights to small random values.
2. Choose a encoded distance and direction pattern X_k from the training set T_p and apply it to the input layer.
3. Propagate the distance and direction signal forwards through the network.
4. Compute the difference between anticipated encoded location and actual output at the output layer (δ error).
5. Use the error calculated in step (4) to compute and update weight change between pairs of layers.
6. Update all weights of the network in accordance with the changes computed in step (5).
Hidden to output layer weights
 $w_{jh}(k+1) = w_{jh}(k) + \Delta w_{jh}(k)$ (7)
Input to hidden layer weights
 $w_{hi} = w_{hi} + \Delta w_{hi}(k)$ (8)
Where $w_{jh}(k)$ and $\Delta w_{hi}(k)$ are weight changes computed in step (5).
7. Repeat step (2) through (6) until the global error falls below a predefined threshold (0.001).

Fig. 3. Backpropagation algorithm.

successive list of connections are the better alternative to prune the data entries. This pruning ensures that the location of mobile client always represents one of the possible access points in the network.

In typical mobile networks, mobile client exhibits some degree of regularity in the mobility pattern. By exploring these regularities in mobility pattern, we can predict the future location of mobile client. Since privacy is an issue that may arise when tracking of the mobile clients, it is better to use generalized movement patterns. The generalized patterns better handle local as well as irregular movement patterns. In this paper, such generalized historical movement patterns are used to train the neural network using backpropagation algorithm (Fig. 3) to predict the future location of mobile client [22][23].

IV. DATA PREFETCHING

Prefetching is concerned with improving the system performance. In mobile database system, prefetching is used as an extension to caching to improve the performance. In LBS, prefetching is the only way to avoid the need of refreshing with location change [15][27]. It is an attractive solution to reduce access latencies perceived by mobile users. Prefetching implies the prediction of information needed by the user in future. In order to effectively limit and filter potential prefetched information into the most likely future location context, one should consider users, moving speed, moving direction, habits, preferences, interests and available bandwidth [14][15][16]. In this paper, prefetching is extended with a cache-miss-initiated prefetch (CMIP) [14] scheme. As prefetching improves MDS performance with intelligent data caching and good replacement policy, this paper proposes an ANN based location prediction, CMIP based prefetching and valid scope invalidation based complete NNMPCR policy in the mobile environment.

Caching frequently accessed data on the client improves MDS performance. If cache miss, the client has to send an

expensive uplink request to fetch the queried data item. To minimize expensive uplink bandwidth prefetching may be used frequently [14]. As prefetching consumes system resources, a good prefetching scheme for LBS and mobile environment must accomplish [15]:

- Data items related to the mobile clients current location must be prefetched first.
- The movement direction and speed of mobile client must be taken into account.
- Association between data items must be taken into account.
- A good cache invalidation scheme to create free space from non-relevant data items.
- Prefetch only relevant information to save uplink bandwidth.

From above requirements, predicting and prefetching right data items in which mobile clients are interested in the future is important. The CMIP scheme satisfies this requirement by prefetching the highly associated data items. Association rule mining is well researched techniques of data mining introduced by Rakesh Agrawal [30]. It aims to extract frequent patterns, associations among sets of data items in the transaction databases and widely used in telecommunication networks. It finds association rules that satisfy the predefined minimum support and confidence [14][30]. Minimum support is used to find the frequent itemsets and minimum confidence is used to generate association rules from these frequent itemsets.

The user habits, preferences and interest results into a subset of data items which are frequently accessed. This frequently accessed subset of data items fulfills the first requirement of association rule. It has been observed that the set of data items requested over a period of time is related to each other and satisfies the second requirement of association rule [30]. In CMIP scheme, the miss-prefetch set is created with closely related subset of data items using association rules. So when cache misses, many highly associated data items from the miss-prefetch set will be fetched with missed data item. This will reduce future cache misses as well as uplink bandwidth [14].

