ICCGI 2012

The Seventh International Multi-Conference on Computing in the Global Information Technology

ISBN: 978-1-61208-202-8

June 24-29, 2012

Venice, Italy

ICCGI 2012 Editors

John Terzakis, Intel, USA

Mirela Danubianu, "Stefan cel Mare" University of Suceava, Romania

ICCGI 2012

Foreword

The Seventh International Multi-Conference on Computing in the Global Information Technology [ICCGI 2012], held between June 24-29, 2012 - Venice, Italy, continued a series of international events covering a large spectrum of topics related to global knowledge concerning computation, technologies, mechanisms, cognitive patterns, thinking, communications, user-centric approaches, nanotechnologies, and advanced networking and systems. The conference topics focus on challenging aspects in the next generation of information technology and communications related to the computing paradigms (mobile computing, database computing, GRID computing, multi-agent computing, autonomic computing, evolutionary computation) and communication and networking and telecommunications technologies (mobility, networking, bio-technologies, autonomous systems, image processing, Internet and web technologies), towards secure, self-defendable, autonomous, privacy-safe, and context-aware scalable systems.

This conference intended to expose the scientists to the latest developments covering a variety of complementary topics, aiming to enhance one’s understanding of the overall picture of computing in the global information technology.

The integration and adoption of IPv6, also known as the Next Generation of the Internet Protocol, is happening throughout the World at this very moment. To maintain global competitiveness, governments are mandating, encouraging or actively supporting the adoption of IPv6 to prepare their respective economies for the future communication infrastructures. Business organizations are increasingly mindful of the IPv4 address space depletion and see within IPv6 a way to solve pressing technical problems while IPv6 technology continues to evolve beyond IPv4 capabilities. Communications equipment manufacturers and applications developers are actively integrating IPv6 in their products based on market demands.

IPv6 continues to represent a fertile area of technology innovation and investigation. IPv6 is opening the way to new successful research projects. Leading edge Internet Service Providers are guiding the way to a new kind of Internet where any-to-any reachability is not a vivid dream but a notion of reality in production IPv6 networks that have been commercially deployed. National Research and Educational Networks together with internationally known hardware vendors, Service Providers and commercial enterprises have generated a great amount of expertise in designing, deploying and operating IPv6 networks and services. This knowledge can be leveraged to accelerate the deployment of the protocol worldwide.

We take here the opportunity to warmly thank all the members of the ICCGI 2012 Technical Program Committee, as well as the numerous reviewers. The creation of such a high quality conference program would not have been possible without their involvement. We also kindly thank all the authors who dedicated much of their time and efforts to contribute to ICCGI 2012. We truly believe that, thanks to all these efforts, the final conference program consisted of top quality contributions.

Also, this event could not have been a reality without the support of many individuals, organizations, and sponsors. We are grateful to the members of the ICCGI 2012 organizing committee for their help in handling the logistics and for their work to make this professional meeting a success.

We hope that ICCGI 2012 was a successful international forum for the exchange of ideas and results between academia and industry and for the promotion of progress in the area of computing in the global information technology.

We are convinced that the participants found the event useful and communications very open. We also hope the attendees enjoyed the charm of Venice, Italy.

ICCGI 2012 Chairs:

Constantin Paleologu, University Politehnica of Bucharest, Romania
Tibor Gyires, Illinois State University, USA
Luc Vouligny, Institut de Recherche d'Hydro-Québec - Varennes, Canada
John Terzakis, Intel, USA
Yasushi Kambayashi, Nippon Institute of Technology, Japan
Mirela Danubianu, "Stefan cel Mare" University of Suceava, Romania
Constandinos Mavromoustakis, University of Nicosia, Cyprus
José Rouillard, Université Lille Nord, France
Beata Czarnacka-Chrobot, Warsaw School of Economics, Poland
Bernhard Freudenthaler, Software Competence Center Hagenberg GmbH, Austria

ICCGI 2012

Committee

ICCGI Advisory Committee

Constantin Paleologu, University Politehnica of Bucharest, Romania
Tibor Gyires, Illinois State University, USA
Luc Vouligny, Institut de Recherche d'Hydro-Québec - Varennes, Canada
John Terzakis, Intel, USA
Yasushi Kambayashi, Nippon Institute of Technology, Japan
Mirela Danubianu, "Stefan cel Mare" University of Suceava, Romania

ICCGI Special Area Chairs

Knowledge/Cognition
Constandinos Mavromoustakis, University of Nicosia, Cyprus

e-Learning/Mobility
José Rouillard, Université Lille Nord, France

Industrial Systems
Beata Czarnacka-Chrobot, Warsaw School of Economics, Poland
Bernhard Freudenthaler, Software Competence Center Hagenberg GmbH, Austria

ICCGI 2012 Technical Program Committee

El-Houssaine Aghezzaf, Gent University, Belgium
Werner Aigner, FAW, Austria
Areej Al-Wabil, King Saud University - Riyadh, Saudi Arabia
Cesar Alberto Collazos, Universidad del Cauca, Colombia
Panos Alexopoulos, IMC Technologies SA - Athens, Greece
Ali Alharbi, The University of Newcastle, Australia
Fernando Almeida, University of Porto, Portugal
Plamen Angelov, Lancaster University, UK
Josep Arnal Garcia, Universidad de Alicante, Span
Ali Barati, Azad University - Dezful Branch, Iran
Reza Barkhi, Virginia Tech - Blacksburg, USA
Tristan Barnett, University of South Australia, Australia
Hatem Ben Sta, University of Tunis, Tunisia
Jorge Bernardino, Institute Polytechnic of Coimbra - ISEC, Portugal
Atteet Bhalla, NRI Institute of Information Science and Technology - Bhopal, India
Mihai Boicu, George Mason University - Fairfax, USA
Eugen Borcoci, University ‘Politehnica’ of Bucharest, Romania
Daniela Briola, University of Genova, Italy

Xiaoqiang Cai, The Chinese University of Hong Kong, Hong Kong
Ani Calinescu, Oxford University, UK
Maiga Chang, Athabasca University, Canada
Emmanuel Chaput, IRIT-CNRS, France
Savvas A. Chatzichristofis, Democritus University of Thrace, Greece
Chi-Hua Chen, National Chiao Tung University - Taiwan, R.O.C.
David Chen, University of Bordeaux, France
Shu-Ching Chen, Florida International University - Miami, USA
Wen-Shiung Chen (陳文雄), National Chi Nan University, Taiwan
Zhixiong Chen, School of Liberal Arts, Mercy College - Dobbs Ferry, USA
Albert M. K. Cheng, University of Houston, USA
Dickson Chiu, Dickson Computer Systems, Hong Kong
Francesco Colace, DIEII - Università degli Studi di Salerno, Italy
Beata Czarnacka-Chrobot, Warsaw School of Economics, Poland
Mirela Danubianu, "Stefan cel Mare" University of Suceava, Romania
Bernard De Baets, Gent University, Belgium
Vincenzo Deufemia, Università di Salerno - Fisciano, Italy
Kamil Dimililer, Near East University - Nicosia, Cyprus
Alexandre Dolgui, Ecole des Mines de Saint-Etienne, France
Ludek Dolihal, Masaryk University - Brno, Czech Republic
Juan Carlos Dueñas, ETSI Telecomunicación, Universidad Politécnica de Madrid, Spain
Chanaka Edirisinghe, The University of Tennessee - Knoxville, USA
Hans-Dieter Ehrich, Technische Universität Braunschweig, Germany
Nabil El Kadhi, Ahlia University, Kingdom of Bahrain
Javier Dario Fernandez Ledesma, Universidad Pontificia Bolivariana - Medellín, Colombia
Joerg Fliege, The University of Southampton, UK
Rana Forsati, Shahid Beheshti University - Tehran, Iran
Panagiotis Fotaris, University of Macedonia - Thessaloniki, Greece
Rita Francese, Università di Salerno - Fisciano, Italy
Bernhard Freudenthaler, Software Competence Center Hagenberg GmbH, Austria
Matjaz Gams, Jozef Stefan Institute - Ljubljana, Slovenia
David Garcia Rosado, University of Castilla-La Mancha, Spain
Joseph Andrew Giampapa, Carnegie Mellon University, USA
Antonio Gómez-Corral, Complutense University of Madrid, Spain
Gustavo González, Mediapro Research - Barcelona, Spain
Feliz Gouveia, University Fernando Pessoa, Portugal
Francesco Guerra, University of Modena and Reggio Emilia, Italy
Nalan Gulpinar, University of Warwick - Coventry, UK
Tibor Gyires, Technology Illinois State University, USA
Maki K. Habib, The American University in Cairo, Egypt
Petr Hanácek, Brno University of Technology, Czech Republic
Sven Hartmann, TU-Clausthal, Germany
Wladyslaw Homenda, Warsaw University of Technology, Poland
Wei-Chiang Hong, Oriental Institute of Technology, Taiwan
Jun Hu, Eindhoven University of Technology, The Netherlands
Larisa Ismailova, National Research Nuclear University "MEPhI" - Moscow, Russia
Kyoko Iwasawa, Takushoku University - Tokyo Japan
Mehrshid Javanbakht, Azad University - Tehran, Iran

Thurasamy Ramayah, Universiti Sains Malaysia, Malaysia
Marek Reformat, University of Alberta - Edmonton, Canada
Luis Paulo Reis, University of Minho/LIACC, Portugal
Agos Rosa, LaSEEB, Portugal
José Rouillard, Université Lille Nord de France
Pawel Rózycki, University of Information Technology and Management - Rzeszow, Poland
Ozgur Koray Sahingoz, Turkish Air Force Academy, Turkey
Manuel Filipe Santos, Universidade do Minho - Guimarães, Portugal
Ana Šaša, University of Ljubljana, Slovenija
Isabel Seruca, Universidade Portucalense - Porto, Portugal
Marc Sevaux, Université de Bretagne-Sud, France
Hossein Sharif, University of Portsmouth, UK
Ashok Sharma, Mahindra Satyam, India
Mei-Ling Shyu, University of Miami - Coral Gables, USA
Patrick Siarry, Université Paris 12 (LiSSi) - Creteil, France
Spiros Sirmakessis, Technological Educational Institution of Messolongi, Greece
Martin Stanton, Manchester Metropolitan University, UK
Kathryn E. Stecke, University of Texas at Dallas - Richardson, USA
Renate Strazdina, Ernst&Young Baltic SIA, Latvia
Vadim Strijov, Computing Centre of the Russian Academy of Sciences - Moscow, Russia
Ryszard Tadeusiewicz, AGH University of Science and Technology, Poland
John Terzakis, Intel, USA
Maria Tortorella, University of Sannio, Italy
Guglielmo Trentin, National Research Council - Genoa & University of Turin, Italy
Theodoros Tzouramanis, University of the Aegean, Greece
José Braga de Vasconcelos, University Fernando Pessoa - Porto, Portugal
Maria-Esther Vidal, Universidad Simon Bolivar - Caracas, Venezuela
Ante Vilenica, University of Hamburg, Germany
Patravadee Vongsumedh, Bangkok University, Thailand
Luc Vouligny, Institut de Recherche d'Hydro-Québec - Varennes, Canada
Mihaela Vranic, University of Zagreb, Croatia
Chieh-Yih Wan, Intel Corporation, USA
Gerhard-Wilhelm Weber, METU - Ankara, Turkey
Viacheslav Wolfengagen, Institute "JurInfoR-MSU", Russia
Min Wu, Oracle Inc., USA
Mudasser F. Wyne, National University - San Diego, USA
Farouk Yalaoui, University of Technology of Troyes, France
Yugang Yu, The University of Science and Technology of China, China
Xuying Zhao, University of Notre Dame, USA

Copyright Information

Table of Contents

E-business and Strategic Management: E-valuation Quality Performance based on ADAM Methods

Alexandra Lipitakis
Kent Business School , University of Kent
Canterbury, Kent CT2 7PE, England 1
aial2@kent.ac.uk

Evangelia A.E.C. Lipitakis
Kent Business School,University of Kent
Canterbury, Kent CT2 7PE, England
eael2@kent.ac.uk

***Abstract*— E-business and strategic management problems in Digital Information Management (DIM) methodologies, Smart Environments (SE) and e-services can be efficiently solved by using a class of adaptive algorithmic modeling (ADAM) procedures. A class of implicit and explicit relationships and attitudes in strategic management and performance in DIM is presented by considering the proposed adaptive algorithmic approach. The adaptability and compactness of the proposed algorithmic schemes combined by the proper choice of singular perturbation parameters allow the (near) optimum solution of a wide class e-business and strategic management problems in Digital Information Management and e-services environments. This research work is based on key-field concepts of four interrelated sciences, i.e., Computer Science (adaptive algorithmic theory), Applied Mathematics (singular perturbation theory and partial differential equations), Management Science (strategic management and e-business) and Economic Science (performance). The proposed adaptive algorithmic approach has been applied to a class of case studies of characteristic e-business and strategic management problems and their corresponding dynamical algorithmic schemes have been presented. The ADAM technique has been used for constructing the basic questionnaire containing the “Phillips-Lipitakis” model, methodology and assumed hypotheses.**

Keywords - adaptive algorithms; digital information management; e-business; e-services; quality performance evaluation; strategy management; the ADAM methods.

I. Introduction

In recent years, extensive research has been directed to the dynamic fields of e-services computing, digital information management and context management for smart environments [3, 6, 11, 14, 20, 29, 32, 34, 35]. Intense investigations have been also focused in various challenging issues related to the above fields, resulting from the usage and application of other important technological and scientific branches, such as e-business, strategic management, knowledge management, algorithmic theory and computational methodologies, in order to identify novel concepts, theories, methods and techniques that enable their efficient application in the new emerging hybrid technological and scientific environments [2, 7, 10, 13, 16, 29, 31, 40, 42, 43].

In this research study, we present an alternative approach for evaluating certain e-business performance and strategy management methodologies based on a class of Adaptive Algorithmic Modeling (ADAM) procedures.

Specifically, we consider the so-called “Philips-Lipitakis” (PL) model for e-business strategy planning and performance measurement [22] and the corresponding basic algorithmic procedure (in the form of a special questionnaire) for the PL methodology and assumed hypotheses. The PL model, in conjunction with the ADAM methods, by using the fundamental principles of adaptivity (adaptive algorithms) and uncertainty (singular perturbation parameters), is presented in a parameterized dynamic algorithmic form, which for a suitable selection of the considered parameters leads to an (near) optimum solution of the e-business performance evaluation problem. In comparison to other related research methodologies, it should be noted that the main advantages of our proposed e-valuation quality performance methodology, in conjunction to ADAM methods, allow the efficient applicability for e-business strategic planning and performance measurement problems in e-business and e-service environments [22].

The basic principles of ADAM methods with their advantages and limitations have been presented in [20, 23]. A synoptic general description of the functionality and applicability of these methods is given in the next sections.

II. The Adaptive Algorithmic Approach and Singular Perturbation Concepts

The basic concepts of the adaptive algorithmic approach on e-business problems and strategic management methodologies have been presented in [19, 20, 21, 23]. It is known that an algorithm can be simply defined as a finite set of rules which leads to a sequence of operations for solving specific types of problems, with the following 5 important characteristic features: *Finiteness, Definiteness, Input, Output and Effectiveness,* as described in [16, 20, 39]. The algorithms can be easily presented by the so-called *‘pseudo-algorithmic*’ form, that is also called '*pseudoalgorithm*' or '*pseudocode*' in the case of preparation computer programs/codes.

The usage of the pseudoalgorithms, with the additional important features of compactness, adaptiveness and incorporation of singular parameters that allow the computation of (near) optimum solutions, can be extended for the efficient solution of e-business and Strategic Management (SM) problems and in a wide area of corresponding applications, a part of which is presented in the references and bibliography section [2, 10, 17, 19, 25, 38, 41]. It should be noted that the stated second basic feature of the algorithm, i.e., the *definiteness* as described by Knuth [16], in the case of our proposed algorithmic approach for solving e-business problems can be *'loosely'* interpreted, i.e., certain computational steps can be described in an approximate way. This is particularly useful for the qualitative nature of certain variables and procedures involved in the given e-business problems.

The concept and application of algorithms (algorithmization) can be defined as 'The Art of Designing and Practicing Algorithms', while the term '*adaptive algorithm*' denotes here an algorithm which changes its behavior based on the available resources. The adaptive algorithm functions provide a way to indicate the user's choice of adaptive algorithm and let the user specify certain properties of the algorithm. The concept of adaptive algorithm is closely related to the corresponding concept of *'adaptive'* or *'variable'* choices of the parameters (input elements) of the algorithm. Adaptive algorithms are algorithms that adapt to contention, often have adjustable steps that repeat (iterate) or require decisions (logic or comparison) until the task is completed, are simple and easy to program. The general technique of adapting simple algorithmic methods to work efficiently on difficult parts of complex computational problems can be a powerful one in the algorithm design, evaluation and practice [18, 19, 20, 21, 23].

According to a recent McKinsey Global Institute research study [26], today's business leaders need to incorporate algorithmic decision-making techniques to successfully run their organizations. This exploratory study attempts to investigate and bridge the gap between e-business strategy and algorithmic procedures.

In the last decade, research has been directed in the study of a class of initial/boundary-value problems and the behavior of the approximate solutions of the resulting linear systems by considering a small positive perturbation parameter, affecting the derivative of highest order [42, 24].

The singular perturbation (SP) parameters have been firstly used by Tikhonov [42] for solving numerically certain classes of initial/boundary value problems. Following this approach, a class of generalized fully parameterized *singularly perturbed* (sp) non-linear initial and boundary value problems can be considered and the way that the SP parameters variation affects their numerical solution can be studied [24].

In this article, adaptive algorithmic modeling (ADAM) methods are used for providing proper data input for algorithmic modeling verification of implicit/explicit relationships and attitudes in strategic management and performance in Digital Information Management (DIM) methodologies and e-services environments. Furthermore, it can be shown that suitable adaptive algorithmic procedures can be efficiently used for solving a wide class of complex computational problems, including e-business and strategic Management problems [22]. We also point out that the selective usage of both algorithmic and SP-concepts, i.e., the usage of ADAM methods in combination with the dynamical choice of SP-parameters, can lead to (near) optimized solutions. The applicability of the proposed adaptive algorithmic approach and SP-concept methodology is demonstrated by considering a characteristic case study in e-business and strategic management applications.

The main advantage of the proposed algorithmic approach for solving e-business and strategic management problems is twofold. Firstly, the adaptive algorithms can be efficiently used for solving a wide class of e-business and SM problems [19, 21, 21A, 22A]. Secondly, the dynamical choice of the SP-parameter values, which can be related to both quantitative and qualitative nature of the input parameters (data) of the given problem, can lead to (near) optimum solutions. The efficient application of such adaptive procedures requires extensive numerical experimentation [20, 22].

III. On Certain Advances in E-Business and Strategic Management

A. *E-business concept*

The term e-business was initially used by IBM marketing and Internet teams in 1996 referring to strategic focus of business with emphasis on several functions that occur using electronic capabilities. E-business is the conduct of business on-line including global communication media (Internet or other electronic networks). In general e-business can be defined as the transformation of key business processes through the use of Internet technologies and is concerned with the application of information and communication technologies in support of all activities of business [1, 4, 27].

E-business can be also defined as the administration of conducting business via the Internet, including the buying and selling goods and services with providing technical or customer support through the Internet, collaborating with business partners on sale promotions and doing jointly research with business partners. It should be noted that e-business refers exclusively to Internet businesses, but also refer to any business that use Internet technologies in order to improve productivity and profitability [8, 12, 37].

B. *E-business adoption*

The e-business adoption can be defined as the *readiness* of the organization by having appropriate attitudes, skills, knowledge and technology to facilitate e-business operations. The e-business adoption can be achieved through the TOP (technology-organization-people) dimensions, which are closely inter-related in such a way that a change to one of these components could have (great) effect on the others [36]. A value chain model has been developed by Porter [34] with independent activities in business, where competitive strategies can be best applied and information systems are likely to have strategic impact. Porter pointed out that the important issue is how to deploy e-business to take advantage of the Internet technology (Porter [35]) and the competitive advantage requires building on proven principles of effective strategies either by operational effectiveness or strategic positioning. In his latter paper Porter characteristically mentioned: *'…the next stage of the Internet evolution will involve a shift in thinking from e-business to business, from e-strategy to strategy. Only by integrating the Internet into overall strategy will this powerful new technology become an equally force for competitive advantage'.*

C. *Quality measures of e-business*

The assessment of quality in e-business is a challenging subject of research studies. In a recent paper by Mohanty et al., [28] the existing quality measures of e-business are reviewed to include the emerging success dimensions of service quality, work group impact and provide comprehensive methods for organizing various measures. A comprehensive set of quality assessment measures are presented that could provide managers with the guidance necessary to develop their own assessment systems. The authors indicate that such assessment systems have the potential to provide the required feedback for competitiveness' enhancement of the e-business.

The study presents a review of quality dimensions in e-commerce, in order to provide the following:

- Better understanding about the value of quality at a conceptual level,
- Deep insight to quality management processes and practices in e-commerce,
- The basic relationships between quality and e-commerce in such a way that it will enable future research for constructing developments.

It is pointed out that quality is a multi-dimensional construct and the contributions of different quality dimensions to business performance are related to customer's satisfaction. Note that quality has properties which are measurable or non-measurable.

The study presents a review of quality dimensions in e-commerce and their possible measures. Specifically, the following class of different quality dimensions with their attributes and corresponding measures are considered [28]:

- Reliability, Time and timeliness, Transcendence, Serviceability,
- Security and System Integrity, Accessibility, Accuracy/Clarity,
- Responsiveness, Courtesy/empathy, Communication & Feedback, Retrievability,
- Structure, Reputation/Brand, Website Intractability, Customization,
- Usability/Navigability, Availability of Website, Integrity & Trust, Functionality & Features,
- Performance, Service Differentiation, Objectivity and Flexibility.

It is also stated that the process of building an integrated e-commerce quality strategy is a dynamic, relentless and iterative procedure [28].

D. *Metrics and e-Business*

Researchers often require metrics in order to build analytical models and conduct empirical research studies on the impact of e-business strategies on organization and firm performance. Since the importance of metrics in all fields of studies is increasingly accepted, researchers rely on accepted metrics to construct analytical models of the impact of managerial strategies on organization performance and to validate empirical field research on specific managerial tactics

In the same lines managers engaged in net-enabled business planning look for appropriate metrics to help them analyze the success of their business investments. Managers are also relying on established metrics to validate several assumptions about their business environments and to evaluate the results of the related managerial practices.

It has been reported that recent development of metrics in several scientific fields, such as finance, management information systems, marketing, human resources, accounting etc., has been mainly motivated by the traditional management saying *'You cannot manage what you do not measure'* (International Center for Information Technology [13], Kaplan et al., [14, 15], Hauser et al. [11], Straub et al., [40]).

Managers, in the framework of the new net-enabled business strategic planning approach, are using metrics in order to easily analyze the success of their online initiatives and the old traditional saying has been accompanied by a recent report related comment '*You cannot measure what you do not define'* (NetGenesis [29]. This particular comment actually is closely related with the ancient Greek mathematical and geometrical theories about measurements, metrology and definition of measure unit [22].

A thorough examination of the existing e-commerce and e-business literature reveals the fact that it relies almost exclusively on several selected case studies and conceptual frameworks. Few research studies use empirical data to characterize the Internet based initiatives and indicate their impact on organization performance (Brynjolfsson et al., [3], Zhu et al., [44]). Furthermore, there is a lack of theory to guide the empirical work (Wheeler [43]), and an increasingly number of researchers argue that the literature is weak in making the linkage between theory and measures apart from subjecting proposed measures for empirical validation for reliability and validity (Straub [40]).

Towards these lines and in order to make such a linkage between theory and measures, a new algorithmic treatment of e-business, e-commerce and strategy management in digital information management methodologies, based on an adaptive algorithmic approach, has been proposed by Lipitakis [19, 20, 21], Lipitakis and Phillips [23].

Recent advances in e-business applications and technologies offer many opportunities for contemporary businesses to redefine their basic strategic objectives, transform services, products, markets and improve their work processes, business communications etc. In a recent research study by Coltman et al., [7] the drivers of the e-business strategy and performance are examined by considering the perspective of strategy content and the perspective of strategy process. By integrating these two perspectives the authors explain why, when and how certain firms are successful with e-business systems, while others remain unwilling or unable to change. A class of modeling techniques is used to show that the considered variables are heavily influenced by the unobservable heterogeneity across firms.

It is noted that a single model cannot explain the relationships between critical e-business factors such as structure, environment, feasibility, managerial beliefs and performance [6, 7].

A general model of e-business performance is used and the basic questions why and how the adoption of e-business should lead to operational and competitive advantage are explained. Four basic hypotheses have been developed and their importance has been tested by using a survey of 293 organizations and field interviews. A cross-sectional survey of senior managers has been conducted and this survey was mailed to 2,000 organizations selected from a random sample of firms across main industry sectors, such as business services (including IT and telecommunications), financial services, manufacturing, primary industries, transport/ distribution, government. Experimental results and statistical analysis for the proposed class models on business performance are presented [7].

An extended literature review on recent advances in e-business and strategic management can be found in Lipitakis [22].

IV. Adaptive Algorithms for E-business Problems and Digital Information Strategy Management Methodologies

In the following sections, we consider a class of certain characteristic case studies concerning the investigation of implicit and explicit relationships and attitudes in strategic management and performance in Digital Information Management methodologies and e-services environments [22].

Case Study 1: *a research survey of strategic management of e-business and performance*

In the framework of a research study in e-business strategy and performance [22] a pilot survey, including a dynamic adaptive algorithmic procedure, has been conducted aiming to investigate implicit and explicit relationships and attitudes in strategic management and performance. This case study is the main basis of the survey formulation that has been conducted in a related doctoral research programme [22]. In the following, we present the corresponding adaptive algorithmic procedure (in pseudo-algorithmic form) of the considered questionnaire containing certain basic information, according to our proposed model, methodology and assumed hypotheses [21, 22, 23]:

```
Algorithm ALQUE-1 (QOI, BUS, LOA, FOR, THO, PAR, SOP, PER, QUE)
Purpose: This algorithm constructs the corresponding Questionnaire to the proposed model, methodology and assumed hypotheses.
Input: Questionnaire's optional information (QOI), Business sector (BUS), level of online activities (LOA), Formality (FOR), Thoroughness (THO), Participation (PAR), Sophistication (SOP), Performance (PER)
Output: Finalized form of Questionnaire (QUE)
Computational Procedure
  Step1: Set up Questionnaire's optional information (QOI):
    Step1.1: Company/Organization Name,
    Step1.2: Answering Person's Name,
    Step1.3: Position in Company,
    Step1.4: Postal Address,
    Step1.5: Telephone/Fax number,
    Step1.6: E-mail
  Step2: Define the Sector the Company/Organization does business in (BUS),
    Step2.1: Business sector
      Step2.1.2: Banking/ Financial services
      Step2.1.3: Tourism/ Hospitality
      Step2.1.4: Publishing
    Step2.2: Industry sector
    Step2.3: Other Public or Private sector
  Step3: define the level of online activities by using a five-point scale (LOA)
    Step3.1: determine the Business to Business (B2B) activity
    Step3.2: determine the Business to Customer (B2C) activity
    Step3.3: determine the Business to Government (B2G) activity
  Step4: Compute Operational measures by using a five-point scale

    Step4.1: Determine Formality (FOR)
    Step4.1.1: usage of Total Quality Management (TQM) programs
    Step4.1.2: determine level of company's formality by using a five-point scale
        Step4.1.2.1: explicit goals
        Step4.1.2.2: written long term plan
        Step4.1.2.3: assign implementation responsibility
        Step4.1.2.4: commit to long range plans
  Step4.2: Determine Thoroughness by using a five-point scale (THO)
    Step4.2.1: utilize experience
    Step4.2.2: employ a number of external & internal source
    Step4.2.3: follow appropriate time scheduled for plan's development
    Step4.2.4: utilize a umber of organization & motivational factors
  Step4.3: Determine Participation by using a five-point scale (PAR)
    Step4.3.1: Senior Manager participation in planning & implementation of strategy
    Step4.3.2: personal participation in planning & implementation of strategy
    Step4.3.3: evaluation of personal participation in planning & implementation of strategy
  Step4.4: Determine Sophistication by using a five-point scale (SOP)
    Step4.4.1: company as informal planner
    Step4.4.2: company as operational planner
    Step4.4.3: company as long-range planner
    Step4.4.4: determine company's strategic planning activities by using a five-point scale
        Step4.4.4.1: short range profit plan
        Step4.4.4.2: final plans era accepted by the responsible persons
        Step4.4.4.3: coordinator person or group
        Step4.4.4.4: planning effort supported by top management
        Step4.4.4.5: decision of what business the company follows is taken by top management
        Step4.4.4.6: judgment of managerial performance by company's plan
        Step4.4.4.7: usage of Strengths-Weakness/Limitations, Opportunities and Threats (SWOT) analysis
        Step4.4.4.8: usage of benchmarking techniques
        Step4.4.4.9: usage of investment appraisal techniques
    Step4.4.5: determine degree of competitive method
Stage-1
        Step4.4.5.1: pricing below competitors
        Step4.4.5.2: new product development
        Step4.4.5.3: broad product range
        Step4.4.5.4: extensive customer service capabilities
        Step4.4.5.5: efforts for highly experienced personnel
        Step4.4.5.6: product quality procedures
Stage-2
        Step4.4.5.7: concern for lower cost per unit
        Step4.4.5.8: high inventory levels
        Step4.4.5.9: narrow limited range of products
        Step4.4.5.10: build brand identification
        Step4.4.5.11: refine existing products
Stage-3
        Step4.4.5.12: influence over channels of distribution
        Step4.4.5.13: effort for availability of raw materials
        Step4.4.5.14: expenditure on production process oriented R&D
        Step4.4.5.15: serve specific geographic markets
        Step4.4.5.16: promote advertising expenditures
Stage-4
        Step4.4.5.17: manufacture of specialty products
        Step4.4.5.18: build reputation within industry
        Step4.4.5.19: innovation in manufacturing process
        Step4.4.5.20: products in higher priced market segments
        Step4.4.5.21: products in lower priced market segments
        Step4.4.5.22: Innovation in marketing
Stage-5
        Step4.4.5.23: Innovation in marketing
        Step4.4.5.24: expectations in return on investment (after tax)
        Step4.4.5.25: expectations in return on assets (after tax)
        Step4.4.5.26: expectations in return on equity (after tax)
        Step4.4.5.27: expectations in return on sales (after tax)
Step5: Compute Organization's Performance by using a five-point scale (PER)
  Step5.1: determine Profitability
    Step5.1.1: return on investment
    Step5.1.2: return on assets
    Step5.1.3: return on equity
    Step5.1.4: return on sales
  Step5.2: determine Growth
    Step5.2.1: expectations in terms of market share
    Step5.2.2: expectations in terms of sales
    Step5.2.3: expectations in terms of cost of transactions with customers
    Step5.2.4: expectations in revenue growth
Step6: Finalize the requested Questionnaire (QUE) and apply the data to your model for verification
```

Note that the e-business strategic planning key-variables of our proposed model, namely (Formality, Thoroughness, Participation, Sophistication), the business performance main components (Profitability and Growth) and the 5 stages for the determination of the competitive method are indicatively marked in the above pseudo-algorithm with bold characters. A single execution of the above algorithm can express the corresponding views of each participant of the research survey on the computation of the operational measures of the considered e-business and the organizational performance by determining both profitability and growth factors. Furthermore, on the completion of the required execution of the ADAM methods combined with our proposed model for all the research survey participants the e-valuation quality performance of e-business can be achieved.

The algorithm ALQUE-1 provides a research survey (questionnaire) for the proposed model and strategic methodology under the assumed conditions and hypotheses, Lipitakis [22].

It should be noted that, in the framework of our proposed adaptive algorithmic approach, the given pseudo-algorithms describe the corresponding successive algorithmic steps in a general descriptive form and each of the used input (output) parameter variable names, depending on the complexity of the original considered problem, could be a (complex) computational procedure or a set of such computational procedures, which in turn may contain several other related computational modules and submodules. The algorithmic scheme can be further refined including several iterative and control computational procedures according to the specifications of the original problem.

V. Towards Optimized Adaptive Algorithmic Schemes for E-business and DIM Methodologies

A special class of the ADAM methods, i.e. optimized adaptive algorithmic procedures for the research survey of strategic management of e-business and performance leading to optimized questionnaire forms, can be obtained by using the singular perturbation concept with appropriate values of SP-parameters [20, 21].

Case Study 2: *an optimized research survey of strategic management of e- business and performance*

A (near) optimized adaptive algorithm can be designed by calling the corresponding dynamic algorithmic procedure OALQUE-1 in the following manner:

Algorithm OALQUE-1 (ε_{QO} QOI, ε_{BU} BUS, ε_{LO} LOA, ε_{FO} FOR, ε_{TH} THO, ε_{PA} PAR, ε_{SO} SOP, ε_{PE} PER, ε_{UF} QUE)

where ε_{QO} , ε_{BU} , ε_{LO} , ε_{FO} , ε_{TH} , ε_{PA} , ε_{SO} , ε_{PE} are singular perturbation parameters applied respectively to the input parameters and ε_{UF} the uncertainty factor parameter applied to the output of the algorithm ALQUE-1 of section 3. These SP-parameters are determined in the course of the computational procedure. The computational procedure of the algorithm OALQUE-1 can be easily constructed by following the corresponding part of the algorithm ALQUE-1

The values of the singular perturbation parameters affecting the corresponding input variables of the optimized algorithm OALQUE-1 can be determined experimentally or approximately from corresponding appropriate mathematical model. Note that the algorithm ALQUE-1 of Section 3 can be considered as a special case of the algorithm OALQUE-1 for the choice of SP-parameters

$$\varepsilon_{QO} = \varepsilon_{BU} = \varepsilon_{LO} = \varepsilon_{FO} = \varepsilon_{TH} = \varepsilon_{PA} = \varepsilon_{SO} = \varepsilon_{PE} = \varepsilon_{UF} = 1$$

The selection of the appropriate SP-parameters leading to (nearly) optimized solutions is dependent on the nature of the considered problem and often requires extensive numerical experimentation [20, 24].

Finally, it should be noted that in the framework of e-valuation quality performance by ADAM methods the dynamical choice of the SP-parameter values, which can be related to both quantitative and qualitative nature of the input parameters (data) of the given problem, can lead to (near) optimum solutions of a wide class of e-business and strategic management problems.

VI. Conclusions and Future Work

In the framework of the application of basic sciences (computer science, applied mathematics, economic science) by combining several of their key-field topics (adaptive algorithmic theory, singular perturbation theory, performance) in certain important topics of the general management science (e-business performance and strategic planning) and in the search of new efficient methodologies and techniques of improving e-business strategy planning and performance management, we consider the usage of innovative extendable models incorporating certain independent variables and measures in combination with the ADAM methods.

The proposed adaptive algorithmic approach has been applied to a class of case studies of characteristic e-business and strategic management problems and their corresponding dynamical algorithmic schemes have been presented. The proposed algorithms with the main advantages of their compactness, adaptability [16, 20] and by incorporating the proper singular perturbation parameters that allow the efficient computation of (near) optimum solutions can be extended for solving efficiently a wide spectrum of e-business and strategy management problems and related applications in DIM and SE [22].

Future research work is focused on the dynamical choice of the singular perturbation parameter values of adaptive algorithmic schemes representing a wide area of e-business problems in Digital Information Management applications. This choice that is closely related to both quantitative and qualitative nature of the input parameters (data) and computable variables/ procedures/ modules, can lead to (near) optimum solutions of e-business problems and strategic planning management methodologies by optimized ADAM methods.

REFERENCES

[1] D. Amor, The e-business (r)evolution, Upper Saddle River, Prentice Hall, 1999

[2] I.H. Ansoff (1980), "Strategic Issue Management", Strategic Management Journal, vol. 1, pp.131-148.

[3] P. Beynon-Davies., E-Business, Basingstoke, 2004

[4] E. Brynjolfsson (1993), "The Productivity Paradox of Information Technology", Comm. of ACM , vol. 36, pp. 67-77

[5] E. Brynjolfsson and L. Hitt (1996), "Paradox lost? Firm-level evidence on the returns to information systems spending", Management Science , vol. 42, pp. 541-558

[6] T.R. Coltman, T.M. Devinney and D.F. Midgley (2005), "Strategy Contents and Process in the Context of e-business Performance", vol. 22, pp. 349-386, JAI Press, NY,

[7] T.R. Coltman, T.M. Devinney and D.F. Midgley (2007), "E-business strategy and firm performance: a latent class assessment of the drivers and impediments to success", Journal of Information Technology, vol. 22, pp. 87-101

[8] A. Farhoomand, Managing (e)Business Transformation: A global perspective, Palgrave Macmillan, Basingstoke, 2005

[9] L. Gerster (2002), Who says Elephants Can't Dance? Inside IBM's Historic Turnaround, pg 172.

[10] S.G. Green., M.B. Gavin and L. Aiman-Smith (1995), "Assessing a multidimensional measure of radical technological innovation", IEEE Transactions in Engineering Management, vol. 42, pp. 203-214

[11] J. Hauser and G. Katz (1998), "Metrics: You are what you measure!", European Management Journal, vol. 16, pp. 516-528

[12] Jelassi, Tawfik and A. Enders., "Strategies for E-Business: Creating Value through Electronic and Mobile Commerce", Pearson Education Ltd, Harlow, 2005

[13] International Center for Information Technology (1988), Measuring business value of Information Technology, ICIT, Washington DC

[14] R.S. Kaplan and D.P. Norton (1992), "The balanced scorecard: Measures that drive performance", Harvard Business Review, vol. 70, pp. 71-80

[15] R.S. Kaplan and D.P. Norton (1992), "The balanced scorecard- Measures that drive performance", Journal of Service Industry Management, vol. 7, pp. 27-42

[16] D.E. Knuth, The Art of Computer Programming-Volume 1: Fundamental Algorithms, Addison-Wesley Publ. co., Reading, Massachusetts, Amsterdam-London-Tokyo, 1968

[17] S.C. Lee, B.Y. Pak and H.G. Lee (2003), Business value of Business-to-Business (B2B) Electronic Commerce: the critical role of inter-firm collaboration, Electronic Commerce Research and Applications.

[18] A. Lipitakis (2003), "Managing E-Commerce aspects of small and medium sized Publishing Houses", HERCMA 2003 Conf. Procs, Vol. 2, pp.788-794, LEA Publishers, Athens, Greece

[19] A. Lipitakis (2005), "On certain Strategic Management methodologies of E-Business in contemporary micro-medium sized Publishing firms", HERCMA 2005 Conference Procs, LEA Publishers, Athens, Greece

[20] A. Lipitakis (2007), "Adaptive Algorithmic Methods and Dynamical Singular Perturbation Techniques for Strategic Management Methodologies", HERCMA 2007 Conference Procs, LEA Publishers, Athens, Greece

[21] A. Lipitakis (2007), "Adaptive algorithmic schemes for e-service strategic management methodologies: Dase studies on Knowledge Management", ICEBE/ SOKM 2007 Conference Procs, IEEE Intl. Conference on E-Business Engineering, October 2007, Hong Kong, China

[21A] A. Lipitakis (2009), "Computational Modelling Methods in e-business and Strategic Banking Management: The case of Banking Sector", 2nd IEEE International Conference on Business Intelligence and Financial Engineering (BIFE 2009 Conference), Beijing, China

[22] A. Lipitakis, E-Business Strategy Planning and Performance: A Comparison Study of the UK and Greece, Doctoral Thesis, KBS, University of Kent, England, 2011

[22A] A. Lipitakis and Ev. Lipitakis (2009), "Adaptive Algorithmic Modeling in e-Business and Strategy Management: The Case of e-Health Services", 14th Intern. Symposium for Health Information Management Research (ISHIMR 2009 Conference), Kalmar, Sweden

[23] A. Lipitakis and P. Phillips (2007), "E- Business Strategies and Adaptive Algorithmic Schemes", HERCMA 2007 Conference Procs., LEA Publishers, Athens, Greece

[24] E.A. Lipitakis, "A universal iterative solver based on the Euclid's algorithm for the numerical solution of general type Partial Differential Equations", in *'Computer Mathematics and its Applications'*, pp. 597-642, LEA Publishers, Athens, 2006.

[25] A.M. Lyles (1981), "Formulating Strategic Problems: Empirical Analysis and Model Development", Strategic Management Journal, vol. 2, pp. 61-75.

[26] McKinsey Global Institute (1998), Driving Productivity and Growth in the UK Economy, London: Author

[27] T.L. Mesenbourg (2004), Measuring electronic business: Definition, underlying concepts and measurements plans, Reseach report, Assistant Director for Economic Programs, Bureau of the Cencus, US Government, Washington DC.

[28] R.P. Mohanty, D. Seth and S. Mukadam (2007), "Quality dimensions of e-commerce and their implications", Total Quality Management, vol. 18, pp. 219-247

[29] NetGenesis Corp. (2000), E-metrics- business metrics for the new economy, Net Genesis, MA (Online) http://www.spss.com/downloads/papers.cfm

[30] P. Phillips, e-Business Strategy: Text and Cases, Maidenhead, McGraw Hill, 2003

[31] P. Phillips (1996), "Strategic planning and business performance in the UK Hotel sector: Results of an exploratory study", International Journal of Hospitality Management, vol. 15, pp. 347-362

[32] P. Phillips and L. Moutinho (2000), "The Strategic Planning Index (SPI): A tool for measuring strategic planning effectiveness", J. of Travel Research, vol. 32, no. 2, pp. 369-379

[33] M. Porter., Competitive Strategy, Free Press, New York, 1980

[34] M. Porter., Competitive Advantage: Creating and Sustaining Superior Performance, Free Press, New York, 1985

[35] M. Porter (2001), "Strategy and the Internet", Harvard Business Review, vol. 79, pg. 62

[36] M. Porter and V.E. Millar (1985), "How information gives you competitive advantage", HB Review, vol. 63, pp. 149-160

[37] C. K. Prahalad and V. Ramaswamy, The Future of Competition, Harvard Business School Press, Boston, 2004

[38] H.J. Sapienza, E. Autio, J. Almeida and T. Keil (1997), Dynamics of growth of new technology based firms: Towards a resource based model, Annual meeting of Academy of Management Conference, Boston, USA

[39] R. Sedgewick, Algorithms, Addison-Wesley Publishing Co., Massachusetts-London, 1984

[40] D.W. Straub, D.L. Hoffman, B.W. Weber and C. Steinfield (2002), "Toward new metrics for net-enhanced organizations", Information Systems Research, vol. 13, pp.227-238

[41] D. Teece., G. Pisano and A. Shuen (1997), "Dynamic Capabilities and Strategic Management", Strategic Management Journal, vol. 18, pp. 509-533.

[42] A.N. Tikhonov and V.Y. Arsenin (1963), Methods for solving ill-posed problems, Doklady AN SSSR, vol. 153, 1

[43] B.C. Wheeler (2002), "NEBIC: a dynamic capabilities theory for assessing Net-enablement", Information Systems Research, vol. 13, pp. 125-146

[44] K. Zhu, K. Ken, X. Sean and D. Jason (2004), "Information Technology payoff in e-business environments – an international perspective on value creation of e-business in the financial service industry", Journal of Management Information Systems (JMIS), vol. 21, pp. 17-54

Comparing Travelling Design Patterns for Mobile Agent Development Using JADE

Nikolaos Karagiannis
Dept. of Informatics
Technological Educational Institute (T.E.I.) of Athens
Athens, Greece
ms09056@teiath.gr

Konstantinos Giannakos
Dept. of Informatics
University of Piraeus
Piraeus, Greece
kon.giannakos@gmail.com

Konstantinos Antonis
Dept. of Informatics and Computer Technology
Technological Educational Institute (T.E.I.) of Lamia
Lamia, Greece
k_antonis@teilam.gr

***Abstract* - Mobile agents are autonomous processes that are used to assign various tasks. Those processes are migrating to several nodes to execute those tasks locally, instead of RPCs. Their migration way may be different according to the application type and it is based on a design pattern. Here, we present comparative results of three different travelling design patterns for mobile agents (Itinerary, Branching and Star-shaped) with the use of an application that we developed. Derived results showed that the branching pattern performs better than the other two in terms of turnaround times, whether we use constant size or variable size of answers to mobile agent requests to servers.**

Keywords - Mobile Agents; Travelling Design Patterns; JADE Application

I. INTRODUCTION

An agent is a special software component providing an interoperable interface to an arbitrary system and/or behaving like a human agent working for some clients in pursuit of its own agenda. Some common tasks for such software components are monitoring of systems, searching for specific information, managing a system, etc. Agent – based methods are becoming more and more popular as time goes by and they are used in many different fields (like economics, e.g., [22]). Some common characteristics of agents are autonomy, sociality, intelligence and mobility [5].

Agents can have a combination of various characteristics. Mobile agents (having as basic characteristic the mobility) are able to migrate from one computer to another autonomously and continue its execution on the destination computer. They are used instead of RPCs (Remote Procedure Calls), exchanging remotely various data. The most basic advantage of mobile agents is the reduction of the used bandwidth, because the agent migrates itself and there is no data exchange between different procedures hosted on different computers (Figure 1) [5].

Also, these procedures are asynchronous and autonomous, so their function depends on network connectivity. That means that there is no need for continuous connection between nodes for data exchange. In the RPC model, if the connection stops when the processes are exchanging data, then the connection has to be restarted. In the meantime, a process waits for a response from the other process aimlessly. But, with the use of a mobile agent a job can be completed locally while the connection is down, and finally waits to migrate until the connection is established again (Figure 2). This is also useful in mobile phones in cases of unstable connections [5].

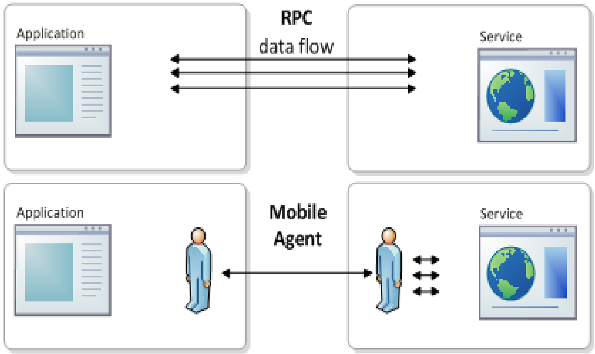


Figure 1. Agent migration vs data migration in RPC.

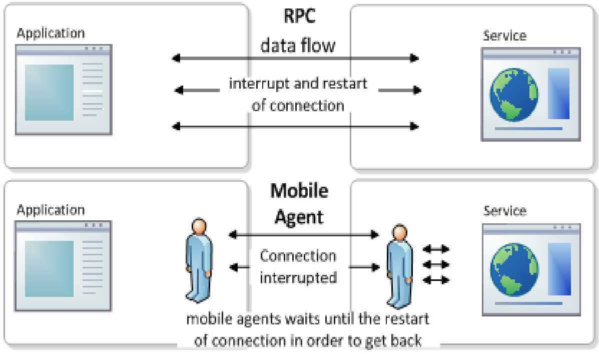


Figure 2. There is no need of continious network connection.

But also, it must also be considered that there are several difficulties/disadvantages of the mobile agent technology application into the internet infrastructures. The most important difficulties are various security problems, the high computation cost required by a server in order to host and serve a lot of mobile agents concurrently and the high difficulty of the creation and the application of those infrastructures.

Multi Agent Systems (MAS) are systems in which many agents interact in order to solve a common problem. The problem is divided into several sub problems distributing each one to different agents in the MAS system. A MAS system works on a set of various computers connected via a specific network (LAN or WAN, etc). A MAS system is very appealing for building open and distributed applications [13]. Various MAS systems can communicate in order to achieve several user needs [5].

It should also be stated that software agents, which bring together the two concepts “process” and “object”, are interesting building blocks for flexible system architectures, even if they are not always mobile. On the one hand, mobile agents provide a novel and useful example for an open and distributed MAS ([14], [15], [16]). On the other hand, static agents (non-mobile) are probably as important as mobile agents: they encapsulate autonomous activities in a stronger way than classical objects, they communicate with other (mobile) agents via the same protocols and interfaces, and they provide (with mobile agents) a uniform way to structure large distributed systems [2].

A variety of design patterns for mobile agents have been proposed in the past organized in different categories [6]. A basic task pattern is the Master-Slave pattern. On this pattern, a master agent delegates a task to be done on a given agency to a slave agent(s). The slave agent visits the indicated agency where it accomplishes the task, and then returns to the source agency carrying the results. The master agent receives the results and then the slave destroys itself. The migration procedure of an agent varies also, creating the category of travelling design patterns [6]. The itinerary pattern (Figure 3) provides a way to execute the migration of an agent, which will be responsible for executing a given task in remote hosts. The agent receives an itinerary on the source agency, indicating the sequence of agencies it should visit. Once in an agency, the agent executes its task locally and then continues on its itinerary. After visiting the last agency, the agent goes back to its source agency [1].

On the branching travelling pattern (Figure 4), the agent receives a list of agencies to visit and clones itself according to the number of agencies in the itinerary. Then, all clones will visit an agency of the received list. Each clone has to execute its corresponding task and notify the source agency when the task is completed. The importance of this pattern is that it splits the tasks that can be executed in parallel [1]. A typical example would be a search agent that sends out slave agents to visit multiple machines in parallel. Of course, mechanisms to control the high degree of dynamism of such agent-enabled parallel computations then become a necessity [2].

So, we can imagine the World Wide Web consisting of servers and clients working on mobile agent platforms exchanging mobile agents. Also, by integrating extra characteristics like intelligence and sociality we will have smarter applications offering high level services (e.g., auto-learning) and achieving better and more specific results.

On the Star-Shaped travelling pattern (Figure 5) the agent receives a list of agencies that it has to visit. So, it migrates to the first destination agency, where it executes a task, going back to the source agency. The agent repeats this cycle until visiting the last agency on its list [1].

In this paper we compare the three above mentioned design patterns with the use of a mobile MAS application that we have implemented. The application contains two static agents representing a web client and a web server (hosted on different machines). Both client and server exchange mobile agents. We developed all those agents using the JADE platform and we execute them in many nodes. A lot of implementations for different design patterns have been proposed in the past (e.g., [1], [6], [8]), but it is not our purpose to present another alternative implementation on the same subject. The contribution of our work is the presentation of comparative results for those patterns in terms of turnaround times, for constant and variable sizes of answers to mobile agent requests from the implemented servers.

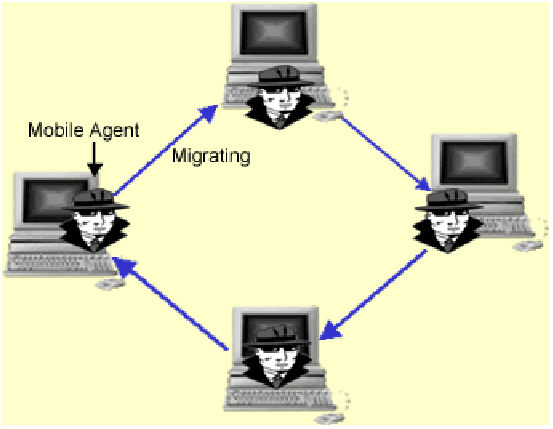


Figure 3. Itinerary pattern.

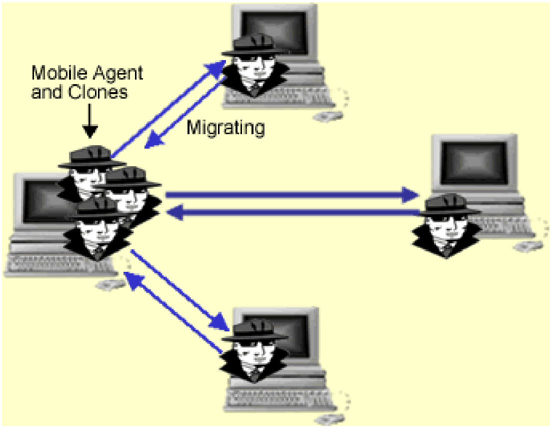


Figure 4. Branching pattern.

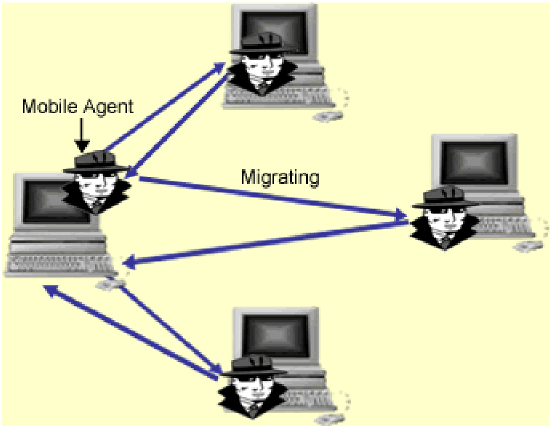


Figure 5. Star-shaped pattern.

II. Related Work

Aridor and Lange [6] reported on several design patterns for mobile agencies classified in three different categories: travelling, task and interaction patterns. Travelling patterns encapsulate mobility management of an agent for one or more destinations. Task patterns are concerned with the breakdown of a task and how these tasks are delegated to one or more agents. Interaction patterns are concerned with locating agents and facilitating their interactions. They also implemented three of them (master-slave, meeting and itinerary) and shared their experiences with it. Eight different agent design patterns are implemented in [1] in JADE: Itinerary, Star-Shaped, Branching, Master-Slave, MoProxy, Meeting, Facilitator, and Mutual Itinerary Recording. The itinerary, branching and star-shaped patterns were proposed in [11].

Kendall et al. [17] present several patterns of intelligent and mobile agents based on a layered architecture considering mobility and intelligence separately. A set of seven patterns related to agent communication mechanisms are discussed in [18], but they do not take into account the intelligence and mobility together.

Eshtay [7] proposed the Hierarchal Traveling design pattern, which is a combination of the itinerary pattern and the depth first algorithm. This pattern was implemented in JADE, too. Wang et al. [9] uses the branching pattern to develop in JADE an agile supply chain management model.

Ojha et al. [12] propose a design pattern for intelligent mobile agents. It helps in efficient mobility of these agents, which are more often fatty. It enables dynamic on-demand behaviour specific to a network host environment. It describes the pattern using a suitable pattern template and reported the results of its implementation (using JADE) in a prototype multi-agent system for e-commerce domain.

Maamar and Labbe [10] described two strategies (servlet and applet) that could enhance the operations of software agents and showed that both strategies could suit them. Agents should be embedded with mechanisms that allow them to make the correct decision: either move or invite.

There are also a lot of design patterns proposed in the past, that they do not consider mobility of agents, but they are based on social and intentional characteristics of an agent (e.g., [19],[20], [21]).

III. Our Application: A Web Agent Explorer

We developed a web application that informs the client about the latest registrations that are added to various e-news sites that interests him. Several mobile agents (each one representing a user) are migrating to various servers to retrieve information that the user is interested in.

The application developed with the use of the JADE platform. It was developed to help us compare the itinerary, star-shaped and branching design patterns under various circumstances and extract useful conclusions. We chose to compare only these three patterns because, one the one hand, they are of the most famous patterns in the research community. On the other hand, our application is too simple and does not involve collaboration or interactions between agents and does not check permissions. So, patterns like the Meeting, or MoProxy, etc, are not suitable for it. The application obeys the master – slave model containing two static agents (masters) representing a web client and a web server. Both client and server exchange mobile agents (slaves) obeying the three above mentioned design patterns, alternatively and we execute them in many nodes. We mention here, that our system is still in prototype level.

IV. *The JADE Platform/Framework*

JADE (Java Agent Development Framework) is a platform supporting agent processes and also offers libraries (framework) for multi agent application development written in JAVA. It is ideal for distributed application development based on multi agent systems. For application development, JADE has an IDE with useful tools and a GUI for platform administration. The platform offers all the necessary services to the agents that they are installed on it. With some of those services, agents can identify and communicate each other, and they can search each other after they have registered on specific platform catalogues [3].

Each platform constitutes a MAS with at least one container. Each container is installed to a computer and it can support agents and offer them all the necessary services. Consequently, a platform can constitute a network of containers (e.g., a LAN) (Figure 6). Two basic services are the AMS (Agent Management System) and DF (Directory Facilitator) directories that are local agents [3], [4].

Agents are java classes inheriting the Agent class of JADE libraries. The actual job, or jobs, an agent has to do is carried out within ‘behaviours’. A behaviour represents a task that an agent can carry out. An agent can execute several behaviours concurrently. The scheduling of behaviours in an agent is not pre-emptive (as for Java threads), but cooperative. This means that when a behaviour is scheduled for execution its method is called and runs until it returns.

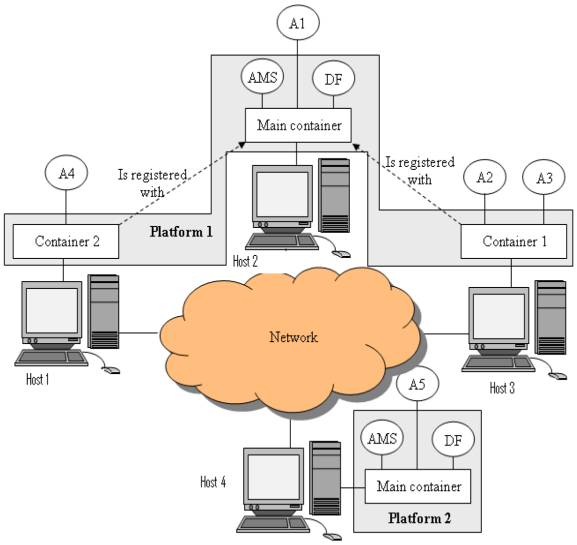


Figure 6. JADE architecture overview.

Therefore, it is the programmer who defines when an agent switches from the execution of a behaviour to the execution of another [3]. Here, we use three different behaviour types:

- OneShotBehaviour: executes once and dies.
- CyclicBehaviour: stays active as long as its agent is alive and is called repeatedly after every event.
- TickerBehaviour: periodically executes some user-defined piece of code.

V. Functional Description

Each user interacts with a local agent via a GUI. User enters to GUI the web domains he is interested in. Consequently, the local agent sends mobile agent(s) according to the mobility pattern that has been set. When the above mobile agent(s) returns having the available pages of each domain, the user is able to select the pages he wants to retrieve the latest updates of them. Next, the user declares the frequency for a mobile agent(s) to visit the selected domains in order to check if the selected pages have been updated. For instance, a user may want the mobile agent(s) to migrate to check those pages every 5 minutes. For the first time, a mobile agent returns the latest registrations from those pages, and only when there have been updates since his latest visit for all the other times. When a mobile agent arrives at a server, it is served by a local agent that "lives" there.

VI. Architecture

Client and Server processes are hosted in different JADE platforms and each one constitutes a different MAS (JADE platforms). Until now, JADE does not support agent mobility between containers that belong to different platforms (inter platform mobility). We used the IPMS addon that provides this feature [23]. The client corresponds to one container (computer) and the server may be distributed to many containers, but in our approach it is constituted by one container.

1) *Client Components Description:* The client side container hosts a local agent except the basic JADE agents (like AMS) (Figure 7). When the application starts, AMS creates an object of this agent. This local agent is static and provides the user a GUI to enter his preferences and receive the results. It also checks and manages the information that mobile agents return. So, according to the Master-Slave pattern, the local agent forwards tasks to a mobile agent(s). It is able to save data to disk, while a mobile agent is not. The mobile agent sends data to a local agent. We created those mechanisms for system security reasons. The static agent has to be capable to authenticate the returned mobile agent preventing from malware attacks. The local agent creates different types of mobile agents classes for each migration pattern. When a mobile agent returns back to the client host, it sends the data collected during migration to the local agent and then destroys itself.

2) *Server components description (Figure 8):* The server side container hosts a local agent too, except the basic JADE agents (like AMS and DF). When the application starts, AMS creates an object of this agent. The local agent publishes in DF the services provided. Those services correspond to the pages that mobile agents can retrieve updates. The local agent is static and serves the incoming mobile agents. The arrived agents send requests to the local server about the info a user is interested. So, the role of this agent is to provide an interface between the arriving agents and the database and to protect the system from malware attacks.

VII. Experimental Results

We compared the three travelling patterns which were mentioned before using our application. We set up one client and three servers into a LAN network. The client was set up on a four core 64bit CPU at 3.3 GHZ with 8GB RAM. Servers were created as 4 virtual machines. One of the servers was set up in a virtual machine of the client's machine having 512 MB RAM, while the others were set up on another machine containing a CPU at 2 GHZ with 512MB RAM.

Throughout the paper we use the term task to refer to the trip of a mobile agent to the 4 servers mentioned above in order to collect data. Specifically, the client agent keeps a hash table, where each key corresponds to the name of a site the user is interested (e.g., e-news.com, games.com, e-shop.gr). Each key indexes a data structure containing the specific pages inside a site, as well as the date and time of the last visit. All data exchanged between a server and a client, are encrypted with SSL. An instance of a mobile agent is created in the client side for each task, containing the appropriate records from this table. The server receives those records and registers the data collected into a message, along with other information like the structure of the page, formatting of the page, etc. Since the information exchanged is in text format, there is no significant overhead for the server to serve a mobile agent.

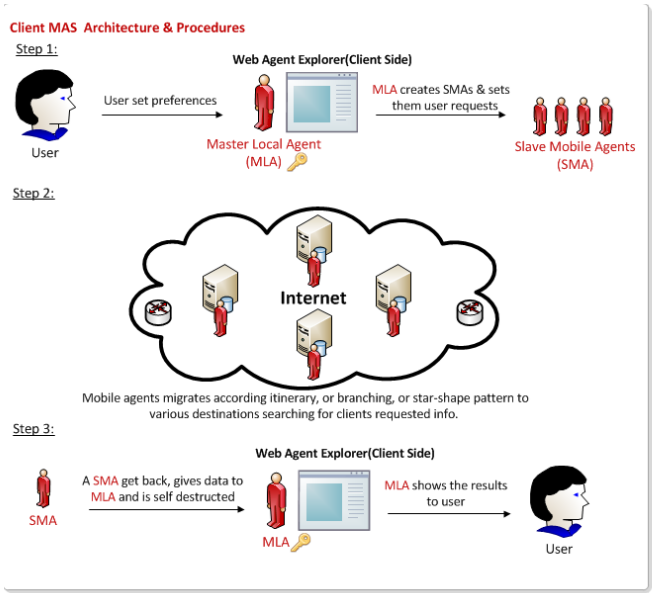

Client MAS Architecture & Procedures
Step 1:
Web Agent Explorer(Client Side)
User set preferences
User
Master Local Agent (MLA)
MLA creates SMAs & sets them user requests
Slave Mobile Agents (SMA)
Step 2:
Internet
Mobile agents migrates according itinerary, or branching, or star-shape pattern to various destinations searching for clients requested info.
Step 3:
Web Agent Explorer(Client Side)
A SMA get back, gives data to MLA and is self destructed
SMA
MLA
MLA shows the results to user
User

Figure 7. Client MAS Architecture.

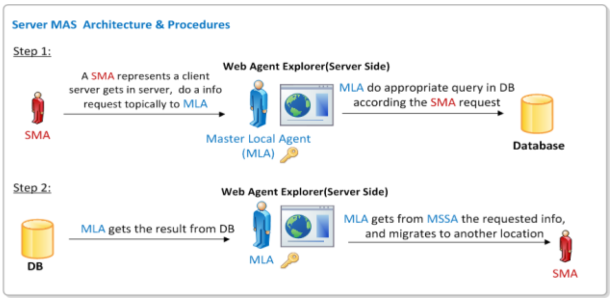

Server MAS Architecture & Procedures
Step 1:
Web Agent Explorer(Server Side)
A SMA represents a client server gets in server, do a info request topically to MLA
SMA
Master Local Agent (MLA)
MLA do appropriate query in DB according the SMA request
Database
Step 2:
Web Agent Explorer(Server Side)
MLA gets the result from DB
DB
MLA
MLA gets from MSSA the requested info, and migrates to another location
SMA

Figure 8. Server MAS Architecture.

Using itinerary and star-shaped patterns we have one mobile agent for each task. Each task starts when a mobile agent is created and ends after having visited all the nodes and returning to client sending the results to the Master Agent. The turnaround time of an agent depends on the workload of the server and the number of the exchanged messages. In the branching pattern case, a mobile agent is sent in parallel for each task. The turnaround time of the task ends when the last mobile agent returns back. The size of the answer to a query imposed by a Server Agent, corresponding to a mobile agent request, is small because their content is in text format. We consider constant and variable size of an answer to measure the turnaround time (in seconds) for each task for the three travelling patterns (itinerary, branching and star-shaped).

A. Constant Size of Answers

We consider 20 tasks arriving sequentially. Upon the arrival of a task, a mobile agent(s) migrates automatically, from the client to the above servers and retrieves 28 database records of the same size from each server. We run the application for each pattern separately.

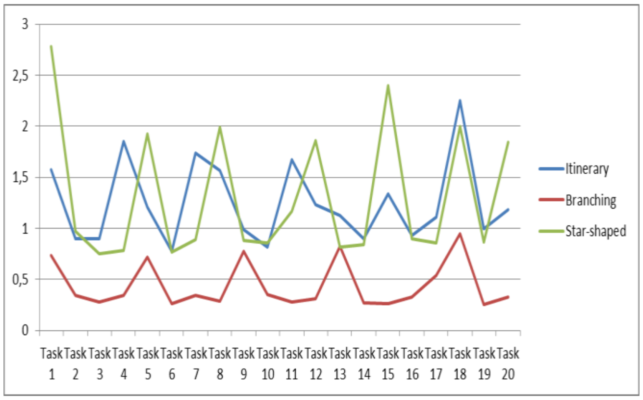


Figure 9. Comparing the three design patterns with constant size of answers.

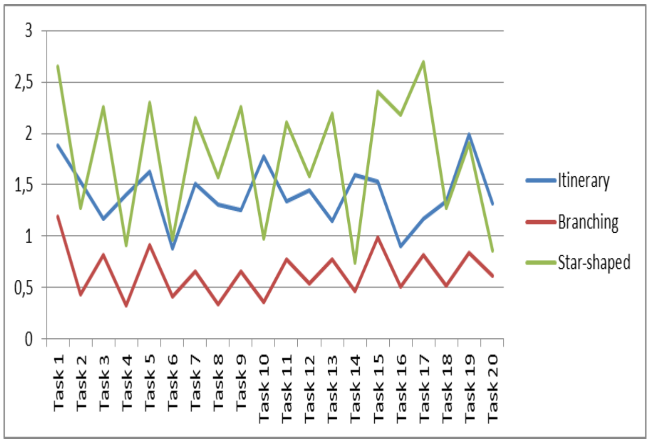


Figure 10. Comparing the three design patterns with variable size of answers.

Figure 9 presents the derived results of turnaround times per task for constant size of answers. Results show that the branching pattern achieves the best turnaround times for all tasks. The itinerary pattern performs a bit better than star-shaped on the average, but results are comparable. This is due to that some servers are hosted on the same machine as different virtual machines, although the star-shaped pattern sends more messages (48 messages vs 42, respectively). This performance difference would be greater if servers could be hosted in different machines.

B. Variable Size of Answers

We consider again 20 tasks arriving sequentially. But in this case we have different sizes of answers in each task. We developed a simple application (News Generator) in JAVA that runs in each server and adds some fake news updates (text format) to the database for various categories (like sports, weather news, politics, etc). The database update frequency varies, but it is less or equal to the task arrival time frequency in any case. We run again the application for each migration pattern separately and Figure 10 illustrates the derived results. Results show that the branching pattern achieves the best turnaround times for all tasks. The itinerary pattern performs better than star-shaped on the average.

VIII. CONCLUSIONS AND FUTURE WORK

A lot of design patterns for mobile agents have been proposed in the past. In this paper, we focus only on travelling design patterns and we try to compare three of the most commonly used of them: itinerary, branching and star-shaped, with the use of a mobile MAS application that we have implemented. The application contains two static agents representing a web client and a web server (hosted on different platforms). Both client and server exchange mobile agents. We developed all those agents using the JADE platform and we execute them in many nodes. The above mentioned design patterns are compared in terms of turnaround times, for constant and variable sizes of answers to mobile agent requests from the implemented servers. The derived results show very clearly that the branching pattern performs better under both circumstances.

In our future work, we intend to investigate hybrid implementations of design patterns and evaluate them on our platform. Derived results will be compared with the evaluation results of the current work and with other results assuming hybrid implementations.

REFERENCES

[1] Emerson Ferreira de Araújo Lima, Patrícia Duarte de Lima Machado, Jorge César Abrantes de Figueiredo, and Flávio Ronison Sampaio, “Implementing Mobile Agent Design Patterns in the JADE framework”, EXP in Search of Innovation,Vol.-3, No.-3, September 2003.

[2] Stefan Funfrocken and Friedemann Mattern, “Mobile Agents as an Architectural Concept for Internet-based Distributed Applications The WASP approach”, Proceedings of the KiVS’99, Springer-Verlag, pp. 32-43, 1999.

[3] Fabio Bellifemine, Giovanni Caire, and Dominic Greenwood, “Developing Multi Agent Systems with JADE”, John Wiley & Sons, Ltd, 2007.

[4] http://jade.tilab.com/doc/tutorials/JADEAdmin/jadeArchitecture.html [retrieved: January, 2012]

[5] Michael Wooldridge, “Introduction to Multi Agent Systems”, The MIT Press, 1999.

[6] Yariv Aridor and Danny B. Lange, “Agent Design Patterns: Elements of Agent Application Design”, In proceeding of the 2th ACM international conference on Autonomous Agents (Agents '98), p. 108-115, 1998.

[7] Mohammed Eshtay, “Hierarchal Traveling Design Pattern for Mobile Agents in JADE Framework”, 4th International Conference On Information Technology (ICIT), 2009.

[8] Ajay Kr. Singh, Ravi Sankar , and Vikram Jamwal, “Design Patterns for Mobile Agent Applications”, Workshop on Ubiquitous Agents on embedded, wearable, and mobile devices, Bologna 2002, http://autonomousagents.org/ubiagents/2002/papers/papers/22.pdf, [retrieved: January, 2012].

[9] Wenjuan Wang, Weihui Dai, Weidong Zhao, and Tong Li, “Research on Mobile Agent Systems for Agile Supply Management”, Journal of Software, vol. 6, no. 8, Agust 2011, pp.1498-1505.

[10] Zakaria. Maamar, and Paul Labbé, “Moving vs. inviting software agents: what is the best strategy to adopt?” Communications of the ACM, Volume 46, Issue 7 (July 2003), pp. 143 – 144.

[11] Yasuyuki Tahara, Akihiko Ohsuga, and Shinichi. Honiden, “Agent system development method based on agent patterns”, In Proceedings of the 21st international conference on Software engineering. IEEE Computer Society Press, 1999, pp. 356-367.

[12] Ananta Charan Ojha, Sateesh Kumar Pradhan, and Manas Ranjan Patra, “Pattern-Based Design for Intelligent Mobile Agents”, 4th International Conference on Innovations in Information Technology (IIT '07), p. 501-505, 2007.

[13] Nicholas R. Jennings, “An Agent-Based Approach for Building Complex Software Systems”, Communications of the ACM, 44(4), April 2001, pp.35-41.

[14] Danny B. Lange and Mitsuru. Oshima, “Seven Good Reasons for Mobile Agents”, Communications of the ACM, 42(3), March 1999, pp. 88-89.

[15] Rahul Jha and Sridhar lyar, “Performance Evaluation of Mobile Agents for E-Commerce Applications”, In Proc. Of the 8th Int. Conf: on High Performance Computing (HiPC 2001), LNCS, vol-2228, 2001, pp. 331-340.

[16] S.R. Mangalwade, K.K Tangod, U.P. Kulkarni and A.R. Yardi, “Effectiveness and Suitability of Mobile Agents for Distributed Computing Applications”, In Proc. Of the 2nd Int. Conf: on Autonomous Robots and Agents, Dec 13-15, New Zealand, 2004.

[17] Elizabeth A. Kendall, P. V. Murali Krishna, Chirag V. Pathak, and C. B. Suresh, “Patterns of Intelligent and Mobile Agents”, In Proc. of the 2nd Int. Conf: on Autonomous Agents, ACM Press USA, 1998, pp. 92 - 99.

[18] Nuno Meira, Ivo Conde e Silva, and Alberto Silva, “A Set of Agent Patterns for a More Expressive Approach”, In Proc. of the 5th European Conf: on Pattern Languages of Programs (EuroPLoP2000), 2000, pp. 331-346.

[19] Sylvain Sauvage, “Design Patterns for Multiagent System Design", In Proc. of 3rd Mexican Int. Conf: on Artificial Intelligence (MICAI 2004), Mexico City, LNCS, Springer Berlin, 2004, pp.352-361.

[20] T. Tung Do, Manuel Kolp, and Alain Pirotte. “Social Patterns for Designing Multi-Agent Systems”, In Proc. Of the 15th Int. Conf: on Software Eng. And Knowledge Eng. (SEKE 2003), USA, July 2003, pp.103-110.

[21] D. Deugo, M. Weiss, and E. Kendall, “Reusable Patterns for Agent Coordination”, Coordination of Internet Agents: Models, Technologies, and Applications, Springer-Verlag, 2001, pp. 347-368.

[22] Valentas Daniunas, Vygintas Gontis, and Aleksejus Kononovicius, “Agent-based Versus Macroscopic Modeling of Competition and Business Processes in Economics”, The Sixth International Multi-Conference on Computing in the Global Information Technology (ICCGI), Luxembourg, pp. 84-88, 2011.

[23] https://tao.uab.cat/ipmp/files/README [retrieved: January, 2012]

A Simple M2M Overlay Entity Discovery Protocol

Teemu Väisänen
VTT Technical Research Centre of Finland, Oulu, Finland
Teemu.Vaisanen @ vtt.fi

Abstract—**This paper deals with discovering M2M overlay entities in Machine-to-Machine (M2M) service networks. The eXtensible Messaging and Presence Protocol (XMPP) is used as a basic building block for the M2M communication. XMPP does not offer mechanisms for discovering unknown entities from unknown contacts, and this paper's goal is to provide a protocol enabling this. The presented protocol does this by using asynchronous remote procedure calls (RPCs), unicast messages and friend-to-friend type of communication. The paper proposes new XMPP subscription statuses to enable exchange of roster items in the M2M overlay entity discovery protocol without compromising privacy, and presents the protocol for discovering unknown M2M overlay entities from unknown M2M overlay entities.**

Keywords - distributed systems, service discovery, privacy

I. Introduction

Machine-to-machine (M2M) is a buzzword meaning a bagful of technologies that allow devices to communicate with one another using different communication channels. Terms such as M2M, Internet of Things (IoT), Smart Objects (SO), and Web of Objects basically all mean the same, including e.g., remote management of devices. Technologies used commonly in M2M include at least naming and identification of entities, service discovery (SD), security services such as authentication, and communication technologies. End-to-end M2M communication can be established with one or more protocol conversion gateways running between the actual M2M devices, but the trend is to build systems where end-to-end communication is possible for example using IPv6 [1]. Overlay networks in M2M are sometimes called middleware and they are often used on top of other networks to make naming easier, to improve routing, and/or to improve Quality-of-Service (QoS). In M2M service networks, a SD protocol can be thought of as a comprehensive discovery protocol including at least functionalities of M2M device, overlay entity and service discovery mechanisms, and mechanisms for selecting discovered M2M services that are to be used. The selection can be based on discovery order, location, etc.

The M2M service network presented in this paper is based on the eXtensible Messaging and Presence Protocol (XMPP) [2]. M2M overlay entities are identified as XMPP Jabber Identifiers (JIDs) running in XMPP clients. M2M services are running in M2M overlay entities. XMPP's SD mechanisms do not offer the possibility for discovering unknown overlay entities from unknown rosters, which is discussed more in details in Section II. The protocol presented in this paper provides a solution to this problem. Later in this paper this M2M overlay entity discovery protocol is called only service discovery (SD) protocol.

The rest of this paper is organized as follows: Section II gives information about XMPP and existing related SD protocols, Section III introduces the main contribution, the SD protocol and how XMPP is used, Section IV gives evaluation and presents some of the use cases, which were used to test the protocol, while Section V concludes the paper and gives proposals for future work.

II. Related work

XMPP is a set of open XML technologies for presence and real-time communication. It is continuously extended through the standardization process of XMPP Standards Foundation. XMPP was originally created for near-real-time messaging, presence, and request-response services [3][4], but it has been used to build e.g., Smart Grids [5], M2M architectures [6][7], and sensor networks [8]-[13].

XMPP offers built-in publish-subscribe ("pubsub") functionality [14], so polling is not necessarily required between clients and services. The specification of pubsub is long, but the idea is simple: 1) An entity publishes information to a node at pubsub service, and 2) the pubsub service pushes a notification to all entities that are authorized to learn about the published information.

XMPP uses Transport Layer Security (TLS) to secure server-to-server and client-to-server connections [2] and offers end-to-end signing and object encryption [15]. In addition to these, security extensions exist or are proposed, including e.g., privacy [16]. XMPP is distributed; anyone can have their own servers with several clients. Client-server architecture commonly enables connectivity through firewalls, because clients initiate sessions. XEPs such as [17][18] exist to help with firewalls.

XMPP has four presence subscription statuses. In 'none' state the user does not have a subscription to the contact's presence information, and the contact does not have a subscription to the user's presence information. In other words, you are not interested in the item's presence, and the item is not interested in yours. In 'to' state the user has a subscription to the contact's presence information, but the contact does not have a subscription to the user's presence information. In 'from' state the contact has a subscription to the user's presence information, but the user does not have a subscription to the contact's presence information. In 'both' state both the user and the contact have subscriptions to each other's presence information [2]. These statuses do not tell anything about how (if at all) your presence information should be told by your contacts to their contacts or contacts unknown to you and/or to your contacts. This paper discusses how these kinds of situations with existing XMPP statuses should be handled and how new presence subscription statuses could be used to improve privacy wishes.

XMPP creates long-lived sessions between communicating entities. Sessions might shut down because of many reasons and repeating session initialization might be problematic and/or slow. XMPP uses XML stanzas, and XML parsers may become a bottleneck in embedded devices and in networks that have limited bandwidth capabilities. However, Binary XML might become more common in the future: XMPP is presented as one use case in WC3's XML Binary Characterization Use Cases [19] and it has been proposed to be used with Efficient XML Interchange (EXI) [20]. If XMPP is used in mobile devices in the same way as in PCs with fixed power, problems such as battery running out will certainly arise. Because of this, an extension providing knowledge of mobile handset behavior has been described [21]. In addition, other XEPs to decrease bandwidth exist, e.g., Stream Compression [22]. Although XML takes resources, the smallest interoperable XMPP client implementations work in embedded devices such as sensors [8]-[13], and there are client and server implementations for mobile phones.

XMPP has two different types of SD protocols [23][24]. In the basic XMPP SD protocol [23] entities are servers, clients or gateways. The protocol provides methods for 1) discovering entities (disco#items) and for 2) discovering features supported by a given entity (disco#info). An XMPP client knows at least one other entity, its server. The client is able to discover services (multi-user chatrooms (MUC) [25], pubsub, etc.) offered by the server and features that are supported in those services. XMPP's serverless messaging specification [24] defines mechanisms that enable working without servers, e.g., in body area networks, or LANs: Principles of zero-configuration networking (Zeroconf) [26] are used. Zeroconf uses multicast DNS (mDNS) [27] and DNS-Based Service Discovery (DNS-SD) [28]. In XMPP roster item exchange can be done e.g., with service administration [29] or roster-item exchange [30].

In social networks, discovery protocols have been used for finding friends, groups, links, etc. "Google's Search plus Your World" (earlier "Google Social Search") [31] has features that allow your friends to affect your search results. The SD presented in this paper lets the XMPP client answering the SD request to build the answer. Facebook social search [32] includes information about the frequency of clicks on the search results by members of the social network who are within a predetermined degree of separation from the member who submitted the request. This degree can be compared to a hop limit in the presented SD protocol.

III. THE SD PROTOCOL

This section presents the main contribution of this paper, the SD protocol. The purpose of the SD protocol is to discover M2M overlay entities from unknown M2M overlay entities without compromising privacy. Protocol must work without centralized naming services, which keep track of M2M overlay entities, it must not broadcast huge amount of messages and flood the network, and it must work over XMPP, but also in XMPP clients without support for service administration [29] or roster-item exchange [30]. This paper presents a SD protocol that meets these requirements, using ideas coming from the following real life examples:

A tap in Eemil's bathroom has started to leak and has caused moisture problems in the bathroom. Eemil wants to find a person who could fix these problems. If he already knows someone who has fixed bathrooms or taps before, he will probably ask these people to help first. If Eemil does not know anyone who is able to fix his bathroom, he might ask from his friends if anyone has a friend who has fixed bathrooms before. If one of his friends knows such a person, it is likely that he or she gives this information to Eemil. If none of Eemil's friends know such a person, they can ask from their friends. If any of Eemil's friends' friends know such people, they can send this information directly to Eemil (if Eemil's friends have told who is the original requester), or to the last requester (Eemil's friend who asked it) and they can forward the reply to Eemil. When thinking the example further, Eemil can select friends from whom he wants to ask the fixer. If Eemil thinks that some of his friends should certainly know about people working with bathroom fixing, or that some of his friends know more people than an average person, he will probably ask from them first.

Based on this real life example and to enable discovering unknown entities from unknown M2M overlay entities, the hop limit was selected to be two. Hop limit is analyzed in Section IV.

XMPP offers basic building blocks listed in Section II, such as message transmission, naming of M2M overlay entities (nodeid@domainid/resourceid) and rosters, for the protocol. In used M2M service network, every XMPP JID in XMPP client has its own private roster, which is called a private M2M overlay. Rosters are stored to XMPP servers. XMPP servers can be clustered. Roster includes other XMPP JIDs. The overlay can be constructed in several ways: by adding roster items after registration, using different presence subscription statuses [2], or accepting all XMPP client subscriptions from the XMPP server it is registered to. These are mainly implementation and configuration issues. MUC [25] rooms can be thought of as second type of a M2M overlay. Describing usage of MUCs in the SD protocol is out of the scope of this paper. A M2M device runs one or more XMPP clients. An XMPP client has one or more JIDs registered, which can be registered to one or more XMPP servers.

The following list contains required features of a M2M overlay entity, which are not offered by XMPP:

1. The entity can ask from its contacts to parse a string presenting the wanted M2M overlay entity name.
2. The entity can forward the parsing request to its contacts.
3. If the entity finds a wanted string from its roster list, and contacts in the answer accept forwarding their information, the entity answers to the requester, who might be the original requester or forwarder of the parsing request.

4. The entity can parse rosters with additional information such as reputation.

The SD protocol works at the application layer inside an XMPP client implementation. It can be categorized as a unicast protocol, as it sends direct discovery messages to known receivers. Its purpose is to discover M2M overlay entities presented as JIDs.

Our proposal is that when using XMPP's four existing presence subscription statuses “none”, “from”, “to”, and “both”, by default the entities should not advertise the existence or the presence of one another to anyone else.

Because some, but not all, M2M overlay entities or their owners might want to share their JIDs to unknown entities, there is a need for new XMPP presence subscription statuses. They could be such that the entity replying to a SD request knows if the JID in the answer allows giving its name to other entities. At least the following new additional presence subscription statuses to XMPP are needed:

subscription='from-anyone': This means the same as the XMPP state ‘from’ but also that your contact can advertise you and your presence information to anyone. Notice that this does not mean that anyone who tries to subscribe to your presence information is necessarily accepted by you.

subscription='from-contacts': This means the same as the state ‘from’, but also that your contact can advertise you and your presence to any of her contacts.

subscription='both-anyone': You and the contact are interested in each other's presence information and can advertise it to anyone.

subscription='both-contacts': You and the contact are interested in each other's presence information and can advertise it to your own and your contact's contacts.

Adding more new subscription statuses such as advertising presence only to certain server is possible, and instead of using new subscription statuses, new fields can be used to tell about these privacy wishes.

The SD protocol uses one-way RPC. It is a variant of asynchronous RPC in which the client continues immediately after sending the request to the server without waiting for the server's acknowledgement [33]. The SD protocol messages are formatted as JSON-RPC [34]. JSON-RPC is a remote procedure call protocol following the same principles as XML-RPC [35]. JSON-RPC's “notification” provided asynchronous RPC for the SD protocol. Notification is a special request which does not have a response, it has same properties as request object except the id must be null. In the prototype, the SD messages are transmitted and processed as XMPP Instant Messaging (IM) messages in XMPP clients.

Some XMPP clients might have service administration [29] or roster-item exchange [30] support, and they can be used with the SD protocol. For instance, if there is an entity authorized to get all rosters from the server, it can give more comprehensive answers for queries.

The SD protocol's JSON-RPC messages' params field has to include at least information about the searched string. A forwardedParseRoster method call's params field includes also information about the original sender. This enables XMPP clients to reply directly to the original parseRoster message sender. Examples of JSON-RPC formatted notification messages are presented in Table 1.

TABLE 1. JSON-RPC FORMATTED NOTIFICATION MESSAGES

{“method”: “parseRoster”, “params”: [{“value”: “weather”}], “id”: null}
{“method”: “forwardedParseRoster”, “params”: [{“value”: “weather”, “originalsender”: “tempsensor.313@vtt.fi/kaitovayla1”}], “id”: null}
{“method”: “reply”, params”: [{“value”: weatherservice@vtt.fi”}], “id”: null}

In a pseudo code, the SD protocol works as presented in Table 2.

TABLE 2. THE SD PROTOCOL IN PSEUDO CODE

```
STRING wanted_service;
XMPP ROSTER own_roster;
JSON-RPC MESSAGES parseRoster, forwardedParseRoster, reply;

IF (own_roster includes wanted_service) Service is discovered;
ELSE Send parseRoster request to selected contacts in own_roster;

IF (proper reply is received) {
      IF (reply includes wanted_service) Service is discovered;
}

IF(proper parseRoster function request is received) {
      IF (own_roster includes wanted_service) Service is discovered
      and reply is sent to the requester.
      ELSE Send forwardedParseRoster request to selected contacts in
      own_roster;
}

IF (proper forwardedParseRoster function request is received) {
      IF (own_roster includes wanted_service) Service is discovered
      and reply is sent to the original requester;
}
```

IV. EVALUATION

The SD protocol was tested in different use cases with 1-3 modified XMPP clients. Each client had 1-3 XMPP accounts (JIDs) registered to 1-3 XMPP servers. Three unmodified XMPP clients were used for debugging.

A. Hop limit

This section presents two use cases used to test the SD protocol and to get information about the hop limit. Figures 1-2 present use cases, where users A, B, C, D, and E are M2M overlay entities, and boxes next to them represent contacts in their rosters. Each JID has their own different M2M overlay (JIDs in their roster).

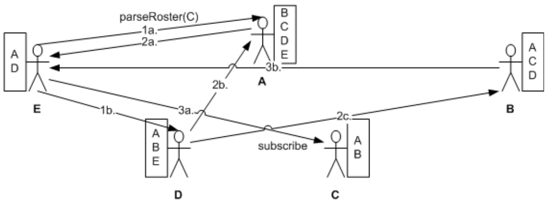


Figure 1. E tries to find C, hop limit is unlimited.

Figure 1 presents an example, in which E tries to find C, or a string that is in C's JID.

1. E sends a roster parsing requests to its contacts A (1a) and D (1b). It could be possible to send them only to the ones who are currently available, not e.g., in Do-Not-Disturb (DND) status.
2. A and D parse their own rosters. A's roster matches and it sends a reply (2a) to E. A and B have already done a presence subscription. Same way D checks its roster and it does not find C. It forwards the request to its contacts A (2b) and B (2c), except E, who is the original requester.
3. B receives the request from D, it parses its own roster, match is found, and it replies to E. The reply includes B's parsed roster and information about C. At this point B could also subscribe to E, and/or exchange roster item [30]. At the same time E already subscribes to C (3a), because it got information about C from A (2a).

Optimizing the hop limit is a complex problem. Using unlimited hop limit was not possible because without proper timeouts it can flood the network and jam devices. One requirement was that the protocol must be able to discover unknown entities from unknown M2M overlay entities. This means that the hop limit must be at least two. If the hop limit is one, the protocol enables discovering unknown entities only from known M2M overlay entities (your contacts).

If thinking about real world, sharing things with or borrowing them to your friends is usually ok. Section III presented an example of Eemil finding a fixer, in which a maximum of two hops was used. Then again, two hops might be too much because borrowing things to or sharing them with friends of you friends might be something most people do not want to do.

When moving these thoughts of human behavior from real-life to M2M and to the SD protocol, it was decided that a limit of two hops would be used: If a wanted M2M overlay entity is not found from contacts, the entity can ask it from its contacts (the first hop). If contacts do not know it, they can forward a message to ask the wanted entity from their contacts (the second hop). If they do not know it, the wanted M2M overlay entity is not found. Selecting two also keeps the protocol as simple as possible but still fulfills the discovery requirement.

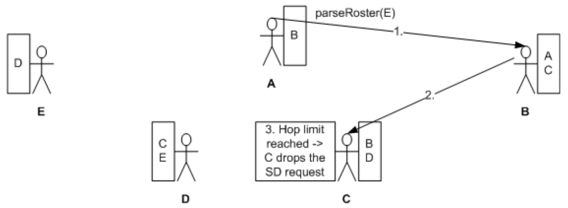


Figure 2. A tries to find E, hop limit is 2.

Figure 2 presents an example in which the M2M overlay entity E is not found, because of the hop limit is two, and because C does not know E. As it can be seen, if M2M overlay entities have only few contacts in their rosters, discovery process is short, may not succeed and only few messages are sent.

B. Replying mechanism

After selecting the hop limit, the amount of transmitted messages was calculated. In the worst case situation, the maximum amount of SD and forwarded SD messages (with considering neither XMPP nor TCP/IP acknowledgements etc.) can be calculated using (1)-(8).

M = *Maximum amount of messages.*
α = *Number of unique contacts in the roster of the SD request message sender A.*
β = *Number of unique contacts in the roster(s) of unique contact(s) of A.*

Case A: All XMPP clients are registered to the one and same XMPP server. Replies are sent directly to the original sender (1) or through the SD request forwarder (2):

$$M = 2\alpha + 4\alpha\beta. \qquad (1)$$

$$M = 2\alpha + 6\alpha\beta. \qquad (2)$$

Case B: The SD request sender XMPP client is the only client registered in the first XMPP server and two other XMPP clients are registered in the one and same XMPP server. Replies are sent directly to the original sender (3) or through the SD request forwarder (4):

$$M = 3\alpha + 5\alpha\beta. \qquad (3)$$

$$M = 3\alpha + 7\alpha\beta. \qquad (4)$$

Case C: 2 XMPP clients (the SD request sender and the first contact) are registered in the one and same XMPP server, and the third XMPP client in another XMPP server. Replies are sent directly to the original sender (5) or through the SD request forwarder (6):

$$M = 2\alpha + 6\alpha\beta. \qquad (5)$$

$$M = 2\alpha + 8\alpha\beta. \qquad (6)$$

Case D: Every XMPP client is registered in its own separate XMPP server. Replies are sent directly to the original sender (7) or through the SD request forwarder (8):

$$M = 3\alpha + 6\alpha\beta. \qquad (7)$$

$$M = 3\alpha + 9\alpha\beta. \qquad (8)$$

Using α=β=16 in (1)-(8) a chart in Figure 4 has been drawn. It presents total amount of SD messages (requests and responses). Y axis presents the amount of transmitted SD messages. X axis presents cases A, B, C, D and (1)-(8). In cases C and D the amount of SD messages are approximate when answering directly to the requester (5) and (7). In fact, (5)/(7) approaches 1 when α and β approach infinity.

The difference between direct replies and sending replies through the forwarder can be calculated by subtracting answers in each case A, B, or C. For example, when using three servers with α=32 and β=32 (7) gives M = 6240 and (8) gives M=9312, so the difference is 3072. When α and β approach infinity, result of (2)/(1) approaches the ratio 1,5. (4)/(3) approaches the ratio 1,4, (6)/(5) approaches the ratio 1+1/3 and (8)/(7) approaches the ratio 1,5.

If communication between clients and servers is not taken into consideration, in the simplest case when α=β=1, only three messages are transmitted: 1) from the SD request sender A to its contact B, 2) B forwarded the SD request to its contact C, and 3) C replies to A (9). If C would answer through B, there would be 4 messages instead (10). These are presented in Figure 5.

$$M = \alpha + 2\alpha\beta. \tag{9}$$

$$M = \alpha + 3\alpha\beta. \tag{10}$$

When α=β=1 and all clients are using only one server, the number of messages increases to 6 (1), with two servers to 8 (3) and (5) and with three servers to 9 (7) respectively. If the contact C would send the reply through B, numbers would be with one server 8 (2), with two 10 (4) and (6) and with three 12 (8). These cases are described in Figures 6-8. S1, S2, and S3 are servers.

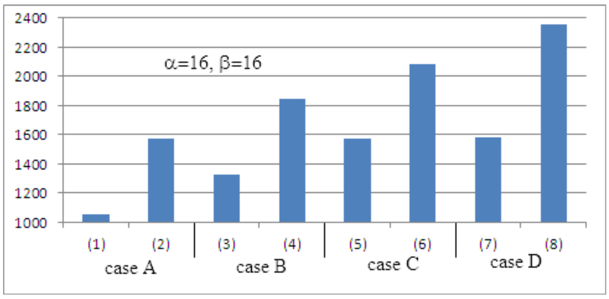


Figure 3. Number of SD messages, α=β=16.

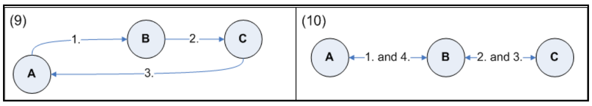


Figure 4. Number of SD messages, no servers, α=β=1.

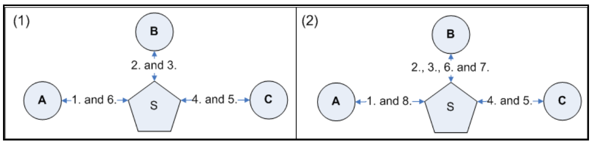


Figure 5. Number of SD messages, one server, α=β=1.

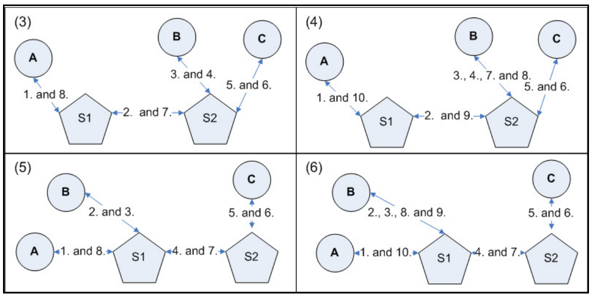


Figure 6. Number of SD message, two servers, α=β=1.

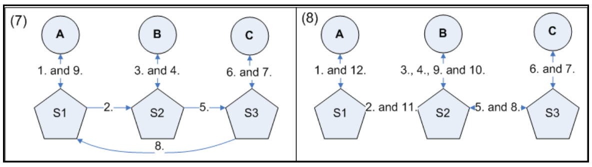


Figure 7. Number of SD messages, three servers, α=β=1.

To save messages and bandwidth, the first approach was selected: XMPP clients reply directly to the original SD requester.

C. *Answer handling*

When the M2M overlay entity tries to find strings that are too commonly used in JIDs, the SD request sender is likely to get answers including several JIDs, from several different senders. In some cases, this can also be thought of as an advantage if entities known by several other entities are thought to be more popular, and as such also more suitable. The SD request message sender could use this suitability information to categorize found JIDs. Currently the JID in the first reply is selected.

D. *Security*

XMPP offers security services for the M2M service network, but this section includes information about new security issues coming from the SD protocol. In one-way RPC the original sender cannot know for sure whether its request will be processed if the reliability is not guaranteed, but this also allows the requester to be able continue its work without the need to wait for the reply. Direct replies to the original SD requester generate threats, related to forging the original sender. This makes flooding and Denial of Service (DoS) attacks possible. Risk of the threats can be decreased in servers, by asymmetric cryptography and processing only certain messages (including proper information or coming from certain domain, for example). In real life, when answering directly to the original requester, request forwarders would not necessarily get information if the answer is found.

V. Conclusion and Future work

XMPP offers several building blocks, such as naming, rosters, message transmission and remote commands, pubsub, SD and security for M2M service networks, but no mechanism for finding roster entities from unknown XMPP entities. Therefore, this paper presented a simple M2M overlay discovery protocol for discovering XMPP JIDs behind several hops, in an XMPP based M2M service network. Hop limit was two, which was selected based on real life examples and in order to keep the protocol simple, but so that it also fulfilled the discovery requirements. The amounts of messages with synchronous or asynchronous RPC were analyzed. Asynchronous one-way RPC and direct replies to original requester were selected to decrease the amount of messages. Each entity or owner of the entity should be able to choose whether its presence and/or JID are shared to unknown entities, or not. The proposed four new XMPP subscription statuses enable describing when information can be advertised only to contacts or to anyone.

Future work includes applying some ideas of the paper to be submitted to XEPs. Mechanisms for handling different JIDs received in different SD replies must be designed and implemented. The SD protocol presented in this paper uses JSON-RPC formatting [34], but the format of the SD messages can be changed to Jabber-RPC format [35][36]. Distributed Hash Tables (DHT) could be used to enable serverless communication between XMPP nodes. Xeerkat [37] is one example implementation of a P2P computing framework that utilizes XMPP as a communication protocol.

Acknowledgment

Author wants to thank UseNet project [38] members and QXmpp [39] and ejabberd [40] developers. In the prototype, QXmpp XMPP client was modified, and Ejabberd was used as an XMPP server.

References

[1] IPSO Alliance’s webpage, http://www.ipso-alliance.org, 16.04.2012

[2] IETF RFC 6120, P.Saint-Andre, “Extensible Messaging and Presence Protocol (XMPP): Core”, March 2011

[3] P. Saint-Andre, K. Smith, and R. Tronçon, “XMPP: The Definitive Guide, Building Real-Time Applications with Jabber Technologies”, O’Reilly, 2009

[4] P. Saint-Andre, "XMPP: lessons learned from ten years of XML messaging," Communications Magazine, IEEE, vol. 47, no. 4, pp. 92-96, April 2009, doi: 10.1109/MCOM.2009.4907413

[5] M. R. Lavelle, (2010), “Micro-Grid Applications - Leveraging XMPP Short Messaging”, http://www.lavelleenergy.com/documents/Micro-grid%20Applications%20Paper.pdf, 16.04.2012

[6] M. Kuna, H. Kolaric, I. Bojic, M. Kusek, and G. Jezic, "Android/OSGi-based Machine-to-Machine context-aware system," Telecommunications (ConTEL), Proceedings of the 2011 11th International Conference on, pp. 95-102, 15-17 June 2011

[7] QEES Open Software Platform website, http://qees.dk/da/services/open-software-platform, 16.04.2012

[8] T. Parkkila, (2005), “Application and platform management of an embedded system”. Smart Systems 2005, Conference Proceedings. Seinäjoki, 3-4 May 2005.

[9] A. Hornsby, P. Belimpasakis, and I. Defee, "XMPP-based wireless sensor network and its integration into the extended home environment," Consumer Electronics, 2009. ISCE '09. IEEE 13th International Symposium on, pp. 794-797, 25-28 May 2009, doi: 10.1109/ISCE.2009.5156807

[10] A. Rowe, M. Berges, G. Bhatia, E. Goldman, R. Rajkumar, L. Soibelman, J Garrett, and J. M. F. Moura, “Sensor Andrew: Large-Scale Campus-Wide Sensing and Actuation”, Carnegie Mellon University, 2008

[11] xbee-xmpp website, http://code.google.com/p/xbee-xmpp/, 16.04.2012

[12] A. Hornsby and E. Bail, "µXMPP: Lightweight implementation for low power operating system Contiki," Ultra Modern Telecommunications & Workshops, 2009. ICUMT '09. International Conference on, pp. 1-5, 12-14 Oct. 2009, doi: 10.1109/ICUMT.2009.5345594A.

[13] P. Saint-Andre, “XEP-xxxx: Sensor-Over-XMPP”, XEP proposal, V 0.0.18

[14] P. Millard, P. Saint-Andre, and R. Meijer, “XEP-0060: Publish-Subscribe”, V 1.13

[15] IETF RFC 3923, P. Saint-Andre, “End-to-end Signing and Object Encryption for the Extensible Messaging and Presence Protocol (XMPP), October 2004

[16] P. Millard and P. Saint-Andre, “XEP-0016: Privacy Lists”, V 1.6

[17] I. Paterson, D. Smith, P. Saint-Andre, and J. Moffitt, “XEP-0124: Bidirectional-streams Over Synchronous HTTP (BOSH)”, V 1.10

[18] I. Paterson and P. Saint-Andre, “XEP-0206: XMPP Over BOSH”, V 1.3

[19] XML Binary Characterization Use Cases, http://www.w3.org/TR/xbc-use-cases/#xmpp, 16.04.2012

[20] P. Saint-Andre, “XEP-xxxx: Stream Compression with Efficient XML Interchange”, XEP proposal, V 0.0.1

[21] D. Cridland, “XEP-0286: XMPP on Mobile Devices”, V 0.1

[22] J. Hildebrand and P. Saint-Andre, “XEP-0138: Stream Compression”, V 2.0

[23] J. Hildebrand, P. Millard, R. Eatmon, and P. Saint-Andre, “XEP-0030: Service Discovery”, V 2.4

[24] P. Saint-Andre, “XEP-0174: Serverless Messaging”, V 2.0

[25] P. Saint-Andre, “XEP-0045: Multi-User Chat”, V 1.25

[26] Zero Configuration Networking (Zeroconf) webpage, http://www.zeroconf.org, 16.04.2012/

[27] S. Cheshire and M. Krochmal, “Multicast DNS”, IETF Internet-Draft: draft-cheshire-dnsext-multicastdns-15, Dec 9, 2011, Expires: June 11, 2012.

[28] S. Cheshire and M. Krochmal, “DNS-Based Service Discovery”, IETF Internet-Draft: draft-cheshire-dnsext-dns-sd-11, Dec 9, 2011, Expires: June 11, 2012.

[29] P. Saint-Andre, “XEP-0133: Service Administration”, V 1.1

[30] P. Saint-Andre, “XEP-0144: Roster Item Exchange”, V 1.0

[31] Google Search Plus Your World website, http://www.google.com/insidesearch/plus.html, 16.04.2012

[32] US Patent 7890501, C. Lunt, N. Galbreath, & J., “Visual tags for search results generated from social network information”

[33] A. S. Tanenbaum and M. van Steen. 2006. “Distributed Systems: Principles and Paradigms (2nd Edition)”. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

[34] JSON-RPC specification, http://json-rpc.org/wiki/specification, 16.04.2012

[35] XML-RPC specification, http://xmlrpc.scripting.com/spec.html, 16.04.2012

[36] DJ Adams, “XEP-0009: Jabber-RPC”, V 2.2

[37] Xeerkat webpage, https://code.google.com/p/xeerkat/, 16.04.2012

[38] UseNet project website, https://usenet.erve.vtt.fi, 16.04.2012

[39] QXmpp website, https://code.google.com/p/qxmpp/, 16.04.2012

[40] ejabberd community site, http://www.ejabberd.im/, 16.04.2012

An Interactive Game Using Wii Remote for Visually Impaired People

Huai-Yong Deng
Department of Applied Information Sciences
Chung Shan Medical University,
Taichung 402, Taiwan, R.O.C
gloria10074@hotmail.com

Hsiao Ping Lee
Department of Applied Information Sciences
Chung Shan Medical University,
Department of Medical Research
Chung Shan Medical University Hospital,
Taichung 402, Taiwan, R.O.C
ping@csmu.edu.tw

Jun-Te Huang *
Department of Applied Information Sciences
Chung Shan Medical University,
Taichung 402, Taiwan, R.O.C
yaya9438@hotmail.com

Abstract*— Number of visually impaired people around the world is increasing year by year. However, companies that develop the games for visually impaired people are very few. For the visually impaired people, we have to create a fair environment, and enhance their entertainment. In recent years, somatosensory device has been introduced. Through the body to control the game is more attractive than the traditional joystick. In this paper, we design an interactive game for the visually impaired people, through somatosensory device and infrared-gun, and combined with Text-to-Speech technology to implement a flash game for the visually impaired people's to enhance entertainment quality of daily life. We invited five subjects to play this game, and complete a questionnaire and interview after the experiment. Our questionnaire divided into three parts: interested, audio guide and novelty of the game. The results showed that most subjects were very interested in interactive games. Through Text-to-Speech technology, most of the subjects were able to smoothly play the game with audio guide. A lot of subjects feel the game is very useful and innovative.

Keywords* - *Visual Impairment People; Somatosensory Device; Interactive Game; Text-to-Speech.

I. INTRODUCTION

Vision is the most important and the natural way for humans to receive the message from the environment. We rely on vision to handle most things in daily life. Not only are the assistive devices, the amount of the games for the visually impaired people, are very few, too. Currently, most entertainment activities for the blind people are static, such as e-book. Moreover, most are have inconvenience user interface. For these reasons, considering the demands of visually impaired people, and design an interactive game for the visually impaired people is very important.

The “Wii” has been introduced by the Nintendo company in 2006 [1, 2]. Different to the traditional games, Wii uses the wireless controller to control it. In this way, it is providing a new way to play the games. Moreover, the wireless controller has a friendly user interface; the player is able to control instinctively. For these properties, the Wii’s controller can be used in other domains, such as education and physical therapy.

In this paper, we are going to develop an interactive game for visually impaired people. The game can assist them to play the game by wireless controller, audio and text-to-speech technology (TTS). In addition, this game will improve and be more interesting for the visually impaired people. For the game testing, we invite five persons to provide us assistance, such as play this game and give some suggestion. Finally, we got some data which about game testing from these people. So far, it is perceived that this game “very interesting”, "have a good voice and sound guidance", and "has novelty" The paper is going to introduce related work about our game and how we implement this game.

II. BACKGROUND AND RELATED WORK

We should to know some sensors and development kit before implement the game. First, the Wii’s controller that named “Wii Remote [3, 4].” In the shape of Wii Remote is look like a bar, it is a small and light handset. It can be divided into three modules: the G sensor [5, 6], CMOS [7] IR receiver and Bluetooth module [8]. The G sensor can calculate X, Y and Z axis acceleration. Next, determine the player's actions on real time. The CMOS IR receiver can receive the infrared from the sensor bar. Then, we find the different position of infrared point to determine controller’s movement. After, Wii or computer handles these data which receive from Bluetooth.

Secondly, if we to drive the components which on the Wii Remote, and communicate with computer, it should use the Wiiflash [9] API to do that. The Wiiflash API is an open source development kit; it is good for designer to write the flash program which linked the computer and Wii Remote. About this development, the kit includes: (1) Wiiflash API library, (2) Wiiflash Server: providing a way and the computer is able to communicate with Flash, (3) examples, and (4) documents.

III. STATE OF ART

Today, the studies for the Wii Remote are more than before. Some people have used the Flash to develop many fun games, and control by Wii Remote [10]. Yi-nung Lin designs a pantomime games for children with cerebral palsy. They wear the simply IR transmitter on their wrist, and lead those children to do rehabilitation with the interactive teaching materials. Children were felt interesting in these teaching materials. So they learn how to play this game really hard.

About the research of the infrared sensor on the Wii Remote, Johnny Lee’s Wiimote whiteboard [11] is the first study. Since the Wii Remote can track the sources of

ISBN: 978-1-61208-202-8

infrared light, Johnny Lee make a pens that have an IR led in the pen’s tip. Wii Remote can combine with IR pens, rear-projected, and notebook, it will become a low-cost interactive whiteboards.

Chien-Yu Lin et al. [12] wore a simple infrared emitter for children who is a cerebral palsy usually. It is very interest for her to use interactive teaching materials, so she works hard to raise her hand and do some exercise. The most current study on the Wii Remote is research in education and rehabilitation. In this paper, we design set of interactive games for the visually impaired people and our purpose is to increase the quality of entertainment for the visually impaired. We combine the Wii Remote with Flash, and designed a game like "Whac-A-Mole". We redesign the mode of operation of the game. We consult the concept of infrared pen, to design an infrared gun, so that the visually impaired people have more immersive in play his game. To allow for the visually impaired can play game alone. Interactive game system that we purpose also has the following function and characteristics:

- Easy to use, visually impaired people can operate independently without others help.
- Use speech in the game to identify the location of hamster through audio.
- Everyone can adjust the scope of the game to increase the applicability of system fitness.