In NNMPCR, the mobile clients access pattern base frequent itemset and association rule algorithms of [14] are implemented to generate the always-prefetch sets and the miss-prefetch sets. The mobile clients access patterns may change from time to time. Change in access patterns may change the association rules and hence prefetch sets. To adapt this change, NNMPCR periodically re-mines the association rules and prefetch sets.

V. CACHE REPLACEMENT POLICY

LBS is spatial in nature. A data item may show different values if it is queried by mobile clients from different locations. Data distance and valid scope area from the mobile clients current location are important parameters for cache replacement. Larger the distance of data item from the clients current position the probability is low that the client will enter into the valid scope area in the near future. Thus, it is better to replace the farthest data value when cache replacement takes

place. If client movement is random, then it is not always necessary that the client will continue to move in the same direction. Therefore replacing data values which are in the opposite direction of client movement but close to the current position of client may degrade the overall performance [11]. In NNMPCR, neural network predict the future location of mobile client based on historical movement patterns, so there is very less probability that the data item with higher access probability related to previous location will get replaced by new data item.

The ANN used for prediction consists of three layers: input, hidden and output. At input layer, we present the input vector. Thus, the number of neurons in this layer is the same as the number of entries in the input vector. The number of input neurons has a great impact on dimensionality and so encoded activation values (distance and direction) are applied at input layer. The encoding of output corresponds to the encoding of the input. The number of neurons in the hidden layer must be chosen suitably to avoid the under-learning or overfitting [21][22] [23]. So to find the number of neurons in hidden layers one can use empirically-derived rules-of-thumb. The commonly used rules-of thumb are:

- The optimal size of the hidden layer is usually between the size of the input and size of the output layers.
- The number of neurons in hidden layer is the mean of the neurons in the input and output layers.
- Set to something near (inputs + outputs) * 2/3 or never larger than twice the size of the input layer.

The neural network of NNMPCR is trained with nine hidden layer neurons.

The direction will be one of eight N, NE, E, SE, S, SW, W, NW directions and computed as shown in Fig-4. The distance traversed by the mobile client for computed direction is calculated by Euclidian distance formula.

The network is trained using backpropagation learning algorithm (Fig-3) in two passes. Execution of the training equations is based on iterative processes and thus is easily implementable. A pair of patterns is presented (X_k, D_k), where

```
Code for future direction of mobile client
int get_direction (double cur_x,double cur_y,
double last_x,double last_y)
{
    if(cur_x == last_x && cur_y == last_y)
        return 0;
    if(cur_x == last_x && cur_y > last_y)
        return NORTH; // north
        .
        .
        .
    if(cur_x <= last_x && cur_y >= last_y)
        return NORTHWEST; // northwest
    if(cur_x <= last_x && cur_y <= last_y)
        return SOUTHWEST; // southwest
}
```

Fig. 4. Code for the future direction of mobile client

X_k is the input pattern (location and direction) and D_k is the desired pattern (distance). The input X_k causes output response at each neuron in each layer and hence an actual output Ok at the output layer. At the output layer, the difference between the actual and the desired outputs yields an error signal (delta- δ). This error signal depends on the values of the weights of the neurons in each layer. So far, the calculations were computed forward (forward pass). Thus, the forward pass is used to evaluate the output of the neural network for the given input in the existing weights. Now, the algorithm reverts one layer and recalculates the weights of the output layer (the weights between the hidden layer and the output layer) so that the output error is minimized. The algorithm continues calculating the error and computing new weight values, moving layer by layer backward, towards the input (backward or reverse pass). Thus, in the reverse pass, the difference in the neural network output (actual output) with the desired output is compared and fed back to the neural network as an error signal to change the weights of the neural network.

The location of the mobile client is primarily thought of as their geographic coordinates. However connections on the move for a mobile client may be very large. So it is better to consider the movement direction of mobile client through the network and the cell address where the calls were made. This limits to the chances of intermittent contact during the movement. This means, the mobile clients location is always one of a finite set of locations representing one of the possible access points in the network [11][23]. It will significantly reduce the state space. In this paper, we consider such pruned data for further analysis.

The MDS is developed with the following data attributes: valid scope, access probability, data size and distance related to LBS. These databases are percolated to MSS in order to make sure that the data relevant in the valid scope is available in the nearest MSS. The access probability is updated in the database with each access to the cached data item. When cache has no enough space to store queried data item then space is created by replacing existing data items from the cache based on neural network and location proximity model. In the worst case, if all locations are in the proximity then replacement will be based on minimum access probability (P_{ai}), maximum distance (d_{si}) and scope invalidation (v_{si}). If data items have same valid scope, then replacement decision is based on minimum access probability. If valid scope and access probability are equal, then replacement decision is based on maximum data distance. In some cases, data size plays an important role in replacement. If fetched data item size is large enough and requires replacing more than three data items from cache then replacement is based on maximum equivalent size with minimum access probability, maximum distance and scope invalidation [11]. The NNMPCR cache replacement policy (Fig. 5) works as follows:

- All mobile clients generalized movement and access pattern data.
- Mobile clients specific data.
- Predict the next location of client by neural network.
- CMIP associated frequently accessed data items.

```
Client-side cache management algorithm:
Mobile Clients current location and direction is given as
input vector to neural network to predict the future
location.
Mobile client request for data item  $D_i$ 
if  $D_i$  is valid and in cache then
    validate and return the data item  $D_i$ 
end
if cache miss for data item  $D_i$  then
    CMIP for data item  $D_i$  and miss-prefetch set
    to server Get data items from server
end
if enough free space in cache then
    Store data items in cache and update  $pa_i$  of
    accessed data items.
else if not enough free space in cache then
    while ( not enough space in cache)
        if (location proximity & single)
            ANN location proximity replacement
        else
            Replacement based on Invalid scope  $vs_i$ ,
            Minimum  $pa_i$  and Large distance  $ds_i$ 
            If (Multiple replacement )
                Large size with scope invalidation.
            end
            Cache data items and update  $pa_i$  of
            accessed data items.
        end
    end
end
```

Fig. 5. Code for the future direction of mobile client

- If the location in proximity apply a neural network for replacement else normal replacement.

VI. SIMULATION MODEL

NS2 is used to simulate the proposed NNMPCR policy. The network is considered as single, large service area. Seamless hand-off from one cell to another is assumed [3][6]. The mobile clients move freely within the service area to obtain location dependent information. The node density is changed by changing the number of nodes between 5 and 50 within fixed service area. The transmission range of 250m is assumed with wireless bandwidth of 2Mbps [3][6][18]. Initially mobile clients are randomly distributed in the fixed service area. The mobile clients move according to the random waypoint mobility model. Each client chooses a random destination and moves towards with a random velocity chosen from $[V_{min} - V_{max}]$.

After reaching the destination, the client stops for duration defined by the pause time parameter. After this duration, it again chooses a random destination and repeats the whole process again until the simulation ends. The query interval follows an exponential distribution. The mobile client does not issue new query unless the pending query is served. The various parameters of NS2 simulation are listed in Table-I. Normally mobile clients follow longer regular paths and only change in direction is occasional.

TABLE 1. SIMULATION PARAMETERS

Range	Default value	Parameter
Database size	200 data items	Each data item have 4-10 different LDD values
Number of Clients	20	5-50
Client cache size	500 KB	200-2000KB
Client Speed ($V_{min} - V_{max}$)	2 m/s	2-20 m/s
Bandwidth	2 Mbps	
TTL	500 s	200-5000 s
Pause Time	10 s	
Transmission range	250 m	25-250 m
Simulation Time	5000 Sec	2000-10000s
Mobility Model	Random waypoint	
Replacement policy	NNMPCR	LRU, FAR, PPRP

VII. PERFORMANCE EVALUATION

NS2 is used to simulate the proposed NNMPCR policy. The performance of NNMPCR is compared with temporal policy LRU, spatial policy FAR and mobility aware policy PPRP. For implementation of NNMPCR, the database is created with different regions with locations and location specific resources. Data for user movement and [data] access is collected. A data set consists of approximately 1200 records with different data size. The assumptions are:

- Data items are updated only at the server.
- A timestamp and broadcast based cache invalidation method [6] for the mobile clients to maintain cache consistency.