IV. System Design

We develop an interactive game for visually impaired people that use the Wii Remote and infrared-gun. For the visually impaired people to have interest in the game, we add the audio and speech.

A. *Experimental Equipment*

1) Wii Remote: Unlike most operate method of Wii Remote general; we use the CMOS sensor camera and Bluetooth to receive the location data from infrared-gun. Then CMOS sensor camera can send the data to the computer.

2) *Infrared-gun:* The process of making infrared-gun:

a) Soldering the wire with the infrared LED's positive pin and negative pin.

b) Add a button in the middle of the negative wire to control the flow of the circuit.

c) Connect the power with wire.

d) Complete. The infrared-gun image is presented in Figure 1.

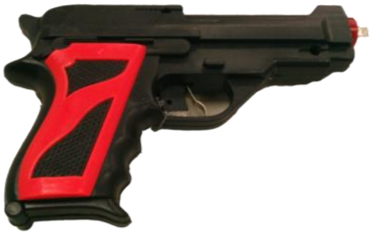


Figure 1. The infrared-gun.

B. *System Architecture*

In this paper, the computer is our data server. When visually impaired people move the infrared-gun, Wii remote will send the location value to data server through Bluetooth technology. Then Wii Flash server can interpret a position value and transmit the value into flash. Flash can process the data and display in computer screen. The position value can determine whether the gun hit the mouse. The system architecture is presented in Figure 2.

C. *Game Flowchart*

We designed a game like "Whac-A-Mole" and use Flash CS3 as our development environment tool. Therefore the code written is ActionScript3.0 [13, 14]. Because of this game is design for visually impaired people, we use audio and speech in game as the main navigation tool to help the visually impaired people to operate the game. The game flow chart is presented in Figure 3.

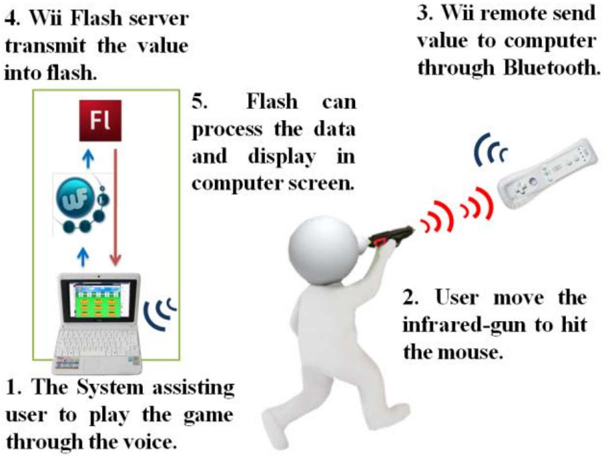

4. Wii Flash server transmit the value into flash.
3. Wii remote send value to computer through Bluetooth.
5. Flash can process the data and display in computer screen.
Fl
2. User move the infrared-gun to hit the mouse.
1. The System assisting user to play the game through the voice.

Figure 2. The system architecture.

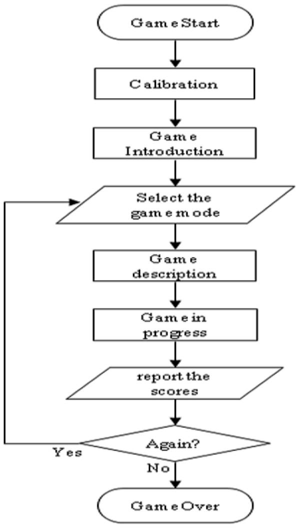


Figure 3. Game flow chart.

The screen of game beginning is presented in Figure 4. After entering the game, that will enter calibration mode first, and assisting visually impaired people to stand the best place and to confirm the game through the speech. Then, let visually impaired people to set the range of the game. When calibration is completed, the system will explain how to play the game by audio message, in this way, visually impaired people can choose the game mode and into the game.

Game screen is divided into nine blocks with squared, is presented in Figure 5. When the hamster appears, the visually impaired people can identify the site of hamster through the different voice, then the visually impaired people use infrared gun to shoot hamster. If successfully shoot, the system will make a sound, and increase the score.

The game were divided notes, such as “Do, Re, Me” and music two modes. In the notes mode, hamster will randomly appear; in the music mode, the system will choose a song, then the hamster will follow the song’s tone appearing. When the visually impaired people shoot hamster successfully, the system will make the next one to appear. When the visually impaired people shoot completely all the tunes, the system will calculate the time it takes and convert into the corresponding scores.

When the game is over, the system will calculate the scores and read out, and then, asked the visually impaired people whether playing again. If the answer is yes, the game will back choose the game mode. The over screen see the Figure 6.

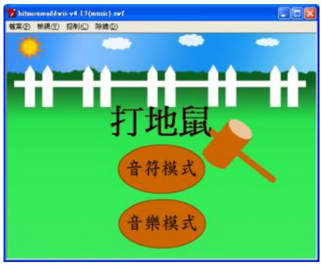

hitmouseaddwii-v4.13(music).swf
檔案(F) 檢視(V) 控制(C) 除錯(D)
打地鼠
音符模式
音樂模式

Figure 4. Start of the game.

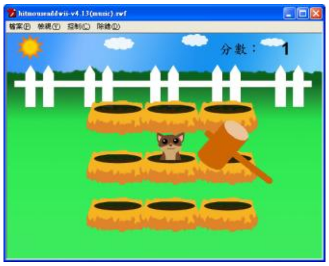


Figure 5. Game screen.

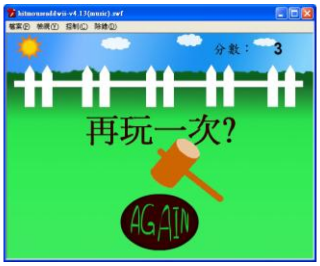


Figure 6. End of the game.

D. *Experimental Design*

Visually impaired people are the main user for this game. In addition, the blind people are harder to interview than the sighted people. When the game testing, we ask the subjects to wear goggles or close the eyes to making simulation. Before the game starting, we tell to user why we designed this game. And then, the user enters and plays the game.

Another important is, we want to know whether the users are able to follow the audio message to play the game or not. Therefore, we record the time when the users starting the game until the game over.

V. RESULTS

In this paper, we use the questionnaire and the user's experiences as a basis for experimental analysis. We invited five persons to be user and test the game. Theirs age from 11 to 50 years old. The average test time is 252 seconds and only one user restart the game, because he forgot the range which he set, and the other users caught a game complete.

A. *Questionnaire*

Table I is the results which investigating by questionnaire from five users. Our questionnaire is using Likert scale and it has three questions. The first one, there are 80 percent users fell that this game is funny and very interesting. The second, about guidance, the game has audio message and voice to guide user, there are 60 percent feel that this game is smooth. And the third, all users are thinking this game is very novelty.

TABLE I. RESULT OF QUESTIONNAIRE

	Strongly agree	**Agree**	**Neither agree nor disagree**	**Disagree**	**Strongly disagree**
Do you think this game can catching your eye	1	3	1	0	0
Do you think the s peech and sound can guidance you play the game smoothly	2	2	1	0	0
Do you think the game is novelty	4	1	0	0	0

B. *Experience*

Besides the questionnaire, we also asked the user about the experience or suggestions in this game.

We found that the most of the users had a high degree of interest in this game. After the end of the experiment, the users will be asked to play again. A few users share the experience with others.

Most of the users believe that this game is novel because they never had an experience to play game like that. However, a few users expressed that they did not find the scope of games when game is started. And they cannot continue the game.

In the future we may refer to these comments, and fix the game problems to remind users who cannot find the scope of game.

VI. CONCLUSION

In this paper, we developed an interactive game for visually impaired people. The game is equipped on the computer, and let the Wii Remote be an IR receiver. The blind people just need taking an infrared-gun target at Wii Remote’s CMOS receiver. The communication between the computer and Wii Remote, the game used WiiFlash API to drive the sensors which on the Wii Remote. And then, we also send the position information that got from CMOS receiver. On the other hand, we have designed a friendly user interface which using text-to-speech and voice to guide the blind user. After the game prototype was completed, we invited five users to play and test game. In the results, 80 percent users think the game is funny and interesting, 60 percent think the audio message and voice guide let game smoothly, and all users think the game is really novelty. In the future, we are going to improve the game’s quality and let it work well.

ACKNOWLEDGMENT

The authors would like to thank the National Science Council of the Republic of China, Taiwan R.O.C, for financially supporting this research under Grants NSC 100-2218-E-040-001.

REFERENCES

[1] Nintendo of America Inc, http://www.nintendo.com/wii, Retrieved 01, 2012.

[2] WII, http://zh.wikipedia.org/zh-hk/Wii, Retrieved 12, 2011.

[3] Farkas, Bart G., The Nintendo Wii Pocket Guide, CA: Addison-Wesley, April 2007.

[4] Elaine Pearson, Chris Bailey, Evaluating the potential of the Nintendo Wii to support disabled students in education, Proceedings ascilite, Singapore, 2007.

[5] Accelerometer, http://en.wikipedia.org/wiki/Accelerometer, Retrieved 01, 2012.

[6] A. Miskam, S. Korakkottil, M. Zaidi and O. Sidek, “Development of a Tilt Measurement Unit Using Microelectromechanical System Accelerometer,” Journal of Applied Sciences, vol. 9, 2009, pp. 2451-2456, doi:10.3923/jas.2009.2451.2456.

[7] CMOS, http://en.wikipedia.org/wiki/CMOS, Retrieved 12, 2011.

[8] Bluetooth, http://en.wikipedia.org/wiki/Bluetooth, Retrieved 01, 2012.

[9] Wiiflash, http://code.google.com/p/wiiflash/, Retrieved 07, 2011.

[10] Perkins, Todd. Nintendo Wii Flash Game Creator’s Guide: Design, Develop, and Share Your Games Online. McGraw-Hill Osborne Media, April 2007.

[11] Low-Cost Multi-point Interactive Whiteboards Using the Wiimote, http://johnnylee.net/projects/wii/, Retrieved 01, 2012.

[12] Chien-Yu Lin, Shu-Hua Chen, Min-Ju Wu, Yi-Shan Liao, Shu-Ling Hsien, and Chian-Huei Guo, “Application of Interactive Interface Design on Rehabilitation for Children with Cerebral Palsy,” Future. Communication. Computing. Control and Management, vol. 142, pp. 361-367, 2012. (references)

[13] Smith, Ben, AdvancED ActionScript 3.0: Design Patterns, CA: Springer-Verlag New York Inc, September 2011.

[14] Rosenzweig,Gary , Gary Rosenzweig's Actionscript 3.0 Game Programming University , CA: Macmillan Computer Pub , August 2007.

Impulsive Control of Chua's Circuits Based on Rule-Wise Linear Computational Verb Systems

Xuyang Lou
Key Laboratory of Advanced Process Control
for Light Industry (Ministry of Education)
Jiangnan University
Wuxi, China
Email: Xuyang.Lou@gmail.com

***Abstract*—In this paper, an impulsive control approach is developed for controlling Chua's circuits that are generalized into rule-wise linear computational verb systems, which are used to approximately segment the dynamics of Chua's circuits into four qualitatively different patterns. Based on the new systems, several theorems are then presented to find conditions under which the chaotic Chua's circuits can be asymptotically controlled to the origin by impulsive control. One example is provided illustrate the effectiveness of the proposed methods.**

***Keywords*-Computational verb system; Chua's circuit; rule-wise linear system.**

I. INTRODUCTION

Since Yang presented computational verb concept in 1997 [1], [2]. computational verb theory has been successfully applied to many industrial products, such as visual card counters [3], visual flame detecting system [4], and so on. The building blocks of computational theory are computational verbs [5], [6], which are applicable to different kinds of control problems [7], [8].

A rule-wise linear computational verb system consists of a set of computational verbs of which the antecedents are conditions specified by using computational verbs and the consequences are linear dynamic systems [6]. In [9], the author presented the structure of verb proportional-integral-derivative (PID) controllers. Robust stability and stabilization of Takagi-Sugeno fuzzy systems that were presented in [10], were analyzed in [6]. Yang designed an asymptotically stable rule-wise computational verb controller for a class of high-order systems based on inverse solutions of Lyapunov equations in [11]. In [8], Tonelli and Yang used a computational verb controller to control the chaotic Chua's circuits based on rule-wise linearization and designed the controller through linear matrix inequalities. However, in their work, the closed-loop control system is constructed by a continuous input control method which is not available for the development of digital control devices.

This paper is devoted to providing an alternative and novel approach that combines computational verb control methodology with impulsive control for controlling a class of chaotic systems. The computational verb rule-wise linear models of Chua's circuits will be used to approximately segment the dynamics of Chua's circuits into those dynamics in the inner region, in the outer region and at boundaries of both regions. Instead of state feedback controllers for each subsystem, impulsive controllers are introduced and may offer a simple and efficient method to deal with systems based on the development of digital control devices which generate control impulses at discrete moments [12]. One numerical example is provided to show the effectiveness of the approach.

The rest of the paper is organized as follows. Section II gives some definitions of computational verb and verb similarity. Section III develops the rule-wise model for the Chua's circuit. Section IV presents the main results for controlled Chua's circuit under computational verb rules. In Section V, an example is provided to show the effectiveness of the main results. Conclusions appear in Section VI.

Throughout this paper, we use the following notations. $\mathbb{R}^n$ denotes the n-dimensional real space. $\mathbb{Z}$ represents the set of positive integer numbers. A^{T} and $A^{\text{-1}}$ denote the transpose and inverse of matrix A, respectively. $\lambda_{\max}(A)$ denotes the maximum eigenvalue of the real symmetric matrix A.

II. THE DEFINITION OF COMPUTATIONAL VERB AND VERB SIMILARITY

A. Computational Verb

As stated [9], the definition of computational verb in [7] is too complex to be operational in the context of engineering applications. Here, a light working definition of computational verb from [6] is giving as follows.

Computational Verb: A computational verb $\mathbf{V}$ is defined by the following *evolving function*

$$\mathscr{E}_{\mathbf{V}} : \mathbb{T} \times \Omega \rightarrow \Omega, \tag{1}$$

where $\mathbb{T} \subseteq \mathbb{R}$ and $\Omega \subseteq \mathbb{R}^n$ are the time and the universe of discourse, respectively.

B. Verb Similarity

The similarity between verbs (verb similarity, for short) is of the essential importance to the inference of verb rules [7].

Since there is no crisp definition of similarity between two dynamic systems, the verb similarity can be defined based on many different concerns as addressed in [7].

Verb Similarity: Given two computational verbs $\mathbf{V}_1$ and $\mathbf{V}_2$, the verb similarity $\mathcal{S}(\mathbf{V}_1, \mathbf{V}_2)$ should satisfy the followings.

1) $\mathcal{S}(\mathbf{V}_1, \mathbf{V}_2) \in [0, 1]$;

2) $\mathcal{S}(\mathbf{V}_1, \mathbf{V}_2) = \mathcal{S}(\mathbf{V}_2, \mathbf{V}_1)$;

3) $\mathcal{S}(\mathbf{V}_1, \mathbf{V}_2) = 1$ if $\mathbf{V}_1 = \mathbf{V}_2$ almost everywhere, where $\mathbf{V}_1 = \mathbf{V}_2$ means both computational verbs have the same evolving function.

III. The Rule-Wise Linear Systems

In this section, we shall control chaotic dynamics of the well-known Chua's circuit [13] whose dynamical behavior is described by

$$\left\{\begin{array}{l} \frac{\mathrm{d}v_1}{\mathrm{d}t} = \frac{1}{C_1}[G(v_2 - v_1) - f(v_1)], \\ \frac{\mathrm{d}v_2}{\mathrm{d}t} = \frac{1}{C_2}[(v_1 - v_2) + i_L], \\ \frac{\mathrm{d}i_L}{\mathrm{d}t} = -\frac{1}{L}[v_2 + R_0 i_L], \end{array}\right. \tag{2}$$

where v_1, v_2 and i_L are the state variables, and $G = \frac{1}{R}$. The characteristic of the nonlinear resistor $f(v_1)$ is taken as the well known piecewise-linear characteristic

$$f(v_1) = G_b v_1 + \frac{1}{2}(G_a - G_b)(|v_1 - E| - |v_1 + E|),$$

where $G_a, G_b < 0$, $E > 0$ is the breakpoint voltage.

Assuming $v_1 \in [-d, d]$, $d \gg E > 0$, we obtain the following sector to bound $f(v_1)$:

$$f_1(v_1) = G_a v_1,$$

$$f_2(v_1) = \Big(G_b + \frac{(G_a - G_b)E}{d}\Big)v_1 \triangleq G_1 v_1.$$

Based on impulsive control, the state equations of the control Chua's circuit is given by

$$\left\{\begin{array}{rcl} \frac{\mathrm{d}v_1}{\mathrm{d}t} & = & \frac{1}{C_1}[G(v_2 - v_1) - f(v_1)],\ t \neq t_k, \\ \frac{\mathrm{d}v_2}{\mathrm{d}t} & = & \frac{1}{C_2}[(v_1 - v_2) + i_L],\ t \neq t_k, \\ \frac{\mathrm{d}i_L}{\mathrm{d}t} & = & -\frac{1}{L}[v_2 + R_0 i_L],\ t \neq t_k, \\ \Delta v_1(t) & = & b_{1k} v_1(t_k),\ t = t_k, \\ \Delta v_2(t) & = & b_{2k} v_2(t_k),\ t = t_k, \\ \Delta i_L(t) & = & b_{3k} i_L(t_k),\ t = t_k, \end{array}\right. \tag{3}$$

where $\Delta v(t)|_{t=t_k} = v(t_k^+) - v(t_k^-)$, $k \in \mathbb{Z}$. Here, $x(t_k^+) = \lim\limits_{h\to 0^+} x(t_k + h)$, $x(t_k^-) = \lim\limits_{h\to 0^+} x(t_k - h)$ with discontinuity instants $t_1 < t_2 < \cdots < t_k < \cdots$, $\lim\limits_{k\to\infty} t_k = \infty$, where $t_1 > t_0$. For convenience, let $t_0 = 0$ and $h > 0$ be sufficiently small. Without loss of generality, it is assumed that $x(t_k) = x(t_k^-) = \lim\limits_{h\to 0^+} x(t_k - h)$. b_{1k}, b_{2k} and b_{3k} are the impulsive control coefficients.

Choose two membership functions:

$$\left\{\begin{array}{rcl} \mu_{\mathrm{in}}(t) & = & \max(0, 1 - |v_1/2|), \\ \mu_{\mathrm{out}}(t) & = & \min(1, |v_1/2|), \end{array}\right. \tag{4}$$

where $\mu_{\mathrm{in}}(t)$ and $\mu_{\mathrm{out}}(t)$ are for 'inner' region and 'outer' region, respectively. Then the state space of Chua's circuit is segmented into an 'inner' region, where $|v_1|$ is not bigger than E, and an 'outer' region, where $|v_1|$ is bigger than E. Denoting $x^{\mathrm{T}}(t) = [v_1,\ v_2,\ i_L]^{\mathrm{T}}$ as the state vector and following the similar design procedure presented in [8], the closed-loop control system based on computational verb rules can be transformed into the following region-wise systems.

*Rule 1: If $|v_1(t)|$ **stays** at the inner region, then*

$$\left\{\begin{array}{rcl} \dot{x}(t) & = & A_1 x(t), t \neq t_k, \\ \Delta x(t) & = & B_{1k} x(t_k), t = t_k, \end{array}\right. \tag{5}$$

*Rule 2: If $|v_1(t)|$ **increases** from the inner region, then*

$$\left\{\begin{array}{rcl} \dot{x}(t) & = & A_2 x(t), t \neq t_k, \\ \Delta x(t) & = & B_{2k} x(t_k), t = t_k, \end{array}\right. \tag{6}$$

*Rule 3: If $|v_1(t)|$ **decreases** from the outer region, then*

$$\left\{\begin{array}{rcl} \dot{x}(t) & = & A_3 x(t), t \neq t_k, \\ \Delta x(t) & = & B_{3k} x(t_k), t = t_k, \end{array}\right. \tag{7}$$

*Rule 4: If $|v_1(t)|$ **stays** at the outer region, then*

$$\left\{\begin{array}{rcl} \dot{x}(t) & = & A_4 x(t), t \neq t_k, \\ \Delta x(t) & = & B_{4k} x(t_k), t = t_k, \end{array}\right. \tag{8}$$

where

$$A_1 = A_2 = \begin{bmatrix} -\frac{G}{C_1} - \frac{G_a}{C_1} & \frac{G}{C_1} & 0 \\ \frac{G}{C_2} & -\frac{G}{C_2} & \frac{1}{C_2} \\ 0 & -\frac{1}{L} & -\frac{R_0}{L} \end{bmatrix}, \tag{9}$$

$$A_3 = A_4 = \begin{bmatrix} -\frac{G}{C_1} - \frac{G_1}{C_1} & \frac{G}{C_1} & 0 \\ \frac{G}{C_2} & -\frac{G}{C_2} & \frac{1}{C_2} \\ 0 & -\frac{1}{L} & -\frac{R_0}{L} \end{bmatrix}, \tag{10}$$

and B_{jk} $(j = 1, \cdots, 4)$ are diagonal impulsive control matrices to be designed.

Therefore, the overall expression of the rule-wise linear computational verb systems (5)-(8) is given by

$$\left\{\begin{array}{rcl} \dot{x}(t) & = & \sum\limits_{i=1}^{4} \mathcal{S}_i(t) A_i x(t) \Big/ \sum\limits_{i=1}^{4} \mathcal{S}_i(t),\ t \neq t_k, \\ \Delta x(t) & = & \sum\limits_{i=1}^{4} \mathcal{S}_i(t) B_{ik} x(t_k) \Big/ \sum\limits_{i=1}^{4} \mathcal{S}_i(t),\ t = t_k, \end{array}\right. \tag{11}$$

where $\mathcal{S}_i(t)$ is the computational verb similarity between waveform $v_1(t)$ and the computational verb in the antecedent of the i^{th} computational verb rule.

To calculate the computational verb similarities $\mathcal{S}_i(t)$ $(i = 1, \cdots, 4)$, it follows from the methods presented in [9] that the canonical forms of verb **become**'s for verb rules in (5)-(8) are given by $\mathbf{become}_i, i = 1, 2, 3, 4,$

respectively, that is,

$$\begin{cases} \textbf{become}_1 & \triangleq & \textbf{become}(\text{inner region}, \text{inner region}), \\ \textbf{become}_2 & \triangleq & \textbf{become}(\text{inner region}, \text{outer region}), \\ \textbf{become}_3 & \triangleq & \textbf{become}(\text{outer region}, \text{inner region}), \\ \textbf{become}_4 & \triangleq & \textbf{become}(\text{outer region}, \text{outer region}). \end{cases} \quad (12)$$

Let T_w be the length of the window and at moment t let us consider the history of dynamics during the period of $[t - T_w, t]$, then the evolving function of **become**$_i$($\texttt{state 1}, \texttt{state 2}$) is given by:
If $\texttt{state 1} \neq \texttt{state 2}$, then

$$\mathscr{E}_{\textbf{become}_i}(\texttt{state 1}, \texttt{state 2})(\tau) =$$

$$\begin{cases} 1 - \dfrac{\tau - (t - T_w)}{T_w}, \ \tau \in \left[t - T_w, t - \dfrac{T_w}{2}\right] \\ 0.5 + \dfrac{\tau - (t - T_w/2)}{T_w}, \ \tau \in \left[t - \dfrac{T_w}{2}, t\right]; \end{cases} \quad (13)$$

otherwise,

$$\mathscr{E}_{\textbf{become}_i}(\texttt{state 1}, \texttt{state 2})(\tau) \equiv 1, \ \ \tau \in [t - T_w, t]. \quad (14)$$

Now, denote $\mathscr{E}_{\textbf{become}}(t) \triangleq \mathscr{E}_{(\texttt{state 1}, \texttt{state 2})}(t)$, and the implementations for verb rules (5)-(8) are given by the following steps.

1) The first half window

$$\begin{aligned} a_1 &\triangleq \int_{\tau = t - T_w}^{t - T_w/2} \mathscr{E}_{\textbf{become}}(\tau) \wedge \mu_{\texttt{state 1}}(v_1(\tau)) \mathrm{d}\tau, \\ b_1 &\triangleq \int_{\tau = t - T_w}^{t - T_w/2} \mathscr{E}_{\textbf{become}}(\tau) \vee \mu_{\texttt{state 1}}(v_1(\tau)) \mathrm{d}\tau. \end{aligned} \quad (15)$$

2) The second half window

$$\begin{aligned} a_2 &\triangleq \int_{\tau = t - T_w/2}^{t} \mathscr{E}_{\textbf{become}}(\tau) \wedge \mu_{\texttt{state 2}}(v_1(\tau)) \mathrm{d}\tau, \\ b_2 &\triangleq \int_{\tau = t - T_w/2}^{t} \mathscr{E}_{\textbf{become}}(\tau) \vee \mu_{\texttt{state 2}}(v_1(\tau)) \mathrm{d}\tau. \end{aligned} \quad (16)$$

3) The balance factor ϖ

$$\varpi = 2 \min\left(\frac{a_1}{b_1 + b_2}, \frac{a_2}{b_1 + b_2}\right). \quad (17)$$

4) The entire window

$$\mathcal{S}(\textbf{become}, v_1(t)) = \frac{a_1 + a_2}{b_1 + b_2} \varpi. \quad (18)$$

Consequently, $\mathcal{S}_i(t)$, $i = 1, 2, 3, 4$, are calculated by

$$\mathcal{S}_i(t) = \mathcal{S}(\textbf{become}_i, v_1(t)) \delta_i(v_1(t)), i = 1, 2, 3, 4, \quad (19)$$

where

$$\begin{aligned} \delta_1(v_1(t)) = \delta_2(v_1(t)) &= \begin{cases} 1, & \text{if } |\mathrm{v_1(t)}| \leq \mathrm{E}, \\ 0, & \text{otherwise.} \end{cases} \\ \delta_3(v_1(t)) = \delta_4(v_1(t)) &= \begin{cases} 1, & \text{if } |\mathrm{v_1(t)}| > \mathrm{E}, \\ 0, & \text{otherwise.} \end{cases} \end{aligned} \quad (20)$$

Note that $\mathcal{S}_i(t)$, $i = 1, 2, 3, 4$, are also computational verb similarities.

IV. Main Results

Now we are arriving at presenting our main theorems for guaranteeing the stability of the impulsively controlled Chua's circuit under computational verb rules. Based on the Lyapunov function $V(t) = x^{\mathrm{T}}(t) P x(t)$ and impulsive control theory, we obtain the following criteria.

Theorem 1: Let $n \times n$ matrix P be symmetric and positive definite, and $Q = PA_i + A_i^{\mathrm{T}} P$, $i = 1, 2, 3, 4$, and λ_1 is the largest eigenvalue of $P^{-1} Q$, λ_2 is the largest eigenvalue of the matrix $P^{-1}(I + B_{ik})^{\mathrm{T}} P (I + B_{ik})$, where $i = 1, 2, 3, 4$, $k \in \mathbb{Z}$, then the origin of impulsive control system (11) is asymptotically stable if there exists a $\xi > 1$ and a differentiable at $t \neq t_k$ and nonincreasing function $K(t) \geq m > 0$ which satisfies

$$-\frac{D^+ K(t)}{K(t)} \leq \lambda_1 \leq \frac{1}{(1 + \varepsilon)\delta_2} \ln \frac{K(\tau_{2k}^+) K(\tau_{2k-1}^+)}{K(\tau_{2k+1}) K(\tau_{2k}) \xi \lambda_2^2} \quad (21)$$

or

$$-\frac{D^+ K(t)}{K(t)} \leq \lambda_1 \leq \frac{1}{\max\{\delta_1, \delta_2\}} \ln \frac{K(\tau_k^+)}{K(\tau_{k+1}) \xi \lambda_2}, \quad (22)$$

where $\delta_1 = \sup\limits_k \{t_{2k+1} - t_{2k}\} < \infty$, $\delta_2 = \sup\limits_k \{t_{2k} - t_{2k-1}\} < \infty$, and for a given constant $\varepsilon > 0$, $t_{2k+1} - t_{2k} \leq \varepsilon(t_{2k} - t_{2k-1})$, $\forall k \in \mathbb{Z}$.

Theorem 2: If there exist a symmetric positive definite matrix P, and positive scalars α_k, β_k, such that the following conditions are satisfied.

$$\mu_M\Big(A_i^{\mathrm{T}} P + PA_i - \frac{\alpha_k}{\delta_k} P\Big) < 0,$$

$$\mu_M\Big(B_i^{\mathrm{T}} P + PB_i - \beta_k P + B_i^{\mathrm{T}} P B_j\Big) < 0,$$

$$\alpha_k + \beta_k \leq 0, \ \beta_k \geq -1,$$

where $i, j = 1, 2, 3, 4$, $\delta_k = t_k - t_{k-1}$, $k \in \mathbb{Z}$, then the origin of impulsive control system (11) is asymptotically stable.

V. Numerical simulations

In this section, we shall provide simulation results to illustrate the proposed method. For simplicity, here we only illustrate the effectiveness of the criterion in Theorem 1. As in [8], the parameters for Chua's circuit (2) are chosen as $R = \frac{10}{7}$, $R_0 = 0$, $C_1 = 0.1$, $C_1 = 2$, $L = \frac{1}{7}$, $G_a = -4$, $G_b = -0.1$, and $E = 1$. With these parameters, Chua's circuit exhibits chaotic dynamics under initial condition $[v_1(0), v_2(0), i_L(0)] = [-2.6, -3.2, 1.1]$ and state $v_1(t)$ will be bounded in the interval $[-15, 15]$; therefore, let us take $d = 15$. To control the Chua's circuit, according to Theorem 1, we set $P = I, K(t) \equiv 1$, $B_{1k} = -0.53I$, $B_{2k} = -0.61I$, $B_{3k} = -0.55I$, $B_{4k} = -0.62I$, and $\xi = 1.1$, which imply that $\lambda_1 = 66.8078$, and $\lambda_2 = 0.2209$, In the simulation, let

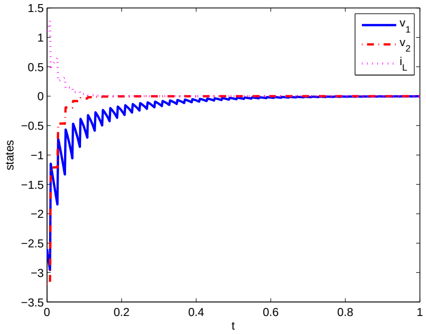


Figure 1. The state trajectories v_1, v_2, i_L of the controlled Chua's circuit

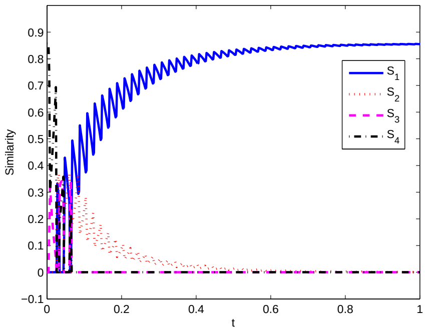


Figure 2. The waveforms of verb similarities $\mathcal{S}_i(t)$ $i = 1, 2, 3, 4$

the impulses be equidistant from each other, that is, $\varepsilon = 1$, thus from (21) we have $0 \leq \delta_1 = \delta_2 \leq 0.0219$. Figure 1 shows the state trajectories of chaotic system (2) under the designed impulsive control and the computational verb controllers for $t_k - t_{k-1} = 0.02$. Clearly, the time series of all the variables of the system converge to zero. Figure 2 shows the waveforms of computational verb similarity $\mathcal{S}_i(t)$ of the i^{th} computational verb control rule for $i = 1, 2, 3, 4$. Note that the similarity $\mathcal{S}_1(t)$ approaches 0.8571 asymptotically while the similarity $\mathcal{S}_2(t)$ approaches zero at the meantime. The third and fourth similarities only had non-zero values at the very beginning and dropped to zero rapidly.

VI. Conclusions

In this paper, an impulsive framework for controlling chaotic Chua system has been proposed through combining impulsive control technique with intelligent rule-wise computational verb methodology. Based on rule-wise linearization, the dynamics of Chua's circuits are approximately segmented into four regions. Two criteria have been proposed to guarantee the global asymptotic stability of the computational verb rule-wise linear systems. The applicability and validity of the proposed control scheme have been illustrated through numerical simulations.

Acknowledgements

The authors would like to expressly thank Professor T. Yang for his constructive suggestions and sharing his codes. This work is partially supported by National Natural Science Foundation of China (No.61174021, No.61104155), and the 111 Project (B12018).

References

[1] T. Yang, "Verbal paradigms-Part I: Modeling with verbs," *Technical Report Memorandum No. UCB/ERL M97/64,* Electronics Research Laboratory, College of Engineering, University of California, Berkeley, CA 94720, 9 Sept. 1997, page 1-15.

[2] T. Yang, "Verbal paradigms-Part II: Computing with verbs," *Technical Report Memorandum No. UCB/ERL M97/66,* Electronics Research Laboratory, College of Engineering, University of California, Berkeley, CA 94720, 18 Sept. 1997, page 1-27.

[3] Yang's Scientific Research Institute LLC and Wuxi Xingcard Technology Ltd, "YangSky-MAGIC Visual Card Counters." http://www.yangsky.us/products/cardsky/cardsky.htm, 2004.

[4] Yang's Scientific Research Institute LLC, "FireEye Visual Flame Detecting Systems." http://www.yangsky.us/products/flamesky/index.htm, 2005.

[5] T. Yang, "Computational verb systems: Verbs and dynamic systems," *Int. J. Comput. Cogn.,* vol. 1, no. 3, pp. 1-50, Sep. 2003.

[6] T. Yang, "Rule-wise linear computational verb systems: dynamics and control," *Int. J. Comput. Cogn.,* vol. 4, no. 4, pp. 18-33, Dec. 2006.

[7] T. Yang, "Computational verb theory: from engineering, dynamic systems to physical linguistics," *volume 2 of YangSky.com Monographs in Information Sciences,* Yang's Scientific Research Institute, Tucson, AZ, Oct. 2002.

[8] R. Tonelli and T. Yang, "Controlling Chua's circuits using computational verb controllers," *Int. J. Robust Nonlin. Contr.,* vol. 18, no. 17, pp. 1622-1636, Apr. 2008.

[9] T. Yang, "Architectures of computational verb controllers: Towards a new paradigm of intelligent control," *Int. J. Comput. Cogn.,* vol. 3, no. 2, pp. 74-101, June 2005.

[10] H.J. Lee, J.B. Park, G.R. Chen, "Robust fuzzy control of nonlinear systems with parametric uncertainties," *IEEE Trans. Fuzzy Syst.,* vol. 9, no. 2, pp. 369-379, Apr. 2001.

[11] T. Yang, "Design computational verb controllers using inverse solutions of Lyapunov equations," *Int. J. Comput. Cogn.,* vol. 6, no. 4, pp. 84-88, Dec. 2008.

[12] Y.P. Zhang, J.T. Sun, "Controlling chaotic L$\ddot{u}$ systems using impulsive control," *Phys. Lett. A,* vol. 342, no. 3, pp. 256-262, July 2005.

[13] L.O. Chua, M. Komuro, and T. Matsumoto, "The double scroll family," *IEEE Trans. Circuits Syst.,* vol. 33, pp. 1072-1118, Nov. 1986.

A Multi-Agent's Modeling for the 4D Model of hiv Dynamics

Toufik Laroum
Université 20 Aout 1955
Skikda, Algeria
laroumtoufik@yahoo.fr

Bornia Tighiouart
Université de Badji Mokhtar
Annaba, Algeria
b_tighiouart@yahoo.fr

***Abstract*—The purpose of the modeling of the biological processes is to understand better the complexity of these phenomena; using various models. The following research is an attempt to study the dynamics of the population of cells intervening during a human immunodeficiency virus (HIV) infection. This problem was mainly studied mathematically by using mathematical models which are based on the differential equations. We will use the approach of modeling Multi-agents to simulate the 4D model of this bio-process and show that the multi-agents model is more close to the real biological phenomenon than a mathematical model without underestimating the mathematical approach. The obtained results are consistent with the biological phenomenon and encouraged us to further improve the model.**

Keywords-Multi-Agents Simulation; dynamic of the populations; the human immunodeficiency virus infection; the virtual community; bio-informatics.

I. Introduction

The mathematical modeling was for a long time used to study the complex phenomena and the efficiency of this approach is not any more to be shown. However, the Multi-agents approach began to be particularly used in the study of the dynamics of the populations relative to the cellular biology so allowing exceeding some limits of the mathematical approach.

This work studies the dynamics of the population of cells concerned by the human immunodeficiency virus infection. There are several mathematical models that treat the dynamics of this phenomenon [3] [5]; the simplest is the 3D model [2] which we modeled with the multi-agents approach in [9]. We are interested in this work by the 4D model which is more complicated than the 3D model because it takes in consideration the dynamics of 4 categories of cells.

This paper begins with a small presentation of the study field; wish is the modeling of bio-process (by using the mathematical modeling). Followed by an explication of the multi-agent modeling approach and the studied biological phenomenon (the human immunodeficiency virus infection).

After that, the multi-agent system is presented with a discussion of the obtained results comparing with the mathematical model and the 3D system.

II. Dynamics of the populations

The dynamics of the population is the science that studies the evolution of the individuals of the population in time and space as well as the interaction between them to understand the global behavior of the population.

The research field is not recent. In 1790, there was a mathematical model of Malthus [10] (the exponential growth of a population), and then, in 1838, the model with logistic growth of Verhulst [7] was proposed. These two models described the evolution of a homogeneous population; but, in 1925 the famous system prey-predator of Lotka-Volterra [1] [11] was the first model describing the evolution of two interacting populations and on which various models were proposed to today. However, the mathematical approach has some limits (complexity of the equations, difficulty in updating the model, abstract models, etc.) that we are trying to overcome by using the multi-agents approach.

III. Multi-agents Modeling approach

The Multi-agents approach is suited well to the study of the complex systems constituted by several entities in interaction. It consists in representing every entity by an agent, then in developing the system with time.

The evolution of different agents with their basic actions and interactions that link them will bring out the dynamic of the studied phenomenon with the appearance of behaviors and unanticipated events [6].

This approach with its low degree of abstraction allows to approach the model from the reality, where every agent moves, reproduces, interacts and reacts with the changes of its environment. The most important is that the agents are different than the others and that every agent is marked and can be followed at any time during its evolution. So, the addition or the retreat of an agent or of a set of agents is an easy operation [8].

IV. Infection by the human immunodeficiency virus

An immune reaction is mainly expressed by the actions of lymphocytes cells called CD4 and CD8. CD4 lymphocyte produced by the Thymus is responsible of the coordination and the activation of cytotoxic lymphocytes CD8. This CD4 cell is an infection subject by HIV virus which considers them as an adequate environment to carry out its cycle of proliferation. So, the destruction of CD4 by the HIV paralyzes the immune defense to its source [4].

The phenomenon of the infection takes place in three stages (see Figure 1):

- Primary infection: lasts from 3 to 8 weeks, it is characterized by a fast diminution of lymphocytes CD4 caused by an increase of the viral load, followed by a decrease of the viral load what allows increasing the number of the cd4.

- The asymptotic phase: its duration is of around 10 years during which the immune system maintains a state of balance (stability) between the number of the CD4 and the viral load.
- AIDS: It is the phase in which the immune system is depressed because of the fast decrease of CD4 lymphocytes (less of 200 / mm3).

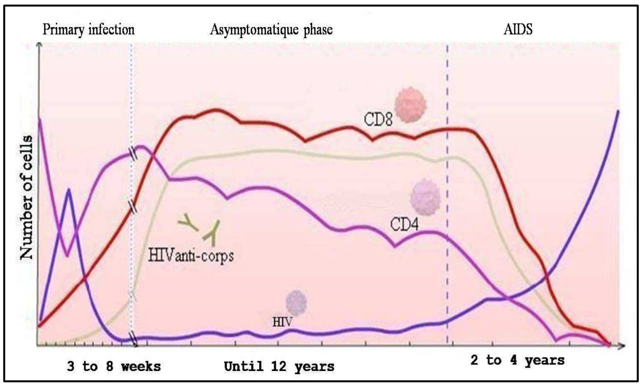


Figure 1. evolution of the biological phenomenon

A. *The 4D Mathematical Model*

We are interested in this study of the 4D model which treats 4 types of cells: the cd4 lymphocytes (T), the viruses HIV (V), the CD4 lymphocytes infected by the viruses (T^{*}) and the CD8 lymphocytes (T_{CTL} for Cytotoxic T-Lymphocyte).

The phenomenon is modeled by the following equations [4], where $T^{'},T^{*^{'}},T^{'}_{CTL}$ and $V^{'}$ indicates respectively the variation rates in density of CD4 cells, infected CD4 cells CD8 cells and the virus populations:

$$\begin{cases} T^{'} = s - \delta T - \beta TV \\ T^{*^{'}} = \beta TV - uT^{*} - qT_{ctl}T^{*} \\ T^{'}ctl = \lambda + aTT^{*}T_{ctl} - \alpha T_{ctl} \\ V^{'} = kT^{*} - cV \end{cases} \qquad (1)$$

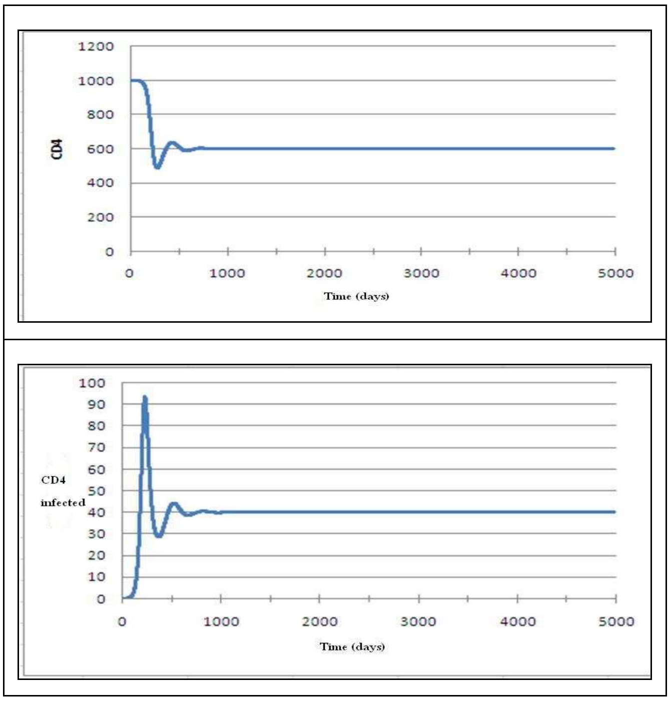
TABLE I. PARAMETERS LIST OF THE 4D MODEL

Parameters	Definition
s	Production of CD4 cells by thymus
δ	Mortality rate of CD4 cells
β	Virus infectivity
u	Mortality rate of infected CD4 cells
q	Cytotoxity of the CD8 against the infected cd4
λ	Production rate of the cd8 by the thymus
a	Rate of proliferation of the cd8
α	Mortality rate of CD8 cells
k	Production rate of virus
c	Mortality rate of virus

CD4 lymphocyte cells are produced by the thymus at a constant rate equal to s cells a day in 1 mm^{3} of blood, and die at a rate of natural mortality equals to δ cells in a day.

The population of CD4 lymphocytes loses also a number of cells which are transformed in infected CD4 cells because of the infection by the virus with a rhythm of βTV where β represents the infectivity of the viruses HIV which is the probability that a contact between CD4 and virus HIV is infectious.

The transformation rate of CD4 cells on infected CD4 is the rate of production of this last one, dying at a natural mortality rate equal to u cells per day. An infected CD4 produces a number of viruses at a rate of k Virus HIV a day, these viruses die at a natural mortality rate equals to c virus a day.



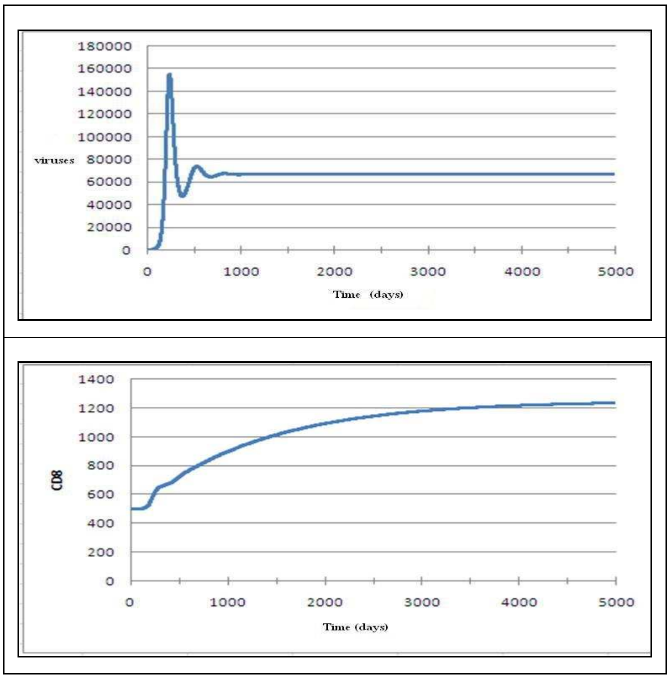


Figure 2. Results of the mathematical model [4].

The CD8 lymphocytes are cells of the immune system, they have a toxic capacity that enables them to play a defensive role to destroy the infected cells cd4 (and foreign objects in general).

The infected CD4 cells are destroyed by the CD8 at a rhythm of qTT_{CTL} where q represents the cytotoxity of the CD8 cells, in other words, the probability that a contact between a cd8 cell and a cd4 infected cell leads to the destruction of this second.

The CD8 cells are produced by the thymus with a constant rate = λ cells per day, and die with a death rate = α cells per day. During their defensive intervention, the cd8 are proliferated with a rhythm = $aTT^{*}T_{CTL}$ proportional to the number of cd4, infected cd4 and the current number of the cd8 cells.

This mathematical model gives the following results (Figure 2), which represents the phase of the primary infection and the asymptomatic phase in the process of the infection.

We notice that the number of the cells CD8 increases during the phase of the primary infection because of the proliferation of the cells, then it starts to stabilize during the asymptomatic phase after the stability of the rates of the CD4 and of the infected CD4 cells.

B. Multi-agents Model

To simulate the phenomenon by a Multi-agents system, we created a virtual environment in which various agents evolve and interact between them. It is an environment in 3 dimensions that corresponds to 1 mm^3 of the blood.

We created four classes of reactive agents feigning the studied cells (The CD4 cells agents, the infected CD4 cells agents, CD8 cells agents and the virus HIV agents (Fig. 3). Each agent reproduces the behavior of a cell; we find the different biological actions of the phenomenon (creation of the cells, movement in the environment, infection, production of the viruses, immune defense, etc.).

In each class, there is a population of agents wish move in the environment. Hiv agents move in the environment, find the closest cd4 agent and infect it if possible.

Infected CD4 agents move in the environment and produce new HIV agents. CD8 agents move in the environment, find the closest CD4 infected agent and destroy it if possible then proliferate to create new agents.

The thymus Agent represents the thymus; his role is the production of CD4 and CD8 agents.

The observer agent is required to execute the system because it provides information's about all agents. For example, to find the closest cd4 agent the HIV agent must calculate the distance from all cd4agents, so he must have the coordinates of all these agents, it is the agent observer who gives this information.

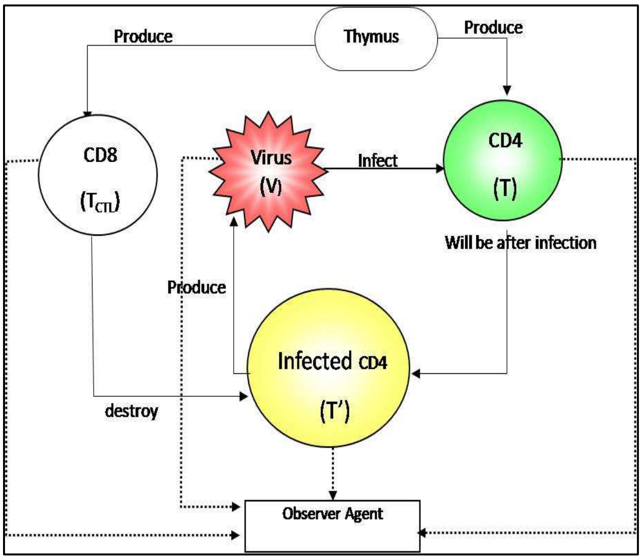


Figure 3. Interactions in the Multi-agent system.

This model Multi-agents is closer to the reality than the mathematical model, which is unable to express the phenomenon of meeting (contact in the biological sense) between a virus and a CD4 cell (the action of the infection) and between a CD8 cell and an infected CD4 cell (the action of defense or the destruction of the infected CD4).

Effectively, in the mathematical model, the number of the produced infected CD4 agents is calculated by multiplying the total number of the possible contacts between the viruses and the CD4 cells (which is equal to TV) by the parameter β which does not describe faithfully the phenomenon.

In other words, by means of the mathematical model, a population of 100 CD4 cells and 10 viruses gives 100*10=1000 infected CD4 cells which are not so exact because in the reality this population produces in maximum 10 infected cells CD4 if we suppose that every virus infects one CD4 cell [9].

The same thing for the number of the infected CD4 cells eliminated by the cd8 cells; the total number is calculated by multiplying the number of the possible contacts between the CD8 and the infected CD4 cells (which is equal to $T_{CTL}T^*$) by the parameter q .

It is obvious that if there are 100 CD8 cells and 10 infected CD4 cells, the number of destroyed cells will be 10*1000 = 1000 cells if we suppose that q =1 while there were only 10 cells at the beginning.

The number of the CD8 is multiplied during the destruction of the infected CD4 cells (by proliferation), consequently it depends on the number of the CD4, infected cd4 and CD8 more exactly depends on the number of the contacts between cd8 and cd4 infected. The mathematical model estimates this value by the multiplication aTT^*T_{CTL}, i.e., all cd8 are proliferated because it does not make a distinction between the cd8 cells contrary to the Multi-agents model where for each contact between cd4-infected and CD8 the two cells (agents) members of this phenomenon are well-known because the agents are distinguished from each other!

That returns because the mathematical approach treats the phenomenon on high-level (consider all the population) contrary to the approach Multi-agents where the treatment is made at the level of the individuals and each contact between cells is treated independently of the others, which gives a more exact representation of the reality.

V. Results

A. Evolution without infection

In the absence of the viruses or foreign dangerous objects, the number of the cells cd8 remains constant. If we suppose that the thymus produces daily 1 CD8 cell, at a mortality rate δ =0,002, that means that the lifespan of cells is 1/ δ =500 days [4]. We can notice that from several random initial states: 0 cells, 1000 cells and 1500 cells with a random initialization of the age of cells (between 0 and 500 days to have a homogeneous population wish is close to the reality), the population of the CD8 will converge on 500 CD8 cells and stabilizes around this value (Fig. 4).

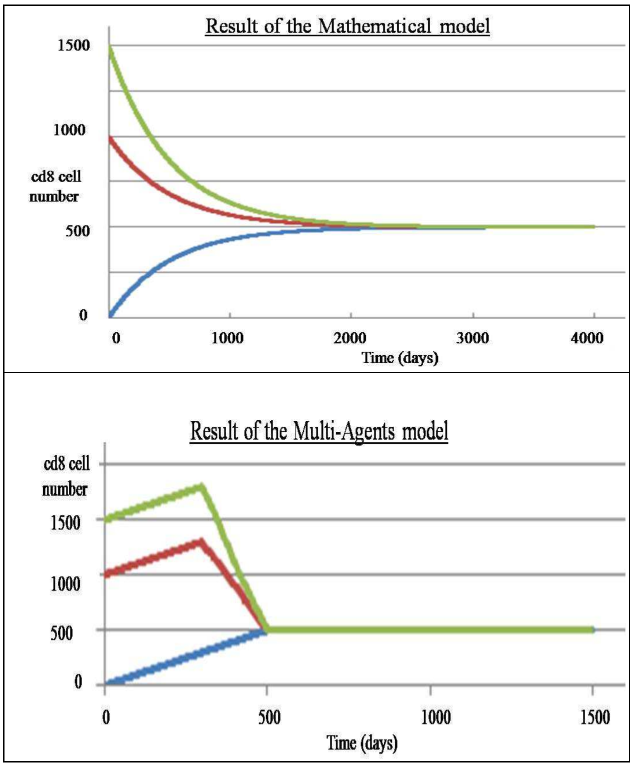
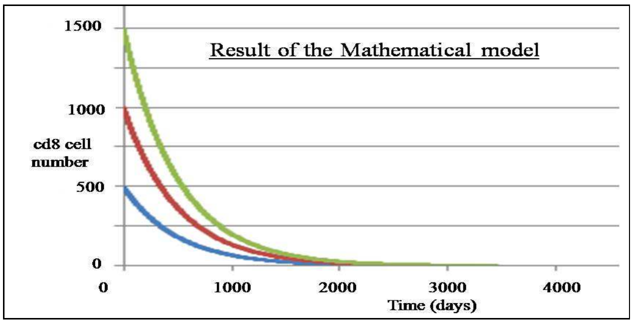


Figure 4. Evolution of the CD8 without infection.

We notice that the model Multi-agents converges more quickly than the mathematical model. The difference between both models appears also in the speed of extinction of an isolated population of CD8 cells, i.e., the rate of production is equal to zero (Fig. 5).



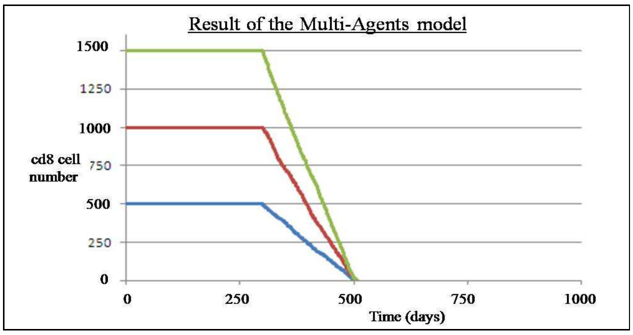


Figure 5. Extinction of the CD8.

The Multi-agents model shows that the population of the CD8 cells disappears immediately after 500 days (which is the lifespan of cells), and that happens independently of the initial number of the cells.

B. *Evolution of the infection*

In an environment, which represents 1 mm^3 of the blood, evolve four categories of cells: infected CD4, CD8, infected CD4 and the viruses HIV which interact between them by feigning the phenomenon of the infection.

In the presence of the infection, the CD8 cells play a defensive role. They exploit their cytotoxic capacity to eliminate the infected CD4 cells and consequently stop the production of the new viruses. The intervention of the CD8 cells is accompanied by a proliferation of these last where new CD8 cells are produced according to the numbers of the other cells (according to the mathematical model).

We notice that the two first phases of the process of the infection are recognizable on the various curves (Fig. 6).The phase of the primary infection is characterized by a growth of the viral population (initially little numerous) which invade CD4 cells (initially numerous). The infection of the healthy CD4 gives infected CD4 cells which are going to produce new viruses able to infect the others CD4. This growth persists until reach a maximal rate with which the reduced number of the population of the CD4 becomes a rare resource, consequently lot of viruses die without being able to infect CD4 and to produce infected CD4. In that case we notice a fall of the viral load and the number of the CD4 infected.

The second phase of the infection is the asymptomatic phase in which a kind of state of balance is established between the rates of the various cells.

Figure 6 shows also the action of the CD8 cells. In the beginning, the production of the CD8 follows a natural production rate, but after the invasion of the viruses we notice that the number of the CD8 agents was increased considerably.

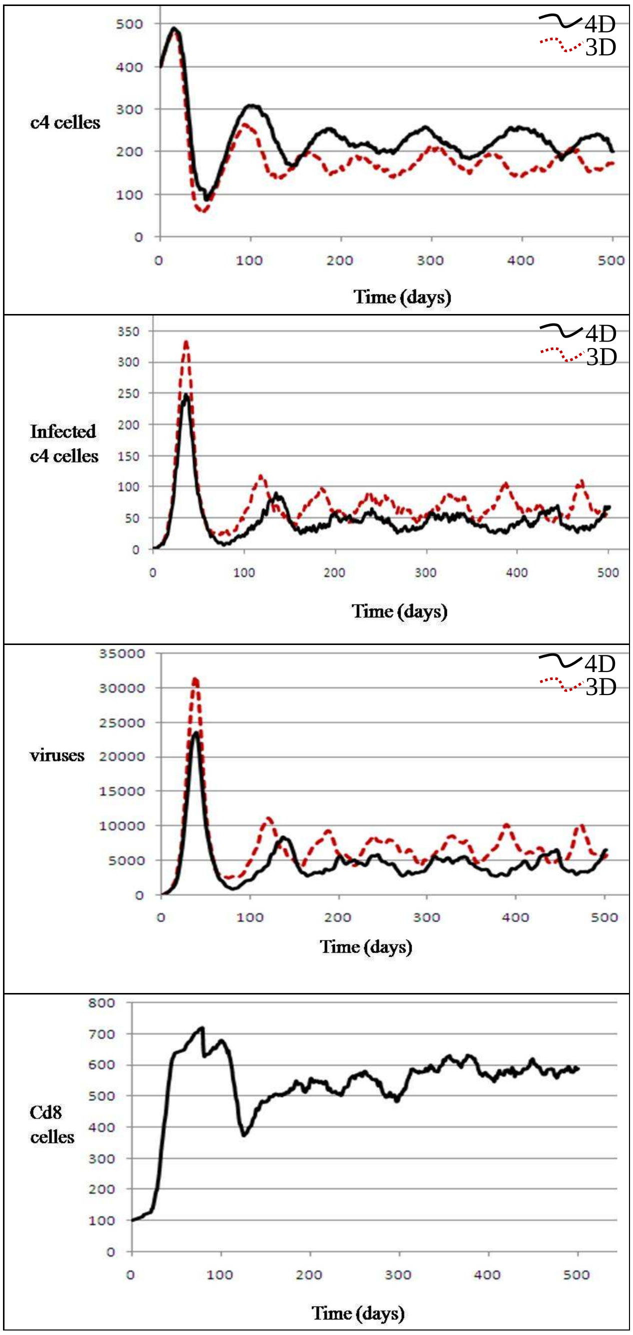


Figure 6. Results of the Multi-Agents model.

The strong increase in the CD8 agents returns to the proliferation of the latter. The action of the CD8 agents is double: destruction of the infected CD4 cells and the proliferation to reinforce the immunity against the infected CD4 cells (The proliferation of the CD8 reaches its maximum when the number of the infected CD4 is max i.e. during the phase of the primary infection).

It is clear in the case of the 3D model (the dotted curve) that the number of the infected CD4 (consequently of the viruses) is more important than the one of the 4D model because there is no resistance from the system like the case of the model 4D.

Figure 7 shows the number of CD8 cells produced by proliferation during both phase of the infection.

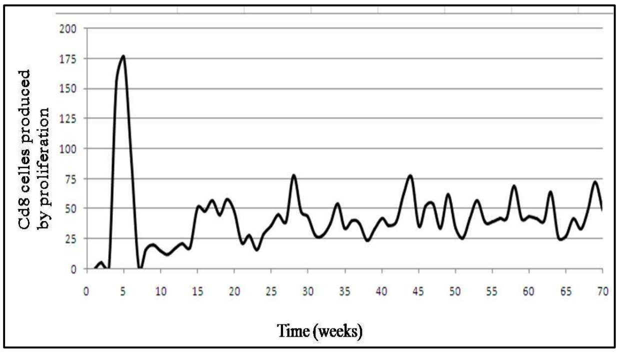


Figure 7. Cd8 proliferation.

If there were not infection of the viruses, the production of the CD8 remains constant. In the primary infection the number of the infected CD4 reached its maximum it is for that we notice the abrupt increase in the CD8 agents because of the proliferation (look at the curve of CD8).

During the asymptomatic phase, the number of the various cells (CD4, infected CD4 and virus) is stable thus we notice that the number of the CD8 agents becomes stable, more exactly; because the number of the CD8 cells agents produced by proliferation becomes more stable and weaker.

VI. CONCLUSION AND FUTURE WORKS

The population of the agents could reproduce the evolution of the biological phenomenon relating to the 4D model. This model enriched the 3D model by the behavior of the CD8 cells which influences the dynamics of the infection in accordance with the expected results from the biological phenomenon. We can see that during the asymptomatic phase the system maintains a rate of the CD4 cells more important than the one which is in the 3D model thanks to the effect of the CD8 cells. Our multi-agent model is extensible, in future work we will add the behavior of other cells involved in this bioprocess; because if we can build a complete multi-agents model for the phenomenon of the infection, we will be able to predict the evolution of the phenomenon and consequently to better direct the treatment.

REFERENCES

[1] A. J. Lotka, “Elements of Physical Biology”, Williams and Wilkins company, Baltimore. (February 1925).

[2] A. S. Perelson and P. W. Nelson, “Mathematical analysis of HIV- dynamics in vivo”, SIAM Review, Vol. 41, No. 1 (Mar., 1999), pp. 3-44.

[3] C.H. Moog, D.A. Ouattara, C. Francois-Brunet, F. Bugnon, V. Ferre, E. Andre-Garnier, F. Raffi, “Mathematical modelling of HIV infection for an aid in the early diagnosis of therapeutical failures”. In XVI International AIDS Conference, Toronto, Canada, August 2006. Ref. CDA0120 on http://www.aids2006.org.

[4] D. A. Ouattara, “Modélisation de l’infection par le VIH, identification et aide au diagnostic”, Thèse de doctorat, Spécialité Automatique et Informatique Appliquée. Université de Nantes (2006).

[5] F. Dubois, VJ. Le Meur. Hervé, and C. Reiss, “Mathematical modeling of antigenicity for HIV dynamics, MathematicS In Action, volume 3, p. 1-35, 2010.

[6] M. Bouzid, “Contribution à la modélisation de l’interaction Agent/Environnement, modélisation stochastique et simulation parallèle”, Thèse de doctorat de l’université Henri Poincaré, Nancy 1 (Spécialité informatique). Laboratoire Lorrain de recherche en informatique et ses applications (2001).

[7] P. F. Verhulst, “Notice sur la loi que la population suit dans son accroissemen, Correspondance mathématique et physique” 109, tome X – ou tome II de la 3e série (1838).

[8] P.Ballet, “Intérêts Mutuels des Systèmes Multi-agents et de l’Immunologie. Applications à l’immunologie, l’hématologie et au traitement d’image”, thèse de doctorat Université De Bretagne Occidentale (2000).

[9] T. Laroum and B. Tighiouart, "A Multi-agent System for the Modelling of the HIV Infection", KES-AMSTA 2011, LNAI 6682, pp. 94–102, 2011.

[10] T. R. Malthus, “An essay on the principle of population”, First Edition, J. Johnson in St Paul’s Churchyard, London. (1798).

[11] V. Volterra, “Variazioni e fluttuazioni del numero d’individui in specie animali conviventi”, In R. N. Chapman : Animal Ecology. McGraw-Hill 1931, New York, 1926.

Applications of RFID in Incident Management

Amir N. Shamdani
IT Department, School of Technology
Purdue University Calumet, USA
ashamdan@purdue.edu

Barbara J. Nicolai
IT Department, School of Technology
Purdue University Calumet, USA
bnicolai@purduecal.edu

***Abstract* – This paper surveys the applications of Radio Frequency Identification (RFID) in incident management to improve the accurate resource tracking and tracing functionalities in disaster situations. The paper started with an initial discussion on what the possible goals working with the active RFID kit (FFID Reader + RFID Tags), to provide a solution to monitor an object for the purpose of identification and tracking using radio waves. Afterwards, Incident Management and its role in specifying the response process and resource management from the Network Centric approach will be determined with an applied approach. It is concluded that what would be the process of monitoring at any time for Dynamic Information Collection to provide a real-time scalable decision support framework built on rapid information collection using RFID technology.**

Keywords – RFID; Incident Management; Command and Control; Network Centric Paradigm

I. INTRODUCTION

Radio Frequency Identification (RFID) is defined as a set of technology that provides network delivered information services dependent on physical object identification captured by radio waves. A typical RFID system consists of four main components, a tag (transponder), a reader (interrogator), an antenna and middleware (Computer) [1]. When the tag is in the electromagnetic field of RFID antenna the tag's presence is detected by the reader. The reader is interfaced to the middleware using a network cable which displays the number of reads on a screen. This entire process can be completed with no human intervention hence RFID system can be truly automatic. Hence this is the most preferred methodology of tracking and tracing objects within and beyond the organization's borders.

In incident management, RFID is the use of an object (typically referred to as an RFID tag) applied to or incorporated into patients, responders, or emergency transport vehicles involved in disasters to identify their locations and status [1]. RFID is a technology that offers huge potential for incident management activities by automating processes and providing accurate, trusted data. Its unique features include giving each physical object a globally unique digital identity read from a distance without requiring line-of-sight capability, and often without using a battery. These features provide new ways of measuring and integrating the real world into information systems. The two most talked-about components of a RFID system are the tag, which is the identification device attached to the item one wants to track, and the reader, which is a device that can recognize the presence of RFID tags and read the information stored on them. The reader can then inform another system about the presence of the tagged items [1], [2].

Disaster response and recovery efforts require timely interaction and coordination of emergency services in order to save lives and property. Decisions about the selection of new technologies such as RFID to track patients, equipment and staff during the response to a disaster require significant investment and can provide a real-time scalable decision support framework built on rapid information collection and accurate resource tracking functionalities. The purpose of this paper is to improve managerial decision making about the adoption of RFID in incident management. It discusses the use of this technology from a managerial viewpoint for disaster managers. This paper contains the following subjects:

At first, the RFID Applications and the necessity of attention to this technology are defined. Then differences in scope and leveling between the terms incident response, incident handling, and incident management are determined. Section IV introduces resource management concepts. The complexity of information age and disaster missions are described in section V. Section VI discusses the RFID tag readability and Section VII concludes the paper.

II. RFID APPLICATIONS

The two major areas of significant where this technology is used are financial services for IT asset tracking and healthcare with more than 60% of the top medical device companies using passive ultrahigh-frequency (UHF) RFID in 2010. RFID use in product tracking applications begins with plant-based production processes, and then extends into post-sales configuration management policies for large buyers. On the other hand, logistics and transportation are important areas of implementation for RFID technology. For example, yard management, shipping and freight and distribution centers are some areas where RFID tracking technology is used. Locomotives and rolling stock are equipped with two passive RFID tags (one mounted on each side of the equipment); the data encoded on each tag identifies the equipment owner, car number, type of equipment, number of axels, etc. Aerospace applications that incorporate RFID technology are being incorporated into Network Centric Product Support Architecture [2]. This technology serves to help facilitate more efficient logistics support for systems maintenance on-board commercial aircraft.

RFID could also be used to mitigate a wide array of logistical challenges such as monitoring evacuees and managing the flow of medical supplies in the immediate aftermath of major disasters, like an earthquake, to help save lives. Researchers found there is a 72 hour 'golden' rescue period following an earthquake during which the efficiency of emergency response procedures is key to the rescue operation [3]. Particularly challenging, is knowing how many people are present in a damaged building or structure that needs to be evacuated. In these scenarios, RFID can facilitate the dispatch of rescue personnel and provide real-time information that could be used to organize search and rescue missions.

A real-world example of the value that RFID can provide in emergency situations was realized immediately following the 7.0 earthquake that struck Port-au-Prince, Haiti on January 12th, 2010. As detailed in an RFID Journal report, the U.S. Department of Defense leveraged its In-Transit Visibility (ITV) network to track shipping containers as they moved to and from the island. Lieutenant Colonel Ralph Riddle, the commander of the 832nd Transportation Battalion, in Jacksonville, FL described the benefits of ITV network: "From a commander's point of view, I'd say that the ITV was critical to the recent aid operations in Haiti. This was a very complex mission, with a rapid deployment. It's something we don't do every day, but we prepare for every day, and the ITV network was absolutely critical to its success" [3].

As an engineering solution, RFID is best understood as technology cluster within the auto ID group of technologies (which also includes, for example, barcode and optical character recognition). Unlike auto ID, however, RFID has no line-of-sight requirement, which means it provides wireless unique identification at the item level that can be seamlessly retrieved at the group level. RFID systems can operate on a stand-alone network of RFID readers and collectors or can be imbedded into responder centers' WiFi.

The following illustration, Fig. 1, shows how the RFID Software modules connect RFID devices and generic sensors with business applications.

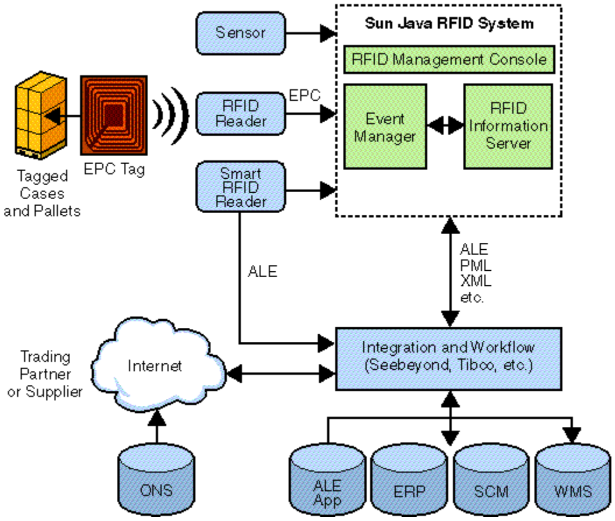


Figure 1. Layers of RFID systems.

The RFID Software consists of the RFID Event Manager, the RFID Management Console and the RFID Information Server modules. The RFID Event Manager gathers information from RFID readers, filters the information, and provides the processed information to the RFID Information Server module or to a third-party ERP system. The integration layer is optional as business applications can obtain RFID sensor events through an integration layer or directly through dedicated connectors [2], [3].

III. INCIDENT MANAGEMENT

Historically, people have used the term "incident response" and "incident handling" to define the activities for tasks and projects of a disaster responder. Consider those phrases also too narrow in scope to adequately address the wide range of work and services a responder might provide. It is shown that although incident handling and incident response are part of that work, the range of work that can be done actually encompasses a larger set of activities that we refer to as incident management [4]. There is a defined difference in scope and leveling between the terms incident response, incident handling, and incident management. We define incident handling as one service that involves all the processes or tasks associated with "handling" events and incidents. Incident handling includes multiple functions:

- **Detecting and reporting:** the ability to receive and review event information, incident reports and alerts.
- **Triage:** the actions taken to categorize, prioritize, and assign events and incidents.
- **Analysis:** the attempt to determine what has happened, what impact, threat, or damage has

resulted, and what recovery or mitigation steps should be followed. This can include characterizing new threats that may impact the infrastructure.

- **Incident response:** the actions taken to resolve or mitigate an incident, coordinate and disseminate information, and implement follow-up strategies to prevent the incident from happening again.

Incident response, as noted in the list above, is one process, the last step in incident handling. It is the process that encompasses the planning, coordination, and execution of any appropriate mitigation and recovery strategies and actions.

The term “incident management” expands the scope of this work to include the other services and functions that may be performed by disaster responders, including vulnerability handling, artifact handling, security awareness training, and the other services outlined in the service management procedures. The definition of this term to include this expanded set of services is important because incident management is not just responding to an incident when it happens. It also includes proactive activities that help prevent incidents by providing guidance against potential risks and threats, for example, identifying vulnerabilities in software that can be addressed before they are exploited. These proactive actions include training end users to understand the importance of computer security in their daily operations and to define what constitutes abnormal or malicious behavior, so that end users can identify and report this behavior when they see it [4].

Usually as part of the wider management process in private organizations, incident management is followed by post-incident analysis where it is determined why the incident happened despite precautions and controls. This information is then used as feedback to further develop the security policy and/or its practical implementation. In the USA, the National Incident Management System, developed by the Department of Homeland Security, integrates effective practices in emergency management into a comprehensive national framework [4].

Fig. 2 illustrates the relationship between the terms incident response, incident handling, and incident management. Incident response is one of the functions performed in incident handling; incident handling is one of the services provided as part of incident management [5].

Information Technology and technological systems provide supporting capabilities essential to implementing and continuously refining the disaster domains. These include voice and data communications systems, information systems (i.e., record keeping and resource tracking), RFID technology and display systems [6]. These also include specialized technologies that facilitate incident operations and incident management activities in situations that call for unique technology-based capabilities. Ongoing development of science and technology is integral to continual improvement. Strategic research and development (R&D) ensures that this development takes place. Each system should rely on scientifically based technical standards that support the nation’s ability to prepare for, prevent, respond to, and recover from domestic incidents [5].

Maintaining an appropriate focus on science and technology solutions as we relate to incident management will necessarily involve a long-term collaborative effort among nation’s partners. Effective communications, information management, and information and intelligence sharing are critical aspects of domestic incident management. Establishing and maintaining a common operating picture and ensuring accessibility and interoperability are principal goals of communications and information management. A common operating picture and systems interoperability provide the framework necessary to [4], [5]:

- Formulate and disseminate indications and warnings.
- Formulate, execute, and communicate operational decisions at an incident site, as well as between incident management entities across jurisdictions and functional agencies.
- Prepare for potential requirements and requests supporting incident management activities.
- Develop and maintain overall awareness and understanding of an incident within and across jurisdictions.

Prior to an incident, entities responsible for taking appropriate pre-incident actions use communications and information management processes and systems to inform and guide various critical activities. These actions include mobilization or pre-deployment of resources, as well as strategic planning by preparedness organizations, multiagency coordination entities, agency executives and jurisdictional authorities. During an incident, incident management personnel use communications and information processes and systems to inform the formulation, coordination, and execution of operational decisions and requests for assistance [5]. Their goal is to restore a normal service operation as quickly as possible and to minimize the impact on responder operations, thus ensuring that the best possible levels of service quality and availability are maintained.

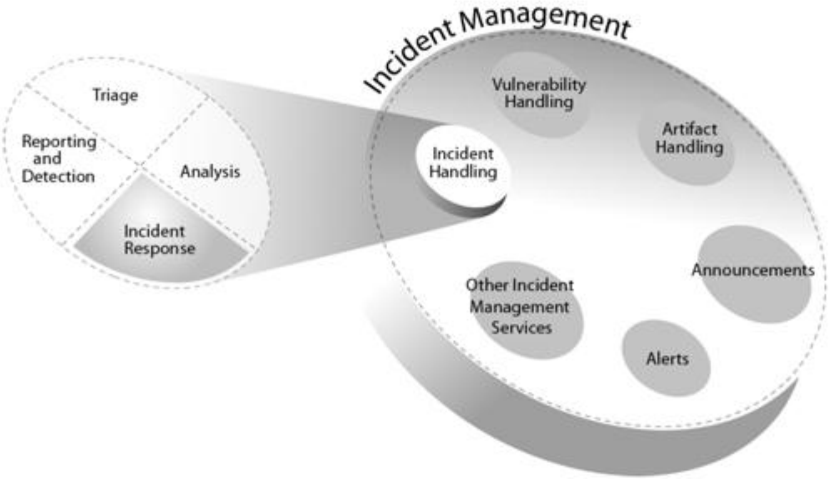


Figure 2. Defining the relationship between incident response, incident handling and incident management.

IV. RESOURCE MANAGEMENT

Resource management involves coordinating and overseeing the application of tools, processes, and systems that provide incident managers with timely and appropriate resources during an incident. Resources include personnel, teams, facilities, equipment, and supplies. Resource management involves four primary tasks:

- Establishing systems for describing, inventorying, requesting, and tracking resources;
- Activating these systems prior to and during an incident;
- Dispatching resources prior to and during an incident; and
- Deactivating or recalling resources during or after incidents.

The basic concepts and principles that guide the resource management processes allow these tasks to be conducted effectively. By standardizing the procedures, methodologies, and functions involved in these processes, resources move quickly and efficiently to support incident managers and emergency responders. The underlying concepts of resource management are that [5]:

- It provides a uniform method of identifying, acquiring, allocating, and tracking resources.
- It uses effective mutual-aid and donor assistance and is enabled by the standardized classification of kinds and types of resources required to support the incident management organization.
- It uses a credentialing system tied to uniform training and certification standards to ensure that requested personnel resources are successfully integrated into ongoing incident operations.

Generally, preparedness organizations work together in advance of an incident to develop plans for managing and employing resources in a variety of possible emergency circumstances [7]. Resource managers use standardized processes and methodologies to order, identify, mobilize, dispatch, and track the resources required to support incident management activities. Resource managers perform these tasks either at a customer request or in accordance with planning requirements. Resources are categorized by size, capacity, capability, skill, and other characteristics. This makes the resource ordering and dispatch process within jurisdictions, across jurisdictions, and between governmental and nongovernmental entities more efficient and ensures that incident management receive resources appropriate to their needs.

V. INFORMATION AGE AND DISASTER MISSIONS

The most effective consequence of Information age paradigms is deep changes in various fields including the disaster environment. Growing complexity and diversity of recent disaster missions, tasks and also methods have affected deeply Command and Control (C2) structure [8]. In fact, various missions in the disaster atmosphere require faster and more flexible plans where the traditional central and hierarchical C2 structure is not suitable. It is obvious that without dynamic information collection and resource tracking there is no guarantee for a smart response to the environment change. Also, there is neither agility nor fast movements and it is hard to plan complex operations in the right place and at the right time. These facts necessitate a new paradigm for the C2 and the main decision maker in the disaster [8], [9]. The Network Centric approach using RFID technology is a good substitution for the traditional C2. In this way, various aspects of disaster environments such as power transferring to the edge, self-similarity, sense making, agility and effectiveness can be achieved more easily [10], [11]. Concurrent planning and execution is one of the most fundamental subjects in which there is always the opportunity to change, modify and/or heal the plans, therefore complex missions can be done. Active RFID is now being used to track and trace victims in a disaster situation. Each tag generating a message each time when passing a reader may be a desired outcome. Some RFID tags can be read from several meters away and beyond the line of sight of the reader. The application of bulk reading enables an almost-parallel reading of tags. Data can be collected in real time and made available to emergency workers immediately, saving precious hours. Crisis management teams, hospitals and other emergency personnel have access to the information via a computer database. Hospitals, for instance, can start planning for the arrival and treatment of disaster victims. The combination of these components will result in the creation of a mobile, scalable tool that can be rapidly deployed at a disaster scene to enable an offsite commander to visualize the location and triage condition of the casualties as well as the available resources. This information will improve the coordination of the response to better match supply (care providers, ambulances, medical equipment) with demand (number of patients, level of acuity) [12], [13].

VI. RFID TAG READABILITY AND OPTIMIZATION OF RFID SYSTEMS

For an incident commander, keeping track of resources, equipment and products at the scene of an emergency is vital. Current performance of RFID systems is highly application dependent: tag-reader combinations behave differently for different target applications, as well as under variations of the environment for a given application. In general, successful data transaction between tags and a reader with a maximum reading range, and tag read rates (as specified by the existing standards EPC Class 0 & 1, ISO 18000-6A/B, ISO 15693, etc.) with unrestrictive tag orientations are the key aspects that measure performance of an RFID system [14]. Currently, low frequency and high frequency technologies perform reliably, but for ultra-high frequency RFID the deployment needs careful tag, reader, protocol and environment selection to achieve acceptable tracking reliability. The main factors that effect the tag readability are type of antenna, height of tag from the ground, displacement of tag and distance of tag from antenna. RFID system implementation always requires that the system be optimized to every application. For optimizing the RFID system experiments need to be performed to identify the placement, location, orientation of tag antenna and other factors which interact with the system. The importance of optimization is explained by Ammu [14]. He identifies the effect of distance between tag and antenna and performs design of experiments to analyze the resulting data. The main measurement is the number of reads for every experiment and its repetitions. The difference between the levels of each factor is also analyzed using statistical analysis software. The placement and location of the tag is identified using the factorial approach. This approach gives an application specific optimization. The RFID system is used as a trigger mechanism to a vision system which determines whether the object is moving towards or away from a particular position. Another similar invention is the integration of video surveillance system with the RFID tracking system. The calibration of RFID tracking system is enhanced by the use of information provided by the video surveillance system. Moreover, the calibration of the vision system is enhanced by the information of the RFID system. RFID systems are calibrated by placing RFID tags at visually apparent locations to determine appropriate correction factors for use in subsequent RFID locations. This invention provides the advantages of RFID tracking system and the video surveillance system to overcome the disadvantages of either systems or both [14].

VII. CONCLUSION

The demand for effective and expediently-made decisions is always in vogue. This is not surprising since making correct decisions is essential for successful operations in many places such as disaster environments. Decisions require data to be processed for quality, concept and context [13]. The goal of information gathering and processing is focused on existing or arising problems. The main conclusion of this paper is expanding the network-centric paradigm allows for access to additional, previously unreachable sources of information in addition to physical and informatics scopes [15]. One the main points in this conclusion is the development of security as well as quality of services rendered by trust networks and reliability systems using RFID technology. RFID is a non-line-of-sight (capable of communicating remotely even when obscured) and contact-less (without direct contact between the transacting elements) automatic identification technology. The identification data is stored on chips that can be attached or embedded into products, animals or even humans. The tag can be active (with on-board power source) or passive (with no power source). These enable robots and humans to use passive RFID tags and GPS devices to map out a disaster area and send information to a command center. While there is a benefit of getting more information, the time spent to weigh information for quality, to fuse information into concepts, and to package for contextual relevance is also increasing.

VIII. REFERENCES

[1]. Takizawa, O., “RFID-based Disaster-relief System”, National Institute of Information and Communications Technology, Japan, 2005.

[2]. Kitayoshi, H. and Sawaya, K., “Technical Conditions for High-power Passive Tag Systems Using the 950 MHz Band (partial report from Information and Communications Council)”, Ministry of Internal Affairs and Communications, Tokyo, Japan, 2004.

[3]. Shibayama, A. and Hisada, Y., “An Efficient System For Acquiring Earthquake Damage Information In Damaged Area”, The 13th World Conference on Earthquake Engineering, No.1121, Vancouver, Canada, August 2004.

[4]. Chertoff, M., “National Incident Management System (NIMS)”, Federal Emergency Management Agency, U.S. Department of Homeland Security, Washington DC, USA, March 2010.

[5]. Alberts, C., Dorofee, A., Killcrece, G., Ruefle, R., and Zajicek, M., “Defining Incident Management Processes for CSIRTs: A Work in Progress”, USA, October 2004.

[6]. Gadh, R. and Prabhu, B. S., “Radio Frequency Identification of Katrina Hurricane Victims”, USA, 2010.

[7]. Atkinson, S. R. and Moffat, J., “The Agile Organization: From Informal Networks to Complex Effects and Agility”. Washington D.C, USA, CCRP 2005.

[8]. Alberts, D. S. and Hayes, R. E., “The future of command and control”, Washington DC, USA, CCRP 2007.

[9]. Alberts, D. S. and Hayes, R. E., “Power to the Edge: Command Control in the Information Age”, Washington DC, USA, CCRP 2003.

[10]. David, A. S., Garstka, J. J., and Stein, F. P., “Network Centric Warfare: Developing and Leverage Information Superiority”, Washington DC, USA, CCRP 1999.

[11]. Dekker A. H., “A taxonomy of Network Centric Warfare Architectures”, Systems Engineering/Test and Evaluation Conference, Brisbane, Australia, November 2005.

[12]. Ling, M. and Selvestrel, M., “An Organisation-Oriented Agents Approach to Modelling Network-Centric Warfare”. SimTecT 2004, Canberra, Australia, 2004.

[13]. Shamdani, A. N., “Intelligent Net Centric Command and Control Architecture Using Cognitive Approach”, The World Congress on Engineering and Computer Science (WCECS 2008), San Francisco, USA, 22-24 October 2008.

[14]. Ammu, A., “Effect of Factors on RFID Tag Readability – Statistical Analysis”, USA, 2009.

[15]. Shamdani, A. N., “The Effect of Avicenna’s Philosophy on the Development of Cognitive Architecture for the Network Centric Command and Control”, The International MultiConference of Engineers and Computer Scientists (IMECS 2010), Hong Kong, 17-19 March 2010.

A Timed Colored Petri-Net-based Modeling for Contract Net Protocol with Temporal Aspects

Djamila Boukredera
Laboratoire des mathématiques appliquées
Université Abderrahmane Mira
Béjaia, Algérie
boukredera@hotmail.com

Ramdane Maamri
LIRE Laboratory
Université Mentouri
Constantine, Algérie
rmaamri@yahoo.fr

Samir Aknine
GAMA Laboratory
Université Claude Bernard, Lyon 1
Lyon, France
samir.aknine@univ-lyon1.fr

***Abstract*—Contract Net Protocol (CNP) is a high level communication protocol. It is one of the most widely used in multi-agent system (MAS) to resolve decentralized task allocation problem. The main aim of the protocol is to facilitate contract negotiation between a manager agent and many contractor agents. A lot of works have been done for the verification of the protocol and its extensions, but there still lacks a formalism for representing temporal interaction aspects which are an essential parameter in the protocol modeling. This paper proposes to use Timed Colored Petri Nets (TCPN) to model correctly and formally this temporal dimension often defined as interaction duration and message deadlines. We will verify by means of simulation techniques and state space analysis important properties namely model correctness, deadline respect, absence of deadlocks and livelocks, absence of dead code, agent terminal states consistency, concurrency and validity.**

***Keywords*-Negotiation protocols; Contract net protocol; Multi-agent systems; Timed Colored Petri Nets.**

I. Introduction

A multi-agent negotiation protocol is a specification of the rules that govern interaction among negotiation agents. Formal modeling as well as validation and verification of such protocols are of crucial importance in the design of automated negotiation systems. Based on FIPA standards [2], the CNP, originally proposed by Smith [8], is one of the most popular interaction protocols used in diverse negotiation contexts. Developed to resolve decentralized task allocation, the CNP represents a distributed negotiation model based on the notion of call for bids. In this protocol, agents can dynamically take two roles: manager or contractor (initiator or participant according to FIPA terminology). CNP is currently used as the basis for developing more complex agent negotiation protocols, that is why it is important to analyze this protocol and to verify that it satisfies various key properties before implementation. Several formal models were proposed in the literature [1], [4], [12]–[14], [17], [18], but few works tackled the modeling of temporal interaction aspects which are specified by FIPA.

This paper addresses this issue and proposes to use Timed Colored Petri Nets (TCPN) to model formally the CNP with two temporal constraints:

- Deadlines: it is a time constraint for message exchange. They denote the time limit by which a message must be sent. Once the deadline expires, the manager starts the evaluation of the received proposals. All proposals which arrive after the due time will be considered to be invalid and consequently ignored.
- Duration: it is the interaction activity time period. It represents the time elapsed between the sending of a request message and the reception of the response. Duration includes two periods: transmission time and response time (task duration).

To model these issues, we adopt TCPN models because, besides their simplicity, they are particularly suitable in the modeling, simulating and analyzing of timed concurrent systems and, moreover, they enhance powerful tools for validation and verification. Our model proposes the modeling of these time constraints as well as the interaction sequence following the contracting phase. Our work contributes to the formal design of the temporal interaction aspects for negotiation systems. This contribution can be enumerated as follows: firstly, we present and we implement the proposed model using CPN Tools. We analyze it by means of the simulation and the state space techniques for various values of the protocol parameters namely the deadline and the number of participants. Secondly, we prove that the above mentioned key properties of the protocol are satisfied.

The rest of this paper is structured as follows: Section II introduces the contract net protocol. In Section III, the temporal interaction constraints are described and a formal definition of TCPN is presented. Section IV shows how the CNP enriched by temporal aspects is modeled in terms of TCPN. We verify this model in Section V. Lastly, Section VI concludes the paper and gives some perspectives.

II. The Contract Net Protocol

In CNP as described by FIPA [2], a manager and participants interact with one another to find a solution for a problem through a four-stage negotiation process. The manager initiates the negotiation process by issuing a Call

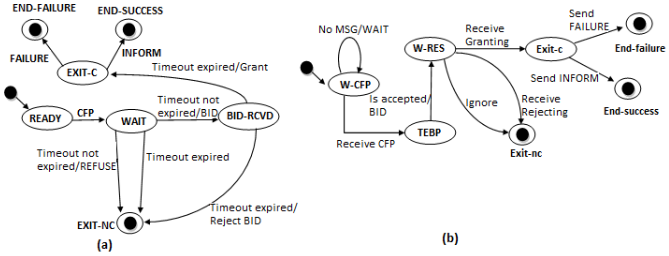


Figure 1. Internal behavior of the manager and the participant agents.

Table I
REPRESENTATION OF STATES.

Manager (Initiator)	Participant
READY (READY to send a CFP)	W-CFP (Waiting for CFP)
WAIT (Waiting for bids or for time-out)	TEBP (Task evaluation and bid preparation)
BID-RCVD (Bid received)	W-RES (Waiting for result)
EXIT-NC (EXIT with no contract)	Exit-nc (exit with no contract)
EXIT-C (EXIT with contract)	Exit-c (exit with contract)
END-SUCCES (END of negotiation with SUCCESS)	End-success (end of task execution with success)
END-FAILURE (END of negotiation with FAILURE)	End-failure (end of task execution with failure)

For Proposals (CFP) announcing the task specification to a number of potential participants. The CFP includes a deadline by which the participants must respond with bids. Participants evaluate the CFP and decide whether to answer with a refusal message or with a proposal to perform the task. Once the deadline expires, the manager evaluate all the received proposals (in due time) and, in turn, awards the contract to the most appropriate participant which becomes a contractor. The manager ignores any proposal that arrives beyond the deadline. The contractor performs the task and sends to the manager an informing message, which can be an error one in the case of a failure. Consequently, the negotiation process includes several scenarios depending on whether the bid process ends with or without a contract, and as the execution of the task ends with or without a success. Therefore, the manager and the participants can reach various states during this process. We suggest to represent the internal behavioral of both types of agents by means of AUML2 statesharts diagrams [3] . These diagrams define the different states that will be later used in the TCPN model of the protocol. Figure 1 (a) and Figure 1 (b) illustrate respectively the internal behavior of the manager and the participant agents. Table I summarizes the various states and their semantics.

III. MODELING TEMPORAL ASPECTS OF INTERACTION

Two temporal interaction aspects are specified by FIPA [2]: duration constraint and deadline constraint. The first one is the interaction activity time period which includes the two periods: transmission time and response time. In our model, we have assumed that the transmission time is infinitesimal and can consequently be ignored. On the other hand, the response time would depend on the defined deadline and hence we would propose a function to estimate it. The second temporal aspect, deadlines, is a time limit for the message exchange. The manager sets a time constraint (timeout) on the CFP message and participants must respond within this time limit, otherwise the response will be ignored. It is a synchronous communication with a limited waiting time. The expiration of the deadline implies the execution of other alternatives, that is why we propose to model this constraint by a timeout mechanism. We adopt using TCPN techniques to represent these temporal interaction aspects. In doing so, we assume a global clock.

A. Timed CPN

The concept of time was not explicitly provided in the original definition of Petri nets. As described in [10], we distinguish three basic ways of representing time in CPN: Firing Durations (FD), Holding Duration (HD) and Enabling Duration (ED). Choosing one of these three techniques depends strongly on the system to be modeled and its specifications. We should note, however, that it is natural to use HD technique in modeling the most processes as transitions represent operation event which, once starts, it does not stop until it ends. It is exactly the case of the system we are modeling. In HD technique, there are two types of tokens: available and unavailable. Available tokens can enable transitions whereas unavailable ones cannot. When a transition, which is assigned a duration, fires, removing and creating tokens are done instantaneously. However, the created tokens are not available to enable new transitions until they have been in their output place for the time specified by the transition which created them. For more details concerning these three techniques of time modeling, the reader can refer to [10]. CPN versions which use HD technique define implicitly the notion of tokens's unavailability by attaching to these tokens a timing attribute called a timestamp.

B. Formal definition of TCPN with Holding Durations

To represent tokens with timestamps we adopt the notation given by [11]. Each token carries a timestamp preceded by the @ symbol. For instance, 2 tokens with timestamp equal to 10 are noted 2@10. The timestamp specifies the time at which the token is ready to be removed by an occurring transition. Timestamps are values belonging to a Time Set TS which is equal to the set of non negative integers N+. The timed markings are represented as collection of timestamps, there are multi-sets on TS: TS_{MS}. The formal definition of TCPN using holding durations is as follows: TCPN = (Σ, f, M_0) where:

- Σ is a colored PN as described in [11]
- **f**: T →TS represents the transition function which assigns to each transition t ∈ T a non negative determinist duration

Table II
REPRESENTATION OF MESSAGES IN THE TCPN MODEL.

Messages issued by the manager	Messages issued by the participant
CFP (Call For Proposals)	BID (BID)
GB (Grant Bid)	REFUSE (REFUSE CFP)
RB (Reject Bid)	FAILURE (task Execution FAILURE)
CB (Cancel Bid)	INF-DONE (INForm-Done)
	INF-RES (INForm-RESults)

- **M**: P $\rightarrow TS_{MS}$ is the timed marking, M_0 represents the initial marking of TCPN.

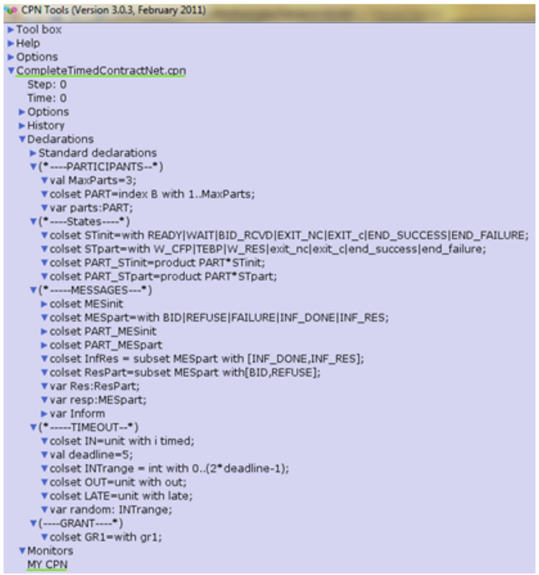
To determine whether tokens are available or unavailable, we define functions over the marking set M. So, For a marking M and the given model time (global clock), we have:
m: $P \times M \times TS \rightarrow$ N which defines the number of available tokens and n: $P \times M \times TS \rightarrow$ N which defines the number of unavailable tokens for each place of the TCPN model at a given instant k where k and the model time belong to TS. There are several computer tools which perform automatic validation and verification of Petri net models. Nevertheless, only CPN Tools permits, besides time representation, the modeling of high level petri nets particularly colored and hierarchical ones.

IV. TCPN MODEL OF THE CONTRACT NET PROTOCOL

When modeling a protocol, there are several design requirements and key characteristics that this protocol should satisfy. Authors in [13] have summarized these issues in 5 factors: state set, role set, rule set, action set and message set. By analogy with our case, study Table I describes the various states that negotiation process should reach and Table II defines messages exchanged between the manager and the participants. This section highlights our contribution and presents how Contract Net Protocol extended with the temporal aspects described in section II can be modeled as TCPN using CPN Tools. When creating the model, we have assumed some assumptions such as the reliability of the communication channel, and that participants have to reply to the CFP. Moreover, when modeling the interaction following the contracting phase, we should not take into consideration task duration, given that this work focuses on temporal interaction aspects. The manager starts evaluating bids after deadline expiration and lastly, the details of messages exchanged are excluded for an abstraction concern.

A. Declarations

Being inspired by [1], our TCPN model is readable and has a compact structure: For each type of agents, we use a single place which would store all its possible states. Similarly, we distinguish two places which represent a reliable channel for both directions of the communication. Figure 2, taken directly from CPN Tools, shows all the declarations used in the model.



```
CPN Tools (Version 3.0.3, February 2011)
▶Tool box
▶Help
▶Options
▼CompleteTimedContractNet.cpn
    Step: 0
    Time: 0
  ▶Options
  ▶History
  ▼Declarations
    ▶Standard declarations
    ▼(*----PARTICIPANTS--*)
      ▼val MaxParts=3;
      ▼colset PART=index B with 1..MaxParts;
      ▼var parts:PART;
    ▼(*----States----*)
      ▼colset STinit=with READY|WAIT|BID_RCVD|EXIT_NC|EXIT_c|END_SUCCESS|END_FAILURE;
      ▼colset STpart=with W_CFP|TEBP|W_RES|exit_nc|exit_c|end_success|end_failure;
      ▼colset PART_STinit=product PART*STinit;
      ▼colset PART_STpart=product PART*STpart;
    ▼(*-----MESSAGES---*)
      ▶colset MESinit
      ▼colset MESpart=with BID|REFUSE|FAILURE|INF_DONE|INF_RES;
      ▶colset PART_MESinit
      ▶colset PART_MESpart
      ▼colset InfRes = subset MESpart with [INF_DONE,INF_RES];
      ▼colset ResPart=subset MESpart with[BID,REFUSE];
      ▼var Res:ResPart;
      ▼var resp:MESpart;
      ▶var Inform
    ▼(*-----TIMEOUT--*)
      ▼colset IN=unit with i timed;
      ▼val deadline=5;
      ▼colset INTrange = int with 0..(2*deadline-1);
      ▼colset OUT=unit with out;
      ▼colset LATE=unit with late;
      ▼var random: INTrange;
    ▼(----GRANT----*)
      ▼colset GR1=with gr1;
  ▼Monitors
    MY CPN
```

Figure 2. Declarations for the TCPN model of the CNP.

B. Model structure

Figure 3 shows the TCPN diagram of CNP. The manager with the timeout mechanism is modeled in the left, the participants in the right. They communicate via a reliable not ordered channel represented by the two places INIT2PART and PART2INIT. The place INIT2PART only contains messages issued by the manager to the participants. Respectively, PART2INIT only contains messages of the participants to the manager. In this model the timed messages carry timestamps indicating when they should be available. Initially, the manager is in the state READY with respect to all the participants. Whereas, all the participants are in the state W_CFP. The place GRonly1 contains one token GR1 and all the other places are initially empty.

V. VERIFICATION OF THE MODEL

Verification is a method to exhaustively examine a design and check to make sure certain predefined key properties are met. There are several software tools to automate this task, however, CPN Tools [9] is currently the most used tool for high level Petri nets particularly for the timed colored ones (TCPN). This tool helps us to assess the correctness of the model.

A. Simulation

Using CPN simulator, we have conducted several automatic and interactive simulations which help us to identify and resolve several omissions and errors in the design. In addition to that, these simulations show that the protocol always seems to terminate in the desired coherent state. That is, it works correctly. Simulation also shows that the characteristics such as concurrency and validity are satisfied. This makes it likely that the protocol works

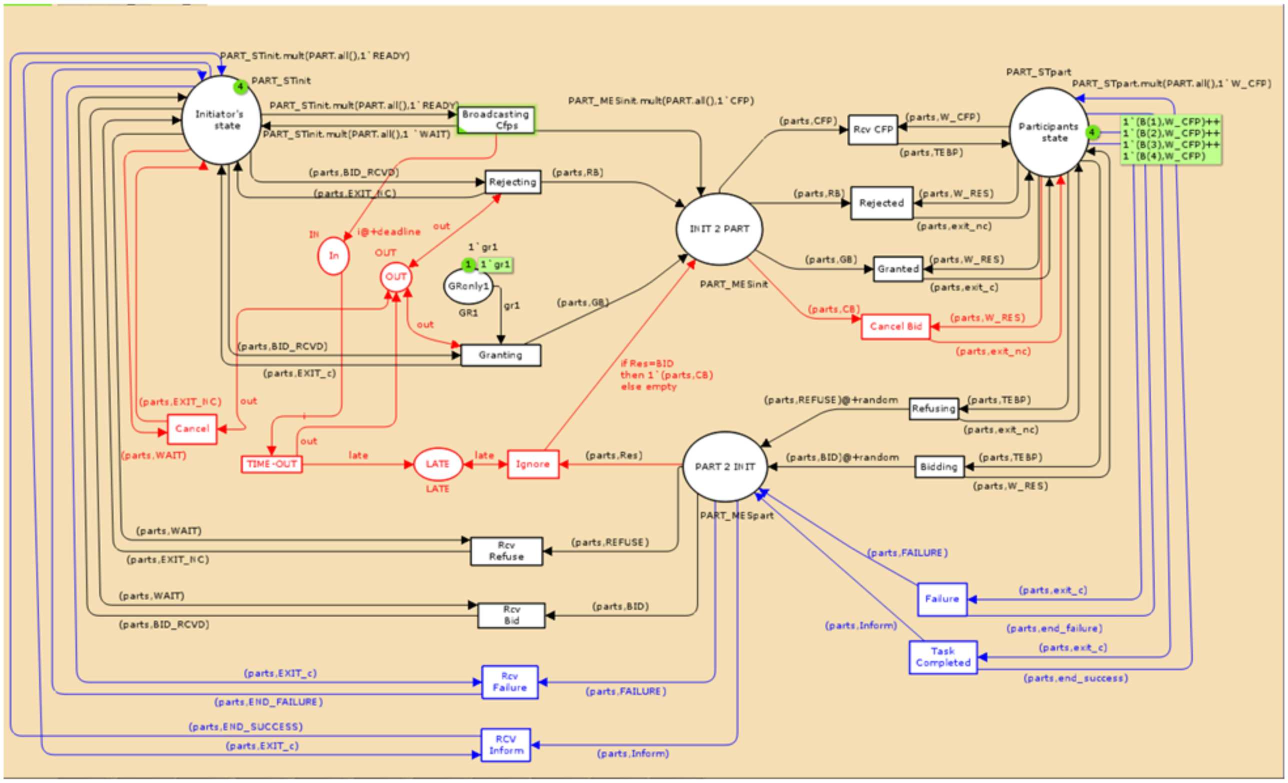


Figure 3. TCPN diagram of the contract net protocol.

correctly but it cannot guarantee that simulation covers all possible executions. That is why simulation cannot be used to verify other functional and performance properties such as the absence of deadlocks and others. However, State space analysis techniques allow us to verify if the system satisfies these behavioral properties.

B. State space analysis

With regard to untimed CPN models, calculating timed state space is a non trivial task and can be quite difficult and time consuming. This is because the reachability graph is too large and can be infinite even if the state space of the corresponding untimed CPN model is finite. This is due to the fact that several timed markings including global clock and timestamps can be different even if the corresponding untimed markings are identical. That is why we have to use some CPN ML queries to verify some properties.

Model Correctness. In this section, we verify the absence of deadlocks and the consistency in beliefs between the manager and the participants. Table III presents the state space analysis results. It shows the properties of the state space obtained by varying the parameter MaxParts from 1 to 4 and the parameter deadline from 1 to 5. The analyzing of the property DeadMraking allows us to verify the model correctness. Each dead marking corresponds to a terminal state of the negotiation protocol. All dead markings are obtained after the deadline expiration, ie, from t=d to t=2*d-1 (proposed estimation for the participants response time), for each discrete value of t belonging to this interval. For any value of MaxParts, one of the dead markings corresponds to an end of negotiation without a contract. In this marking, all the participants are in the state exit_nc and the manager in the state EXIT_NC with respect to all the participants. This is illustrated by the marking 14 in Figure 4. The description of this node shows that the place GRonly1 has still the token GR1 implying that none bid had been granted. The place In is empty, signifying that the deadline has expired and the timeout has fired. This particular dead marking is acceptable because the manager may reject all the bids or may not receive any bid in the due time. Among the rest of the dead markings, we distinguish those calculated at t=d and those obtained at t>d:

At t = d and for any values of MaxPArts: besides the particular dead marking mentioned above, the dead markings calculated at this time corresponds to the end of negotiations where a contract has been awarded to one participant (i=1..MaxParts) while the rest of negotiation with the rest of participants has ended without a contract. Therefore, P_i changes state to exit_c, performs the task which can ends by a success or a failure. P_i can, then, be in the state end_success or end_failure respectively. At the same time, the manager which was in the state

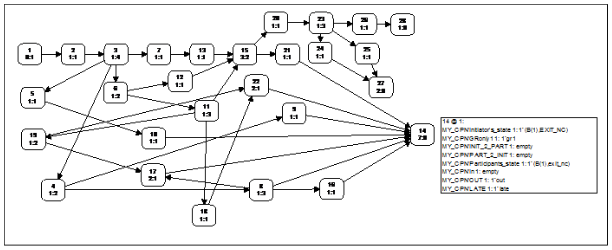

14 @ 1:
MY_CPN'Initiator's_state 1: 1`(B(1),EXIT_NC)
MY_CPN'GRonly 1 1: 1`gr1
MY_CPN'INIT_2_PART 1: empty
MY_CPN'PART_2_INIT 1: empty
MY_CPN'Participants_state 1: 1`(B(1),exit_nc)
MY_CPN'In 1: empty
MY_CPN'OUT 1: 1`out
MY_CPN'LATE 1: 1`late

Figure 4. State space for (MaxParts = 1 et d = 1).

EXIT_C with respect to P_i (and EXIT_NC with respect to the rest of the participants) changes to END_SUCCESS or END_FAILURE with regard to P_i. All the other participants P_j (j≠i) are in the state exit_nc. Thus, we can deduce that at t=d and for any value of MaxPArts we have:

NumberDeadmarkings = (2*MaxParts +1).

The rest of the dead markings is calculated at t>d which correspond to scenarios after the fire of the timeout where at least one participant is not in the due time. Two cases can be distinguished: a particular case of a single participant (MaxParts=1) and a general case of several participants (MaxParts $>$ 1):

t $>$ d and MaxParts = 1: this is particular because the single participant may miss the deadline and, consequently, changes state to exit_nc because of the canceling of its late response. The manager is in the state EXIT_NC with respect to this participant. This corresponds to the end of negotiation without a contract caused by the deadline overrun. This dead marking is reached for any discrete value of t where d<t>=2*d-1, ie, (d-1) times and thus we deduce:

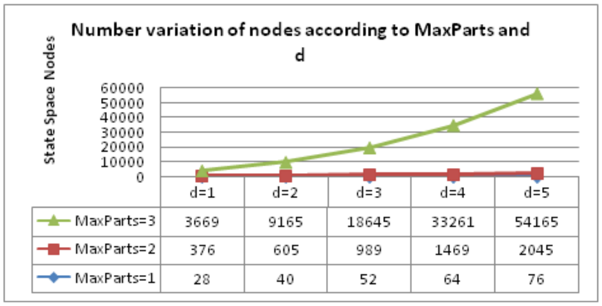
$$NumberDeadmarkings = 2 * MaxParts + d \qquad (1)$$

which is equal in this case to (2+d).

t $>$ d and MaxParts $>$ 1: all the dead markings calculated after the timeout and for each discrete value in the interval (d..2*d-1) are similar to those obtained at t=d. The only difference is that the global clock values and the timestamps of the tokens differ. Thus, these are equivalent timed markings. Consequently, we obtain (d-1) times the same number of dead markings , ie, (d-1)* (2*MaxParts +1) and, therefore, we deduce:

$$NumberDeadmarkings = (2 * MaxParts + 1) * d \quad (2)$$

All these dead markings are desired terminal states of the protocol. This discussion justifies that the protocol works correctly and the beliefs between the manager and the participants are consistent. Also, it should be noted that if for a given marking two or more transitions are enabled, then the choice of the transition to fire is non-determinist. This means that our system satisfies concurrency and non-determinism which are key characteristics. About the communication channel, we note that at the end of negotiations, the places PART2INIT and INIT2PART are empty, signifying that there is no unprocessed messages in the network, proving, hence, that the property of cleaning the network from late messages is satisfied.



	d=1	d=2	d=3	d=4	d=5
MaxParts=3	3669	9165	18645	33261	54165
MaxParts=2	376	605	989	1469	2045
MaxParts=1	28	40	52	64	76

Figure 5. Number variation of the reachability graph nodes according to Maxparts and the deadline.

Absence of livelocks and correct termination. Table III shows that the size of the state space increases exponentially with the number of participants and the value of the deadline. This is illustrated by the graph of the Figure 5. The large number of nodes and particularly of dead markings is essentially caused by the increasing value of the deadline. The reason for this is that the timing information makes more markings distinguishable and contributes to the presence of more nodes in the state space leading to several equivalent timed markings. To verify that all the dead markings for all the values of MaxParts specified in Table III form a home space, we have used the CPN ML function HomeSpace (ListDeadMarkings()) which evaluates to true. This confirms that there is no livelocks and the system will always terminate correctly. Table III also shows that, for all values of MaxParts examined, the number of nodes and arcs in the SCC graph always remains the same as that of the state space, this implies that there is no cyclic behavior in the system, which is expected. From Table III, we conclude that there is no live transitions because of the presence of dead markings.