For evaluation, results are obtained when the system has reached the steady state so that the warm-up effect of the client cache is eliminated.

The primary performance metric to evaluate the performance of CRP is a cache hit ratio. Other performance metrics can be derived from it. The ratio shows the number of queries answered locally without sending a request to the server. Thus, if higher the cache hit ratio, higher is the local data availability and less uplink costs. We have conducted experiments by varying the mobile client speed, query interval and cache size. Query interval is the time interval between two consecutive client queries. The cache hit ratio of NNMPCR and other

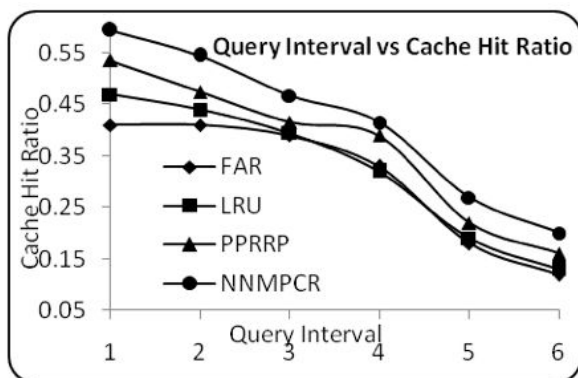


Fig. 6. Cache hit ratio vs Query Interval (sec).

replacement policies with moderate speed is shown in Fig 6. The cache hit ratio decreases with the increase in query interval

since lesser number of queries is executed at each location. Small query interval means more local queries and allows mobile client to fill its cache with more local information quickly, resulting in more cache hits. The LRU policy gives good cache hit ratio with less speed. As speed increase, the temporal locality of access pattern decreases, resulting drop in performance of LRU. The NNMPCR start giving better cache hit ratio by predicting mobile clients future location when making cache replacement decisions. The average performance improvement of NNMPCR is 29.84% as compared to temporal and 14.73% that of mobility aware scheme.

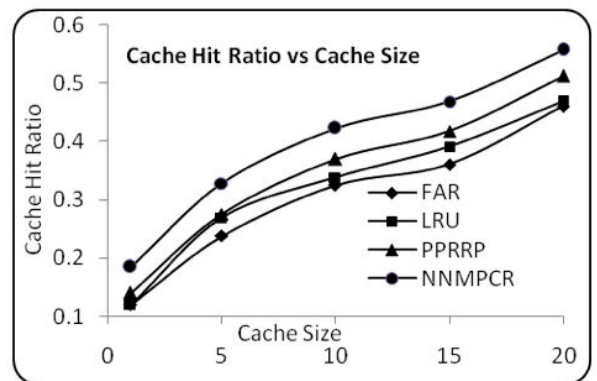


Fig. 7. Cache hit ratio vs Cache size

For small cache size, the average performance improvement of NNMPCR is 16.64% better than the mobility aware scheme and around 8.67% better for large cache size (Fig. 7). With large cache size and query interval the replacement becomes less frequent. This is because the cache can hold a large number of data items which increases the probability of getting cache hit. In such cases what does matter is how accurate our predictions are. So with almost 95% accurate prediction and associative prefetch, hit ratio maintains the consistency. In certain situations, if future location prediction and so prefetching is incorrect then it hampers the performance. Still the overall cache hit ratio is around 8-10% as compared to other mobility aware replacement policies.