Absence of dead code. A dead code corresponds to a dead transition. According to table III, there is no dead transitions in the system for all values of MaxParts examined, this implies that all the specified actions are executed.

Channel bound. Table III shows that the communication channel is bounded by the MaxParts value examined, this confirms that the manager issues a single message to each of the participants and then MaxParts messages. Similarly, each participant issues, at a given moment, one message to the manager justifying the limit of MaxParts responses.

VI. Conclusion and perspectives

In this paper, we have proposed a TCPN model of the contract net protocol with temporal aspects. We have used

Table III
STATE SPACE ANALYSIS RESULTS AS A FUNCTION OF THE PARAMETERS MAXPARTS AND DEADLINE (D).

Properties	**MaxParts=1**					**MaxParts=2**					**MaxParts=3**					**MaxParts=4**	
	d=1	d=2	d=3	d=4	d=5	d=1	d=2	d=3	d=4	d=5	d=1	d=2	d=3	d=4	d=5	d=1	d=2
State Space Nodes	28	40	52	64	76	317	605	989	1469	2045	3669	9165	18645	33216	54164	42337	140513
State Space Arcs	38	53	68	83	98	801	1357	2081	2973	4033	14113	30143	55863	93817	146549	221393	619193
Time (seconde)	00	00	00	00	00	00	00	00	01	02	07	33	161	404	831	1298	16119
SCC nodes	28	40	52	64	76	317	605	989	1469	2045	3669	9165	18645	33216	54164	42337	140513
SCC Arcs	38	53	68	83	98	801	1357	2081	2973	4033	14113	30143	55863	93817	146549	221393	619193
Dead Markings	3	4	5	6	7	5	10	15	20	25	7	14	21	28	35	9	18
HomeSpace	true	true	true	true	true	true	true	true	true	true	true	true	true	true	true	true	true
Dead Transition Instances	None	None	None	None	None	None	None	None	None	None	None	None	None	None	None	None	None
Live Transition Instances	None	None	None	None	None	None	None	None	None	None	None	None	None	None	None	None	None
Channel bound	1	1	1	1	1	2	2	2	2	2	3	3	3	3	3	4	4

the simulation and the state space analysis techniques to verify some key properties of the proposed model for different values of both parameters MaxParts and deadline. In addition to have proved that the deadline is always respected, we have also proved the beliefs consistency between the manager and the participants and that the protocol works and ends correctly. The properties namely concurrency, absence of livelocks and absence of dead code were verified too. Furthermore, we have shown how the number of dead markings (terminal states) is related to both MaxParts and deadline parameters. The channel bound is, however, related to only the MaxParts parameter. As perspectives, we would like to use advanced state space reduction methods [15] like equivalence classes [5], [7] to alleviate the impact of the state explosion problem which is most accentuated for timed models. In doing so, we would verify the model for wider values of MaxParts and deadline. We would also like to model real time contract net [6], [16], [18] where, besides interaction aspects, time constraints related to task execution would be considered. These extensions would concern more complex versions of CNP. On the other hand, we would like to model a fault tolerant CNP so that the manager provides a fault tolerant behavior if ever the contractor crashes during task performing.

REFERENCES

[1] J. Billington and A. Gupta, *Effectiveness of Coloured Petri Nets for Modelling and Analysing the Contract Net Protocol*, Proc. Eighth Workshop and Tutorial on Practical Use of Coloured Petri Nets and the CPN Tools, Aarhus, Denmark, 2007, pp. 49-65 (ISSN 0105 8517).

[2] *FIPA*, Foundation for intelligent physical agents 2003. *FIPA Modeling Area: Temporal Constraints*. Retreived May 10, 2012, from *http://www.fipa.org*

[3] *Agent Unified modeling language, AUML*. Retreived May 15, 2012, from *http://www.AUML.org*

[4] S. Aknine, S. Pinson, and M. F. Shakun, *An Extended Multi-Agent Negotiation Protocol, Autonomous Agents and Multi-Agent Systems* 8(1), pp. 5-45 (2004).

[5] W.M.P van der Aalst, *Interval timed coloured petri nets and their analysis*, In Application and Theory of Petri Nets 1993, Proc. 14th International Conference, volume 691, pp. 453-472, Chicago, (USA), 1993. Springer-Verlag, Lecture Notes in Computer Science.

[6] L. Qiaoyun , L. Jiandong, D. Dawei, and K. Lishan, *An extension of contract net protocol with real time constraints*. Wuhan University Journal of Natural Sciences. Wuhan University Journals Press. ISSN:1007-1202 (Print) 1993-4998, Volume 1, Number 2 / juin 1996, pp. 156-162.

[7] B. Berthomieu, *La méthode des Classes d' Etats pour l'Analyse des Réseaux Temporels - Mise en Oeuvre, Extension à la multi-sensibilisation, Modélisation des Systèmes Réactifs*, In Proc. of MSR'2001, pp. 254-263, Toulouse, France, 2001. Hermes Sciences.

[8] R. G. Smith, *The Contract Net Protocol: High-Level Communication and Control in a Distributed Problem Solver*, IEEE Trans. Computers 29(12): 1104-1113 (1980)

[9] *CPN tools homepage*. Retreived May 10, 2012, from *cpntools.org/*.

[10] F.D.J. Bowden, *A brief survey and synthesis of the roles of time in Petri nets*, Mathematical and Computer Modelling 31 (2000) pp. 55-68.

[11] K. Jensen and L. M. Kristensen, *Coloured Petri Nets - Modelling and Validation of Concurrent Systems*, Springer, July 2009.

[12] J. Shujuan , Q. Tian and Y. Liang, *A Petri-Net-Based Modeling Framework for Automated Negotiation Protocols in Electronic Commerce*, Lecture Notes in Computer Science, 2009, Volume 4078/2009, pp. 324-336.

[13] F. S. Hsieh, *Automated Negotiation Based on Contract Net and Petri Net*, Lecture Notes in Computer Science, vol. 3590, pp. 148-157, 2005. (SCI).

[14] J. Billington, A. K. Gupta, and G. E. Gallasch, *Modelling and Analysing the Contract Net Protocol - Extension Using Coloured Petri Nets*, Lecture notes in computer science, 2008, NUMB 5048, pp. 169-184, Springer-Verlag.

[15] M. A. Piera and G. Music, *Coloured Petri net scheduling models: Timed state space exploration shortages*, Mathematics and computer in simulation 82, (2011), pp. 428-441, Elsevier.

[16] N. Dragoni, M. Gaspari, and D. Guidi, *An ACL for specifying Fault-Tolerant protocols*, AI*IA 2005: Advances in Artificial Intelligence 9th Congress of the Italian Association for Artificial Intelligence, Italy 2005 pp. 237-248.

[17] L. Changyou and W. Haiyan, *An Improved Contract Net Protocol Based on Concurrent Trading Mechanism*, iscid, vol. 2, pp. 318-321, 2011 Fourth International Symposium on Computational Intelligence and Design, 2011.

[18] W. L. Yeung, *Behavioral modeling and verification of multi-agent systems for manufacturing control*, Expert Systems with Applications, 38(11):13555-13562, 2011.

State Complexity of Hidden Markov Model

Jamil Ahmed
Department of Computer Science
University of Western Ontario
London, Ontario, Canada N6A 5B7
jahmed6@uwo.ca

Stephen M. Watt
Department of Computer Science
University of Western Ontario
London, Ontario, Canada N6A 5B7
watt@uwo.ca

Sherjil Ahmed
Department of Software Engineering
Venture Chest
Shahra-e-Faisal, Karachi, Pakistan
sherjil.ahmed@venturechest.com

***Abstract*—A Classification problem can be viewed as a problem of assigning a category or class to a given input. The Hidden Markov model is a well known stochastic model that is used to solve classification problems and has been widely exploited in diverse computing applications ranging from speech, acoustics, gesture recognition to part-of-speech tagging, cryptography to Google page rank and the list goes on. State Complexity of Deterministic Finite automata is now a well established research area. State complexity of Deterministic Finite Automata defines the total number of states in the minimal Deterministic Finite Automata. State complexity, if known, of a given automata helps to realize how expensive the application would be that will exploit that automata. Similarly, if known, the state complexity of the Hidden Markov Model will help to know the complexity of a computing application that exploits that Hidden Markov Model. In this paper, we have explored several, yet unpublished, important facts about the Hidden Markov Model including the state complexity of Hidden Markov Model and the diagram of 2nd order Hidden Markov Model (Fig. 3). Our discussion of the Hidden Markov Model is unique in the sense that we present a complete diagram of the 1st order Hidden Markov Model (Fig. 2) with all estimated parameters for a given input sequence. We explicitly define a generalized rule to give "Dimension of Transition probability matrix of HMM" which is also not available in the literature yet. We present a generalized rule to draw the Mth order Hidden Markov Model diagram for M greater than 1. We present the generalized state complexity of the Mth Order HMM, the state complexity of the diagram for the "Training of Mth Order HMM" and also present the diagram for second order Hidden Markov Model.**

Keywords – State complexity; Hidden Markov Model.

I. Introduction

The Hidden Markov Model (HMM) has states, input symbols and transitions much like a deterministic finite automata (DFA). Transitions between states of HMM are labeled with the probabilities, unlike DFA, defining how likely that transition is to take place. States of HMM actually represent the classes or categories that we intend to assign to the symbols of input sequence. How it works is that first we define the classes or categories that will be represented by the states of HMM. In order to assign categories or states to input symbols, we process the input sequence on the HMM. The sequence of states assigned to as input sequence is called the output sequence. Many recent techniques exploit the Hidden Markov Model as in [10] [11][12].

Section II defines the HMM as well as the related terminology. State Complexity of Deterministic Finite automata is now a well established research area [8]. In Section III, we define the state complexity of the 1st order HMM. In Section IV, we define the rule for drawing the Mth order HMM diagram for M>1. In Section V, we define the transition probability matrix dimensions of Mth-Order HMM. In Section VI, we define the state complexity of the Mth-Order HMM for M>1. In Section VII, we define the 2nd order HMM Example. In Section VIII, the diagram of the 2nd Order HMM is presented. Section IX concludes the paper.

II. Hidden Markov model

We present an overview and terminologies for the HMM.

A. Formal Definition

A HMM is a tuple (S, ∑, ∏, A,B)

- A set of states: $S=\{S_1,S_2,\ldots S_m\}$
- A set of input symbols: $\sum = \{O_1,O_2\ldots.O_k\}$
- Initial state probability: $\prod = \sum_{i=1}^{m} \prod \quad = 1$
- Transition probability matrix: $A=\{a_{i\,j}\}$
- Emission probability matrix: $B=\{b_{i\,j}\}$

In order to show the working of the HMM, usually, we add two additional states "Start" and "End" states. Dimensions of the transition probability matrix and emission probability matrix, above, are for the 1st order HMM and are contingent to the order of the HMM. Transition probability is termed as "$P(t_i \mid t_{i-1})$" as mentioned in Section V.

B. Training of HMM

Processing of the input sequence on a given HMM develops a HMM training diagram, which we call "HMM with all estimated parameters". Those parameters include transition/emission probabilities. One such but partial HMM training diagram is given in [1]. "HMM with all estimated parameters" is also called "training of HMM" [9]. This "training of HMM" gives all possible sequences of states or categories which can be assigned to the input sequence. All these possible sequences of states or categories are also called hidden states sequence because these sequences are not known unless we train the original HMM for a given input sequence. Each hidden state sequence of states or categories assigned to the input sequence is also called the output sequence. Brute Force expansion of the HMM is usually intractable for most real world classification problems, as the number of possible hidden state sequences is extremely high and scales exponentially with the length of input sequence.

All output sequences that are assigned to a given input sequence have probability of likelihood. This probability defines how likely the output sequence is as an appropriate assignment for the input sequence. These probabilities of likelihood are calculated with the help of transition probabilities and emission probabilities of the original HMM. For a given input sequence, HMM chooses the state sequence that maximizes in the following formula.

P(input symbol/state) * P(state/previous 'M' states)

In the above formula, Transition probability = P(state/previous 'M' states) and Emission probability = P(input symbol/state) . Transition probabilities (TPs) of each state of HMM either depend on one previous state or more than one previous state. If they depend on one previous state (i.e., M=1), this HMM is called a 1st order HMM. If all TPs of states of HMM depend on two previous states (i.e., M=2), this HMM is called a 2nd order HMM and so on.

C. 1st Order HMM Example

Below, we have given an example HMM of two states, cold and hot. Major components of the HMM are also mentioned in the section ahead as well as the dimensions of the transition probability matrix.

We simulate a real world phenomena in Fig. 1 related with the 'weather of a day'. We suppose we can have either a Hot or Cold day. We know the probability (transition probability) of the next day is either cold or hot depending upon the weather of the previous day. This 1st Order HMM represents how many ice creams servings, (e.g., 8, 7, 6 etc), a person is likely (emission probability) to eat on a given day depending upon the "weather of that day". Further we show the training diagram of the 1st order HMM diagram in Section F.

We can see in Fig. 1 that each state of 1st Order HMM directly corresponds to one single category, e.g., Cold or Hot in this example. This is also shown in [4].

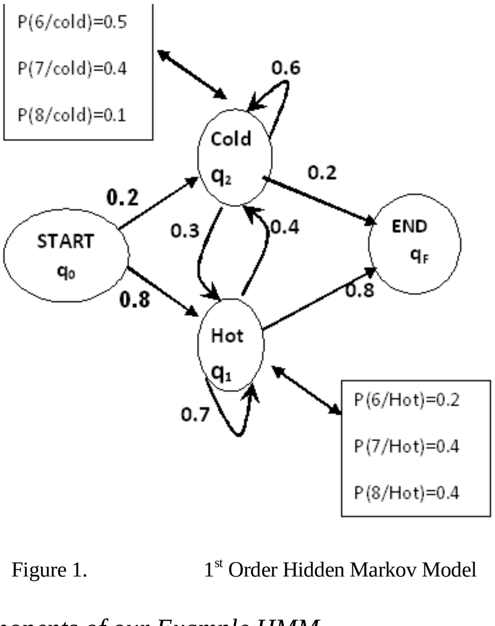


Figure 1. 1st Order Hidden Markov Model

D. Components of our Example HMM

Components of our HMM given in the previous Section are shown in the tabular form below.

TABLE I. Components of Hidden Markov Model

Q-Set of states	Q= {Hot, Cold} $=\{q_1,q_2\}$
Transition probability Matrix [2-dimensional 2*2]	a_{11} = 0.7 , a_{12}=0.3, a_{21}= 0.4 , a_{22}= 0.6
Vocabulary of Inputs	V= {6,7,8,9}
Input Observation under consideration for training of this HMM	(8 7 6), i.e., O_1 =8 ,O_2= 7, O_3=6
Emission Probability: $b_1(6)$=P[6/Hot],$b_1(7)$=P[7/Hot], $b_1(8)$=P[8/Hot],$b_2(6)$=P[6/Cold], $b_2(7)$=P[7/Cold],$b_2(8)$=P[8/ Cold]	$b_1(6)$=0.2 , $b_1(7)$=0.4, $b_1(8)$=0.4,$b_2(6)$=0.5, $b_2(7)$=0.4, $b_2(8)$=0.1
Start state and End state	q_0 , q_F
Transition probability from start state to q_1 and q_2	a_{o1} = 0.8 , a_{02}=0.2
Transition probability to End state from q_1 and q_2	a_{1F} = 0.8 , a_{2F}=0.2

E. Dimension of Transition probability matrix of 1st Order HMM

Dimensions of the transition probability matrix are based on the simple principle, i.e., Transition probability matrix shows probabilities of all the transitions from each state to all other states as well as to itself.

If 'S' is the number of states and 'M' is the order of HMM then the Transition probability matrix of the HMM is 'M+1' dimensional such as S^i *….* S^{M+1}. The superscript on 'S' represents the dimensions, i.e., how many times 'S' should appear.

In case S=3 and M=1, i.e., the 1st order HMM then the transition probability matrix is 2 dimensional, i.e., 3*3. So, if S=3, there are three states say Q=$\{q_1,q_2,q_3\}$, the two dimensional (3*3) matrix will be as follows:

a_{jk}= probability of transition from state j to k, $\forall$ j,k $\in$ Q

F. 1st Order HMM Training Diagram

We show a training of the HMM of the 1st Order for the input sequence "8 7 6". We process this input sequence on the HMM and try to find the most likely sequence of "weather of days" corresponding to the given "number of ice creams" a person eats each consecutive day.

The terminologies we have used are such that the equation "[CHH] $X_3(1)$=0.0032 * 0.14 = 0.000448" from Fig. 2 represents that this probability is the likelihood for assigning state sequence [CHH] to the input sequence "8 7 6" and it is finally calculated in the 3rd column for State q_1(HOT) by multiplying two other probabilities [CH] $X_2(1)$ and P(H/H) * P(6/H) because of "(1)" of Section H.

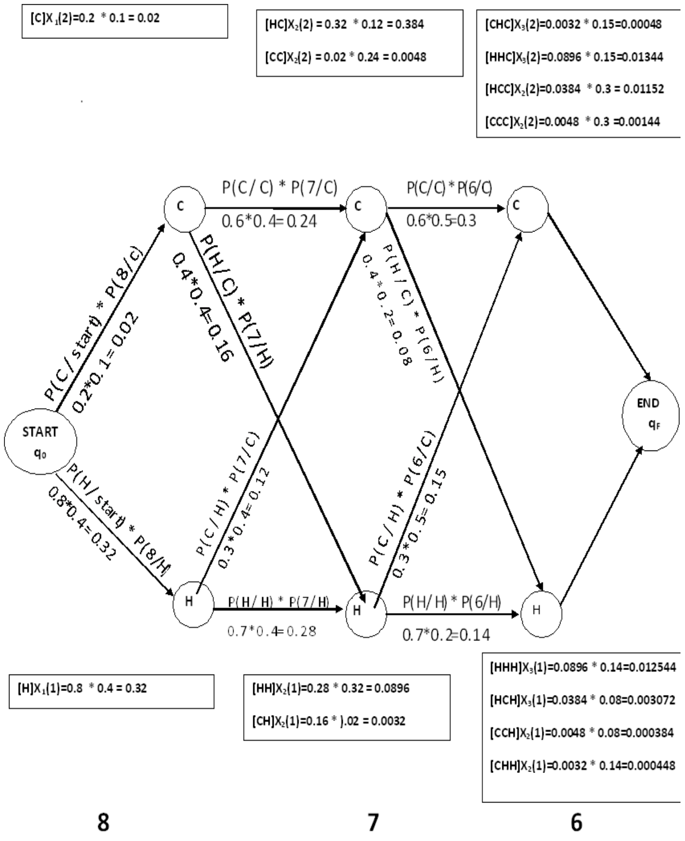

[C]X1(2)=0.2 * 0.1 = 0.02
[HC]X2(2) = 0.32 * 0.12 = 0.384
[CC]X2(2) = 0.02 * 0.24 = 0.0048
[CHC]X3(2)=0.0032 * 0.15=0.00048
[HHC]X3(2)=0.0896 * 0.15=0.01344
[HCC]X2(2)=0.0384 * 0.3 = 0.01152
[CCC]X2(2)=0.0048 * 0.3 =0.00144
P(C / C) * P(7/C)
0.6*0.4=0.24
P(C/C) * P(6/C)
0.6*0.5=0.3
C
C
C
P(C / start) * P(8/C)
0.2*0.1=0.02
P(H / C) * P(7/H)
0.4*0.4=0.16
P(H / C) * P(6/H)
0.4*0.2=0.08
START
q0
END
qF
P(H / start) * P(8/H)
0.8*0.4=0.32
P(C / H) * P(7/C)
0.3*0.4=0.12
P(C / H) * P(6/C)
0.3*0.5=0.15
H
H
H
P(H / H) * P(7/H)
0.7*0.4=0.28
P(H / H) * P(6/H)
0.7*0.2=0.14
[H]X1(1)=0.8 * 0.4 = 0.32
[HH]X2(1)=0.28 * 0.32 = 0.0896
[CH]X2(1)=0.16 *).02 = 0.0032
[HHH]X3(1)=0.0896 * 0.14=0.012544
[HCH]X3(1)=0.0384 * 0.08=0.003072
[CCH]X2(1)=0.0048 * 0.08=0.000384
[CHH]X2(1)=0.0032 * 0.14=0.000448
8
7
6

Figure 2. 1st Order HMM Training diagram

[CHH] is one possible output state sequence of the "weather of days" corresponding to the given input sequence "8 7 6" of the "number of ice creams" a person eat each consecutive day.

G. *All Possible Output Observations*

This table shows all possible output observations that can be assigned to the input observation along with their likelihood for the given input observation.

Table II. Output Observation Probabilities

Output Observation	Probability
H H H	0.012544
C C C	0.00144
H C C	0.01152
H H C	0.01344
C H H	0.000448
C C H	0.000384
H C H	0.003072
C H C	0.00048

We conclude that "H H C" is the most likely output observation for the input sequence "8 7 6" because its probability is found to be highest, i.e., 0.01344 using "(2)"of Section H.

H. *Training of 1st Order HMM*

Transition probabilities (TPs) of each state of the HMM either depend on one previous state or more than one previous state. If they depend on one previous state (M=1), the HMM is called 1st order HMM. If all TPs of states of HMM depend on two previous states (M=2) then, the HMM is called 2nd order HMM and so on. We gave a HMM in Section C and then trained that model for the input sequence "8 7 6" in Section F. The trained model shows all calculated parameters. Transition probabilities of the 1st order model (i.e., M=1) are represented as below:

$$P(q_i \mid q_{i-1}) \qquad \forall\ q_1 \ldots q_S$$

For our example model we have given these probabilistic parameters along with the transitions of Fig. 1 in Section C. The likelihood of each possible output observation is calculated from the following formula.

$$[\,\Pi\ P(t_i \mid t_{i-1})P(w \mid t_i)] \qquad (1)$$

The most likely output observation is calculated from the following formula.

$$\underset{q_1 \ldots q_S}{\mathrm{Argmax}} \quad [\,\Pi\ P(q_i \mid q_{i-1})P(w \mid q_i)] \qquad (2)$$

The 1st Order HMM training diagram shows all possible output observations and their likelihood probability estimates by a brute force approach. The viterbi algorithm calculates only the most likely output observation sequences. For each input observation, we calculate its likelihood for both states, i.e., for both cold and hot. So, for three input observations we have shown three columns. Each column has state q_1(cold) and q_2(Hot).

III. State complexity of 1st Order HMM

Training of the HMM is a state diagram that shows how a given input observation sequence is processed by the underlying HMM. This Training model gives us all possible output observations or all possible classifications for a given input observation sequence.

If 'H' is a 1st order HMM, 'S' is the number of states of 'H'. 'X' is an input observation sequence. 'L' is the training diagram of 'H' for 'X'. 'K' is the number of observations in 'X'. 'M' is the total number of hidden states of 'L' which are trained for 'X'.

M= total number of hidden states of 'L'
i.e., M= state complexity of 'L'
Then, M = S * K

So, we can see the state complexity of the training diagram of the 1st order HMM depends on S (number of states of underlying HMM) and K (number of input observation of input sequence) such that for each input observation ('K' in total), we calculate its likelihood for all states ('S' in total). As we can see in Fig. 2, for each of the three input observations, we have to find the likelihood for both states, i.e., (K=3 and S=2, so M=2*3=6). There are 6 likelihoods calculated in total represented by 6 states of Fig. 2.

Two additional ('start' and 'end') states are also added to complete the Fig. 2. This state complexity is of the brute Force expansion, which is usually intractable for most real world classification problems, as the number of possible hidden states sequences is extremely high and scales exponentially with the length of the input sequence.

IV. Rule to draw Mth order HMM diagram for M>1

First of all, we have to define 'S' basic categories or states we intend to assign to the input symbols. In the case of the Mth order HMM, each of these 'S' categories will depend on 'M' previous categories.

Then, we will find all possible combinations for ‘M’ previous categories. The M^{th} order HMM diagram will have two kinds of states.

(1) The ‘S’ original states (or categories) that we intend to be ultimately assigned to input symbols.

(2) The states that represent all possible combinations of ‘S’ states of length ‘M’. (i.e., S^M).

Each of these ‘S’ distinct states and S^M distinct combinations of states will be combined to form M^{th} order HMM as shown in Fig. 3 for M=2.

Example1: if we have two basic categories for “Weather of the Day”, i.e., Cold (C) or Hot (H), i.e., S=2 and M=2. The total possibilities for ‘M’ previous categories are as follows:

Table III. Possible Previous States for M=2

Possibility for M=2 previous state	Remarks
YZ= HH, CH, HC, CC	Last category is Z, second last is Y. Each of the Y and Z represent either H or C

Since, the total possibilities of M previous categories are $S^M=2^2=4$.

Example2: if we have two basic categories for “Weather of the Day”, i.e., Cold (C) or Hot (H), i.e., S=2 and M=3. The total possibilities for ‘M’ previous categories are as follows:

Table IV. Possible Previous States for M=3

Possibility for M=3 previous state	Remarks
XYZ=HHH,HHC,HCH,CHH, HHC, HCC, CHC,CCC	Last category is Z, second last is Y and third last is X Each of the X, Y and Z represent either H or C

Since, the total possibilities of M previous categories are $S^M=2^3=8$.

V. Transition probability matrix Dimensions of M^{TH} Order HMM

The M^{th} order HMM means that the probability of transiting to each state depends on last ‘M’ states.

If ‘S’ is the number of states (categories) and ‘M’ is the order of HMM, then the Transition probability matrix of the HMM is ‘M+1’ dimensional such as S^i *….* S^{M+1}. The superscript on S represents the dimensions, i.e., how many times S should appear.

In case S=2 and M=2, i.e., 2nd order HMM then the transition probability matrix is 3 dimensional, i.e., 2*2*2.

Further, $A=\{a_{ijk}\}$ such that $\sum a_{ijk} = 1$

Transition a_{ijk} represents the transition probability to state ‘k’ depending upon the last two categories j & i respectively as shown in Fig. 3.

VI. State Complexity of M^{TH} Order HMM for M>1

Theorem:
State complexity of M^{th} Order HMM = $S+S^M$ and
State complexity of M^{th} Order HMM training diagram= N*K.
∀ M>1.

Proof:

Let ‘S’ represent the total number of basic categories that we intend to assign to input symbols and M is the order of the HMM. Let ‘N’ be the state complexity of the M^{th} order HMM. Recall our basic idea of the higher order HMM, that each state (or category) depends on previous ‘M’ states. The M^{th} order training diagram of the HMM will have two kinds of states.

(1) The ‘S’ original states (or categories) that we intend to assign to input symbols.

(2) The states that represent all possible combinations of ‘S’ states of length ‘M’. (i.e., S^M)

Therefore, combining both the above kinds of states for the M^{th} order HMM will give us a total number of states or state complexity of the M^{th} Order HMM as below:

N=State complexity of M^{th} Order HMM = $S+S^M$.

For completeness, we add ‘2’ as well for the “start” and “end” state.

$$N= 2+S+S^M \qquad (3)$$

State complexity of its training model will depend on the number of observations of the input sequence similar to what is discussed in Section F.

Let ‘X’ be an input sequence. Let ‘K’ be the number of input observations of ‘X’. Let ‘P’ be the total number of states of training diagram of the M^{th} order HMM for ‘X’. Then,

$$P=N*K \qquad (4)$$

Equations “(3)” and “(4)” above prove the theorem.

All possible combinations of states, i.e., S^M is calculated in [7] as well. We have now given a generalized formula “(4)” for the state complexity of the higher order HMM diagram.

Indeed, $P(t_3|t_2,t_1)$ represents the transition probability of the 2nd order model because it says the probability of “t_3”depends on the last two (i.e., M=2) states “t_2” and “t_1”. Similarly, we know that $P(t_2|t_1)$ is the transition probability of the 1st order model because it says the probability of “t_2” depends on the last single (i.e., M=1) state “t_1” and this is mentioned in [2] as well.

Law and Chan [4] also defined that, for the M^{th} order HMM, all possible combinations of states should be S^M but do not explain how each of these S^M states will be connected and any higher order HMM diagram is also not illustrated.

VII. 2ND order HMM Example

For M=2, transition probabilities should depend on the last two states. We represent each transition probability by ‘a_{ijk}’.

This means that a_{ijk} = transition probability to state k depending on the last two states j & i respectively.

$$a_{ijk} = P(q_k \mid q_j,q_i)$$

In the 2nd order HMM, the transition probability matrix is three dimensional, i.e. a_{ijk}, as the probability of transiting to a new state depends on the last two previous states as mentioned in [5] as well. Hence the dimension of the transition probability matrix is S*S*S, where ‘S’ is the number of states.

Let’s suppose for our example diagram of 2nd order HMM, the number of states, i.e., S=2. The transition probability for each state will depend on all possible “pairs of states”. If we have S=2 states then the total number of pairs of states = S^2 = 4

The dimension of the transition probability matrix is S*S*S, i.e., 2*2*2 , which means, in total, 8 transitions represented in general by a_{ijk} .

Each a_{ijk} = the transition probability to state ‘k’ such that last two states are j & i respectively.

We write the following formula using “(3)”.

Total number of states of the 2nd order HMM = $2+S+S^2$

$= 2+2+2^2 = 8$

VIII. Diagram of 2nd Order HMM

We draw in this section the 2nd Order HMM diagram of the example discussed in Section VII. Let’s suppose we have S=2 states. The first state is called ‘Cold’ represented by C_1 and the second state is ‘Hot’ represented by H_2. The diagram of this example second order HMM is given in Fig. 3 with two additional start and end states. The transition probability matrix will contain 8 transition probabilities, i.e., $a_{ijk}=\{a_{111},a_{112},a_{121},a_{122},a_{211},a_{212},a_{221},a_{222}\}$. The other transition probabilities in the diagram q_{ik} and q_{ijk} are just for completion of the diagram and for the calculation of actual transition probabilities, a_{ijk}.

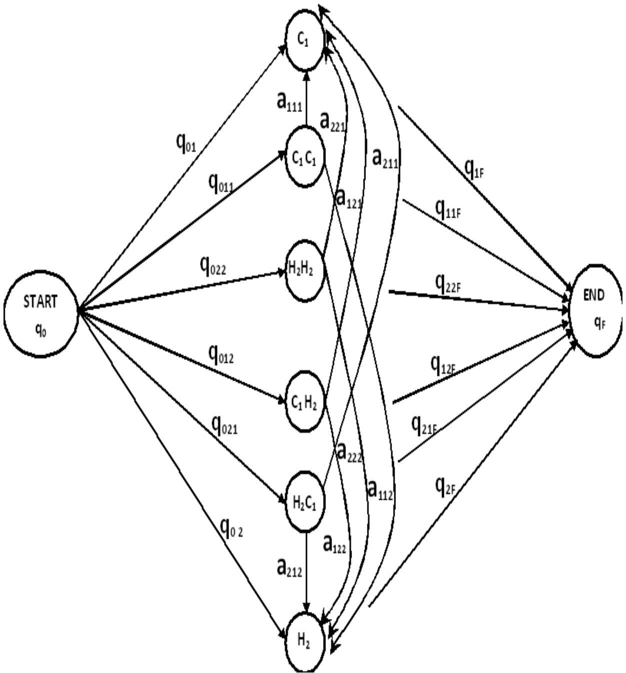


Figure 3. 2nd Order Hidden Markov Model diagram

We describe the transition probabilities, a_{ijk}, for the Fig. 3 in Table V.

Table V. Transition probability Description for 2nd Order Hidden Markov Model diagram

Transition Probability	Transition From State	Transition To State	Description
a_{111}	C_1C_1	C_1	The transition represents the probability of occurrence of day C_1 based on the last two days C_1 & C_1 respectively.
a_{112}	C_1 C_1	H_2	The transition represents the probability of occurrence of day H_2 based on the last two days C_1 & C_1 respectively.
a_{121}	C_1 H_2	C_1	The transition represents the probability of occurrence of day C_1 based on the last two days H_2 & C_1 respectively.
a_{122}	C_1 H_2	H_2	The transition represents the probability of occurrence of day H_2 based on the last two days H_2 & C_1 respectively.
a_{211}	H_2 C_1	C_1	The transition represents the probability of occurrence of day C_1 based on the last two days C_1 & H_2 respectively.
$a_{\ 212}$	H_2 C_1	$H_{\ 2}$	The transition represents the probability of occurrence of day $H_{\ 2}$based on the last two days C_1 & H_2 respectively.
a_{221}	H_2 H_2	C_1	The transition represents the probability of occurrence of day C_1 based on the last two days H_2 & H_2 respectively.
a_{222}	H_2 H_2	H_2	The transition represents the probability of occurrence of day H_2 based on the last two days H_2 & H_2 respectively.

IX. Conclusion And Future Work

In this paper, we presented the state complexity of the Mth Order HMM. We also presented the state complexity of the diagram for “Training of Mth Order HMM”. We explicitly defined HMM of different orders along with its training in a clear cut way. We presented a complete diagram of the “1st order training HMM” in which all possible sequences of categories are mentioned along with their probabilities of assigning these output observations to input observations. We explicitly defined a generalized rule to give the “dimension of Transition probability matrix of HMM”, which is also not available in the literature yet. We defined a generalized rule for drawing the Mth order HMM diagram for M>1 and also drew the HMM diagram for M=2, which is not available in the literature yet. Our study of the state complexity of HMM will thus help researchers to analyze the complexity of implementation of their proposed HMM based scheme. This paper has introduced the terminology, i.e., “State complexity of HMM” that will lead researchers to explore the state complexity of different kinds of HMM and similar stochastic models. Diagrams for Mth Order HMM for M>2 can also be explored. Since the complexity of computing applications that exploit HMM depends on the complexity of HMM used. Therefore, our analysis of state complexity of HMM will help to analyze the complexity of those computing applications that exploit HMM.

Acknowledgment

We acknowledge the suggestions of the Late Prof. Dr. Sheng Yu and Prof. Bob Mercer for this paper. This work is funded by the Higher Education Commission, Pakistan and the University of Western Ontario, Canada.

REFERENCES

[1] Daniel Jurafsky and James H. Martin, “Speech and Language Processing an Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition”, 2nd Edition 2008, pp. 146- 148, 177.

[2] Scott M. Thede and Mary P. Harper, “A Second-Order Hidden Markov Model for Part-of-Speech Tagging”, In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics (ACL '99). Association for Computational Linguistics, Stroudsburg, PA, USA, 1999, pp. 175-182, doi:10.3115/1034678.1034712.

[3] Thorsten Brants, “TnT: A Statistical Part-of-Speech Tagger”. In Proceedings of the Sixth Conference on Applied Natural Language Processing (ANLC '00). Association for Computational Linguistics, Stroudsburg, PA, USA, 2000, pp. 224-231, doi:10.3115/974147.974178.

[4] Hubert Hin-Cheung Law and Chorkin Chan, “N-th Order Ergodic Multigram HMM for Modeling of Languages Without Marked Word Boundaries”. In Proceedings of the 16th Conference on Computational Linguistics - Volume 1 (COLING '96), Vol. 1. Association for Computational Linguistics, Stroudsburg, PA, USA, 1996, pp. 204-209, doi:10.3115/992628.992666.

[5] Zaid Md Abdul Wahab Sheikh and Felipe Sanchez-Martinez, “A Trigram POS Tagger for Apertium Free/Open-Source Machine Translation Platform”, In Proceedings of the First International Workshop on Free/Open-Source Rule-Based Machine Translation, Alacant, Spain, November, 2009, pp. 67-74

[6] Fahim Muhammad, Hasan, Naushad Uzaman, and Mumit Kahn “Comparison of Different POS Tagging Techniques (N-gram, HMM and Brill’s Tagger) for Bangla”, Bangladesh, 2006, pp. 31-37

[7] Mohammed Albared, Nazlia Omar, and Mohd. Juzaiddin Ab Aziz, “Developing a Competitive HMM Arabic POS Tagger Using Small Training Corpora”, In Proceedings of the Third International Conference on Intelligent Information and Database Systems - Volume Part I (ACIIDS'11), Ngoc Thanh Nguyen, Chong-Gun Kim, and Adam Janiak (Eds.), Vol. Part I. Springer-Verlag, Berlin, Heidelberg, 2011, pp. 288-296.

[8] Sheng Yu, “State Complexity of Regular Languages”, Journal of Automata, Languages and Combinatorics”, Volume 6 Issue 2, May 2001, pp. 221-234.

[9] Greg Kochanski, “Markov Model, Hidden and Otherwise”, UTC, March, 2005, pp. 1-11

[10] Shaojun Zhao, “Named Entity Recognition in Biomedical Texts Using an HMM Model”. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications* (JNLPBA '04), Association for Computational Linguistics, Stroudsburg, PA, USA, 2004, pp. 84-87.

[11] Szymon Jaroszewicz, “Interactive HMM Construction Based on Interesting Sequences”. In Proc. of Local Patterns to Global Models (LeGo'08) Workshop at the 12th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'08), Antwerp, Belgium, 2008, pp. 82-91.

[12] Yu-Shu Chen and Yi-Ming Chen, “Combining Incremental Hidden Markov Model and Adaboost Algorithm for Anomaly Intrusion Detection”, In Proceedings of the ACM SIGKDD Workshop on CyberSecurity and Intelligence Informatics (CSI-KDD '09), ACM, New York, NY, USA, 2009, pp. 3-9, doi:10.1145/1599272.1599276.

[13] Rahul Khanna and Huaping Liu, “Distributed and Control Theoretic Approach to Intrusion Detection”. In Proceedings of the 2007 International Conference on Wireless Communications and Mobile Computing (IWCMC '07). ACM, New York, NY, USA, 2007, pp. 115-120, doi:10.1145/1280940.1280965

[14] Mohammed J. Zaki, Christopher D. Carothers, and Boleslaw K. Szymanski, “VOGUE: A Variable Order Hidden Markov Model With Duration Based on Frequent Sequence Mining”, Journal ACM Transactions on Knowledge Discovery from Data, Volume 4, Issue 1, Article 5, January 2010, 31 pages, doi:10.1145/1644873.1644878

The Application of Hierarchical Linear Model in the Study of Tumor Progression

Shuang Li
College of Electronic Information and Control Engineering
Beijing University of Technology
Beijing, China
690736804@qq.com

Pu Wang, Li Yingfang, Pei Yuchen
College of Electronic Information and Control Engineering
Beijing University of Technology
Beijing, China
{wangpu, fangliying}@bjut.edu.cn
530795037@qq.com

***Abstract*—This paper presents how to analyze the tumor progression data using hierarchical linear model. In the recent 20 years, cancer has been a serious threat to the human health and life; therefore, how to analysis the longitudinal data of patients effectively has become an urgent problem to be solved. In this article, we describe the principle of the hierarchical linear model and its application in the longitudinal study. Take a group of followed-up data of lung cancer patients for example; the output is the estimation of various parameters. I discussed the meaning of the various parameters in the hierarchical linear model and find out the influence of different fixed individual characteristics and the time-varying factors on the tumor progression. At the same time, we made a reasonable analysis of the strengths and weaknesses of hierarchical linear model.**

Keywords - hierarchical linear model; longitudinal data; cancer; time-varying.

I. INTRODUCTION

Longitudinal data, more specifically, can be considered as measurements on several variables for the same (groups of) individuals on a number of consecutive points in time. In this paper, we will be concerned with the analysis of individual growth as well as the average growth trend of a group of subjects [1]. For example, a longitudinal medical study offers a multifaceted way to analyze factors leading to a certain disease. This paper makes a further study for the physical tests and the physical symptoms of the subjects who are tested in frequent and irregular intervals for a long period of time.

Cross-sectional and simple time series approaches do not make full use of data available. Longitudinal medical data has both cross-sectional and time series characteristics. They are cross-sectional because they include many physical symptoms; we say that these data are differentiated across ‘space’ [2]. They are time series data because they represent many points in time. Longitudinal data methods are appropriate for these types of data.

There are two aspects concerning longitudinal data study; first, describing the individual development trend and the differences of the development trend between individuals, and second, explaining the development trend and the reason; the prediction variables can be unstable factors with time, and also can be fixed individual characteristics factors.

In the recent years, the techniques of longitudinal data analysis had made great progress; the hierarchical linear model had been widely used in longitudinal research. The term of the hierarchical linear model was firstly presented by Lindley and Smith in 1972 [3]; however, the traditional parameter estimation method (OLS) did not apply to the model because of the limitation of computing technology. In 1980s, iteratively reweighted generalized least squares and other methods had been used to estimate parameter, and there were some calculation software such as HLM (hierarchical linear model) [4]. In this article, we use HLM to analyze the medical followed-up data.

The remainder of this paper is organized as follows. First, we present the limitations of traditional statistical techniques and the advantages of hierarchical linear model in Section 2. In Section 3, we discuss the theory of hierarchical linear model. The application of hierarchical linear model in the study of tumor progression is discussed in Section 4. Concluding remarks are given in Section 5.

II. THE ADVANTAGES OF THE HIERARCHICAL LINEAR MODEL

A. The assumptions of the homogeneity of variance and the independence of random error

In the traditional statistical techniques, we always use variance analysis and multiple regression analysis to handle the longitudinal data. However, both of the two methods were used at the assumptions of the homogeneity of variance and the independence of random error, which were difficult to guarantee [5]. There were similarities between multiple tracking data from the same individual, and systematic errors might exist in the observations at the same time point, these issues made the assumption of random error independence hard to meet. Meanwhile, dependent variable occurred with a regular increase or decrease with time, which caused the increase or decrease of the variance. Those changes had a great effect on the homogeneity of variance. Therefore, traditional statistical techniques might lead to unreasonable or even wrong conclusions. Moreover, hierarchical linear model does not require the assumptions of the homogeneity of variance and the independence of random error, so it was more suitable for longitudinal studies.

ISBN: 978-1-61208-202-8

B. Problems of missing values and unequal measurement interval

Longitudinal studies needed to do repeatedly tracking observations of the same individual; it was prone to the loss of the sample. In the traditional statistical tools, there were two methods to handle the lost, one was removing the observed object which had missing values, another was fitting the missing values, and the former caused the waste of information, while the latter reduced the precision. While the HLM allowed the existence of missing values, it also made full use of the existing information. In addition, the traditional statistical technique required that all objects were observed at the same time interval. The HLM not only allowed the measurements of different time intervals, but also allowed the observed objects to have different observation schedules, this feature enabled researchers to have more convenience and flexibility [6].

C. The ability of dealing with the hypothesis

HLM allowed researchers to put forward different assumptions in different levels, such as whether it had a significant increase or decrease? Whether the different individuals had the same change rate? Which factor could predict the difference of change rate between different individuals? It also could create multiple development models and select the assumption which is the most consistent with observation data through the square test.

III. Hierarchical Linear Model

Hierarchical linear model in different research fields had different terms, such as Multilevel Statistical Model, Mixed Model, Random Coefficient Model and so on, the diversity of the term reflected that this method had been widely used in different fields. No matter what the term was, the core content of the method was same, mainly to handled the nested structure data. For repeated tests, these data had nested structure that the measurement nested with the individual [7]. This method could both solve the individual development trend and the differences of the development trend between individuals, also it could directly deal with the unequal measuring time interval, and at the same time it could analysis the missing value of data reasonably.

We used the hierarchical linear model to analyze the longitudinal data; data were found in different hierarchies. At first, established a regression equation for the first level variables, in which the tracking results that came from different observation times were the first layer and the invariant individual characteristics or the dispose that had been accepted were the second layer data. Through the processing, it could explore the effect of different levels on the dependent variables. In the first floor of the data structure, the track observation result was considered as the dependent variable.

$$Yij = \beta_{0i} + \beta_{1i} X_{ij} + \varepsilon_{ij} \tag{1}$$

As in "(1)", subscript "0" means intercept, subscript "1" means slope, subscript "i" means the i-th observation object, Subscript "j" indicates the j-th observation time." $\beta_{0\mathrm{i}}$ " is the intercept of the equation, it indicates the average of the i-th observed objects. " $\beta_{1\mathrm{i}}$ "is the regression coefficient, it indicates the changing rate of the i-th observation object." X_{ij} " means the values of the variable X when the i-th observed object is in the j-th observation time, " ε_{ij} " means residual, the implication is that the measured value Y of the i-th object in the j-th observation time that cannot be explained by the independent variable X.

Equation (1) is similar with the general regression equation, the only difference is, intercept and slope are not constant [8]. Different observation object have different intercept and slope, they may subject to the affect of variable of the second layer. In the second layer of the data structures, the intercept and slope are used as the dependent variable in (1), and the individual characteristics or the dispose that have been accepted are considered as independent variables, then we create two regression equations for the second layer:

$$\beta_{0i} = \gamma_{00} + \gamma_{01} W_{1i} + \mu_{0i} \tag{2}$$

$$\beta_{1i} = \gamma_{10} + \gamma_{11} W_{1i} + \mu_{1i} \tag{3}$$

In the above two equations, each parameter has two subscripts, if the first subscript is "0", this is the parameter that relates to the intercept of (1). If the first subscript is "0", this is the parameter that relates to the slope of (1). If the second subscript is "0", it means the intercept part of the second layer equation, if the second subscript is "1", it means the slope part of the second layer equation.

γ_{00} is the intercept of (2), it can be understood as the average of the dependent variable Y when the independent variable W1 is 0.

γ_{01} is the regression coefficients of the variables W1 in (2), it can be understood as the impact of the variable W1 to the initial value of the dependent variable Y.

γ_{10} is the intercept of (3),it can be understood as the changing rate of the observed object when the variable W1 is 0.

γ_{11} is the regression coefficients of the variable W1 in (3), it can be understood as the effect of the variable W1 on the changing rate.

μ_{1i} is the residual of (3), it can be understood as the changing rate of dependent Y that can't be explained by the variables W1 [9]. If the variance statistical test is significant, then the model needs to introduce new variables to explain the variation of the rate.

To simplified the problem, the second layer of the above regression equations only contains one variable W1, if there are multiple independent variables, accordingly, the slope part of (2) and (3) are needed, respectively, add $\gamma_{02} \cdot \gamma_{03} \cdot \gamma_{12} \cdot \gamma_{13}$ and so on.

IV. The Application of Hierarchical Linear Model in the Study of Tumor Progression

A. *Data description*

In order to study the condition trend of patients with the non-small cell lung cancer (NSCLC), treatment data of 52 patients, including six treatment data which interval is one month, were selected and two SPSS files were established. One file contains the patients' ID and the variable of "type" which is used as the number of measurement. The score of fact and symptom of each inspection are also included in it. Fact score is used as the output, the symptom and type are used as the predictor variables. Another file contains the id and sex as well as whether smoking in the initial state of the patients. Two SPSS files were read-in using HLM. The lengthways variation trend was analyzed respectively through two models as following.

B. *Unconditional growth model*

The second layer of equations does not contain any independent variable in the unconditional mean model. The function of the model is to describe the variation trend of the total observed object, and to make decision of that weather to introduce the second layer of the explanatory variables. Then set the following two-layer equations in this model.

The first layer equation：

$$Fact = \beta_0 + \beta_1（type）+\beta_2（symptom1）+ \varepsilon \quad (4)$$

The second layer equations：

$$\beta_0 = \gamma_{00} + \mu_0 \quad (5)$$

$$\beta_1 = \gamma_{10} + \mu_1 \quad (6)$$

For the variable type,0,1,2,3,4,5 are used to express it respectively, for the variable of symptom, In the Chinese traditional medicine, -1 indicates the asthenia syndrome, 0 indicates the compounding of the excess and deficiency syndromes, 1 indicates the old trauma sthenia. Use the software HLM to estimate the parameter, the result is showed in Table 1:

TABLE I. The Parameter Estimation1

fixed effects	coefficient	Standard Error	T-ratio
γ_{00}	43.78	2.17	20.16
γ_{10}	-0.35	0.44	-0.79
γ_{20}	1.23	2.76	0.45
Random Effect	**Variance Component**	**df**	**Chi-square**
μ_0	118.15	20	51.77
μ_1	2.01	20	30.52
μ_2	153.54	20	44.17

The analysis indicates that in the first measurement, the average of the fact score of all patients is 42.64. Every month, the average of the fact score declined 0.35 points; at the same time, the fact score increase 1.23 points with the one level rising of the symptom.

In the following, we use the figures to explain the effects of the predictor variables (symptoms) on the development trends. For the patients whose symptoms are the compounding of the excess and deficiency (symptom1=0), the trend is 43.78-0.35*type, which is showed in figure 1.The figure2 shows that the patients whose symptoms is 0 of the previous three times, -1 of the fourth time, 0 of the fifth time, and -1 of the sixth time.

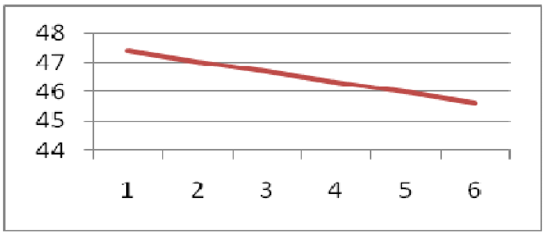


Figure 1. The variation trend1

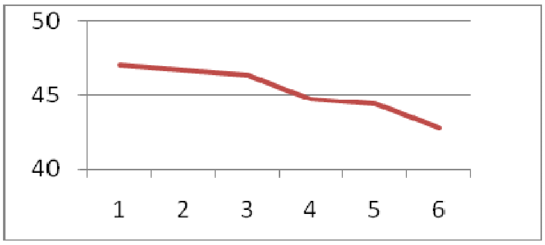


Figure 2. The variation trend2

It can be seen from the figures, if the time-varying predictor variables are included in the model, the model curve may vary based on the predictive value of different time points. The residual variation in (5) and (6) are significant that; indicating large differences in the fact score and rate of change. And thus the variables of second layer are needed to be introduced to get a better explanation.

C. *Combined model*

In order to better explain the individual differences of the current score and the rate of change in (4), two second-level variables are introduced: gender (male is 0, female is 1) and whether smoking in the initial state (smoker is 1, nonsmoker is 0), independent variables are included in this two layers.

The first layer equation：

$$Fact = \beta_0 + \beta_1（type）+\beta_2（symptom1）+ \gamma \quad (7)$$

The second layer equations：

$$\beta_0 = \gamma_{00} + \gamma_{01}(sex1) + \gamma_{02}(smoke) + \mu_0 \quad (8)$$

$$\beta_1 = \gamma_{10} + \gamma_{11}(sex1) + \gamma_{12}(smoke) + \mu_1 \quad (9)$$

ISBN: 978-1-61208-202-8

$$\beta_2 = \gamma_{20} + \gamma_{21}(\text{sex1}) + \gamma_{22}(\text{smoke}) + \mu_2 \qquad (10)$$

TABLE II. THE PARAMETER ESTIMATION2

fixed effects	coefficient	Standard Error	T-ratio
β_0			
γ_{00}	44.79	5.60	8.00
γ_{01}	4.06	5.68	0.71
γ_{02}	-5.74	5.66	-1.01
β_1			
γ_{10}	-0.82	1.06	-0.77
γ_{11}	0.12	1.12	0.11
γ_{12}	0.83	1.11	0.74
β_2			
γ_{20}	3.57	7.15	0.50
γ_{21}	-5.44	7.29	-0.75
γ_{22}	-0.48	7.26	-0.07
Random Effect	**Variance Component**	**df**	**Chi-square**
μ_0	107.40	18	37.61
μ_1	1.87	18	28.50
μ_2	149.01	18	43.59

The estimation results of fixed parameters shows that, for the first level intercept γ_{00} ,it means that the average of the fact sores is 44.79 when the type equals 0 and symptom equals 0; there is a significant difference between female and male in the initial state(γ_{01} =4.06 , se=5.68,t=0.71), the female patients have a higher score than men have; smoking in the initial state have a negative predictive effect on the average of the score(γ_{02} =-5.74,se=5.66,t=-1.01). Smoking patients, have a lower initial fact score. For the slope β_1 of the first level, the fact score has a significant descending trend with the increase of the check number (γ_{10} =-0.82, se=1.06,t=-0.77), over time, the gender has little effect on the rate of descending, the fact scores of smoking patients has a fast rate of descending than the nonsmokers'.

the slope β_2 of the first level, the fact score has a significant ascending trend with the increase of symptom, with the symptom changes, the gender has effect on the rate of ascending, female has a slower rate, (γ_{21}= -5.44, se=7.29, t=-0.75).Weather smoking in the initial time has a little effect on the ascending trend, the smoker has a slower rate (γ_{22} =-0.48 , se=7.26, t=-0.07).

The parameter estimation results of the random effects show that, comparing with the unconditional growth model, taking into account of the influence by the gender and smoking, the variance of the intercept and the slope have decreased, indicating that the two new independent variables explained a part of the intercept and slope variance, but from the random effects of the test results, the remaining unexplained variance is till significant; therefore, it is also needed to introduce a new variable to explain individual difference [10].Meanwhile, when increase the number of the parameters, the running time has no significant increase.

V. CONCLUSION AND FUTURE WORK

According to the analysis of tumor progression data by the hierarchical linear model, the model not only describes the individual dependent trend and the differences between individuals, but also gives the explanation. The model is specifically used with the medical data; it can study the influence of various factors on the fact score, and can study the influence by the fixed individual characteristics or the unstable time-varying factors. In the future work, importing more variables to analyze the impacts on the output and getting more favorable medical conclusions will be taken into consider.

Hierarchical linear model can handle the relationship between the tested data and the time variable data, and it can make a valid estimation of the parameters of the non-equilibrium measurements. So by using the hierarchical linear model to deal with the repeated measurements data, it needn't require the individuals to have the same number of observation times, and does not require the same time interval between the different individual tests. However the model has its drawbacks, firstly, compare with traditional estimation methods, its more complicated [11]. Secondly, using the hierarchical linear model to analyze data requires more than three times of tracking data, most of the data is hard to meet the requirement. Finally, the outcome variable in level 1 must be equivalent in each test. If all tests use the same test tool, the equivalence of the outcome variables can basically be assured, which means the measurement results are comparable.

REFERENCES

[1] Laaksonen, M., Simell, B., Salakoski, T., and Simell, O. Integrated data management and analysis environment for medical longitudinal research with machine learning based prediction models. 2009 WRI World Congress on Computer Science and Information Engineering, CSIE, 31 March-2 April 2009, pp. 552-556.

[2] Frees, E W., and Miller, T W. Sales forecasting using longitudinal data models. International Journal of Forecasting, 2004.

[3] Fearn, T. Towards a Bayesian package[C]. Wien, Austria: Physica-Verlag, 1978, pp. 473-479.

[4] Van, D L R., Vrijburg, K., and De, L J. Review of two different approaches for the analysis of growth data using

longitudinal mixed linear models: comparing hierarchical linear regression (ML3, HLM) and repeated measures designs with structured covariance matrices (BMDP5V)[J]. Computational Statistics and Data Analysis. 1996, pp. 583-605.

[5] Liu, H Y., and Meng, Q M . A Review on Longitudinal Data Analysis Method and It's Development. Advances in Psychological Science 2003, pp. 586-592.

[6] Liu, H Y. How to Abstract Developmental Variations:Latent Growth Mixed Model[J]. Advances in Psychological Science. 2007, pp. 539-544.

[7] Ren, X P., and Pan, C X. On the Study of Forecast of the Foodstuff Output Based on the Input Factors—The Longitudinal Data From 1978 -2005 of Henan Province. China Machine Press.

[8] Singer, J D., and Willett, J B. Applied Longitudi nal Data Anal ysis: Modeling Change and Event Occurrence. New York: Oxford University Press, 2003.

[9] Gai, X S., and Zhang, X H. The Application Of Multilevel Model In Longitudinal Research. Psychological Science 2005 , pp. 429-431.

[10] Molenberghs, G., and Verbeke, G M. Models for discrete longtitudinal data , new York:springer, 2005.

[11] Liu, L., Ma, J Z.,and Johnson, B A. A multi-level two-part random effects model, with application to an alcohol-dependence study B-4968-2009. Statistics in medicine. 2008, pp. 3528–3539.

Influence of Nutrient Limitation on Bacterial Patterns: Applying a new Bacterial Cellular Automaton Growth Model

Jean-Denis MATHIAS
IRSTEA
Laboratoire d'Ingénierie pour les Systèmes Complexes
24, avenue des Landais - BP 50085
63172 AUBIERE CEDEX, FRANCE
jean-denis.mathias@irstea.fr

***Abstract*— The nutrient concentration greatly influences the formation of various colony patterns generated by bacterial populations. We consider a 2D cellular automaton growth model of bacterial colony in the case of nutrient limitation. The present cellular automaton simulates the growth process in order to obtain these patterns in the case of random inoculation. We show that numerical patterns are close to those experimentally observed in the literature.**

Keywords-Cellular automaton; bacterial colony; nutrient limitation.

I. INTRODUCTION

Bacteria predominantly live in surface-associated communities [1]. They develop at any interfaces that are suitable for microbial growth. Important examples where bacteria develop are teeth [2], waste water treatment [3], problem of biocorrosion [4]. It is well known that these biological systems are combination of several behaviors of interacting individuals. These interacting individuals are able to produce higher-level patterns especially in the case of in-plane expansion of colonies.

In the case of random inoculation, the collaboration of bacteria has been recently studied in the case of *Pseudomonas Aeruginosa* [5]. This study shows the existence of important links between patterns and the competition between growth and nutrient access. By varying the nutrient access, several patterns have been observed such as: dense, labyrinth, worm-like, spots or small and big holes (see Figure 1). The experimental setting consisted in cultivating a biofilm on glass coverslip submerged in inoculated liquid medium. This study investigated how evolutionary competition among individuals affects colony patterns. The main contribution was to provide a formal link between higher level patterning and the potential for evolutionary conflict in social systems. The "worm-like" configuration is obtained at the beginning of the experience when cells begin to colonize the surface. The nutrient competition between cells is very important due to a limited substrate. The colony growth is therefore limited and small colonies form the biofilm. If we increase the substrate concentration, nutrient competition is less important between bacteria, growth becomes heterogeneous in space, circular colonies deform due to fingering [6] [5] and leads to the "labyrinth" configuration. Conversely, if the nutrient concentration is saturated, all bacteria have a nutrient access, the nutrient competition is lower, the colony growth is fast until reaching the "dense configuration".

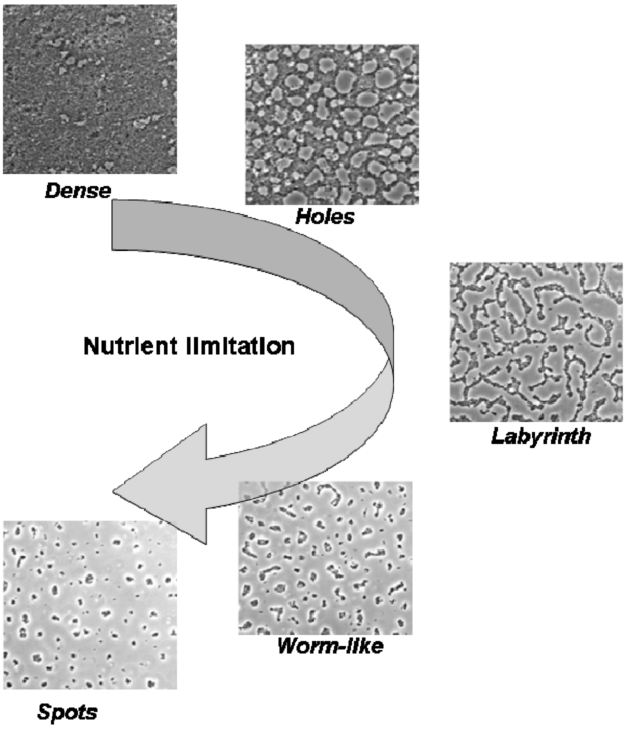


Figure 1. Different colony patterns depending on nutrient concentration [5].

Cellular automata are now commonly used in ecology research. This modeling approach is appealing because it represents directly the individuals and their behavior, making easier the link with scientific expertise about the ecosystem than in more abstract models. They show interesting spatio-temporal patterns that can be compared with observations. Several individual-based models have been used to model bacterial colony patterns in the case of in-plane expansion.

The aim of the present study is to include the competition phenomenon in the growth process and to highlight the advantages and the limitations of a such model. For this purpose, a simple cellular automaton based on the numerical and experimental observations is proposed in the first section by focusing on nutrient competition aspects. This model is then explored in the case of random inoculation and numerical patterns are compared to patterns experimentally observed in the literature. Results are analyzed so as to give new means about competition between growth and nutrient access and explanations on the emergence of higher-level patterns.

II. CELLULAR AUTOMATON OF BACTERIAL COLONY

Discrete modeling of bacterial dynamics has been developed using individual-based models or cellular automata [7]. They have been already used in several and various domains in order to simulate patterns and constitute an additional approach to the differential equation approach. Since 10 years, a lot of individual-based models of bacterial biofilms have been developed, mainly by the Delft team [7] [8] [9]. Individual-based model and cellular automaton have been developed in order to model bacterial patterns in the case of in-plane expansion. Here, we propose a very simple model of bacteria expansion that focuses on the growth process including a competition behavior with cooperation in order to obtain some of the patterns described in the introduction. The main originality lies in the fact that the nutrient dynamics is not modeled.

A. Spatial distribution of bacteria

In the cellular automaton growth model, each bacterium is only represented by its spatial coordinates (in 2 dimensions). Let *n* be the number of bacteria. The distribution of bacteria in the 2D space is given by the list $(x_i)_{1\le i\le n}$ of 2D positions.

B. Bacteria dynamics

Bacteria dynamics only includes a growth process. We assume that each bacterium cell has a constant probability *b* to produce a daughter cell during a time interval *dt*. Moreover, we had a competition function $C(d(x'))$ in order to take into account the competition behavior. For this purpose, we consider a bacterium *i*, located at *xi*, that gives birth to a new bacteria, located at x’ such as to maximize the probability $B(x_i, x')$ as follows:

$$B(x_i, x') = b\omega_1(x' - x_i) \times C(d(x')) \qquad (1)$$

This process represents the probability to a bacteria, located at x_i, to give birth to a new bacteria, located at x', that is dispersed following the kernel ω_1. This dispersion kernel allows us to model the cell spreading on a surface which is commonly observed in the literature [5] [10]. The second term $C(d(x'))$ represents a competition function. The aim is to propose a relevant competition function $C(d(x'))$ based on the numerical observations performed in Section 2. This competition function depends on the local bacterial density $d(x')$ of the new bacteria, defined as follows:

$$d(x') = \sum_{i=1}^{N} \omega_2(x_i - x') \qquad (2)$$

where ω_2 represents a competition kernel. The ω_i -function is based on circular uniform kernels parameterized with σ_1 and σ_2 as follows:

$$\begin{cases} \omega_i(x'-x) = \dfrac{1}{(\pi\sigma_i)^2}, |x'-x| \le \sigma_i \\ \omega_i(x'-x) = 0, |x'-x| > \sigma_i \end{cases} \qquad (3)$$

The function $d(x')$ takes into account the influence of other bacteria in the growth process. The function ω_2 enables us to model a nutrient competition due to a low diffusion coefficient through the value of σ_2. Indeed, if we consider a reaction-diffusion model, the distance between two bacteria affects the bacteria growth, especially in the case of a low diffusion coefficient of the nutrient. The influence of the nutrient diffusion coefficient process can be represented by the value of σ_2. In an experimental point of view, this competition process is observed for example in the DLA-like patterns [11]. The idea is to model this competition term with a simple function. This competition function $C(d(x'))$ has to have the following properties:

$$\begin{cases} C(0) = 1 \\ \lim_{d(x')\to 1} C(d(x')) = 0 \end{cases} \quad (4)$$

Furthermore, the function has to have a parameter that enables us to tune the intensity of the competition due to the nutrient concentration. Different functions can be used such as exponential or polynomial We have chosen the following polynomial function:

$$C(d(x')) = [1 - d(x')]^n \quad (5)$$

where *n* enables us to control the intensity of the competition process. This function has been plotted on Figure 2.

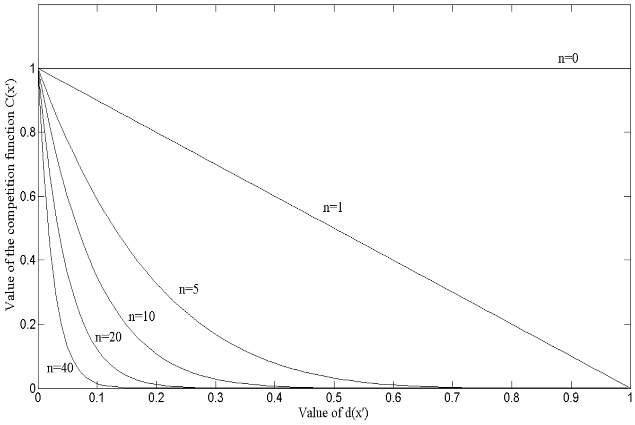


Figure 2. Value of the competition function $C(d(x'))$ following the number of bacteria in the local environment $d(x')$.

In the case of *n* = 0, the competition term $C(d(x'))$ is equal to 1, leading to no competition. It leads to have a classical growth process where the nutrient concentration is very important (limit case). For *n* > 0, the competition term $C(d(x'))$ tends toward 0 for a high bacterial density in the local environment of the potential new bacteria. In this case, the bacteria become "`non active"'. When there is no bacterium in the local environment of the potential new bacteria, the competition term $C(d(x'))$ is equal to 1, that is to say there is no competition. As explained above, *n* enables us to tune the intensity of the competition (see Figure 2). For high values of *n*, the competition is very important. It leads to slow down the growth of the bacterial population and to simulate bacterial behaviours observed in Figure 1.

Finally, the lower the value of the competition function $C(d(x'))$, the higher the competition, the lower the value of the probability $B(x_i, x')$. It leads that new bacteria has a higher probability to be located in a new place where there are few bacteria.

The cooperative aspect of bacterial biofilm has been described in [5] [12]. This cooperation is simply taken into account through the maximization of $B(x_i, x')$: bacteria mechanically push the new bacteria where the environment is the most favourable. It enables us to model the mechanical pressure with a preferential direction (where the competition is the lowest). Losses in biomass are not considered because we consider that this phenomenon can be neglected in this phase of growth. Some simulations have been done (not reported in this paper) and have shown that losses in biomass influence the density of bacteria and the dynamics but not the obtained patterns in a qualitative point of view. The aim here is to propose a new growth process (including competition behavior) and to show that we can obtain patterns observed in the literature with the proposed growth process.

C. Implementation

The current model depends on 4 variables: *b*, σ_1, σ_2 and *n*. A cellular automaton has been implemented using a 200x200 µm2 grid. The model was implemented in Matlab (7.2) for Windows with the following operations:

1. initialization of a population of *N* bacteria randomly located. The distribution of bacteria in the 2D space is given by the list $(x_i)_{1\le i\le n}$ of 2D positions. Go to step 2.1;

2. growth process:

2.1. initialization of index *i* (*i*=1). Go to step 2.2;

2.2. the value of $B(x_i, x')$ is calculated for all possible locations x' following the process defined above. Go to step 2.3;

2.3. the location x_m that maximizes the probability $B(x_i, x')$ is chosen. If there are several locations that maximize $B(x_i, x')$, the location x_m is randomly chosen. Go to step 2.4;

2.4. a random number α is chosen in the range [0,1]. If the value of α is inferior to the value of $B(x_i, x_m)$, go to step 2.5. Otherwise, go to step 2.6;

2.5. a bacteria located at x_m is added;

2.6. if $i < N$, advance bacterial index i and go to step 2.2. If $i = N$, go to step 3;

3. advance time and go to step (2) with the updated bacteria distribution.

III. RESULTS

In this section, we have computed the spatial patterns for different values of the parameters. However, we have decided to fix the value of σ_1 (σ_1=1µm). It seems to be unrealistic to have a higher value of σ_1. Moreover, if we increase this value, spatial structures tends to be uniform. We have chosen b = 1 bacteria.t^{-1}, here. Note that for low values of b, spatial structures don't change but execution times increase. In the following, we started simulations from an initial state that represented a uniform inoculation with individuals, placed at random locations. The initial density of individuals is equal to 1% of the domain. We have tested different values of n and σ_2 and highlighted their influence on the competition process: n = (1, 5, 15, 25) and σ_2 = (4, 6, 8) µm. Simulations have been stopped when the increase of individuals between two time steps is inferior to 0.1% in order to have a quasi-steady state. Results are plotted on Figure 3. Comments are:

- influence of n: n is directly linked to the competition between individuals. The n-parameter increases this competition and leads to a decrease of the number of individuals. When n=0 (not reported here) or 1, the competition is weak and the distribution of the bacteria is uniform. Then, when n increases, labyrinth appears and for high values of n we have worm-like configurations. These results are concordant with the observed patterns (see Figure 1): no competition leads to a dense and uniform configuration; important competition leads to worm-like patterns and the labyrinth pattern corresponds to the intermediate case;

- influence of σ_2 : σ_2 influences the competition distance. When the value of σ_2 is low, the competition between bacteria is not important. It leads to labyrinth with small voids. On the contrary, when the value of σ_2 is high, large voids are observed within the labyrinth;

- influence of the competition on the bacterial density ρ: we can see that configurations are directly linked to the final bacterial density. Indeed, the steady state depends on the competition between individuals and on the number of individuals, that is to say the bacterial density. When the competition is not important (n=5), bacteria can grow and we have a high steady bacterial density. On the contrary, when the competition is important, the growth is slowed down by the competition leading to a lower steady density. From a qualitative point of view, we have compared on Figure 7 the final density calculated from the simulated patterns and the density calculated from the experimental patterns. It clearly shows that we have a same qualitative correlation between densities and patterns;

- main models enable us to converge to a steady state using losses in biomass. The current model leads to a convergence of the bacterial density with the use of a competition term in the growth process. We can see that patterns are directly linked to the final bacterial density (see Figure 5).

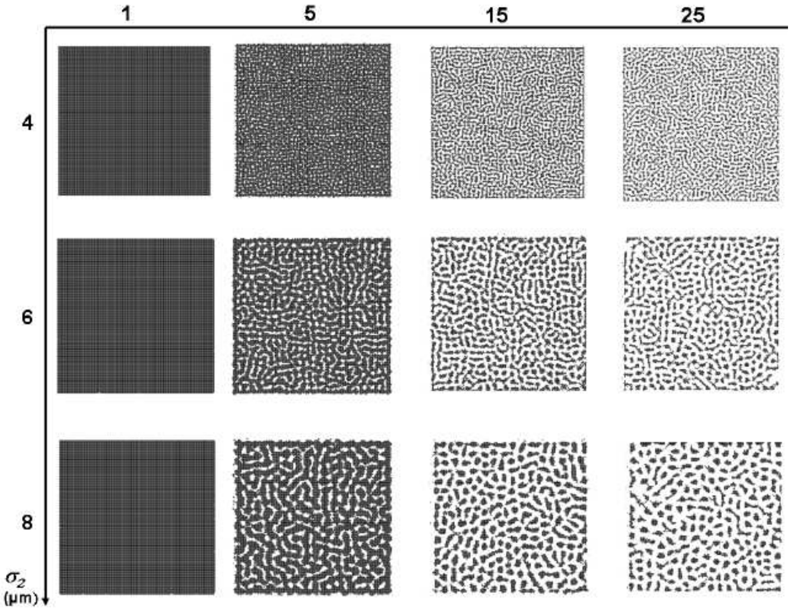


Figure 3. Colony patterns for different values of *n* and σ_2 .

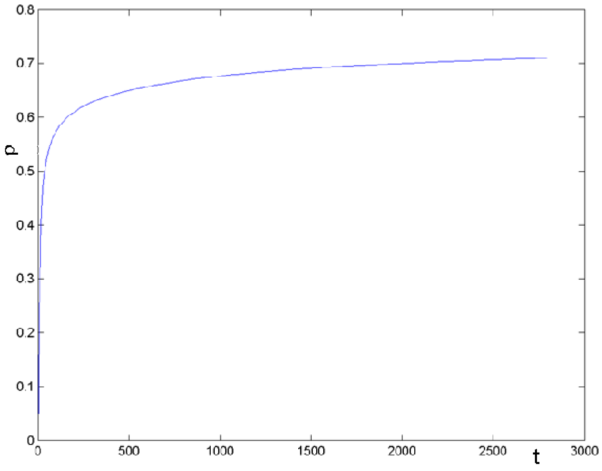


Figure 4. Evolution of the density ρ with respect to the time.

IV. Conclusion

Since twenty years ago, patterns of expansion produced by bacterial populations have experimentally been highlighted in the literature. Different models have been proposed in the literature, mainly based on differential equations, in order to simulate these patterns. A growth process including competition has been proposed in this paper that has been implemented in a cellular automaton. The competition aspect is taken into account by the calculation of a local bacterial density that is weighted by a polynomial function. Results have shown that this model enables us to obtain observed patterns in the case of random inoculations. This model leads to steady states with the use of the competition term in the growth process. Results have also shown that the obtained patterns are linked to the final bacterial density. Finally, this growth process (with a competition term) can be used in more complex models so as to take into account competition.

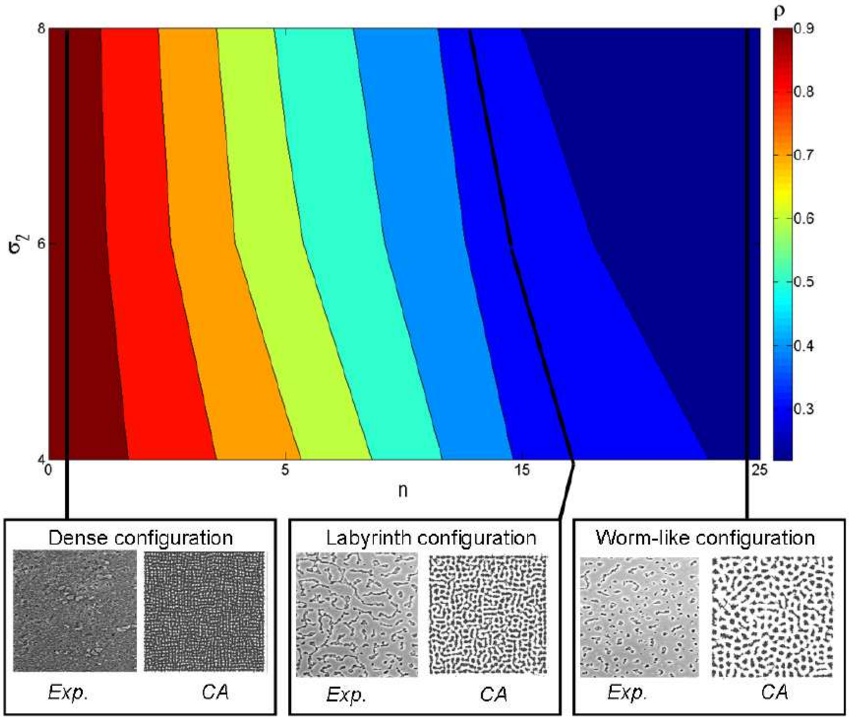


Figure 5. Steady state density ρ following *n* and σ_2: "`Exp"' corresponds to the pattern obtained experimentally in [5]; "`CA"' corresponds to the pattern obtained with the cellular automaton.

References

[1] J. Costerton, Z. Lewandowski, D. Caldwell, D. Korber, and H. Lappin-Scott, “Microbial biofilms”, Annual Review of Microbiology, vol. 49, 1995, pp. 711-745.

[2] S.S. Socransky and A.D. Haffajee” Dental biofilms: Difficult therapeutic targets”, Periodontology 2000, vol. 28 (1), 2002,pp. 12-55.

[3] H. Daims, P. Nielsen, J. Nielsen, S. Juretschko, and M. Wagner, “Novel nitrospira-like bacteria as dominant nitrite-oxidizers in biofilms from wastewater treatment plants: Diversity and in situ physiology”, Water Science and Technology, vol. 41 (4-5), 2000, pp. 85-90.

[4] I. Beech and J. Sunner, “Biocorrosion: Towards understanding interactions between biofilms and metals”, Current Opinion in Biotechnology, vol. 15 (3), 2004, pp. 181-186.

[5] J. Xavier, E. Martinez-Garcia, and K. Foster, “Social evolution of spatial patterns in bacterial biofilms: When conflict drives disorder”, American Naturalist, vol. 174 (1), 2009, pp. 1-12.

[6] J. Dockery and I. Klapper, “Finger formation in biofilm layers”, SIAM Journal on Applied Mathematics, vol. 62 (3), 2002, pp. 853-869.

[7] J.-U. Kreft, G. Booth, and J. Wimpenny, “Bacsim, a simulator for individual-based modelling of bacterial colony growth”, Microbiology, vol. 144 (12), 1998, pp. 3275-3287.

[8] J.-U. Kreft, C. Picioreanu, J. Wimpenny, and M. Van Loosdrecht, “Individual-based modelling of biofilms”, Microbiology, vol. 147 (11), 2001, pp. 2897-2912.

[9] C. Picioreanu, J.-U. Kreft, and M. Van Loosdrecht, “Particle-based multidimensional multispecies biofilm model”, Applied and Environmental Microbiology, vol. 70 (5), 2004, pp. 3024-3040.

[10] M. Matsushita, J. Wakita, H. Itoh, K. Wanabe, T. Arai, T. Matsuyama, H. Sakaguchi, and M. Mimura, “Formation of colony patterns by a bacteria cell population”, Physica A, vol. 274, 1999, pp. 190-199.

[11] F. Hiramatsu, J. Wakita, N. Kobayashi, Y. Yamazaki, M. Matsushita, and T. Matsuyama, “Patterns of expansion produced by a structured cell population of serratia marcescens in response to different media”, Microbes and Environments, vol. 20 (2), 2005, pp. 120-125.

[12] T. Matsuyama, and M. Matsushita, “Population morphogenesis by cooperative bacteria”, Forma, vol. 274, 1999, pp. 190-199.

Modeling User Experience

An integrated framework employing ISO 25010 standard

Maissom Qanber Abbasi, Philip Lew, Irfan Rafique, Jingnong Weng and Yunhong Wang
School of Computer Science and Engineering
Beijing University of Aeronautics and Astronautics
Beijing, P.R. China
e-mail: maissom@cse.buaa.edu.cn, philip.lew@buaa.edu.cn, irfan@cse.buaa.edu.cn, wengjn@buaa.edu.cn and yhwang@buaa.edu.cn

Abstract— The concept of user experience has been given much importance in the contemporary human computer interaction research. However, modeling user experience requires quality evaluation schemes that are not restricted to the traditional concepts of usability only, where requirements have generally been task oriented. On the contrary, in addition to modeling of usability (or task oriented) requirements, comprehensive methodologies to model subjective user needs should be put forward. In this paper, we discuss and relate various facets of user experience in order to lay foundation for engineering user experience requirements. In doing so, we propose a model to capture temporal requirements of user experience. We further employ this model and integrate it with the existing ISO 25010 standard to build a comprehensive and flexible user experience modeling framework. The usefulness of the proposed framework is also demonstrated by outlining a general guideline for specifying and evaluating user experience requirements.

Keywords-user experience modeling; user experience temporal requirements; quality in use; user satisfaction.

I. INTRODUCTION

User Experience (UX) is an evolving concept to the extent that we find lack in consensus for its definition [1]. ISO defines UX as a person's perceptions and responses that result from the use or anticipated use of a product, system or service [2]. As established in [3], the fulfillment of user's task-oriented goals (pragmatics) is not the only thing that users seek; rather there are certain underlying hedonic needs that they look and expect the product to fulfill. While pragmatics (or "do-goals") focus towards achievement of user needs that are objective in nature, e.g., task performance, effectiveness, etc., hedonics (or "be-goals") on the other hand, focus on the accomplishment of user needs such as satisfaction, stimulation, evocation, etc. [1]. Also, UX is not restricted to a momentary or instantaneous interaction with a certain product or application. The boundary of UX is wider than a mere user-product interaction, spreading from anticipation of use to actual use and further motivation to use. It is over time, that users adopt certain products [4], retain their usage [4] and then bond themselves with the product/product brand [5]. UX can therefore be investigated during and after, even long after, any interaction [6]. Time dimension or temporal aspects together with the environment (or the context of use) are, therefore, among key factors that influence UX [7].

The "user" part of UX is the key driver towards achieving UX; although, the product itself has to be well-designed to enable the user achieve his pragmatic and hedonic goals. Thus, modeling user experience calls for user-centered designs (UCD); designs that take into account traditional user needs as well as those that are abstract and subjective in nature. Bevan [8] highlights this very approach and discusses that despite the fact that the UCDs have been put into practice, they still lack consistency in their application. He further emphasizes that the UCD processes need exclusive UX professions that should involve teams covering aspects such as ergonomics, cognitive sciences, information quality, etc.

The recent ISO 25010 [9] standard outlines two perspectives of quality: *product* and *in use*. The product quality perspective relates to the core *product* design (internal and external characteristics), while the *in use* aspect of quality relates the user interaction with a product in a specified context of use. Recall, that the concept of UX bounds itself to the user-product interaction in a certain environment (context) as well as pre and post user-product interactions and therefore evolves over time. ISO 25010 therefore can be potentially utilized for modeling UX, by employing product quality (PQ) and quality in use (QinU) for modeling respectively "product" and "user-product interaction" entities of UX. However, if ISO intends usability to cover the whole UX, it needs to encompass all of its aspects [10]. Therefore, as a first step towards modeling UX by employing ISO 25010, we need to assess the extent to which the current standard captures all dimensions of UX.

Although the current ISO 25010 standard does cover under QinU the pragmatics and hedonics aspects of UX and product requirements under PQ, in order to completely capture and model UX, there is a need to integrate the temporal aspects [6] of UX along with its core pragmatics and hedonic dimensions. Using this as our motivation, we propose to define a UX temporal requirements (UXTR) model and integrate it with the existing ISO 25010 quality perspectives to develop a comprehensive and flexible UX modeling framework. The proposed framework and its models, i.e., PQ (P), QinU (pragmatics (P) and hedonics (H)) and temporal (T) (2PHT, for short) will represent a complete picture of UX requirements and thus can be put to use for instantiating different models for understanding and evaluation purposes. The proposed scheme is compliant with the current ISO standard for quality and is in alignment with recent related research contributions as well.

Ultimately, the contributions of this research are: (a) modeling temporal aspect of UX and (b) devising an integrated and flexible framework for modeling UX.

The rest of this paper is organized as follows: we review the related work in Section II. Sections III and IV, respectively, specify our integrated UX modeling framework and its practical significance for evaluating UX for software applications. We draw our conclusions in Section V.

II. Related Work

UX has gained much attention in the field of human computer interaction (HCI) in recent years. Even though there is a lack of consensus on a unified definition of UX, we still find in contemporary research, various approaches in defining and modeling UX. In an earlier classification [11], *an experience* is understood as something with a definitive beginning and end, with whatever happening in between constituting the UX. According to Bevan [12], user experience can be conceptualized as:

- An elaboration of the satisfaction component of usability.
- Distinct from usability, which conventionally focuses on user performance.
- Broader term for all the user's perceptions and responses, subjective or objective in nature.

In one of our earlier works [13], we have listed and categorized various UX definitions into actors and scenarios, where actors represent the UX touch points that include user, product, designer (organization) and environment, and scenarios represent the interaction phase (interacting, pre-interacting, design, post-interacting) of the UX. In the same research a complete UX evolution lifecycle framework (UXEL) was presented in order to understand the diverse UX dynamics. In doing so, UX building blocks were identified, explaining how they interact in three evolution stages of UX namely: Designed UX, Perceived UX and Actual UX.

As established in Section I that UX involves aspects of both product (PQ) and in-use (QinU) perspectives of quality standard put forward by ISO, it is worth analyzing how the two perspectives have been addressed in contemporary research. For example, Lew et al. [14] draw relationships among usability, information quality (IQ), QinU, and UX. In doing so, they integrate the concepts of PQ, QinU, Actual usability and Actual UX (2Q2U) in a flexible modeling framework to evaluate and improve QinU of web applications (WebApps). Similarly the current ISO 25010 [9] standard divides the concept of system/software PQ into eight characteristics and QinU into five characteristics as shown in Tables I and II respectively. However, in the older version of the standard (ISO/IEC 9126-1), the concepts of PQ and QinU were respectively classified into six and four characteristics. New characteristics and sub-characteristics have been added and/or renamed in the recent version, to enhance descriptiveness.

TABLE I. ISO 25010 PQ MODEL

(Sub)Characteristics	Availability
1. Functional Suitability	Fault tolerance
Functional completeness	Recoverability
Functional correctness	**6. Security**
Functional appropriateness	Confidentiality
2. Performance efficiency	Integrity
Time behavior	Non-repudiation
Resource utilization	Accountability
Capacity	Authenticity
3. Compatibility	**7. Maintainability**
Co-existence	Modularity
Interoperability	Reusability
4. Usability	Analysability
Appropriateness recognizability	Modifiability
Learnability	Testability
Operability	**8. Portability**
User error protection	Adaptability
User interface aesthetics	Installability
Accessibility	Replaceability
5. Reliability	
Maturity	

Hassenzahl et al. [3] model UX in terms of user's pragmatic (or do-goals) and hedonic (or be-goals) goals. Pragmatic goals or pragmatic quality refers to the user's perception about the product quality in its ability to support carrying out certain tasks, for example completing an online transaction, and focuses on the product's usability in making the user achieving do-goals. Hedonics, on the other hand, focus towards accomplishment of user's be-goals, i.e., how happy or satisfied the user feels after achieving his do-goals through using the product. They further state that it is the fulfillment of be-goals over time that the users strive for and that do-goals are a pre-requisite in achieving user's hedonic goals.

Given the current state of research and notions established on UX, it is clear why UX has become the most sought after quality aspect in modern day products. Not only do we expect them to help our tasks done, at the same time we also expect them to be enjoyable to use and make us feel satisfied. In light of [3], we can correlate the current ISO 25010 QinU model with the two dimensions of UX, i.e., pragmatics (do-goals) and hedonics (be-goals). For example, *satisfaction* characteristic can be correlated with hedonic goals (be-goals) of UX and measures of *effectiveness* or *efficiency* can be correlated with the fulfillment of pragmatics (do-goals). But since UX also involves a

TABLE II. ISO 25010 QINU MODEL

(Sub)Characteristics
1. Effectiveness
2. Efficiency
3. Satisfaction
Usefulness
Trust
Pleasure
Comfort
4. Freedom from risk
Economic risk mitigation
Health and safety risk mitigation
Environmental risk mitigation
5. Context Coverage
Context completeness
Flexibility

temporal dimension [7], it is important that while modeling UX requirements, we not only consider its pragmatic and hedonic dimensions, but also take into account its longitudinal aspect.

III. INTEGRATED UX MODELING FRAMEWORK

The aim of our study is twofold: first, modeling the temporal aspect of UX and second: integrate the proposed UXTR model together with the existing ISO 25010 quality model to build a complete UX modeling framework (as shown in Fig. 1). The proposed framework can then be used flexibly to instantiate models for achieving specific objectives.

A. *Modeling temporal aspect of UX*

Regarding modeling UXTR, Fig. 1.c shows the following two characteristics that collectively constitute our proposed model for longitudinal aspect of UX:

1) Appeal: or "appealingness" (as Hassenzahl [1] calls it) is defined as the degree to which a user gets motivated to get engaged with a certain product. It involves phases where the user associates certain anticipations and expectations from the product use and is not necessarily restricted to the direct intercation with the product. Appeal is further sub-characterized into:

a) Adoption: Defined as the scale that indicates how many users start using a certain product or application in a given time period [4].

b) Retention: Defined as the scale that indicates how many of the users from a given time frame are still using a certain product or application in some later time period [4].

2) Brand Association: Defined as personal liking or attachment with a certain service provider or an organization that manufactures certain product(s), and has certain popularity rate in the market and among various user groups.

Subsections III-B and III-C below, further develop understanding of the two characteristics described above and present reasoning for their inclusion in our proposed UXTR model.

B. *Appeal Characteristic*

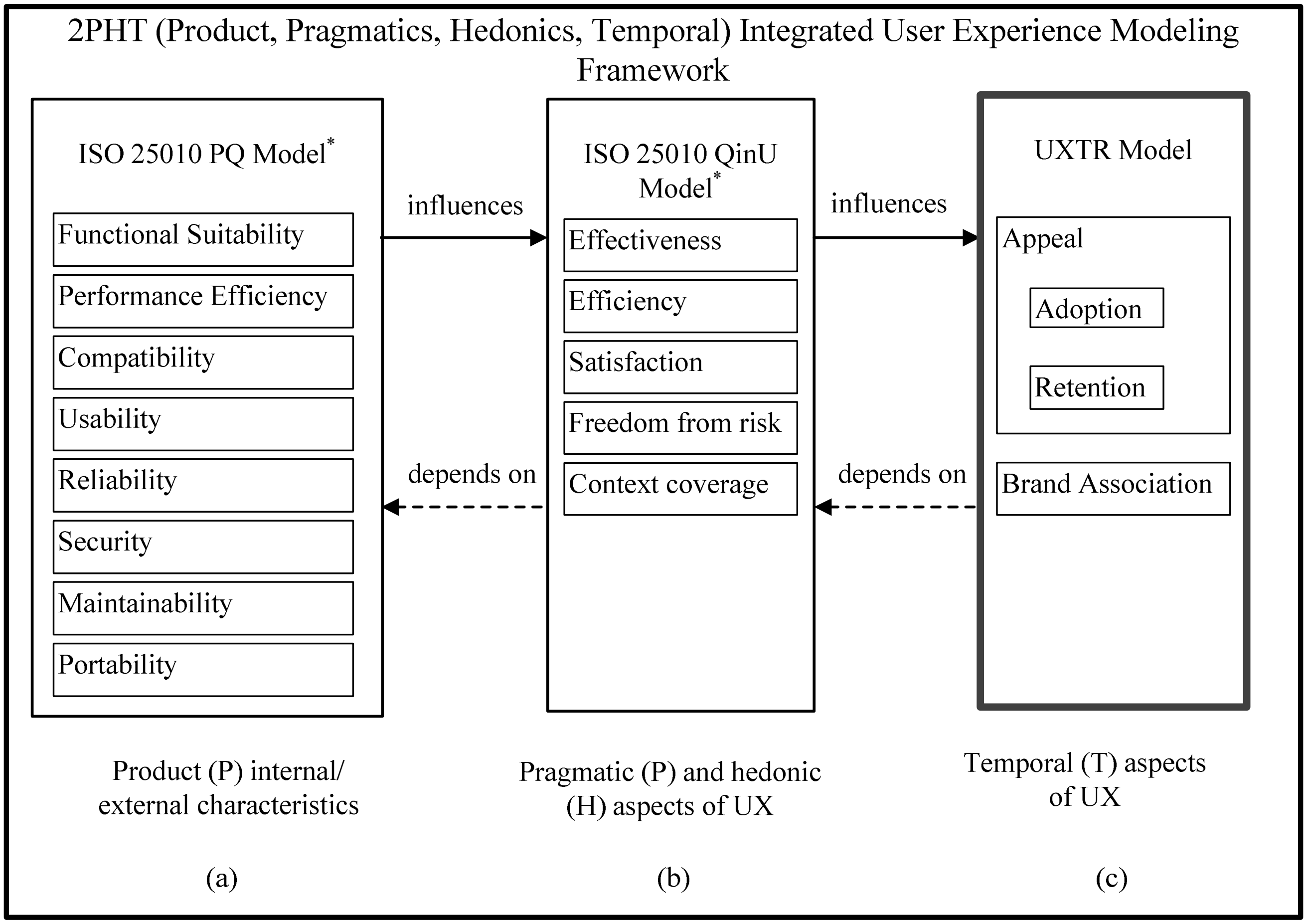
Appeal is a product's attribute that makes it attractive to the user as described above. Product characteristics such as "user interface aesthetics", "soft feel", etc. among others, contribute towards making the product appealing to the end user. On the other hand, appeal has an *in use* aspect which is triggered when the user is actually interacting with the product in a specific context. Achieving a certain task-oriented goal (do/pragmatic goal), for example, that satisfies the user, can attract the user for exploring the product further. Note that appeal here does not refer to the "visual appeal", which has more to do with the outlook or aesthetics of the product. On the contrary, appeal here refers to the desire for further interaction with the product.

A third perspective of appeal is temporal in nature and is beyond *product* and *in-use* aspects. Take for example, a web-based radio application that is to be launched in near future. There has been a lot of advertisement regarding its potential success among the listeners and fans of music, and this has led to individuals having anticipations about their interaction with the application as soon as it is launched and setting certain expectations in the form of pragmatic and hedonic goals. This pre-interaction phase is still motivating the user towards *adopting* a certain product, although the interaction has not begun yet. This form of appeal is still making the user go through an experience. Similarly, user's post-interaction scenario with the application may involve recounting the earlier experience over a time period and therefore, making him feel compelled to interact and use certain features, thus *retaining* his usage with the application. This example explains how the dynamics of *appeal* (or appealingness) are governed over time, thus making it an integral part of UXTR. This example also explains our classification of *appeal* into *adoption* and *retention* as we have proposed in our UXTR model.

Our proposed addition of *appeal* characteristic to the UX temporal requirements model is in alignment with Hassenzahl's [1] notion that UX changes over time; e.g., a product that was stimulating in the beginning might become less appealing with the passage of time or vice versa. Further, Hassenzahl classifies appealingness into *motivating* and *inviting* (among others). Our sub-characterization of *appeal* into *adoption* and *retention* is based upon this rationale.

C. *Brand Association characteristic*

Products or services that users interact with are not stand alone entities. Each product or service is designed by a certain organization and targeted for a certain user base. A user-product interaction is not confined to a user's



* For sub-characteristic level details for the characteristics of PQ and QinU models, refer to Tables I and II. Further details can be found in [9].

Figure 1. Proposed 2PHT UX modeling framework

engagement with the product or service, but implicitly encompasses a hedonic relationship (a bond or *association*) between the user and the creator of the product (the *brand*), hence the name *brand association*. Each time a product is interacted with, an unconscious engagement with its brand is there. Likewise, when a certain brand is heard of or a brand name is seen somewhere, an abstract (unconscious) interaction with one of the brand's products takes place. In either case, an *experience* is triggered.

Association with a brand is not only on a moment by moment basis. Loyalty with the brand evolves over time and is therefore, temporal in nature. Since *brand association* results in product bonding (and further product usage), thus affecting the overall UX, this characteristic is included as one of the characteristics of our UXTR model.

D. Integrating UXTR model with ISO 25010 Quality models

As established in the first section, modeling UX involves specification of *product* requirements, the *in-use* requirements covering the pragmatic and hedonic aspect of UX, and finally the *temporal* requirements. As shown in Fig. 1, all the three constituents of modeling UX requirements scheme are presented, whereby, our proposed UXTR model (Fig. 1.c) is integrated with ISO 25010 PQ model (Fig. 1.a) (covering product requirements specification part) and ISO 25010 QinU model (Fig. 1.b) (covering the pragmatic and hedonic dimensions of UX). Together, the three models form our proposed 2PHT UX modeling framework, representing a complete picture of UX requirements.

ISO 25010 states that there exist relationships between the PQ and QinU views of quality whereby the former *influences* the later and likewise the later *depends* on the former (refer to Fig. 1). The same approach is extended towards our proposed UXTR model. We argue that a good QinU will *influence* the temporal aspect of UX which in turn *depends* on the QinU perspective of the UX (as shown in Fig. 1).

Our proposed 2PHT integrated framework is also in line with our earlier work [13] where we define three phases of UX evolution lifecycle (UXEL), namely Designed UX, Perceived UX and Actual UX. The first phase (Designed UX) involves determining UX requirements and involves requirements elicitation processes leading to a UCD. This phase, therefore, relates to the first part of 2PHT, i.e., the PQ (Fig. 1.a). The second phase (Perceived UX) involves the product specific expectations and anticipations based on the advertisements, brand association, peer reviews etc. and therefore relates to the temporal aspects of UX (Fig. 1.c).

The third and last phase of UXEL (Actual UX) involves user interaction with a product in a specific context of use resulting in accomplishment of pragmatic and hedonic goals. The in-use and context specific aspects of this phase relate to the pragmatic and hedonic perspective (QinU) of UX (Fig. 1.b).

IV. Instantiating 2PHT Framework for Specifying and Evaluating UX requirements

The purpose of our proposed UX modeling framework is to consistently evaluate UX from its three perspectives, namely, PQ, QinU and temporal. Through our integrated approach, different non-functional requirements related to UX can be specified in order to meet specific evaluation needs for improving UX. In this section we outline a general guideline for instantiating our proposed 2PHT framework for specifying and evaluating UX requirements.

A. *Specifying UX requirements employing 2PHT framework*

Utilizing our proposed 2PHT UX modeling framework, we choose "user interface aesthetics" (UIA) (a sub-characteristic of "Usability", refer to Table I) as an example and specify its requirements from PQ and QinU perspectives of UX. Requirements specification from both views (i.e., PQ and QinU) is in alignment with ISO which categorizes a quality construct into a collection of related sub-characteristics providing a convenient breakdown of a quality concept [9].

1) *Specifying PQ UIA Requirements*

For specifying UIA requirements from the product perspective, the PQ model (Fig. 1.a) of our proposed 2PHT framework can be employed to instantiate the UIA sub-characteristic of *usability*. In light of Pham's [15] categorization of aesthetic design principles, a complete requirement tree for PQ UIA can be specified. For the purpose of demonstration, a generic breakdown (sub-characterization) of the PQ UIA sub-characteristic is shown in Table III.

2) *Specifying QinU UIA Requirements*

For specifying UIA requirements from the user (or in-use) perspective, the QinU model (Fig. 1.b) of our proposed 2PHT framework can be employed. Further, the QinU model can be supplemented with "Aesthetic Appeal" sub-characteristic under "Pleasure", which is a sub-characteristic of "Satisfaction" (refer to Table II). The reason to add *aesthetic appeal* under the Pleasure sub-characteristic is that aesthetics affects the pleasure and harmony that users experience while interacting with a product [16] and is a strong determinant of user satisfaction [17]. Based on this rationale, a generic breakdown of QinU Aesthetic Appeal is shown in Table IV.

B. *Evaluating UX requirements Employing 2PHT Framework*

In this sub-section, taking the same example as in subsection IV-A, we outline general principles for practically evaluating UX requirements based on the requirements specified in subsection IV-A. We further lay foundation to observe the resultant PQ and QinU evaluations effect on the longitudinal aspect of UX utilizing our proposed UXTR model (Fig 1.c) of the 2PHT framework. Further, QinU evaluation can involve subjective surveys asking users questions about their interaction with a particular product, whereas the PQ evaluation can be done through manual inspection.

TABLE III. PQ UIA Requirements Tree

2PHT.PQ UIA requirements
1 **User Interface Aesthetics** (UIA)
1.1 sub-characteristic 1
1.2 sub-characteristic 2
1.2.1 *attribute 1*
…

1) *UIA evaluation from QinU perspective*

As per the ISO premise, PQ influences QinU (refer to Fig. 1). Therefore, at first, the current state of UIA of an application can be evaluated during real time user interaction. In order to carry out the subjective evaluation for UIA, the QinU requirement tree specified in Table IV can be mapped with standard subjective questionnaires for usability testing. Users response, for example on a 7-point Likert [18] scale with responses varying from "strongly disagree" to "strongly agree" scale labels, can be used to evaluate the corresponding characteristic/sub-characteristic of *aesthetic appeal*. The overall rank of *aesthetic appeal* can then be calculated by aggregating the scores of its constituent sub-characteristics.

2) *UIA evaluation from PQ perspective*

For evaluating the UIA from the PQ point of view, different metrics can be developed for objectively quantifying UIA sub-characteristics and attributes. For the purpose of demonstration, we define a metric for "object clarity" (where object can represent text, image, or animation on the UI) that can be treated as a sub-characteristic of PQ UIA requirements outlined in Table III. This metric classification is shown in Table V.

C. *PQ and QinU evaluation analysis*

Evaluation results for both PQ and QinU perspectives of UIA will set the stage for improvement considerations.

TABLE IV. QinU UIA Requirements Tree

2PHT.QinU UIA Requirements
1 Satisfaction
1.1 Pleasure
1.1.1 **Aesthetic Appeal**
1.1.1.1 sub-characteristic 1
1.1.1.2 sub-characteristic 2
1.1.1.2.1 *attribute 1*
…

TABLE V. METRIC CLASSIFICATION FOR PQ UIA EVALUATION

UIA PQ Evaluation metric item	Details
Characteristic/Sub-Characteristic	User Inteface Aesthetics (UIA)
Attribute name	Object clarity
Metric name	Object clarity level
Objective	Determine if the objects on the UI (such as text, image, animation, etc) are visually identifiable
Measurement method	The UI is inspected to determine the object clarity level rating on a scale of 0-3. Observers observe whether objects on the UI are visually identifiable.
Scale	Numerical percentage ratio
Allowed Values	(0) none of the objects on the UI are visually identifiable; (1) few of the objects on the UI are visually identifiable; (2) most of the objects on the UI are visually identifiable; (3) all of the objects on the UI are visually identifiable.

Based on the QinU UIA evaluation, improvement recommendations from the design perspective can be deduced. Improving the design on the basis of improvement recommendations will call for another round of PQ and QinU evaluations to see if the improvement from the PQ perspective also resulted in improvement in the QinU aspect of UX.

D. Evaluating temporal aspect of UX

Once the recommended improvements have been performed on the PQ side and the desired level of QinU has been achieved, we can assess the resultant effect of the improvement on the temporal aspect of UX.

For example, we can examine our proposed UXTR model to specify the temporal requirements. This requirement specification is shown in Table VI. Based on our proposed UXTR model, *adoption* and *retention* sub-characteristics can be measured intrusively or as outlined in [4], to evaluate the *appeal* requirement of the instantiated model. Similarly, the *brand association* characteristic can also be evaluated using subjective surveys. Collectively, the evaluation measures for *appeal* and *brand association* can give a measure for the temporal aspect of UX. Note that, since the temporal aspect evolves over time, evaluating UXTR will span over a specific time period, consisting of multiple rounds of intrusive evaluations focusing on the same group of users.

V. CONCLUSION AND FUTURE WORK

In this paper we have related three perspectives of UX namely: PQ, QinU (Pragmatics and Hedonics) and longitudinal (temporal). In doing so, we have developed an integrated framework called 2PHT for modeling UX by proposing a UXTR model and integrating it with the current ISO 25010 standard. We have provided reasoning for our proposed UXTR model in which we have introduced two concepts of *appeal* and *brand association* as the characteristics defining the longitudinal dimension of UX. We have also characterized the concept of *appeal* into *adoption* and *retention* sub-characteristics and described their importance in light of the current research. A demonstration for a specific requirement tree instantiation and evaluation, based on the proposed framework is also given. The three constituent models (Fig. 1.a, 1.b and 1.c) of our proposed UX modeling framework are intended for modeling UX requirements for software products as a whole and can therefore be used to evaluate and improve UX aspects for different types of software, such as WebApps, for example.

TABLE VI. MODEL COMPOSITION REPRESENTING UXTR

2PHT.UXTR
1. Appeal
1.1. Adoption
1.2. Retention
2. Brand association

Based on this research, our future work focuses on devising a thorough strategy that will involve experience requirements elicitation and recommendation processes combined with our 2PHT UX modeling framework for evaluating and improving UX of software applications with focus on geographic information systems and digital earth applications.

ACKNOWLEDGMENTS

Authors gratefully acknowledge support from the lab of Digital earth and GIS, School of Computer science and engineering, Beijing University of Aeronautics and Astronautics (BUAA/BeiHang), P.R. China.

REFERENCES

[1] M. Hassenzahl, "The thing and I: understanding the relationship between user and product", in Funology: From Usability to Enjoyment, M.A. Blythe, A.F. Monk, K. Overbeeke, P.C. Wright (Eds.) Kluwer Academic Publishers, 2003, 1-12.

[2] ISO DIS 9241-210. "Ergonomics of human system interaction - Part 210: Human-centred design for interactive systems", 2008.

[3] M. Hassenzahl and V. Roto, "Being and doing - A perspective on User Experience and its measurement", Interfaces, vol. 72, 2007, pp. 10-12.

[4] K. Rodden, H. Hutchinson, and X. Fu, "Measuring the user experience on a large scale: user-centered metrics for web applications", Proc. CHI 2010, Atlanta, Georgia, USA, April 10–15, doi: 10.1145/1753326.1753687.

[5] P. Ketola and V. Roto, "Exploring User Experience Measurement Needs", Proceedings of the International Workshop on Meaningful Measures: Valid Useful User Experience Measurement (VUUM), Reykjavik, Iceland, 2008, pp. 23-26.

[6] E. L.-C. Law, V. Roto, M. Hassenzahl, A. P. O. S. Vermeeren, and J. Kort, "Understanding, scoping and defining user experience: a survey approach", Proceedings of the 27th international conference on Human factors in computing systems, Boston, MA, USA, 2009, pp. 719-728, doi: 10.1145/1518701.1518813.

[7] M. Hassenzahl and N. Tractinsky, "User experience - a research agenda", Behaviour & Information Technology, vol. 25 (2), 2006, pp. 91-97, doi: 10.1080/01449290500330331.

[8] N. Bevan, "Creating a UX Profession", Proc. CHI 2005, Portland, Orgeon, USA, 2005, pp. 1078-1079, doi: 10.1145/1056808.1056820.

[9] ISO/IEC 25010 Systems and software engineering — "Systems and software Quality Requirements and Evaluation (SQuaRE) — System and software quality models", 2011.

[10] N. Bevan, "UX, Usability and ISO Standards", Proc. The 26th Annual CHI Conference on Human Factors in Computing Systems, 2008, pp. 1-5.

[11] J. Forlizzi and S. Ford, "The building blocks of experience: an early framework for interaction designers", Proceedings of the 3rd conference on Designing interactive systems: processes, practices, methods, and techniques, New York City, New York, United States, 2000, pp. 419-423, doi: 10.1145/347642.347800.

[12] N. Bevan, "What is the difference between the purpose of usability and user experience evaluation methods?", UXEM'09 Workshop, INTERACT 2009, Uppsala, Sweden, 2009.

[13] M. Q. Abbasi, P. Lew, I. Rafique, and L. Zhang, "User experience evolution lifecycle framework", Proc. International conference on Information, Systems and Engineering (ICISE) 2012, Zurich, Switzerland, 2012, January 15-17, pp. 947-952.

[14] P. Lew, L. Olsina, and L. Zhang, "Quality, Quality in Use, Actual Usability and User Experience as Key Drivers for Web Application Evaluation", Proc. LNCS 6189, Springer, 10th Int'l Congress on Web Engineering (ICWE2010), Vienna, Austria, 2010, pp. 218-232, doi: 10.1007/978-3-642-13911-6_15.

[15] B. L. Pham, "Design for aesthetics: interactions of design variables and aesthetic properties", of SPIE IS&T/SPIE 11th Annual Symposium - Electronic Imaging '99, 1999, pp. 364-371.

[16] G. Lindgaard, "Aesthetics, Visual Appeal, Usability and User Satisfaction: What Do the User's Eyes Tell the User's Brain?", Australian Journal of Emerging Technologies and Society, vol. 5 (1), 2007, pp. 1-14.

[17] T. Lavie and N. Tractinsky, "Assessing dimensions of perceived visual aesthetics of web sites", International Journal of Human-Computer Studies, vol. 60, 2004, pp. 269-298, doi: 10.1016/j.ijhcs.2003.09.002.

[18] T. Tullis and B. Albert, Measuring the User Experience. Collecting, Analyzing and Presenting Usability Metrics., 2008.

Advancing Disaster Response Systems

Implementing Biometric Technologies as Demographic Identifiers

Omorodion Steve Eguasa
School of Technology
Purdue University Calumet
Hammond, USA
oeguasa@purduecal.edu

Barbara Nicolai
School of Technology
Purdue University Calumet
Hammond, USA
bnicolai@purduecal.edu

***Abstract*— This study covers multiple findings from the origin of biometric technology, its application in the modern world, and new ideas that have made this technology practice very useful and popular. It also provides an assessment of how future developments of Natural Disaster Response Systems can benefit from utilizing parts of this technology during search and rescue situations to administer emergency medical care for disaster-stricken victims that may be unresponsive and without identification. Using biometric technologies such as fingerprint identification and iris recognition software on handheld devices will allow responders to scan fingerprints or the iris of unresponsive victims to gain emergency medical records that their healthcare professionals use to treat them.**

Keywords-biometric technology; natural disaster; response systems; fingerprint; iris; handheld; emergency medical care

I. Introduction

Agencies on the local, state, or federal level all face different challenges and continue to encounter issues when it comes to rescuing and administering emergency medical assistance to stricken victims. The first Emergency responders that arrive on the scene - firefighters, police, coast guards, paramedics, etc. - can only apply emergency care within their realm. Agencies that manage disaster situations however, have been taught the proper response procedures. In terms of technology, emergency responders have limited options for managing and/or delegating tasks. Despite these limitations, disaster situations occur more frequently than they have historically, spurning the need for advancements in how responders apply medical assistance and what resources are available to help them do this in the most efficient way. Potential drawbacks of this advancement are: security breaches, accuracy, consistency, cost, privacy and legal issues, adaptation to newer technologies (in an environment where new technologies emerge often) etc. Nonetheless, with a rapidly changing world and considering natural and man-made disasters of the past couple decades, these two factors alone provide sufficient cause to make more resources available to target such technologies that could save more lives in the wake of such events.