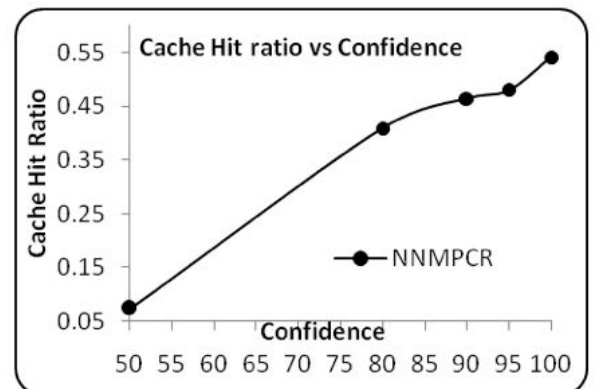


Fig. 8. Cache hit ratio vs Confidence

Fig.8 shows the impact of confidence on cache hit ratio. As expected, cache hit ratio increase with the increase in

confidence. With small confidence, the unassociated data items becomes part of cache-miss set and will be prefetched resulting in very low cache hit ratio.

VIII. CONCLUSION

In this paper, we presented a cache replacement policy using neural network prediction model and cache-miss-initiated prefetch scheme. The neural network predicts the future location of a mobile client based on generalized history of movement patterns. The ANN is designed and trained for single and multiple moves. The NNMPCR considers the temporal and spatial properties of mobile clients access patterns to improve caching performance. CMIP the client-side prefetching scheme is used to prefetch the right data items to reduce future cache misses. The neural network future location prediction, CMIP scheme and location proximity for replacement improves cache hit ratio by manifold. In certain cases if location proximity false then valid scope, data size, data distance and access probability of data item based replacement will take place. Whenever single (data item) storage results into multiple (three or more than three) replacements, we have considered this situation as critical and handled differently. Simulation results for query interval and cache size show that the NNMPCR has significant improvement in performance than the LRU, FAR and PRRP.