Currently, fingerprints and iris recognition are two biometric identifiers used within various industries as tools to enhance security. Although very few states attempt to implement biometric technology within their disaster response models, more widespread use can help deliver specialized emergency response to victims who may have special needs (e.g., may have preexisting conditions that responders don't know of, etc.). The use of these technologies to acquire emergency medical records of victims to aid in life-saving treatments may be the next step in improving disaster response, bridging care that would be received at hospitals or other medical facilities.

The significance of the problem began after Hurricane Katrina pounded the Gulf Coast and most of the 1 million people displaced by the storm were left with no medical records, making it difficult, if not impossible, for emergency responders working on scenes, medical centers and community hospitals to treat them [1]. The patient medical histories remained unknown and medical responders made vital treatment decisions with incomplete information.

In the case of the Haiti earthquake, the Haitians were under-served and under-represented by the lack of medical knowledge and health services. Patients didn't had medical records to make available for emergency medical responders but just a sticky note from the triage tent with a chief complaint, age, and sex on their chests [2]. If by any chance any of the patients were transferred to another medical station, a piece of paper with critical information listed would be taped to their chest.

Biometrics consists of methods for uniquely recognizing humans based upon one or more intrinsic physical or behavioral traits. In computer science, in particular, biometrics is used as a form of identity access management and access control. It is also used to identify individuals in groups that are under surveillance. Biometrics comprises of two parts; the physical and behavior aspects. Features measured include face, fingerprints, hand geometry, handwriting, iris, retinal, vein, and voice [3]. The fingerprint identification is one of the most well-known and publicized biometrics because of their uniqueness and consistency over time, fingers have been used for identification for over a century, more recently becoming automated. Iris recognition is the process of recognizing a person by analyzing the random pattern of the iris. The automated method of iris is relatively young, existing in patent only since 1994 [4]. The iris is a muscle within the eye that regulates the size of the pupil, controlling the amount of light that enters the eye.

This study will further cover the statement of the problem which describes includes key dependent and independent

ISBN: 978-1-61208-202-8

variables, the research question and a hypothetical account if such technologies were to be implemented, assumptions, limitations such issues hindering the use of biometric technologies to acquire victim information; and delimitations that should be taken into account for this concept to be eventually become reality.

II. Statement of the Problem

There is a need for these responders in the frontlines of natural and man-made disasters to have access to victims emergency medical records which in turn allows them to efficiently provide the best medical services to those with or without unique health situations. In the cases of Hurricane Katrina and the earthquake in Haiti, providing medical assistance to victims proved to be very tough without the presence of effective medical care information. Unresponsive victims either died from not been given proper medical treatments or experienced further complications from allergic reactions to improper treatments.

The utilization of fingerprint and iris recognition as demographic identifiers in a disaster response system will allow first responders and professional disaster workers to efficiently provide the best medical services to those with or without unique health situations. The key variables of the problem will help answer the question about the influences, the factors contributes to the presumed effect of saving more lives in disaster situations. These variables include:

Dependent

- Electronic medical records databases
- Developed software web application
- Handheld mobile devices equipped with fingerprint and iris scanner

Independent

- Disaster response command center
- Medical care centers
- Emergency mobile units

The process of advancing current response systems would start from the development of electronic medical records of inhabitants in each community regardless of whether there are known or unknown issues. This will help populate a National Biometric ID database and also provide to baseline to giving responders adequate knowledge in providing response care if needed. It can be updated at the request of a prospective client to their assigned physicians. For example, proper information about a client would help avoid a responder treating an unresponsive victim with antibiotics that they are allergic to.

An efficiently ran command center has to be able to translate received data into valuable information for responders in the field to properly utilize. With information been passed down, responders on-scene can quickly determine at the side of a responsive and unresponsive victim if they administer medical care right away or be transported to a nearby medical center. With appropriate mobile devices on hand that are equipped with federal regulated finger and iris recognition software that will be developed, a responder can receive real-time medical information about victims and how best they can be treated to avoid further health complications besides the shock from experiencing the event. Mobile units including ambulances, medical helicopters and even foot responders have to be properly dispatched with these mobile devices in order to shorten the response time in these disaster situations. The process in Figure 3 displays a conceptual framework of the prospective concept's flow of research and development.

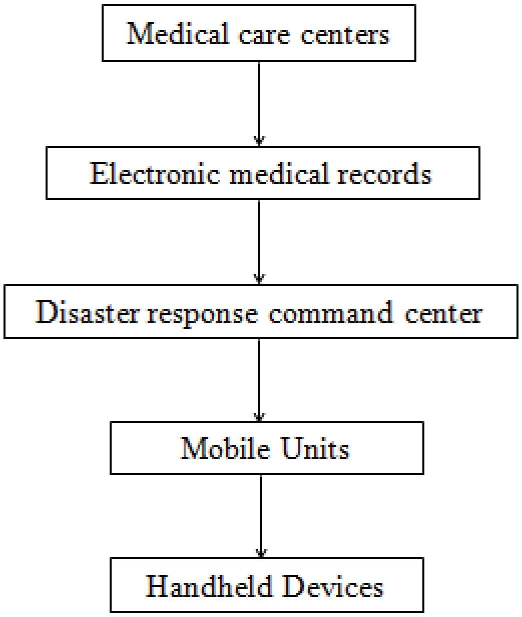


Figure 1. Conceptual framework.

An investigation will be conducted to help analyze and interpret data, sources and results by employing a hypothesis testing approach. Considering the type of research problem stated earlier, a well suggested experiment to employ will be based on both a lab and a field approach as described below:

Lab

- Done in an artificial or contrived environment.
- To control other factors that might contaminate the discovery of the cause-and-effect relationship.
- To manipulate the independent variable to establish the extent of its causal effect.

Field

- Done in the natural environment where activities take place regularly.
- It may not be possible to control all the contaminating factors.
- The use control groups.

A conducted lab experiment will allow for the recording and studying of real time request and response times during the use of these handheld devices. The experiment will also track the use of the fingerprint and iris recognition capabilities. This will allow for the creation of guidelines to follow when interacting with responsive and unresponsive clients. It will also allow me to measure the difference between traditional interaction between responders and clients before medical care is administered. Reducing that time and administering the right medical will make the biggest impact on whether the implementation of that technology will be successful. This process will validate the testing internally due to a controlled environment that will be simulated.

There will be dual types of useful data in this situation, some primary and secondary sources of data. The primary sources will be individual emergency responders and a focused group of emergency responders (two different townships); and the secondary source will be analyses gathered from the industry's emergency decision makers. In collecting data from individual emergency responders, an assessment can be made on how they personally feel it improves their performance in assisting patients in a lab or field environment with the presence of obstructive or non-obstructive measures. This data will also ascertain if there can be possible changes in the system processes from start to finish. With a focus group of respondents, interpretations and opinions can be collected on how effectiveness and efficiency evolves within a team in a controlled environment. It will display how quickly a respondent can end a task and assist in another to safeguard more lives. In terms of figuring out how a secondary source like analyses collected from the industry of emergency respondents will be utilized, it will go to show more than just in-field data like other influences involved in the business of saving lives. It will also show how saving time and costs in effectively managing man hours, legal and non-legal measures, community acceptance etc. can either encourage the project’s implementation or not.

A structured set of survey questions will be utilized and observations will be closely monitored during experiments. The survey will be designed to fit the lab and field tests individually since there are specifically influences to both. Questions for the respondents will curtail to field operations and the other set of questions relating to the industry’s analyses will be referred to agency managers. Both sets of questions will be arranged properly and accompanied by clear instructions, guidance and good alignment to minimize biases so as to acquire the best assessment from both parties. Observation will only occur during the lab and field testing to collect data on how reducing the time and administering the right medical can make the biggest impact on whether using biometric technologies such as fingerprints or iris recognition to access medical data from secure data servers.

Based on the data collection methods which involves the selection of emergency responders (Paramedic personnel), conclusions should be drawn based on how many samples are collected. Samples collected will be examined closely because the results will almost mirror how real-time performances will occur. To keep the sample simple and random, there should be participation from fifty different paramedic agencies who can allow the involvement of two personnel each. This will allow for close observation of partners working together to attend to a patient in a fast and efficient manner while utilizing the technology. Training sessions will take place initially to give the responders ample time to get familiar with the device. There will be no bias in the selection process because if the device is successfully lunched, every responder will be required to partake in training so as to keep procedures at an even standard.

After the data sampling information are recorded, they will need to be analyzed and interpreted properly to become useful to the project. The Analysis process will include the interpretation of the results from the experiments conducted. There are steps that will be followed before analysis which include:

- Editing data
- Handling blank responses
- Coding
- Categorizing
- Entering data

Once these steps have been followed carefully, getting a feel for the data collected from paramedics would follow. Also testing the goodness (stability and validity) and the hypothesis of the data would follow suit to ensure that mistakes are omitted and the process works.

Like all other technologies, they will have performance standards or metrics which must be evaluated first before implementing them in disaster recovery systems. Biometric News wrote that these performance standards, or metrics, are widely used by the biometric industry in order to gauge the effectiveness of the various biometric technologies. These standards are not particular to any specific biometric technology; they apply to all of the technologies [5]. These standards are:

- The False Acceptance Rate;
- The False Rejection Rate;
- The Equal Error Rate;
- The Failure to Enroll Rate;
- The Ability to Verify Rate.

This could pose to be very disastrous in scenes because everything going on is already hectic and such systems are meant to bring some sought of stability in the areas of emergency medical care. Equipping a handheld disaster response device a fingerprint scan and iris recognition can end up saving more lives of disaster struck victims and it should be able to rely on emergency medical information that can be accessed by the response team onsite.

There are various benefits and disadvantages of implementing biometric technologies as identifiers. The government would play the largest role in its use in order to monitor security for its citizens just like the United Kingdom that already uses a National Biometric ID. There are few reasons why it’s been slow to be implemented in the U.S due to survey responses conducted concerning cost, risks of information misuse, privacy of personal data, and security [6]. Figures 2 & 3 below show the graphical illustrations from an earlier study.

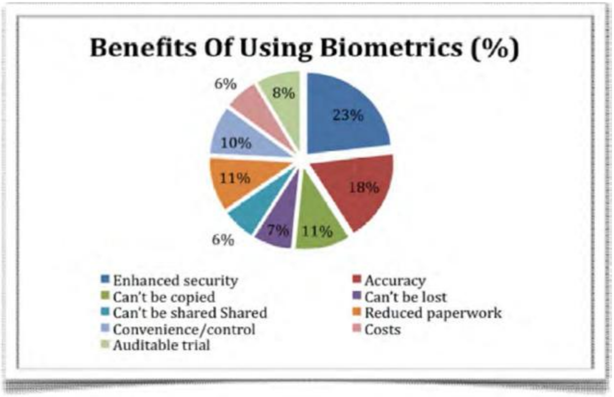


Figure 2. Benefits of using biometrics (%).

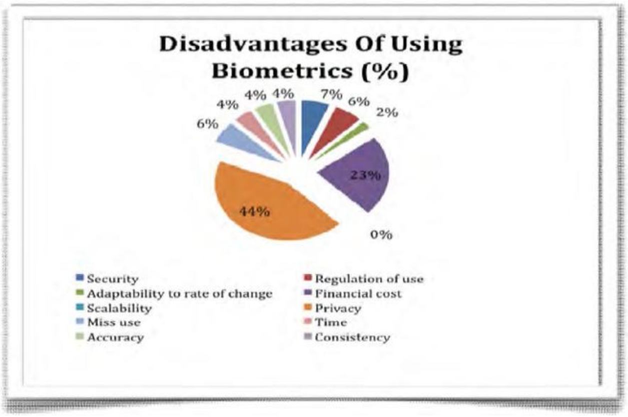


Figure 3. Disadvantages of using biometrics (%).

V. Chu, and G. Rajendran (2009), "Use of biometrics".

III. RESEARCH QUESTION/HYPOTHESIS

Research Question

Can the implementation of biometric technologies such as fingerprint and iris scan recognitions in disaster response systems be utilized as useful demographic identifiers in providing emergency medical responses to disaster stricken victims?

Hypothesis

Since natural or man-made disasters do not happen as often, it will be adequate to employ the help of paramedics who are healthcare professionals that work in emergency situations regularly to responsive and unresponsive patients. Paramedics provide advanced levels of care for medical emergencies and trauma. The majority of paramedics are based in the field in ambulances, emergency response vehicles, or in specialist mobile units such as cycle response. They provide out-of-hospital treatment and some diagnostic services, although some may undertake hospital-based roles, such as in the treatment of minor injuries. They have to perform at high expectations in every giving situation just as a disaster responder would, including following ethical and work related guidelines. With this approach, the paramedics will be able to utilize the proposed handheld devices that will run the web application software that handles fingerprint and iris recognition to retrieve medical information related to the clients.

This process validates the testing externally due to the fact that we will be aware of how and when it will be utilized. It will record real-time response of data request from the server to the handheld device and how effectively it assists the paramedic in applying medical care. A request will have to be made to the client to allow for the testing. A survey will be administered to clients that received treatments from the participating paramedics about how they felt about the medical assistance they got. It will studied to see if using the device saved ample time spent in trying to ask about medical information from the clients rather than spending it on actually treating them, not to mention the unresponsive ones.

IV. ASSUMPTIONS

Emergency medical records (EMRs) or Electronic health records (EHRs) can play big role in helping emergency responders during disasters because that would be direct access to important data that can cut down the increase in fatalities. Digitized records provide a timely, cost-effective way to share patient information. If physicians aren't using them in their private practices, they lose those benefits, as do the hospitals they work with [7]. If physicians can utilize this this system effectively, why can't emergency responders do the same? EMR systems can integrate evidence-based recommendations for preventive services (such as screening exams) with patient data (such as age, sex, and family history) to identify the ones that need specific services.

A group of researchers at Weill Cornell Medical College and the University of California, Davis, where they predict improvements in patient outcomes after a major earthquake through more effective use of information technology. They insist a control tower-style telemedicine hub to manage electronic traffic between first responders and remote medical experts could boost the likelihood that critically injured victims will get timely care and survive, according to the team's computer simulation model [8]. Currently the use of biometric technology is being used to track responders on scene by an emergency command center. With a smart mobile computer in the hands of responders, they can bring real-time mobile data to the point of activity. On site at an incident, such solutions can help rapidly screen, check in, and track first responders. Motorola developed mobile devices using mobility solutions that help improve the efficiency and effectiveness of first responders. However, with firefighters, police, rescue, EMS, and other first responders flooding the scene, keeping track of personnel, victims, and assets is no easy task. When every second counts, the paperwork burden and its associated productivity loss can translate into lost lives [9].

Furthermore, with manual systems, critical information often resides on clipboards at the scene and is inaccessible to offsite command centers, evacuation sites, hospitals and other agencies. Complete electronic record documents are made available when responders enter and exit an incident scene which will provide National Incident Management

System compliance (NIMS) which mandates visibility into available resources to best prepare for, respond to and recover from an incident as well as efficient communications and information access during an incident.

The Emergency Operations Center (EOC) is the physical location where various organizations come together under the direction of Emergency Operations Management (EOM) during an emergency to coordinate response and recovery actions and resources. These centers may alternatively be called command centers, situation rooms, war rooms, crisis management centers, or other similar terms [10]. During an incident, there may be a need to confirm the identity and medical credentials of an individual clinical care provider when they request permission to enter the scene. At this step in the flow, the incident control personnel at the scene could request confirmation of the medical credentials in one or more possible ways, including:

- Via a request made to EOC systems. That center could either confirm the credentials from their own internal information sources (e.g. EMTs on staff for at center), or via a query to the Health Information Service Provider. In the latter instance, the response to the query could potentially include additional information that may be used to identify and authenticate the individual, as well as information which describes the role(s) which that individual is authorized to perform (as defined in steps I and III).
- Potentially via a field-deployable authentication device (e.g. identification card reader or biometric device) which could transmit information directly to a remote authentication service and receive authentication confirmation and authorization information in return.
- There may be additional mechanisms available for querying to confirm the credentials of a medical provider in the field. For example in the future, it may be possible to make this query directly from the field without the need for an intermediary (e.g. a query sent directly from the requestor to a registry service).

This is an on-going trend that wasn’t seen as a possibility in the area of emergency response that has continued to gain popularity. Emergency response officials in Tallahassee, Florida, for example, must pass through a biometrics-enabled security system to access their Emergency Operations Center (EOC) as well as their daily office space [11]. There fingerprint scanners at the center where they go up to the doors and put their fingers on to be scanned. It has to be done to be able to access the building.

Federal Emergency Management Agency (FEMA) did a case study in 2008 on responder authentication which stated that advancements in biometric technology and the development of biometric tools for the public safety realm have begun to provide solutions to identity verification issues [12]. Such technologies, when integrated into emergency management plans and processes, can be a powerful tool for emergency response organizations in both meeting day-to-day operational needs and disaster response.

V. LIMITATIONS/DELIMITATIONS

There are several weaknesses that could revolve around implementing these types of technologies currently for disaster response which could possibly include security breach issues, accuracy, consistency, cost, privacy and legal issues, adaptation to newer technologies etc. Some of these limitations carry more effect than others but should all be considered as relevant challenges to this study. The major concerns for the general public's acceptance of the voluntary use of biometrics identification will be: privacy, necessity and identity protection. Many individuals will be concerned that information collected about them could be used against them such as medical records preventing them from the ability to get health or critical illness insurance. Major restrictions that will not be able to be addressed include policies by governing bodies on how medical records are assessed and the absence of a national biometric identification.

Just as the credit bureau keep track of how well we pay our bills and manage our credit, so does the Medical Information Bureau (MIB) on everyone that applies for health insurance and they are required to follow the same rules as the credit bureau because they are considered by the government to be a consumer reporting agency, its services must adhere to the US Fair Credit Reporting Act and the Fair and Accurate Credit Transactions Act. The purpose of MIB is to provide a vehicle for each of its members to contribute health and medical information obtained in connection with the underwriting of, and payment of claims made under, life and health insurance policies, and receives this same type of information contributed to MIB by other members [13]. MIB operates under certain but not all rules and regulations under the Health Insurance Portability and Accountability Act (HIPAA) of Privacy Rule as also stated in the article. This allows them to collect important medical data and share it to its members.

The Health Insurance Portability and Accountability Act (HIPAA) Privacy and Security Rules protects the privacy of individually identifiable health information; the HIPAA Security Rule, which sets national standards for the security of electronic protected health information; and the confidentiality provisions of the Patient Safety Rule, which protects identifiable information being used to analyze patient safety events and improve patient safety [14]. Although this agency would be considered one of the biggest hindrances to the concept of using biometric technologies during disaster response, it would make sense to have state legislatures approve the development of an emergency medical record database that would strictly assist emergency responders on the frontlines of disasters. Policymakers can sponsor universal electronic medical records (EMRs) and propose incentives for “meaningful use” of EMRs. They also state that even though emergency responders are particularly sensitive to the benefits and unintended consequences of EMR adoption, surveillance has been limited [15]. This just means that if further research can be approved by government officials who oversee emergency response agencies, the ability to display to the public that collecting these medical data can be advantageous.

Although this concept may seem out of reach in reaching reality, disaster response continues to lack focus from the U.S government and the past few disasters that occurred showed how much more response processes needs to be addresses and improvements to be made to save more lives.

ACKNOWLEDGMENT

The authors would like to thank RS2 Technologies LLC, a technology-driven manufacturer of cutting edge access management hardware and software, for sponsoring the submission of this study.

REFERENCES

[1] D. Mann, "Katrina shows need for electronic health records," 2005, in press. [Retrieved: April, 2012] Available:
http://www.foxnews.com/story/0,2933,170146,00.html

[2] D. Barry, "Providing medical care to haiti earthquake victims: A nurse practitioner's experience. College health in action," vol. 49, pp. 26. 2010, in press. [Retrieved: April, 2012] Available:
http://www.acha.org/Promotional_Opportunities/docs/ACTION_v49n4.pdf

[3] The Biometric Consortium. "Introduction to biometrics," unpublished. [Retrieved: February, 2012] Available:
http://www.biometrics.org/introduction.php

[4] K. Smith, J. Matey, R. Lazerick, and J. Cambier, "Iris recognition." Biometrics, pp. 1. 2006, in press. [Retrieved: February, 2012] Available:
http://www.biometrics.gov/Documents/irisrec.pdf

[5] BiometricNews.net, "Business and technical factors to be taken into consideration before implementing a biometric system at your place of business," publications, pp. 2-3. [Retrieved: February, 2012] Available:
http://www.biometricnews.net/Publications/Biometrics_Article_Business_Technical_Factors.pdf

[6] V. Chu, and G. Rajendran, "Use of biometrics". TechCast, pp. 1, 3- 4. 2009, in press. [Retrieved: February, 2012] Available:
http://www.techcast.org/Upload/PDFs/634122830612738824_Biometrics-VivianandGayathrilo-res.pdf

[7] M. McGee, "Electronic health records: Time to get onboard. InformationWeek," July 2010, in press. [Retrieved: February, 2012] Available:
http://reports.informationweek.com/abstract/105/3613/Healthcare/electronic-health-records-time-to-get-onboard.html

[8] ScienceDaily, "Health information technology control tower could improve earthquake response," March 2011, in press. [Retrieved: February, 2012] Available:
http://www.sciencedaily.com/releases/

[9] Motorola, "Improve the efficiency and effectiveness of your first responders with mobility,' pp. 1, 2007, unpublished. . [Retrieved: February, 2012] Available:
http://www.motorola.com/web/Business/Products/_Documents/_Static%20files/AB-FirstResponder_1007.pdf

[10] Department of Health & Human Services, "Emergency responder electronic health record," Detailed use case, pp. 32-33, 40, 2006, in press.

[11] D. Bates, "Fingerprints and more: New biometric tools help first responders secure facilities and incident scenes," Homeland1, May 2006, in press. [Retrieved: February, 2012] Available http://www.homeland1.com/homeland-security-products/biometrics-facial-recognition-finger-print-identification/articles/349715-fingerprints-and-more-new-biometric-tools-help-first-responders-secure-facilities-and-incident-scenes/

[12] Federal Emergency Management Agency (FEMA), "Utilizing biometrics to identify Responders in the National Capital Region," 2008, in press. [Retrieved: February, 2012] Available: http://www.fema.gov/pdf/emergency/nims/Alexandria_ANSI_INCITS_398.pdf

[13] B.T. Casey, "Health Insurers' Relationship with the Medical Information Bureau: The Impact of HIPAA Privacy Regulations," riskVue. 2003, in press, [Retrieved: March, 2012] Available: http://www.riskvue.com/articles/rb/rb0307c.htm

[14] United States Department of Health and Human Services, "Health Information Privacy", unpublished. [Retrieved: March, 2012] Available:
http://www.hhs.gov/ocr/privacy/

[15] B.P. Geisler, J.D. Schuur, and D.J. Pallin, "Estimates of Electronic Medical Records in U.S. Emergency Departments," PLoS ONE, vol 5, pp. 1, 2010, in press.

Solving Shortest Hamiltonion Path Problem Using DNA Computing

Hala Mohammed Alshamlan
Computer Science Department
King Saud University
Riyadh, Saudi Arabia
halshamlan@ksu.edu.sa

Mohammed El Bachir Menai
Computer Science Department
King Saud University
Riyadh, Saudi Arabia
menai@ksu.edu.sa

Abstract— **Deoxyribonucleic acid (DNA) computing fundamentally being similar to parallel computing provides a nice way to make trillions of similar calculations in less than moment. Moreover, DNA computing has ability to solve main NP-complete problems such as Hamilton Path Problem, 3-SAT Problem, and Maximum Clique Problem. In this paper, we apply DNA computing to solve Shortest Hamiltonian Path Problem using two steps. First one, determine all Hamiltonian path from specific weighted graph, and then, in second step, we select the shortest one and return it as solution of our problem.**

Keywords-Shortest Path Problem; Hamilinion Shortest Path Problem; DNA Computing; NP Hard Problems.

I. Introduction

DNA computing is a form of computing which uses DNA, biochemistry, and molecular biology, instead of the traditional silicon-based computer technologies [10]. DNA computing, or more generally molecular computing, is based on manipulations with DNA strands using some basic biological transformations. Being very similar to parallel computing, DNA computing promises to solve many NP-complete problems, much faster than modern silicon-based computers do [13].

Leonard Adleman of the University of Southern California initially developed this field. In 1994, Adleman demonstrated a proof-of-concept use of DNA as a form of computation, which solved the seven-point Hamiltonian path problem (HPP) [2]. HPP is to find an air flight path from given cities such that each city is visited once and only once. Therefore, HPP is NP-complete Problem [1][[2][9][12]. Particularly, since Adleman solved a small instance of the Hamilton path problem successfully, the DNA computing has become a new focus in the scientific areas of nanotechnology, biology, mathematics, medicine and information science [4][5]. Compared with electronic computers, which often need exponential time, DNA computing has its own advantages. With its huge parallelism for computing, almost 10^{18} information data can be proceeded in parallel [2].

After that, a major goal of subsequent research is how to use DNA manipulations to solve P and NP-hard problems, especially 3-SAT problems [17]. 3-SAT is one of NP-complete problems, and it is as hard as all the other NP problem, which is to search for a model (or solution) of a set of clauses with each clause composed of no more than three literals, where a literal is a variable (or an atom) or its negation. Various solutions were tried to solve the 3-SAT problem, Lipton [6] proposed DNA experiments on test tubes to solved a satisfiability problem based on DNA sticker computing model. Later, Ouyang used short linear dsDNA molecules and DNA restriction enzymes to solve maximal clique problem [7]. After that, another way for DNA computing was developed, In 2000, Liu et al. [8] introduced a new simple case and method to solve a 3-SAT problem, in which the feasibility of DNA surface computing was verified and also proved that the fluorescence could be used accurately in DNA computing.

Moreover, there are many weight encoding method have been studied in literature like in [14][15][16][17]. Thus, further study of the DNA encoding of weight is very important. Notably, the weight encoding methods are mostly used to Finding simple shortest path problem, which is P problem. Based on the existing literature, we show many methods and algorithms proposed to solve shortest path problem [14] [15][17], but these methods can only solve a specific examples and there are some limitations and concentrations in these methods .In other hand, Hu et al. in [16] proposed an effective and new IMCE encoding method based on Incomplete Molecule Commixed Encoding and use this method to find the shortest path of a seven vertex weighted graph.In order the shortest path problem is a variant of the Hamiltonian path problem in that it asks for the shortest route/path between two given nodes, and because the methods proposed in [16] is very effective comparing to other proposed methods in literature, we apply the Incomplete Molecule Commixed Encoding (IMCE) encoding method proposed in [16] in weighted directed graph to investigate the solution of the shortest Hamilton path problem using DNA computing.

The rest of this paper organized as follow: Since the DNA computing required understanding of biological structure and operation of DNA, we will present general background about this issue in Section 1. In Section 2, we define our problem in formal way. After that, in Section 3, we will illustrate how we use DNA stand to encode shortest Hamiltonian Path Problem. In Section 4, proposed DNA algorithm will be presented. Then, our result and analysis will be demonstrated in Section 5. Finally, our conclusion and contribution appear in last section.

II. DNA Biological Background

In order to understand the DNA computing application, and because the DNA molecular is little bit consider as sophisticated for computer scientist reader, we give the reader foundation about the molecular biology. In this section, we present the DNA biological structure and the main DNA operation.

A. DNA Structure

The (deoxyribonucleic acid) DNA stand is encodes the genetic information of cellular organisms. It consists of polymer chains, commonly referred to as DNA strands [1]. Each strand may be viewed as a chain of nucleotides, or bases, attached to asugar-phosphate "backbone". An n-letter sequence of consecutive bases is known as an n-mer or an oligonucleotide (commonly abbreviated to "oligo") of length n. Strand lengths are measured in base pairs (b.p.). These complex molecules are composed of basic blocks called nucleotides, nucleic acid bases A (adenine), G (guanine), C (cytosine), and T (thymine) [or U (uracil) in RNA] [5]. Each strand, according to chemical convention, has a 5' and a 3' end, thus, any single strand has a natural orientation. The two pairs of bases form hydrogen bonds between each other, two bonds between A and T, and three between G and C. Each base has a bonding surface, and the bonding surface of A is complementary to that of T, and that of G is complementary to that of C. This complementary rule is called Watson–Crick complementary. A single DNA strand can pair with another strand when their sequences of bases are mutually complementary and the chains have opposite polarity [1]. Here is an example of a double stranded DNA chain.

5‘ CCCAATGAACCCCATTT 3’
3’ GGGTTACTTGGGGTAAA 5’

B. DNA Operations

Some (but not all) DNA-based computations apply a specific sequence of biological operations to a set of strands. These operations are all commonly used by molecular biologists. In this section we describe the basic and more important of them in more detail.

1) Synthesis

The synthesizer is supplied with the four nucleotide bases in solution to obtain randomly all possible solutions, which are combined according to a sequence entered by the user. The instrument makes millions of copies of the required oligo and places them in solution in a small vial [18][20].

2) Denaturing, annealing, and ligation

Double-stranded DNA may be dissolved into single strands (or denatured) by heating the solution to a temperature determined by the composition of the strand [2]. If a molecule of DNA in solution meets its Watson-Crick complement, then the two strands will anneal that is, twist around each other to form the famous double helix. In other hand, annealing is the reverse of melting, whereby a solution of single strands is cooled, allowing complementary strands to bind together.

In double-stranded DNA, if one of the single strands contains a discontinuity (i.e., one nucleotide is not bonded to its neighbor), then this may be repaired by DNA ligase. This particular enzyme is useful for DNA computing, as it allows use to create a unified strand from several strands bound together by their respective complements. [9][18]. DNA ligase is used by the cell to repair breaks in DNA strands.

3) Separation of strands

This operation mainly use when we need select specific solution or in filtering step. For this, we may use a "molecular sieving" process known as **affinity purification**. If we want to extract from a solution single strands containing the sequence x, we may first create many copies of its complement, x. We attach to these oligos biotin molecules (a process known as "biotinylation"), which, in turn, bind to a fixed matrix [18].

4) Gel electrophoresis

The contents of a test tube can be separated by increasing length. This is achieved by gel electrophoresis, whereby longer strands travel more slowly through the gel. Electrophoresis is the movement of charged molecules in an electric field [20]. Since DNA molecules carry a negative charge, when placed in an electric field, they tend to migrate toward the positive pole. The negatively charged DNA molecules move toward the anode, with shorter strands moving more quickly than longer ones [4]. Hence, this process separates DNA by length

5) PCR

PCR employs polymerase to make copies of a specific region (or target sequence) of DNA that lies between two known sequences. In order to amplify template DNA with known regions (perhaps at either end of the strands), we first design forward and backward primers (i.e. primers that go from 5' to 3' on each strand. We then add a large excess (relative to the amount of DNA being replicated) of primer to the solution and heat it to denature the double-stranded template. Cooling the solution then allows the primers to anneal to their target sequences. We then add the polymerase, which extends the primers, forming an identical copy of the template DNA. Thus, if we then repeat the cycle of heating, annealing, and polymerising, it is clear that this approach yields an exponential number of copies of the template (since the number of strands doubles after each cycle) [18][20].

III. problem Definition

In this research paper, our main problem is how we identify shortest Hamiltonian path from specific weighted graph using Incomplete Molecule Commixed Encoding (IMCE) model. Before we describe our proposed algorithm that is based on IMCE model to solve the shortest Hamiltonian path problems, we give some description about incomplete molecule. In IMCE computing model or in other word incomplete molecules used in this scheme like the domino shown in Figure 1. The biological operations of these molecules are similar with sticker model.

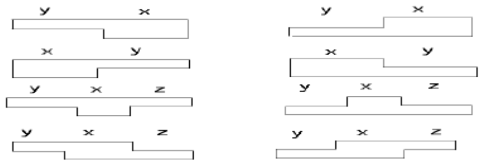


Figure 1. The molecular structure of domino

Sticker model is a universal computing system, which is used by Adelman [1][9] and Lipton [6]. In DNA biological operation, Restriction Enzymes or Nucleases are used to cut and stick the strands. The principle of this model is shown as Figure 2, the logic of the sticker model presented in [21], which is based on the paradigm of Watson–Crick complementarity. In short, the model involves a long single memory strand and a number of sticker strands or stickers as indicated in Figure 2. A memory strand is a single-stranded DNA with n bases. It is divided into k non-overlapping substrands, each of which has m bases, and therefore, n = km. Each sticker has m bases and complementary to exactly one of the k substrands in the memory strand. During a course of computation, each substrand is identified as a Boolean variable and is considered “true” or “false” as its corresponding sticker is annealed or not.

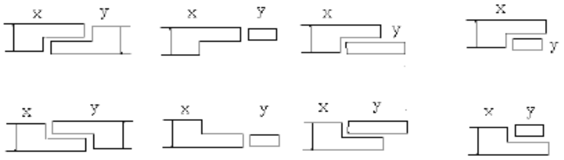
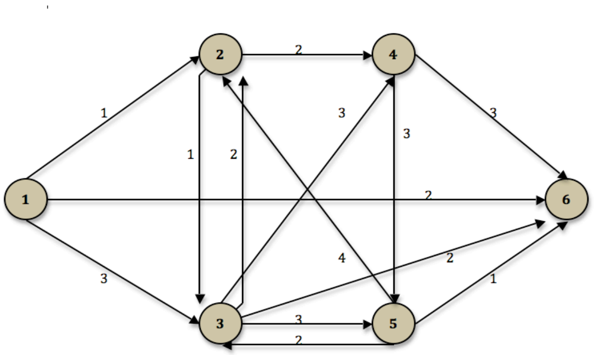


Figure 2. The structure of Sticker model.

Later on, in this paper, we will describe in detail how the IMCE Encoding Scheme solves the shortest Hamiltonian path problems. Therefore, we will define the meaning of abbreviations used in IMCE scheme as follows:
VE(*Vi*): vertex encoding of *Vi*.
WE(*Wi*): weight encoding of *Wi*.
EE: edge encoding

To solve our problem, first, we will determine all Hamiltonian paths from specific weighted graph G, and then we will select the shortest one and return it as solution of this problem. For clarification, given a weighted graph G=(V, E), the vertex set is V, the weight set is W, the Edge set is E, where w ϵW (see Figure 3).



Fighure 3 . A weighted Graph G

In Graph G, there are many Hamiltonian paths presented in Table 1. Each one of them has specific weight, and based on these weights we can determine that the last path is the shortest Hamiltonian path in graph G, which is has the weight equal 9.

TABLE 1. IDENTIFY SHORTEST HAMILTONIAN PATH FROM WEIGHTED GRAPH G

Hamilton Path	Weight
1 → 2 → 4 → 5 → 3 → 6	1+2+3+2+2=10
1 → 3 → 2 → 4 → 5 → 6	3+2+2+3+1=11
1 → 2 → 3 → 4 → 5 → 6	1+1+3+3+1 = **9** **(The Shortest Hamilton Path)**

Adelman [1] applied five steps to solve Hamiltonian path problem (HPP) using DNA computing. Our problem in this study is similar to HPP with additional two steps and update one-step. The first new step is weight encoding and representation. Notably, we consider this step with vertexes and edges encoding as first step, and we must do it before the path construction. Because the weight of each edge is varying, in this problem we cannot determine exactly which paths was visited exactly n vertexes like HPP. For this reason, we update step four to become more suitable in this problem. In addition, the second new step is to select the shortest Hamiltonian path.

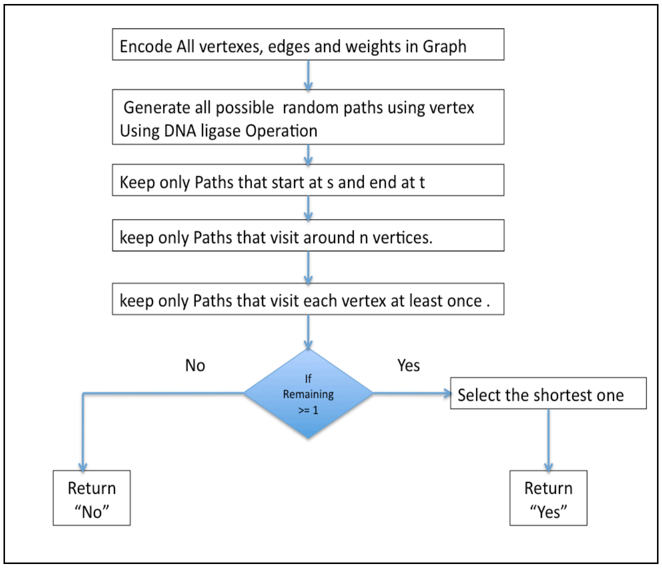

Encode All vertexes, edges and weights in Graph
Generate all possible random paths using vertex Using DNA ligase Operation
Keep only Paths that start at s and end at t
keep only Paths that visit around n vertices.
keep only Paths that visit each vertex at least once .
No
Yes
If Remaining >= 1
Select the shortest one
Return "No"
Return "Yes"

Figure 4. our proposed methodology

Moreover, in this study, we proposed seven main steps that demonstrate our methodological description to solve shortest Hamilton path problem (see Figure 4), as follows:

1. Encode the vertexes, edges and weights.
2. From step1, generate random paths using vertex
3. From all paths created in step 2, keep only those that start at s and end at t.
4. From all remaining paths, keep only those that visit around n vertices.
5. From all remaining paths, keep only those that visit each vertex at least once.
6. If the path remaining more than 1 then From all remaining paths, keep only the shortlist path.
7. If any path remains return, "yes" with the path ;otherwise, return "no".

IV. DNA ENCODING

In this section, we illustrate in detail how we can encode and representing vertexes, weights, and edges in direct graph. In our problem the vertex encoding is similar to Hamiltonian path problem presented in [1][9]. Therefore, we need to investigate specific encoding technique to encoding and representing the weight and then representing the edges.

A. *Vertex Encoding*

Basically, we adopt the length of single strand is 20 mer, because this length is widely applied by many researcher in DNA computing field like by Adelman [1][9] and Lipton [6]. Each encoding of the vertex is unique. For the instance shown in Figure 3, the encoding of vertex V1, V2 and V6 are as follows:
VE(*V1*)= ATGCAAGGTC TGACGGTTCA
VE(*V2*)= GATCGGTAAC GACTGGTTAC
VE(*V6*)= TACGGTTACA TCGATTGAAA

B. *Weight Encoding*

There are important definitions that must be illustrated before we introducing the method of weight encoding, we give as follows:

• *Definition 1*
Weight Set: The set consisted with the weights of the edge in a weighted graph G. For instance, the Weight Set of the graph shown in Figure 3, each edge has different weight; some of them are 1, 2,or 3. Thus, the weight set of graph G is A={1,2,3}.

• *Definition 2*
Minimum function (MIN): Return the minimum value in the Weight Set. For instance, MIN (A)=1.

• *Definition 3*
Complementary strand Mapping function (H): Following the principle of Watson-Crick Complementary, this function will return a complementary base sequence of a sequence. For instance: H (ATTGCA)=TAACGT.

• *Definition 4*
Distribution Ratio: It indicates the disparity between the weights and their average of a given graph, all the elements of the average difference of A is the resolution of graph. Its formula is:

$$\frac{\sum_{j=1}^{n}\sum_{i=j+1}^{n}(W_i - W_j)}{C_n^2 + n}$$

Where Wi and Wj are the weights of edge i and edge j, and they belong to weight set A. Where n is the number of elements of edges in A. For instance, the Distribution Ratio of the graph shown in Figure 3 can be calculated as:

$$=\frac{(2-1)+(3-1)+(3-2)}{C_3^2+3}$$

$$=\frac{(2-1)+(3-1)+(3-2)}{1+3}$$

$$=\frac{(1)+(2)+(1)}{4}$$

$$=\frac{4}{4}=1$$

Now, we begin to illustrate the encoding method of weight. The weight is encoded as double-strand with variable-length. As mentioned previously, in DNA Biological structure, A and T pair form two hydrogen bonds while G and C form three hydrogen bonds. By use of this difference, we use G and C pair to express 1, while A and T pair to express 0. Using this definition, we can translate the weight to binary string. Meanwhile, The binary encoding of 1 = "01" and the binary encoding for 0 = "00". We consider the length of the binary string is variable. Now, the question is: can we use the length of DNA strand in this encoding to distinguish the different length of path of a weighted graph? Let us use an example to illustrate. Supposing from one vertex to another vertex can be directly reached by an edge which weight is 5.It can also be arrived by two edges with weights of 2 and 2. Obviously, 5>2+2. How about the length of the binary string mentioned above? The binary encoding of 5 is “101”. The binary encoding of 2 is “10”. So, the string length of 4 is “1010”. (Here, do not consider the encoding length of the vertex). Obviously, the length of “1010” is larger than “101”. But 4 < 5, so this method is invalid. We cannot use the length of encoding to judge the problem of the shortest Hamiltonian path. In order to solve this problem, we introduce the concept of Distribution Ratio. For the given graph, as shown in Figure 3, the weight set A={1,2,3}, and MIN (A)=1 by the definition above, the Distribution Ratio of A is 1, we can encode based on the minimum value of MIN and Distribution.

So, in our example, the binary encoding of 1 = “01” . Thus, the DNA molecules can be expressed as follows:

C T
WE (1)= G A
2=1+1. The binary encoding is “0101”,
G A G A
WE (2)= C T C T
3=1+1+1. The binary encoding is “010101”,
C T G A G A
WE (3)= G A C T C T

C. *Edge Encoding*

Edge encoding depends on the above vertex encoding and weights encoding methods. We applied incomplete molecular form in our proposed edge encoding method. To represent the edge, first, we put two single-strand (ss) represent the vertex encoding named (*Vi*, and *Vj*) into the test tube .Each single stand divided on two half, for instance *Vi* consist of *Vi’* as first half , and *Vi”* as second half. After that, we put the incomplete double strand (ds) in the same tube, which is represent the weight encoding of edge between vertexes Vi and *Vj* named *Wij*. Figure 5 illustrates the structure of an edge.

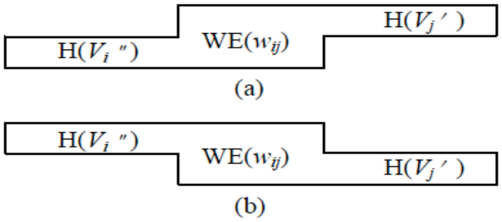


Figure 5. The structure of edge encoding

In Figure 4, the structure of incomplete molecule consists on three parts. The first part H (*Vi*”) is the complementary sequence of *Vi*”, which is the second half of the single strand of vertex *Vi*. The second part is WE (*wij*), which is double-strand represent the edge weight between vertex *Vi* and *Vj*. And the third part H(*Vj’*) is the complementary sequence of *Vj*’, which is the first half of the single-strand of vertex *Vj*. Moreover, Each edge in the graph shown in Figure 3 should be encoded to two incomplete double-strands as described in Figure 4. For example, the encoding of the edge from vertex 1 to vertex 2 with weight 1 in Figure 3 has two incomplete double strands expressed as follow:

EE (*V1-V2*)= H(*V1”*) + WE(1) + H(*V2‘*) =
CTAACTAGCCATTG
ACTGCCAAGT**GA**

And
ACTGCCAAGT**CT**
GAAACTAGCCATTG

Following the above method, we can encode any vertex in graph shown in Figure 3. Notably, that *Vi* is equal (*Vi*‘+*Vi*“). And, we express it by 20bp oligonucleotide fragments, where the length of *Vi’* and *Vi”* are both 10bp (All the direction of encoding is 5′→3′). Then, we can calculate the weight set of the graph and the distribution ratio. After that, we start encoding the weights based on the above algorithm, and then encode the incomplete molecule structure of each edge.

V. DNA ALGORITHM

In this section, we will explain each step in our proposed DNA algorithm to solve shortest Hamiltonian path problem in more details.

Step 1: Encode the vertexes, edges and weights

We already illustrated this step in previous section.

Step 2: Generate Random paths.

In this step, we will mix first tube contain the vertex encoding and the second tube that contain the edge encoding with weight into one tube T. Then in T many ligase operation reactions will take place.

Step 3: keep only those that start at s and end at t.

In this step, we use the PCR operation using the prime of first and end vertexes. Suppose the first vertex is *V2* and the end vertex is *V6*. Thus, in this step we enlarge the reaction for vertex *V2* and *V6* and the number of strands that begin at *V2* and end at *V6* will sharply increase. But, the number of other strands does no change.

Step 4: keep only those that visit exactly n vertices.

The result of this step is approximated, because in this problem we already express the vertex encoding by 20bp oligonucleotide fragments. But, we cannot determine the weight for each edge in the path. Suppose N is the number of vertexes in graph, MDR is the minimum value of MIN and Distribution Ratio, and AVG is the average of weight set A.

In this step, we use the Gel Electrophoresis, and the length L of accepted stand should be as follows:

$$((N+1) * 20) + ((N+1) * |AVG|) < L < (N*20) + (N * |MDR|)$$

Step 5: keep only those that visit each vertex at least once.

In this step, we make strands separation by using sequence of Affinity Purification operation using the compliment of each vertexes in graph many times until reach that each vertex appears exactly one time in each path in tube.

Step 6: keep only the shortlist path.

DNA strands generated from step 5 can be separated in terms of its length by means of gel electrophoresis. The molecules are separated according to their weight, which is almost proportional to their length. Because each edge (i,j) has three parts (H(*Vi"*) + WE (*Wij*) + H(*Vj'*)), and H(*Vi"*) , H(*Vj'*) has equal length which is 10 bp ,we concludes that the different between edges in WE(*Wij*) part. Therefore, the longest path has long molecule stand.

Step 7:Obtaining the Answer.

In the last step, we need to use PCR prime operation and Nucleases (primers) operation to determine the order of each vertexes in shortest Hamiltonian path problem as follow:

Conduct a "graduated PCR" using a series of PCR amplifications. Use primers for the start vertex s and the nth item in the path. So, to find where vertex x lies in the path you would conduct a PCR using the primers from vertex s to vertex x, and by using the following proposed algorithm:

```
// Suppose that L is the length form vertex s to vertex x
1. $\bar{L}$ = Round (L/20).
2. E= ($\bar{L}$ -1)* |AVG|
Where E is Expected weight, and AVG is Average of weight set A.
3. Subtract the expected weight form the length, L=L-E.
4. The order of vertex x = round (L / 20).
```

VI. Analysis and Result

In this section, we proved the efficiency of our proposed algorithm, by present two practical examples.

First example: Suppose that we have graph G that contains two shortest Hamiltonian paths, the first on weight equal 5 and the second one weight 3. The first path weight represented as follows:

GTGTGTGTGT
CACACACACA

So, in this path we required 10 mer. On other hand, the second path weight represented as follows:

GTGTGT
CACACA

Here, in the second path, we required only 6 mer. We observed that the first path is longest than the second path. Therefore, when we put them in the gel electrophoresis we will observe that the second path move faster. However, this technique is based on the fact that DNA molecules are negatively charged. Hence, by putting them in an electric field, they will move towards the positive electrode at different speeds. The longer molecules will remain behind the shorter ones.

Second example: Suppose the length from s the vertex x is 64. So, L = 64 base pairs. (i.e. 4 base for the weight : 1+1= 0101). To determine where vertex x lies in this path, we adopt our proposed algorithm as follow:

$\bar{L}$ =Round (64/20), $\bar{L}$ =3

2. E= (($\bar{L}$ -1) * 4), (i.e. 4= |AVG| =|0101|
   E= 2 *4 = 8
3. Subtract the expected weight form the length (64 - 8 =58)
4. The order = round (58 / 20) nucleotides in the path = 3rd vertex.

Finally, based on previous example we can contribute that our proposed DNA algorithm is visible to solve shortest Hamiltonian path problem.

VII. Discussion and Conclusion

DNA computing is a promising method for unconventional computation, owing to its merits of massive parallelism and efficiency in NP problem solving. One of the most challenging topics in the field of molecular or DNA computing is how to obtain an efficient degree of spatial complexity in ''manufacturing'' the molecules. Here, the word ''manufacturing'' refers to the tasks for preparing or producing the materials by certain technical methods that will be used to build a molecular computer. Therefore, the major problem for Adleman's and Lipton's in DNA computing experiments, is the time involved in extracting and recombining DNA. While DNA processes within the test-tube can take place millions of times per second, extraction processes, whereby individual strands of DNA are manually isolated and spliced, can take several hours and even days, just for the simplest problems. Thus, if we are to apply molecular computing algorithms to the processes of NP- complete problem solving, we really need to obtain a linear order in the space of controlling (i.e., the number of molecules to be controlled) under the condition of linear time complexity. This has led several researchers to conclude that the complexity aspects of DNA algorithms will limit their applicability. However, the research direction in DNA computing field ignores some fundamental biological and computational issues. Such as, the research paper in [3], try to solve this problem by proposed signaling pathways in cells, which is aimed at cutting the cost of building a molecular computer.

Moreover, digital computer provides a way to interact with its processor and memory in such a way that modern programmer simply writes lines of code in some high level language organizing loops, control flow statements and declaration of variables, while silicon-based computers take programmer away from basic operations, DNA computer does not have this ability, to solve a particular problem on DNA molecules one should perform its simplest operations himself spending time in the laboratory.

In this paper, we proposed a new DNA algorithm that solve shortest Hamiltonian path problem. Notably, in literature, there is no research paper proposed to solve this problem yet. For this reasons our proposed solution is original. In other hand, one of the biggest challenges facing the field of DNA computing is that no efficient implementation has been produced for testing, verification, and general experimentation. While Adleman's initial experiment was performed in a lab, many of the subsequent algorithms in DNA computing have never been implemented or tested. For this reason, in future research, we need real experimental DNA algorithm to give measurable and meaningful result.

REFERENCES

[1] Adleman, L., " Molecular Computation of Solutions to Combinatorial Problems." Science. 266: 1021-1024 (Nov. 11, 1994).

[2] Donald B. ,"Molecular Computing", NSF,1995.

[3] Jian Q., and Katsunori S.,"Signaling-pathway-based molecular computing for efficient 3-SAT problem solving ",Information Sciences, Vol.161 2004, pp.121–137.

[4] Suruchi S., Dhiraj B., Yamuna K.,"Transforming bases to bytes: Molecular computing with DNA" , Current trend in science, 2009.

[5] Max H. , and Russell J.,"Biomolecular Computing and Programming", IEEE Transaction on Evolutionary Computation, Vol. 3, No. 3, 1999.

[6] Lipton R., "DNA solution of Hard Combinatorial Problems." Science., Vol. 268, 1995, pp.542-545.

[7] Ouyang Q., Kaplan P. D., Liu S.,"DNA Solution of the Maximal Clique Problem", Science. Vol. 278, 1997, pp.446-449.

[8] Liu Q., "DNA computing on surfaces", Nature, Vol.403, 2000 , pp.175–179.

[9] Adleman L.,"Computing with DNA",Scientific American ,1998.

[10] Martyn A.,"DNA Computing" Invited article for the Encyclopedia of Complexity and System Science, Springer, 2008.

[11] Arita M., Akio N., Hagiya M., " Improving sequence design for DNA computing", Japan society for the Promotion of Science, 2000.

[12] Arita M., Akio S., Hagiya M., "A Heuristic Approach for Hamilton Path Problem with Molecules", Japan society for the Promotion of Science, 1997.

[13] Shalini R. Vijay S., Naveen H ,"Bioological computer Model to Solve NP Complet Problem", International Journal of Information Technology and Knowledge Management , Vol 4, No. 1, 2011, pp. 191-194.

[14] Zuwairie I., Yusei T., Osamu O.,"Direct-Proportional Length-Based DNA Computing for Shortest Path Problem", International Journal of Computer Science & Applications , Vol. I, No. 1, 2004, pp. 46 – 60.

[15] Zhenye W., Zhang Q. , Dang Y., "The Improvement of DNA Algorithm to the Directed Shortest Hamilton Path Problem",IEEE Transaction on Evolutionary Computation, 2009.

[16] Qing W. , Zhi L. , Xiu P. , Qi S., Hong Z.,"DNA algorithm based on incomplete molecule commixed encoding for the shortest path problem", The 1st International Conference on Information Science and Engineering, 2009.

[17] Ajit N., and Spiridon Z., "DNA algorithms for computing shortest paths" University of Exeter press, 1998.

[18] Lovgren S., "Computer Made from DNA and Enzymes". National Geographic. Retrieved ,2009.

[19] Yaakov B., Binyamin G., Uri B., Rivka A., Ehud S.,"An autonomous molecular computer for logical control of gene expression". Nature journal, Vol.429 , 2004, pp.423–429.

[20] Shudong W., Wenbin L., Jin X., "A DNA computing model minimal covering by plasmid", Journal of Huazhong University of Science and Technology, Vol. 32, 2004, pp.59-61.

[21] Roweis S, Winfree E, Burgoyne R, Chelyapov NV, Goodman MF, Rothemund PW, Adleman LM, "A sticker-based model for DNA computation", Vol.5, 1998, pp.615-29.