REFERENCES

- [1] D. Barbara, Mobile Computing and Databases: A Survey," IEEE Transactions on Knowledge and Data Engineering, Vol. 11, No. 1, pp. 108-117, January/February 1999.
- [2] D. Barbara and T. Imielinski, Sleepers and Workaholics: Caching Strategies in Mobile Environments, In the Proceedings of the ACM SIGMOD Conference on Management of Data, Minneapolis, USA, pp. 1-12, 1994. Pearson: Prentice Hall. p. 6.
- [3] A. Kumar, M. Misra and A.K. Sarje, "A Predicated Region based Cache Replacement Policy for Location Dependent Data In Mobile Environment", I. J. Communications, Network and System Sciences. 2008.
- [4] Ajey Kumar, Manoj Misra and A. K. Sarje, A New Cost Function based Cache Replacement Policy for Location Dependent Data in Mobile Environment, The 5th annual inter research institute student seminar in computer science (IRISS 2006).
- [5] Ajey Kumar, Manoj Misra and A. K. Sarje, A Weighted Cache Replacement Policy for Location Dependent Data in Mobile Environments, SAC07, March 11-15, 2007, Seoul, Korea. Copyright 2007 ACM, pp 920-924.
- [6] K. Lai, Z. Tari and P. Bertok, Location-Aware Cache Replacement for Mobile Environments, IEEE Global Telecommunication Conference (GLOBECOM 04), Vol. 6, pp. 3441-3447, 29th November- 3rd Dec'04.
- [7] Mary Magdalene Jane, R. Nadarajan, Maytham Safar, a spatio-temporal Cache replacement Policy for location Dependent Data in Mobile environments, International Journal of Business Data Communications and Networking, 6(3), 31-48, July-September 2010.
- [8] E. ONeil and P. ONeil, The LRU-k page replacement algorithm for database disk buffering, In the Proceedings of the ACM SIGMOD, Vol. 22, No. 2, pp. 296-306, 1993.
- [9] Q. Ren and M. H. Dhunham, Using Semantic Caching to Manage Location Dependent Data in Mobile Computing, In the Proceedings of 6th ACM/IEEE Mobile Computing and Networking (MobiCom), Boston, USA, pp. 210-221, 2000.
- [10] Kakhkashan Tabassum , Mahmood Quadri Syed and A. Damodaram, Enhanced-Location-Dependent Caching and Replacement Strategies in Mobile Environment, IJCSI Issues, Vol. 8, Issue 4, No 2, July 2011.
- [11] Hariram Chavan , Suneeta Sane , H. B. Kekre, A Markov-Graph Cache Replacement Policy for Mobile Environment, International Conference on Communication, Information and Computing Technology, ICCICT 2012 19-20th October 2012.
- [12] Rooma Rathore, Rohini Prinja, An Overview of Mobile Database Caching, white paper.
- [13] Heloise Mnica, Murilo Silva de Camargo, Alternatives for Ccache Management in Mobile Computing , IADIS International Conference Applied Computing 2004.
- [14] Hui Song, Guohong Cao, Cache-miss-initiated prefetch in mobile environments, Computer Communications www.elsevier.com, 28 (2005) pp. 741-753
- [15] El Garouani Said, El Beqqali Omar, Laurini Robert, Data Prefetching Algorithm in Mobile Environments, European Journal of Scientific Research ISSN 1450-216X Vol. 28 No.3 (2009), pp.478-491
- [16] George Pallis, Athena Vakali, Jaroslav Pokorny, A clustering-based prefetching scheme on a Web cache environment, www.elsevier.com/locate/compeleceng , Computers and Electrical Engineering 34 (2008) 309-323.
- [17] Ajey Kumar, Manoj Misra and A. K. Sarje , Strategies for Cache Invalidation of Location Dependent Data in Mobile Environment, Department of Electronics and Computer Engineering IIT Roorkee, Roorkee, India.
- [18] B. Zheng, J. Xu and D. L. Lee, Cache Invalidation and Replacement Strategies for Location-Dependent Data in Mobile Environments, IEEE Transactions on Computers, Vol. 51, No. 10, pp. 1141-1153, Oct'02.
- [19] Alok Madhukar, Tansel O zyer, Reda Alhadj, Dynamic cache invalidation scheme for wireless mobile Environments, Springer Science+Business Media, LLC 2007, pp. 727-740.
- [20] Anil K. Jain, Jianchang Mao, K.M. Mohiuddin, Artificial Neural Networks: A Tutorial, 1996 IEEE, pp. 31-44.
- [21] Jens Biesterfeld, Elyes Ennigrou, Klaus Jobmann, Neural Networks for Location Prediction in Mobile Networks, Institut fr Allgemeine Nachrichtentechnik, Universitt Hannover D-30167 Hannover, Germany.
- [22] B. P. Vijay Kumar, P. Venkataram, Prediction-based location management using multilayer neural networks", J. Indian Inst. Sci., 2002, 82, 7-21, Indian Institute of Science.
- [23] Joe Capka and Raouf Boutaba, Mobility Prediction in Wireless Networks Using Neural Networks, J. Vicente and D. Hutchison (Eds.): MMNS 2004, LNCS 3271, pp. 320-333, 2004.
- [24] Pamler, Hertz, Krogh, Introduction to the theory of Neural Network, pp.3-15,115-161.
- [25] Jean-Marc Francois, Performing and Making Use of Mobility Prediction PhD Thesis 2006-07.
- [26] Tong Chang, Analysis of critical success factors of mobile location-based services, Thesis, June 20, 2009.
- [27] Holger Kirchner, Reto Krummenacher, David Edwards-May and Thomas Risse, A Location-aware Prefetching Mechanism, Fraunhofer IPST, Darmstadt Germany, Euromapping, Seyssinet, France.
- [28] Lucian Vintan, Arpad Gellert, Jan Petzold, and Theo Ungerer"Person Movement Prediction Using Neural Networks ",In First Workshop on Modeling and Retrieval of Context, Ulm, Germany, September 2004.
- [29] Satish Kumar, Neural Networks - A Classroom Approach, TMH, pp. 164-175.
- [30] Rakesh Agrawal, Tomasz Imielinski, Arun Swami, Mining Association Rules between Sets of Items in Large Databases, Proceedings of the 1993 ACM SIGMOD Conference Washington DC, USA, May 1993.