ISBN: 978-1-61208-202-8

Rapid DNA Signature Discovery Using A Novel Parallel Algorithm

Hsiao Ping Lee[*†], Yen-Hsuan Huang[‡] and Tzu-Fang Sheu[§]
[*]Department of Applied Information Sciences, Chung Shan Medical University, Taichung, Taiwan, 40201 ROC
Email: ping@csmu.edu.tw
[†]Department of Medical Research, Chung Shan Medical University Hospital, Taichung, Taiwan, 40201 ROC
[‡]Department of Applied Information Sciences, Chung Shan Medical University, Taichung, Taiwan, 40201 ROC
Email: kevin656504@hotmail.com
[§]Department of Computer Science and Communication Engineering, Providence University, Taichung, Taiwan, 43301 ROC
Email: fang@pu.edu.tw (corresponding author)

***Abstract*—DNA signatures provide valuable information that can be used in various applications in bioinformatics, for example the identification of different species. Rapid signature discovery algorithms are required by biologists to discover signatures. Since more and more computers are equipped with a CPU of many processing cores, parallelism becomes a feasible solution to accelerate the discovery. However, most of the existing signature discovery algorithms are sequential algorithms. Parallel signature discovery algorithms are rare. In this paper, a parallel signature discovery algorithm is proposed. The algorithm discovers hamming-distance-based signatures from DNA databases. The proposed algorithm is a parallel enhancement of an existing discovery algorithm. Through parallel computing, the algorithm accelerates the process of signature discovery. In the experiment on a human chromosome EST database of 88M bases, the proposed algorithm has up to 73.28% less processing time than the existing discovery algorithm when 4 processors are used.**

***Index Terms*—DNA signature, human chromosome EST database, parallel algorithm, unique signature discovery.**

I. INTRODUCTION

Based on the assumptions of the theories of evolution and natural selection, almost all species shared a common ancestor at a point in time. Random mutations in DNAs sometimes lead to differently structured proteins. If such changes give rise to advantages in survival, the DNAs is prevailed in the gene pool. The advantageous mutations are one of the ways that genomes diverge from one another. The result of the evolution is that the different species might own some unique patterns in their DNA sequences, and the species can be identified by the unique patterns. For example, specific oligonucleotides have already been used in a polymerase chain reaction (PCR) method for the identification of 14 human pathogenic yeast species [1].

DNA patterns are referred to as unique signatures if they appear in a DNA database only once, and have some minimum mutation distance from all other patterns in the database. The unique signatures are used in several bioinformatics researches. For example, unique signatures are used to identify HIV-1 subtypes and 28S rDNA sequences from more than 400 organisms [2]; the selected signature probes with microarray analysis are used to identify bacteria [3].

Unique signature discovery is to find all unique signatures in a DNA database. The methods of unique signature discovery have been widely studied, and many related algorithms, tools and applications have been developed [2]–[14]. For example, insignia [6] is a web application for rapidly identifying unique DNA signatures. Zheng's algorithm [12] is a hamming-distance-based unique signature discovery algorithm. The algorithm deals with DNA databases, and discovers unique signatures from the databases. CMD [13] is an algorithm designed to discover all implicit signatures from DNA databases under a discovery condition, where the implicit signatures are the patterns that satisfy the discovery conditions looser than the given discovery condition.

The internal-memory-based unique signature Discovery (IMUS) algorithm [14] improves upon the Zheng's algorithm. The IMUS algorithm deals with DNA databases. The algorithm discovers hamming-distance-based unique signatures. Let l and d be two positive integers, where $d \leq l$. An l-pattern is a string of l characters in the alphabet set $\{\texttt{A, C, G, T}\}$. A pattern P is (l, d)-mismatched to a pattern Q if the length of P and Q is l and the hamming distance, which is the number of mismatches, between P and Q does not exceed d. A pattern P is referred to as a unique signature under the discovery condition (l, d) if and only if no other pattern Q exists in the given DNA database such that P and Q are (l, d)-mismatched. The IMUS algorithm is designed for efficiently discovering the unique signatures under the discovery conditions of signature length l and mismatch tolerance d.

The underlying idea of the IMUS algorithm is that the unique signatures appear after all of the patterns that are not unique are discarded. Instead of finding unique patterns, the IMUS algorithm focuses on finding non-unique patterns. The IMUS algorithm is based on the observation that if two patterns P and Q are (l, d)-mismatched, then at least one of the two halves of P is $(l/2, \lfloor d/2 \rfloor)$-mismatched to the corresponding part of Q. The IMUS algorithm is a two-phase algorithm. In the first phase, the algorithm divides DNA sequences into patterns of length l. Each l-pattern consists of two consecutive $l/2$-patterns. An index system is built based

ISBN: 978-1-61208-202-8

```
$S \leftarrow$ divide all of the DNA sequences in the input database
into $l$-patterns which comprise two consecutive $l/2$-patterns
$\sigma \leftarrow$ construct an index of $4^{l/2}$ entries, which is based on
the $l/2$-patterns as index keys
for an entry $E$ in $\sigma$ do
  for a pattern $P$ in $E$ do
    for an entry $E'$ in $\sigma$, whose key is $(l/2, \lfloor d/2 \rfloor)$-
    mismatched to $E$'s key do
      compare $P$ to all patterns in $E'$
      if $P$ is $(l, d)$-mismatched to any of the compared
      patterns then
        discard $P$
      end if
    end for
  end for
end for
the remaining patterns is the unique signatures of $(l, d)$
```

Fig. 1. The IMUS algorithm.

on the $l/2$-patterns as index keys, in which l-patterns that contain same index keys are gathered in a single index entry. Assume that E is an entry and its key is K_E. P is an l-pattern in E. Based on the IMUS observation, all of the l-patterns that are (l, d)-mismatched to P are in the entries whose keys are $(l/2, \lfloor d/2 \rfloor)$-mismatched to K_E. In the second phase, P is compared to the patterns that are possibly (l, d)-mismatched to it. P is discarded if it is (l, d)-mismatched to any of the compared patterns. The IMUS algorithm is presented in Figure 1.

Nowadays, CPUs of many processing cores are commonplace, and the prices of the CPUs are in an acceptable range. For example, the price of an Intel Core i7 870 quad-core CPU is around 300 US dollars in November, 2011. Parallel computing technology has been used in several bioinformatics research areas, such as sequence alignment and analyses [15], protein structure prediction [16], [17], and motif finding [18]. Based on our experiments made on a computer with an Intel 2.93GHz CPU, the IMUS algorithm spent about 12.5 hours to discover unique signatures from a database of 88M bases under the discovery condition of signature length 24 and mismatch tolerance 4. However, the IMUS algorithm is a sequential algorithm. The increasing number of processing cores in a CPU would not increase the discovery efficiency of the IMUS algorithm. Therefore, upgrading the IMUS algorithm to a parallel algorithm would further accelerate the signature discovery processes.

In this work, an algorithm that is called parallel internal-memory-based unique signature discovery (PIMUS) algorithm is proposed. The PIMUS algorithm is a parallel enhancement of the typical IMUS algorithm. To improve discovery efficiency, the PIMUS algorithm uses an efficient scheduling heuristic proposed in [13] to generate a reordered processing list. The processing list helps to reduce discovery time to approaching the optimal discovery time for a multi-processor platform. Based on the results from the experiments on human chromosome EST databases of 88.0 and 36.4M bases, the PIMUS algorithm respectively spent about 3.5 hours and 0.62 hours to discover signatures from the EST databases under the discovery condition (24,4) when four processing cores are used. the PIMUS algorithm has up to 71.35% and 72.06% less processing time than the typical IMUS algorithm in the signature discoveries.

TABLE I
A LIST OF 6 PATTERN ENTRIES AND THEIR PROCESSING TIME.

ID	A	B	C	D	E	F
Time	5	8	36	4	13	4

The rest of the paper is organized as follows. The PIMUS algorithm is presented in Section II. The time complexity of the algorithm is analysed in Section III. The results of the performance evaluation about the proposed algorithm are presented in Section IV. Finally, the conclusions of this work are given in Section V.

II. METHODS

The proposed parallel internal-memory-based unique signature discovery (PIMUS) algorithm discovers signatures efficiently from a DNA database that can be entirely loaded into main memory under a certain discovery condition. The PIMUS algorithm improves upon the IMUS algorithm, and accelerates signature discovery by using parallel computing.

An intuitive way to apply parallel computing to the IMUS algorithm is to assign randomly an available processor to process a pattern entries in sequential order. For example, a computer with m processors is used to handle n pattern entries. Initially, processor 1 can be assigned to entry 1, . . .,and processor m can be assigned to entry m. Assume that processor 3 is the first to complete its task; the processor is immediately assigned to the next entry, entry $m + 1$. The next available processor is similarly assigned to the next entry until all of the n pattern entries are completed. The optimal processing time when m processors are used is $1/m$ of the processing time of a single-processor computer.

Table I shows the processing time of six pattern entries. The entries can be treated in 70 time units by a single-processor computer. The optimal processing time is therefore 70/2=35 time units for a two-processor computer. However, in the case of the assignment in sequential order, processor 1 is assigned to entries A and C, and processor 2 is assigned to entries B, D, E and F. The assignment of the entries is presented in Figure 2. The processing time is 41 and 29 time units respectively. Since the processor that takes longest dominates the overall processing time, the overall processing time is 41 time units in this case, which exceeds the optimal processing time.

The order of pattern entries in the processing list influences the overall processing time for parallel discovery. An efficient scheduling heuristic, called the parallel entry list (PEL), is used in the CMD algorithm [13]. Figure 3 presents the PEL

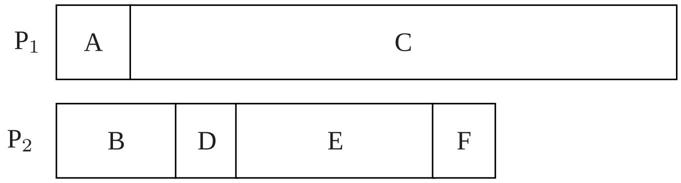


Fig. 2. The assignment of the entries in Table I in sequential order for a two-processor computer.

TABLE II
THE PROCESSING LIST GENERATED BY THE PEL HEURISTIC FOR PROCESSING THE ENTRIES IN TABLE I.

ID	C$_1$	C$_2$	E	B	D	A	F
Time	18	18	13	8	4	5	4

heuristic. The PEL heuristic yields a processing order list for pattern entries in which the entries that involve more patterns are before those that involve fewer. The PEL heuristic is similar to a partial quicksort. Unlike quicksort, the PEL heuristic is iterative, and only operates on the left part of a list in each iteration. The PEL heuristic focuses on all entries in the entry list initially. Assume g is the average number of patterns in each entry within the focused part. The PEL heuristic moves the entries that contain more than g patterns forward to the left part of the entry list in each iteration. In the next iteration, the heuristic focuses on the left part of the entry list, that consists of the entries that contain more than g patterns. The time complexity of the PEL heuristic is $O(n)$, where n is the number of pattern entries in the entry list. The processing list generated by the PEL heuristic for processing the entries in Table I is presented in Table II. Figure 4 presents the assignment of the entries in the processing list in sequential order for a two-processor computer. The overall processing time is 35 time units in this case, which equals the optimal processing time.

Figure 5 presents the PIMUS algorithm. Let l be the desired signature length and d be the mismatch tolerance. The PIMUS algorithm uses a DNA database as an input, and discovers all unique patterns from the database under the discovery condition (l, d). The PIMUS algorithm uses the PEL heuristic to generate a processing order list, and applies parallel computing techniques to process multiple pattern entries simultaneously to accelerate signature discovery processes.

The PIMUS algorithm divides all of the DNA sequences in the input database into l-patterns. Each of the l-patterns comprises two consecutive $l/2$-patterns. An index of $4^{l/2}$ entries is built based on the $l/2$-patterns as entry keys. A multi-level index can be adopted if the index is too large to be fit in main memory. The l-patterns that contain one same key are collected in a single entry. A processing order list of the entries in the index is generated by the PEL heuristic. The reordered entry list makes the number of patterns treated by each of the processors approximately equal. It reduces the overall discovery time to approaching optimal processing time when parallel computing is used.

```
$L \leftarrow$ generate a processing order list that consists of all pattern entries in arbitrary order
$m \leftarrow$ the number of available processors
$n \leftarrow$ the number of entries in $L$
$g, h \leftarrow$ the average number of patterns in each entry in $L$
while $g < mh$ do
  $s \leftarrow 0$, $k \leftarrow 1$
  while $n > k$ do
    while $|L_n| \leq g$ do
      $n \leftarrow n - 1$
    end while
    while $|L_k| > g$ do
      $s \leftarrow s + |L_k|$
      $k \leftarrow k + 1$
    end while
    if $n > k$ then
      exchange $L_k$ and $L_n$
      $s \leftarrow s + |L_k|$
    end if
  end while
  $g \leftarrow s/n$
end while
for $i \leftarrow 1$ to $n$ do
  $Y \leftarrow L_i$
  divide $Y$ into $m$ partitions $Y_1, Y_2, \ldots, Y_m$
  remove $L_i$ from $L$
  put $Y_1, Y_2, \ldots, Y_m$ into $L$
end for
return $L$
```

Fig. 3. The PEL heuristic. L_i is the i-th entry in L. $|L_i|$ is the number of patterns in L_i.

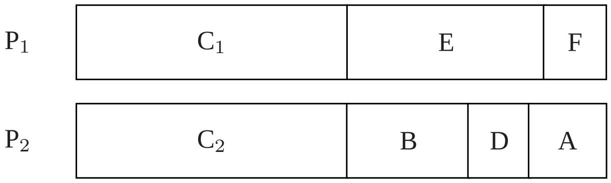


Fig. 4. The assignment of the entries in the processing list in Table II in sequential order for a two-processor computer.

Observation 1 (IMUS Observation). If two patterns P and Q are (l, d)-mismatched, then at least one of the two halves of P is $(l/2, \lfloor d/2 \rfloor)$-mismatched to the corresponding part of Q.

An available processor is assigned to handle the next untreated entry in the processing order list. Two index entries are called similar entries if the number of mismatches between the keys of the two entries is less than or equal to $\lfloor d/2 \rfloor$. Assume that E is an index entry, and P is an l-pattern listed in E. Based on the IMUS Observation, if a pattern Q is (l, d)-mismatched to P, then Q must be in one of the entries

```
$S \leftarrow$ divide all of the DNA sequences in the database into $l$-patterns
$\Gamma \leftarrow$ construct the index of $4^{l/2}$ entries based on $S$
$L \leftarrow$ generate a processing order list of the entries in $\Gamma$ by using the PEL heuristic
for an entry $E$ in $L$ do
    assign an available processor to handle $E$
    for a pattern $P$ in $E$ do
        for an entry $E'$ in $\Gamma$, which is similar to $E$ do
            compare $P$ to all $Q$s, where $Q \in E'$
            if $P$ is $(l,d)$-mismatched to any of the compared $Q$s then
                set the duplication flag of $P$ to true
            end if
        end for
    end for
end for
discard all of the non-unique $l$-patterns
return the remaining $l$-patterns, which are the unique signatures of $(l,d)$ in the database
```

Fig. 5. The PIMUS algorithm.

similar to E. To check the uniqueness of P, P is compared to all patterns in the entries which are similar to E. In each of the comparisons, $l/2$ characters, excluding the key region, are compared. P is not unique if it is (l,d)-mismatched to any of the compared patterns. After all of the entries in the index are treated, the non-unique patterns are discarded. The remaining patterns are the unique signatures under the discovery condition (l,d) in the input database.

III. Mathematical Analysis

Let l be the signature length and d be the mismatch tolerance. The time complexity of the PIMUS algorithm under the discovery condition (l,d) when m processors are used is analyzed.

Assume D is the input database, and $|D|$ denotes the size of the database. Γ denotes the index used in the PIMUS algorithm. Γ consists of 4^{α} pattern entries under the discovery condition (l,d), where $\alpha = l/2$ is the length of the entry keys. Let Γ_i denote the i-th entry in Γ, where $1 \leq i \leq 4^{\alpha}$. $|\Gamma_i|$ denotes the number of patterns in Γ_i. The relationship between $|D|$ and $|\Gamma_i|$ is:

$$\sum_{i=1}^{4^{\alpha}} |\Gamma_i| \quad = \quad 2|D|$$

Assume that E and E' are two entries in Γ. HD(E,E') denotes the hamming distance between E and E', which is defined as the hamming distance between the entry keys of E and E'. A pattern P in an entry $\Gamma_i \in \Gamma$ requires $\sum_j |\Gamma_j|$ string comparisons to check if the patterns that are (l,d)-mismatched to it exist, where $\Gamma_j \in \Gamma$ such that HD$(\Gamma_i,\Gamma_j)\leq \lfloor d/2 \rfloor$. All characters in the l-pattern P, excluding the entry key region, are compared in each of the string comparisons, yielding $l - \alpha = l/2$ character comparisons. Therefore, the number of character comparisons that is used to process all patterns in Γ_i is $(l/2)|\Gamma_i|\sum_j |\Gamma_j|$.

The total amount of character comparisons used in the discovery under the discovery condition (l,d), denoted as $M_{l,d}$, is:

$$M_{l,d} \quad = \quad \sum_{i=1}^{4^{\alpha}} (l/2)|\Gamma_i| \sum_j |\Gamma_j|$$

where $\alpha = l/2$ and $\Gamma_j \in \Gamma$ such that HD$(\Gamma_i,\Gamma_j)\leq \lfloor d/2 \rfloor$.

Assume the input database D is in uniform distribution. In this case, each entry $\Gamma_i \in \Gamma$ should contain $|\Gamma_i| \approx 2|D|/4^{\alpha}$ patterns because of the assumption of uniform distribution. the amount of character comparisons used in the discovery under the discovery condition (l,d), denoted as $\overline{M}_{l,d}$, is:

$$\begin{aligned}
\overline{M}_{l,d} &= \sum_{i=1}^{4^{\alpha}} (l/2)|\Gamma_i| \sum_j |\Gamma_j| \\
&= \sum_{i=1}^{4^{\alpha}} (l/2)(2|D|/4^{\alpha})\kappa(2|D|/4^{\alpha}) \\
&= 4^{\alpha}(l/2)\kappa(2|D|/4^{\alpha})^2 \\
&= 2l\kappa|D|^2/4^{\alpha}
\end{aligned}$$

where $\alpha = l/2$. $\Gamma_j \in \Gamma$ such that HD$(\Gamma_i,\Gamma_j)\leq \lfloor d/2 \rfloor$. $\kappa = \sum_{k=0}^{\lfloor d/2 \rfloor} 3^k \binom{\alpha}{k}$ is the number of all possible permutations that the number of changes does not exceed $\lfloor d/2 \rfloor$ bases in a string of length α.

The time complexity of the PIMUS algorithm when m processors are used, denoted as $\overline{M}^{m}_{l,d}$, is:

$$\begin{aligned}
\overline{M}^{m}_{l,d} &= \overline{M}_{l,d}/m \\
&= 2l\kappa|D|^2/(4^{\alpha}m)
\end{aligned}$$

IV. Experimental Results

The platform that was adopted in the experiments was a personal computer with an Intel Core i7 870 2.93GHz quad-core CPU, 16GB RAM and 1.5TB disk space. The operating system was CentOS release 5.5. The algorithms were implemented in JAVA language, and the programs were compiled by JDK 1.6. The DNA data that were used in the experiments were from the human chromosome 4 and 13 EST databases. The experimental data were denoted as D_4 (human chromosome 4 EST database) and D_{13} (human chromosome 13 EST database) respectively, and their corresponding sizes were approximately 88.0M and 36.4M bases. Before the experiments, the remarks in the databases were removed; all of the universal characters, such as ‘don’t care‘, were replaced with ‘A‘, and DNA sequences that were shorter than 36 bases were discarded. The experiments in this section focused on discovering signatures of length between 24 and 30 with mismatch tolerances of two and four.

TABLE III
THE PERFORMANCE OF THE PIMUS ALGORITHM WHEN USING 4 PROCESSING CORES. THE TIME UNIT IS A SECOND.

(A) database = D_4			
(l, d)	IMUS	PIMUS	Saving(%)
(30,2)	4172.08	1150.29	72.43
(28,2)	6305.20	1710.57	72.87
(26,2)	7324.49	2054.43	71.95
(24,2)	9523.10	2627.35	72.41
(30,4)	8389.32	2252.30	73.15
(28,4)	14113.54	3941.78	72.07
(26,4)	23998.85	6413.65	73.28
(24,4)	44951.49	12878.85	71.35

(B) database = D_{13}			
(l, d)	IMUS	PIMUS	Saving(%)
(30,2)	1048.50	223.12	78.72
(28,2)	945.93	261.86	72.32
(26,2)	1184.62	325.79	72.50
(24,2)	1753.88	458.23	73.87
(30,4)	2333.16	553.16	76.29
(28,4)	2532.26	720.57	71.54
(26,4)	4046.71	1139.96	71.83
(24,4)	7959.96	2224.20	72.06

For reasons of performance and memory consumption, a two-level index was used in the implementation of the IMUS and PIMUS algorithms. The first level of the index comprised 4^{10} direct-accessible entries, and a binary search was used to locate a specified entry in the second level. Since the purpose of our experiments was to evaluate the improvements provided by parallel computing, additional filters, such as the frequency filter that was used in the IMUS algorithm, was excluded from the implementation of the algorithms.

The improvements in discovery performance delivered by the PIMUS algorithm were examined. For 4 processing cores, the performance of the PIMUS algorithm was evaluated by using the algorithm to discover signatures from the experimental databases, D_4 and D_{13}. The percentage time saved is used to evaluate the improvements in the processing time of signature discovery. The time saving is defined as (1-(processing time of the PIMUS algorithm)/(processing time of the IMUS algorithm))*100%. A larger 'saving' means a greater improvement by the PIMUS algorithm. Table III presents the processing time that for the IMUS and PIMUS algorithms under various discovery conditions. The table also presents the time savings delivered by the PIMUS algorithm. The experimental results reveal that the PIMUS algorithm with 4 processing cores requires up to 78.72% less processing time to discover all signatures from D_{13} than the IMUS algorithm under the discovery condition (30,2). Moreover, more than 71.35% of the processing time is saved under every discovery condition in the experiment. Restated, the proposed PIMUS algorithm performs at least 3.49 times faster than the IMUS algorithm when 4 processing cores are used.

To elucidate the benefits of parallel computing for signature discovery, various number of processing cores were used and the PIMUS algorithm was used to discover the signatures of $(l = 30, d = 4)$ from D_4 and D_{13}. Table IV shows the experimental results. The acceleration is the processing time normalized to the processing time when one processor is used. The acceleration values of the PIMUS algorithm increase with the number of processing cores used in the experiment. For example, to discover signatures from D_{13}, the discovery processes that use 2, 3 and 4 processing cores are approximately 2.23, 3.13 and 4.22 times faster than those that use a single processing core respectively.

TABLE IV
THE BENEFITS OF PARALLEL COMPUTING FOR SIGNATURE DISCOVERY. THE TIME UNIT IS A SECOND.

(A) database = D_4		
CPUs	Time	Acceleration
1	8389.32	1.00
2	4239.86	1.98
3	3021.27	2.78
4	2252.30	3.72

(B) database = D_{13}		
CPUs	Time	Acceleration
1	2333.16	1.00
2	1048.36	2.23
3	744.46	3.13
4	553.16	4.22

The PIMUS algorithm treats pattern entries based on their order in a processing list. The influence on discovery efficiency made by the processing list was examined. The PIMUS algorithm that uses the processing list generated by the PEL heuristic is denoted as PIMUS_P and that uses the processing list of pattern entries in the original order in index is denoted as PIMUS_N. To evaluate the improvements in the discovery efficiency of the PIMUS algorithm provided by the PEL heuristic, PIMUS_P and PIMUS_N were respectively used to discover signatures from D_4 and D_{13} in this experiment. Table V presents the improvements in the discovery efficiency of the PIMUS algorithm delivered by the PEL heuristic when 4 processing cores were used. The percentage time saved is used to evaluate the benefits to the PIMUS algorithm made by the PEL heuristic. The time saving is defined as (1-(processing time of the PIMUS_P algorithm) / (processing time of the PIMUS_N algorithm))*100%. A larger 'saving' means a greater improvement delivered by the PEL heuristic. The experimental results reveal that the PIMUS algorithm that uses the processing list generated by the PEL heuristic saves up to 41.35% overall processing time than that uses the processing list of pattern entries in the original order in index. The amount of time used by the PEL heuristic to reorder the processing list is presented in Table VI. All of the processing time used by the PEL heuristic to reorder the processing list are less than 1.03 seconds in the experiment. Compared with the discovery time, the generation time of the reordered processing lists is negligible.

V. CONCLUSIONS

This work proposes a parallel unique signature discovery algorithm called parallel internal-memory-based unique signature discovery (PIMUS) algorithm. The PIMUS algorithm is a parallel enhancement of the existing IMUS algorithm. The proposed PIMUS algorithm discovers hamming-distance-based unique signatures under a certain discovery condition efficiently. For example, when 4 processing cores are used, the PIMUS algorithm can discover the unique signatures of length 30 and mismatch tolerance 2 in 1150 seconds from

TABLE V
THE BENEFITS TO THE PIMUS ALGORITHM MADE BY THE PEL HEURISTIC WHEN 4 PROCESSING CORES WERE USED. THE TIME UNIT IS A SECOND.

(A) database = D_4			
(l, d)	PIMUS_{N}	PIMUS_{P}	Saving(%)
(30,2)	1359.70	1150.29	15.40
(28,2)	2194.71	1710.57	22.06
(26,2)	2585.03	2054.43	20.53
(24,2)	3439.14	2627.35	23.60
(30,4)	2523.20	2252.30	10.74
(28,4)	4554.72	3941.78	13.46
(26,4)	7331.77	6413.65	12.52
(24,4)	14253.49	12878.85	9.64
(B) database = D_{13}			
(l, d)	PIMUS_{N}	PIMUS_{P}	Saving(%)
(30,2)	360.42	223.12	38.09
(28,2)	435.27	261.86	39.84
(26,2)	555.50	325.79	41.35
(24,2)	760.05	458.23	39.71
(30,4)	737.69	553.16	25.01
(28,4)	813.03	720.57	11.37
(26,4)	1315.04	1139.96	13.31
(24,4)	2987.61	2224.20	25.55

TABLE VI
THE PROCESSING TIME USED BY THE PEL HEURISTIC TO GENERATE THE REORDERED PROCESSING LIST. THE TIME UNIT IS A SECOND.

(A) database = D_4		(B) database = D_{13}	
(l, d)	Time	(l, d)	Time
(30,2)	1.03	(30,2)	0.72
(28,2)	0.92	(28,2)	0.65
(26,2)	0.80	(26,2)	0.60
(24,2)	0.51	(24,2)	0.45
(30,4)	1.03	(30,4)	0.71
(28,4)	0.92	(28,4)	0.65
(26,4)	0.76	(26,4)	0.60
(24,4)	0.54	(24,4)	0.46

an EST database of 88M bases. Compared with the typical IMUS algorithm, it saves more than 72% of the discovery time. The PIMUS algorithm can be used to rapidly discover signature data for further analysis, for example finding implicit signatures.

ACKNOWLEDGMENT

The authors would like to thank the National Science Council of the Republic of China, Taiwan, for financially supporting this research under Grants 100-2218-E-040-001.

REFERENCES

[1] B. M. Kiryu and C. P. Kiryu., “Rapid identification of candida albicans and other human pathogenic yeasts by using oligonucleotides in a pcr.” *J. Clin. Microbiol.*, vol. 73, pp. 1634–1641, 1998.

[2] L. Kaderali and A. Schliep., “Selecting signature oligonucleotides to identify organisms using dna arrays.” *Bioinformatics*, vol. 18, no. 10, pp. 1340–1349, 2002.

[3] P. Francois, Y. Charbonnier, J. Jacquet, D. Utinger, M. Bento, D. Lew, G. M. Kresbach, M. Ehrat, W. Schlegel, and J. Schrenzel, “Rapid bacterial identification using evanescent-waveguide oligonucleotide microarray classification,” *Journal of Microbiological Methods*, vol. 65, no. 3, pp. 390–403, 2006.

[4] A. M. Phillippy, J. A. Mason, K. Ayanbule, D. D. Sommer, E. Taviani, A. Huq, R. R. Colwell, I. T. Knight, and S. L. Salzberg, “Comprehensive dna signature discovery and validation,” *PLoS Computational Biology*, vol. 3, no. 5, 2007.

[5] E. K. Nordberg, “Yoda: selecting signature oligonucleotides,” *Bioinformatics*, vol. 21, pp. 1365–1370, 2005.

[6] A. M. Phillippy, K. Ayanbule, N. J. Edwards, and S. L. Salzberg, “Insignia: a dna signature search web server for diagnostic assay development.” *Nucleic Acids Research*, vol. 37, no. 2, pp. 229–234, 2009.

[7] S. H. Chen, C. Z. Lo, S. Y. Su, B. H. Kuo, C. A. Hsiung, and C. Y. Lin., “Ups 2.0: unique probe selector for probe design and oligonucleotide microarrays at the pangenomic/genomic level.” *BMC Genomics*, vol. 4, no. 6, 2010.

[8] R. C. Fry, M. S. DeMott, J. P. Cosgrove, T. J. Begley, L. D. Samson, and P. C. Dedon., “The dna-damage signature in saccharomyces cerevisiae is associated with single-strand breaks in dna.” *BMC Genomics*, vol. 7, no. 313, 2006.

[9] M. W. J. van Passel, E. E. Kuramae, A. C. M. Luyf, A. Bart, and T. Boekhout., “The reach of the genome signature in prokaryotes.” *BMC Evolutionary Biology*, vol. 6, no. 84, 2006.

[10] M. Nicolau, R. Tibshirani, A.-L. Borresen-Dale, and S. S. Jeffrey., “Disease-specific genomic analysis: identifying the signature of pathologic biology.” *Bioinformatics*, vol. 23, pp. 957–965, 2007.

[11] K. C. Bader, C. Grothoff, and H. Meier., “Comprehensive and relaxed search for oligonucleotide signatures in hierarchically-clustered sequence datasets.” *Bioinformatics*, vol. 27, pp. 1546–1554, 2011.

[12] T. J. J. Zheng, T. J. Close and S. Lonardi., “Efficient selection of unique and popular oligos for large est databases.” *Bioinformatics*, vol. 20, pp. 2101–2112, 2004.

[13] H. P. Lee, T. F. Sheu, and C. Y. Tang, “A parallel and incremental algorithm for efficient unique signature discovery on dna databases.” *BMC Bioinformatics*, vol. 11, p. 132, 2010.

[14] H. P. Lee, T. F. Sheu, Y. T. Tsai, C. H. Shih, and C. Y. Tang., “Efficient discovery of unique signatures on whole-genome est databases.” in *Proceeding of the 20th annual ACM Symposium on Applied Computing (SAC2005)*, 2005, pp. 100–104.

[15] Y. Chen, A. Wan, and W. Liu., “A fast parallel algorithm for finding the longest common sequence of multiple biosequences.” *BMC Bioinformatics*, vol. 7, no. 4, 2006.

[16] W. Sun, S. Al-Haj, and J. He., “Parallel computing in protein structure topology determination.” in *Proceedings of 26th Army Science Conference*, 2008.

[17] J. R. Green, M. J. Korenberg, and M. O. Aboul-Magd., “Pci-ss: Miso dynamic nonlinear protein secondary structure prediction.” *BMC Bioinformatics*, vol. 10, no. 222, 2009.

[18] W. N. Grundy, T. L. Bailey, and C. P. Elkan., “Parameme: a parallel implementation and a web interface for a dna and protein motif discovery tool.” *Bioinformatics*, vol. 12, pp. 303–310, 1999.

Does Collarette of Iris Work for Recognizing Persons?

Ren-He Jeng, Wen-Shiung Chen
VIP-CCLab., Dept. of Electrical Engineering
National Chi Nan University
Puli, Nantou, Taiwan
s98323907@mail1.ncnu.edu.tw, wschen@ncnu.edu.tw

Lili Hsieh
Dept. of Information management
Hsiuping University of Science and Technology
Taichung, Taiwan
lily@mail.hust.edu.tw

***Abstract*—In recent years, iris has been extensively discussed in the field of biometrics. An iris recognition system has three main stages such as image preprocessing, feature extraction and template matching. Since eyelid and eyelashes act as a kind of armour that protect the eye from harm, it makes iris localization inaccurate in image pre-processing step. A novel method is proposed to locate iris radius based on collarette of iris muscle. The collarette is the thickest region of the iris, separating the pupillary portion from the ciliary portion. Some researches suggest that the parameters of iris normalization algorithm adopts collarette to replace iris outer radius. However, reducing the normalization radius of iris will deform normalized iris image and result in lose of iris feature information. In this paper, we present our experiments by adopting different iris radii and different normalization algorithms in iris recognition system. We also propose an iris localization method and a collarette localization method. In feature extraction, we adopt the Gabor wavelet filter to extract local texture features from iris images. All experiments are tested on UBIRIS Sessao.1 and CASIA.v1 databases. The experimental results show that the proposed approach has achieved a high accuracy of 96%.**

***Keywords*-Biometrics; Iris Recognition; Iris Localization; Iris Normalization; Collarette;**

I. INTRODUCTION

In biometric-based automatic identity authentication techniques, iris recognition is one of the most reliable and trusted methods. Human iris is a thin circular organ which lies between the cornea and the sclera of a human eye. Literature show that, among all the biometric traits, iris has most rich texture information and very high uniqueness, which has been proved in the first automated iris recognition system developed by Daugman in [1]. In his system, a human iris is localized by using the integro-differential operators and then the cropped iris region is linearly normalized to rectangular image. Following the preprocessing step, 2D Gabor wavelets are used to extract iris codes based on the sign of the phase angle, and the iris codes are matched by using Hamming distance.

Another well-known iris recognition system was proposed by Wildes et al. [2], in which the Hough transform is applied to locate iris and the Laplacian pyramid is used to extract four band-pass components from one iris image as their feature presentation. Typically, the framework of iris recognition systems includes three steps: image preprocessing, feature extraction, and classification/recognition. In fact, the key step is to detect the range of interesting in the image preprocessing part. This key step is sometimes called the normalization step and it effects the system recognition performance drastically. The general normalization algorithm was proposed by Daugman [1]. This algorithm is the most popular and has been widely used in many systems, in which irises are assumed in a homogenous ”rubber-sheet” model. In Daugman’s approach, the annular iris region is linearly mapped or transformed into a fix-sized rectangular block via the following formulas:

$$\begin{cases} x\left(r,\theta\right) = \left(1-r\right)x_p\left(\theta\right) + rx_i\left(\theta\right) \\ y\left(r,\theta\right) = \left(1-r\right)y_p\left(\theta\right) + ry_i\left(\theta\right) \end{cases} \quad (1)$$

where $(x_p(\theta), y_p(\theta))$ and $(x_i(\theta), y_i(\theta))$ are the polar coordinates of the inner and outer boundary points in the direction θ in the original image, (x, y) are the Cartesian coordinates. In 2000, H. J. Wyatt’s [3] proposed a meshwork of ’skeleton’ that can minimize ’wear-and-tear’ of iris as pupil size varies. By following, Yuan and Shi [4] adopted the idea in [3] as a basic model and simplified it, and developed a non-linear iris normalization model algorithm. The modified approach was applied to overcome the non-linear deformation on the iris texture caused by pupil variations. It has been shown that this modified approach achieves a relatively good performance. However, the model needs to solve two simultaneous equations, it is complicated to get the sampling points and the time complexity is high. Moreover, the model is not entirely accurate since it assumes the stretch of iris tissue in radial direction is linear as the pupil size changes. Changing the size of pupil is controlled by iris with sphincter muscle and radial muscle from different expanded level of direction. Therefore, to develop a non-linear normalization method for resolving iris texture deformation is necessary.

Though the eyelid and eyelash protect the iris of a human eye, the blocking from them actually affects the processing of image preprocessing step such that almost all systems are not capable of precisely locating the radius of an iris. Accordingly, some research works focus on how to locate the iris precisely. One of the works is to localize the so-called

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-202-8

”collarette” [5] [6], which is a clear division line between sphincter muscle and radial muscle, as illustrated in Fig. 1. Some researchers believe that the collerate may replace iris outer radius potentially and improve the performance of iris recognition systems. In 2004, Sung *et al.* [5] proposed a framework of iris recognition with collarette detection algorithm for locating boundary using statistical information, and the success rate increases 1.0%. This system was tested in two fields, one is between inner boundary and outer boundary and another is between inner boundary and collarette boundary.

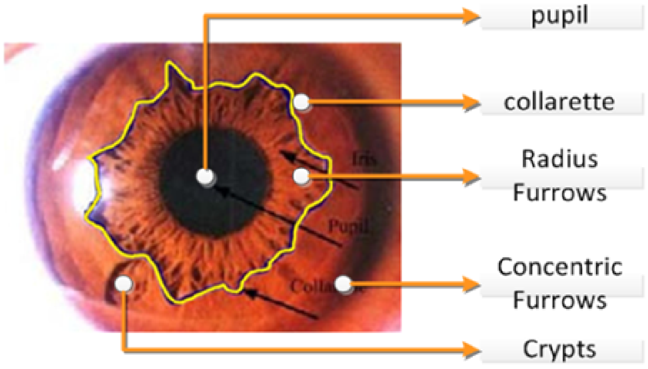


Figure 1. The structure of a human iris.

II. Our Method

A. Iris Localization

The framework of our recognition system is shown in Fig. 2. In the image preprocessing part, there are three processes such as initial iris localization, collarette detection, and normalization. Our proposed collarette detection method has a limited condition for initial iris localization radius due to collarette zone, which is the region between iris outer radius and iris inner radius. Accordingly, the initial iris radius could wrap collarette. If initial iris radius could not provide enough normalization radius, it does not illustrate collarette in unwrapping image.

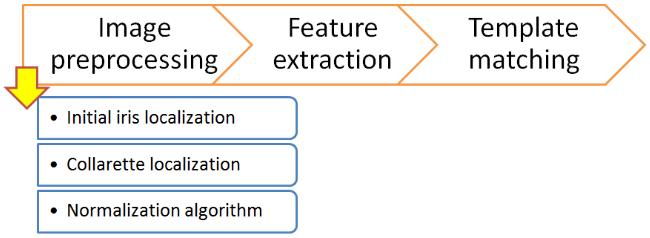


Image preprocessing
Feature extraction
Template matching
• Initial iris localization
• Collarette localization
• Normalization algorithm

Figure 2. The framework of the iris recognition.

1) Initial Radius Localization: Initial radius localization consists of four operations in the following: segmentation, k-means computation, boundary points detection and radius localization, as shown in Fig. 3. First, we must assume a center point, radius and segmentation range for segmenting assumed iris zone, as shown in Fig. 4 (a)(c). Besides, we used pixel intensity of slices to decide initial boundary, as $S_i = \max|\mathrm{P}_i - u_i|$, where S_i is boundary, P_i is pixel intensity, and u_i is the mean value of each slice. A slice means a serial of pixels. Then the initial radius initial is defined to be the mean value of S_i. Following, in the second part, we used k-means algorithm to cluster color features, as illustrated in Fig. 4 (b)(d). Moreover, we find the closest different value with value on index initial of slice, which are radius points r_p, as illustrated in Fig. 4 (e). Finally, the r_p points are mapped onto original eye image, and then the mean values of axis-x and axis-y are the coordinates of the initial radius center point, due to the instinctive symmetry in the shape of irises. In fact, the radius is the mean of distances between center points and r_p.

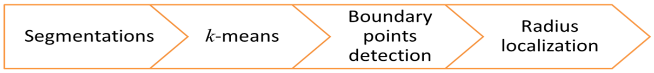


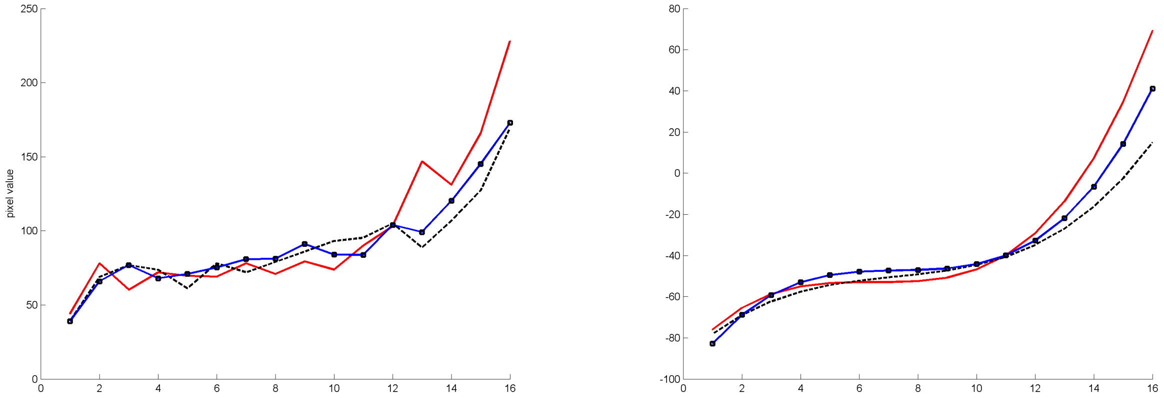
Figure 3. The flow diagram of initial radius localization.

2) Collarette Localization: In the beginning, we refer to ”pushing and pulling” model [7] to develop our proposed initial radius localization. After unwrapping iris by eq. 1 with initial radius, we observe pixel intensity of the slices, as shown in Fig. 5(a), it is an example curve about pixel intensity in three angles. Since the trend of curves are upwards, our goal is to find the points where the intensity changes from flat into upwards in every curves. Figure 5(a) that the curves are not smooth so that it might be located at local parts. Accordingly, we construct a curve in polynomial curve fitting [8], as illustrated in Fig. 5(b), that has the best fit to a series of data points, possibly subject to constraints.

When computing the mean value u_i in i-th curve (L_i), we find the closest point in the curves $P_i = \min|L_i - u_i|$, as illustrated in Fig. 6, and the real-line is the simulated collarette line. And then computing the mean of P_i, we obtain the position index of the collarette in iris unwrapping image. Finally, by re-mapping the pseudo collarette to original iris, the real collarette radius is then detected.

B. Iris Normalization

Since the location of the iris is known, our experiments are tested on UBIRIS.v1 series 1 and CASIA.v1 with linear normalization and non-linear normalization algorithm, respectively. The linear normalization was proposed by Daugman [1], and in this paper we will introduce a non-linear normalization algorithm in the following subsection.



(a)

(b)

Figure 5. The pixel value samples of an iris.

1) Fast Algorithm of Non-Linear Normalization: According to the method in the non-linear normalization model [3], [4] mentioned above, we know that the final goal is to find out the Cartesian coordinates A_x and A_y of the sampled point A, indirectly by first knowing the coordinates of the virtual point A'. If we know the length of $\overline{\text{OA}}$ between the point A and the pupil center O, and the angle $\theta_r(i)$ between $\overline{\text{OA}'}$ and y-axis, the coordinates of the point A may be determined. It is observed from Fig. 7 that the two points A$'$ and A are collinear,so $\overline{\text{OA}}$ and $\overline{\text{OA}'}$ have the same angle $\theta_r(i)$. Obviously, the three points, the point A$'$, the pupil center O and the center o_1 of $arc(\text{P}'\text{I}')$, form a triangle $\triangle\text{A}'\text{O}o_1$, as shown in Fig. 7. Since the lengths of three sides of the triangle are known, the angle $\theta_r(i)$ can be determined according to the law of cosine.

Similarly, the three points, the point A, the pupil center O and the center o_2 of $arc(\text{PI}')$, also from another triangle $\triangle\text{AO}o_2$, as shown in Fig. 7. In this triangle only $\overline{\text{OA}}$ is unknown. According to the law of cosine, the length of $\overline{\text{OA}}$ may be determined from $\theta_r(i)$ which might be obtained from $\triangle\text{A}'\text{O}o_1$. Trivially, the coordinates, A_x and A_y, of the sampled point A are computed by the following equations:

$$\begin{cases} A_x = \Delta_1(i)\sin(\theta_r(i)) \\ A_y = \Delta_1(i)\cos(\theta_r(i)) \end{cases} \tag{2}$$

where

$$\theta_r(i) = \cos^{-1}\left[\frac{y_2^2 + (r_{ref} + \Delta_2(i))^2 - r_2^2}{2y_2(r_{ref} + \Delta_2(i))}\right] \tag{3}$$

and

$$\Delta_1(i) = \left(r_1^2 + y_1^2 - 2y_1 r_1 \cos(\theta_k(i))\right)^{1/2} \tag{4}$$

with

$$\theta_k(i) = \sin\left(\pi - \sin(\theta_r(i)) - \frac{y_1 \sin(\theta_r(i))}{r_1}\right) \tag{5}$$

Finally, we adopt the cartesian coordinates of all of the sampled point A on $arc(\text{PI}')$ to construct a non-linear normalization model directly. The detailed procedure is shown in algorithm

Algorithm 1 Fast Non-Linear Normalization Algorithm

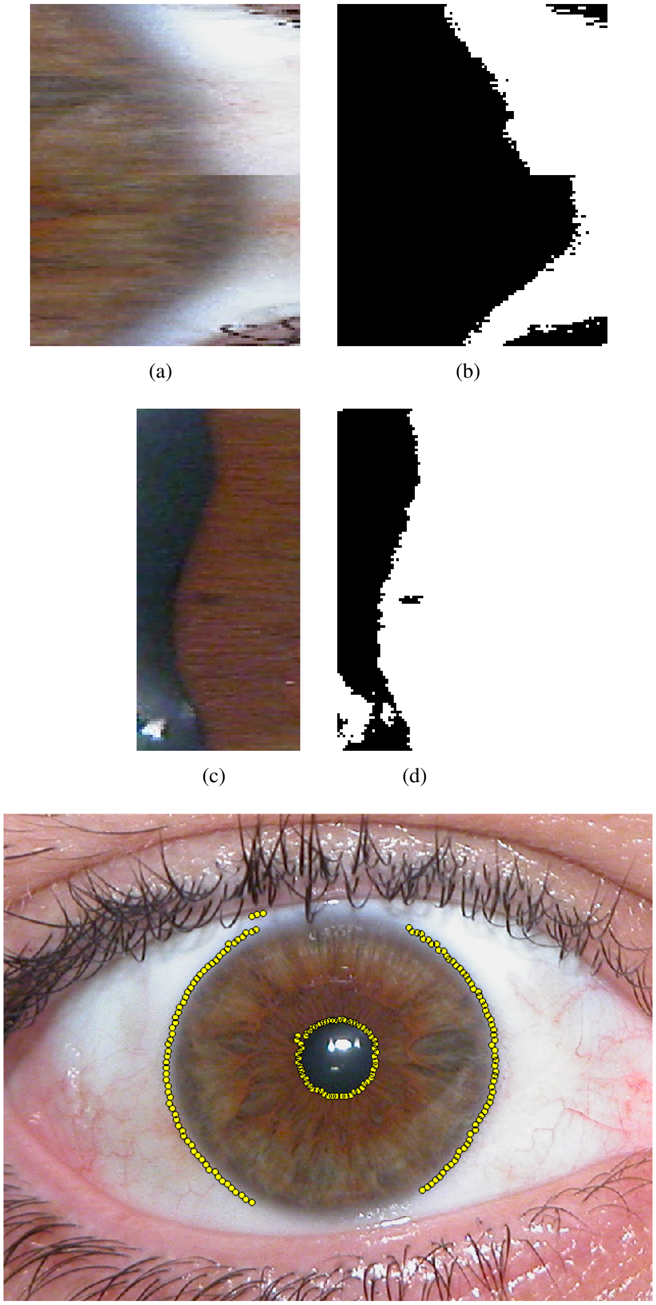
```
FOR $i = 1$ to $m - 1$ do
    Compute $\theta_k(i)$;
    Compute $\Delta_1(i)$;
    Compute $\theta_r(i)$;
    $A_x = \Delta_1(i)\sin(\theta_r(i))$;
    $A_y = \Delta_1(i)\cos(\theta_r(i))$;
END
```

C. Iris Enhancement

First, we use linear normalization and non-linear normalization to transform the detected iris region into 128 × 32 and 64 × 64 rectangular iris image, respectively. Then the mean of pixel is computed for each block of size 16 × 16, and these means are processed by bi-cubic interpolation to estimate the background illumination. After the background illumination factor is reduced, the local histogram equalization is conducted to reveal the details of the iris texture.

D. Feature Extraction

Feature extraction is used to reduce the data size of the image and to extract the key features. Therefore, we adopt 2D Gabor wavelets [1] to extract features. We only store a small number of bits for each iris code, so the real and imaginary parts are each quantized. To perform this task, we



(a) (b)

(c) (d)

(e)

Figure 4. The example result of radius localization.

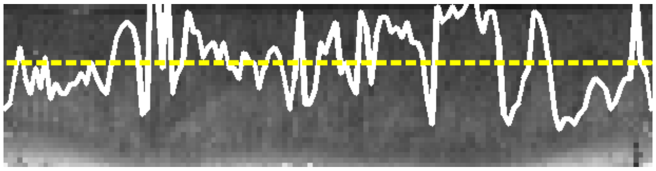


Figure 6. The simulated collarette line.

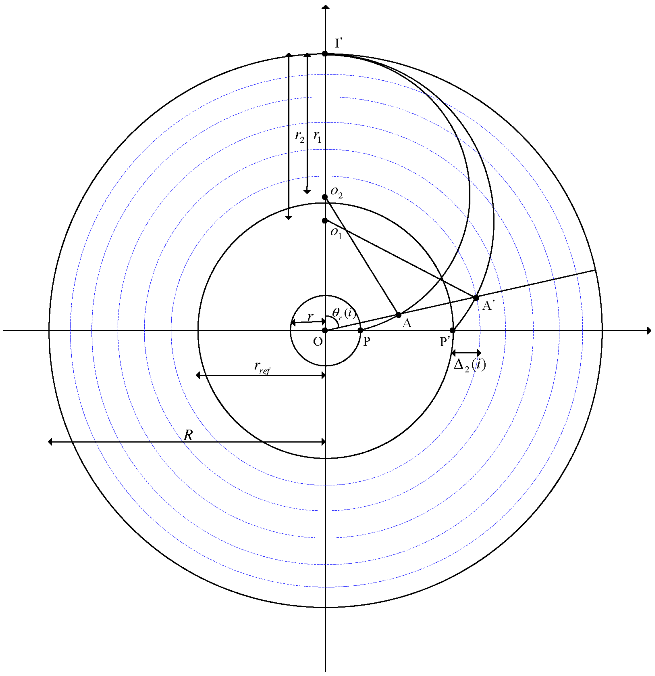


Figure 7. Fast non-linear normalization model.

measure the Hamming distance [1] between two iris codes and then set threshold to recognize them.

III. Experiments and Analysis

A. Iris Database Description

Our experiments for identification will be tested on UBIRIS.v1 Sessao 1 and CASIA.v1 database, respectively. The UBIRIS.v1 Sessao 1 [9] has 241 people for a total 1214 images and at least 5 photos in each class of size 800 × 600, and taken at a moderate distance under visible wavelength light and the primary objective is to reduce the need for cooperation, which means the users would not feel constrained during the process of image acquisition. It is developed by the SOCIA lab (Soft Computing and Image Analysis Group) of the University of Beira Interior. In the CASIA.v1 database [10], there are 756 images for 108 people and least 7 photos in each class of size 320 × 280, and captured by near infra-red camera automatically in room.

B. Comparison of the Sampling Distances with Different Iris Radius Sets

In this paper, we tested iris recognition system with different normalization algorithms, linear normalization algorithm and non-normalization algorithm, and different radius. For the unwrapping part, we tested three fields. The first field is between inner boundary and outer boundary, the second one is between collarette and outer boundary, and the third one is the field between inner boundary and collarette boundary.

As iris radius and collarette were detected, we computed the difference value between the coordinates of *i*-th and (*i*+1)-th sample points, as shown in Fig. 8. Since the image size of eye in CASIA is smaller than UBIRIS.v1 Sessao 1, the sampling distance curves are beating lightly in Fig. 8 (c)(d).

C. Experimental Results

In our experiments, we define three iris radii: iris inner (*ir*), iris outer (*IR*) and collarette (*cr*). In fact, all experiments are tested in three wrapping fields with *ir-IR*, *ir-cr* and *cr-IR*. As mentioned in previous section, we extract iris features by using 2D-Gabor wavelet filter, store a small number of bits for each iris code, and compare them by using Hamming distance. Observing Table 1 and Table 2, that experimental results obtained on UBIRIS.v1 Sessao 1 shows that *cr-IR* is better than *ir-cr*, no matter what linear normalization algorithm or non-linear normalization algorithm are used. Besides, the results for CASIA.v1 normalized in *cr-IR* is better than in *ir-cr* with linear normalization algorithm. However, the results for CASIA.v1 with non-linear normalization are just in reverse, as listed in Table 1. Comparing eye images in two different databases, we found that the UBIRIS.v1 Sessao 1 has much complex furrows besides the collarette, and the texture under collarette is straight muscle. After unwrapping it, the divergence of texture of iris images in UBIRIS.v1 Sessao 1. presents lightly. Moreover, the iris furrows are complex near pupil in CASIA.v1, so the feature representation in *ir-cr* is better than in *ir-IR*.

Table I
EXPERIMENTAL RESULTS WITH LINE NORMALIZATION ALGORITHM (%)

	ir-IR	ir-cr	cr-IR
CASIA.v1	6.9	9.3	4.3
UBIRIS.v1 Sessao	3.5	4.5	3.7

Table II
EXPERIMENTAL RESULTS WITH NON-LINE NORMALIZATION ALGORITHM (%)

	ir-IR	ir-cr	cr-IR
CASIA.v1	3.5	4.3	9.6
UBIRIS.v1 Sessao 1	3.5	3.3	3.2

IV. CONCLUSION

In this paper, we propose a collarette detection method based on pixel intensity variation by using polynomial curve fitting, unlike existing variants of collarette by using a strong classifier. We only extract features with 2D-Gabor wavelet filter and store a small number of bits for each iris code, then compare them by using hamming distance, and set a threshold to classify them. Therefore, the experimental results show the effectiveness of our radius and collarette localization algorithms as well as iris normalization algorithm.

REFERENCES

[1] J. G. Daugman. High confidence visual recognition of persons by a test of statistical independence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(11):1148–1161, 1993.

[2] R. P. Wildes. Iris recognition: An emerging biometric technology. *Proceedings of the IEEE*, 85(9):1348–1363, 1997.

[3] H. J. Wyatt. A ‘minimum-wear-and-tear’ meshwork for the iris. *Vision Research*, 40(16):2167–2176, 2000.

[4] X. Yuan and P. Shi. Iris recognition using collarette boundary localization. In *Proceedings of International Workshop on Biometric Recognition Systems on Advances in Biometric Person Authentication (IWBRS 2005)*, volume 4, pages 135–141, 2005.

[5] H Sung, J. Lim, J. H. Park, and Y. Lee. Iris recognition using collarette boundary localization. In *Proceedings of the 17th International Conference on Pattern Recognition (ICPR 2004)*, volume 4, pages 857–860, 2004.

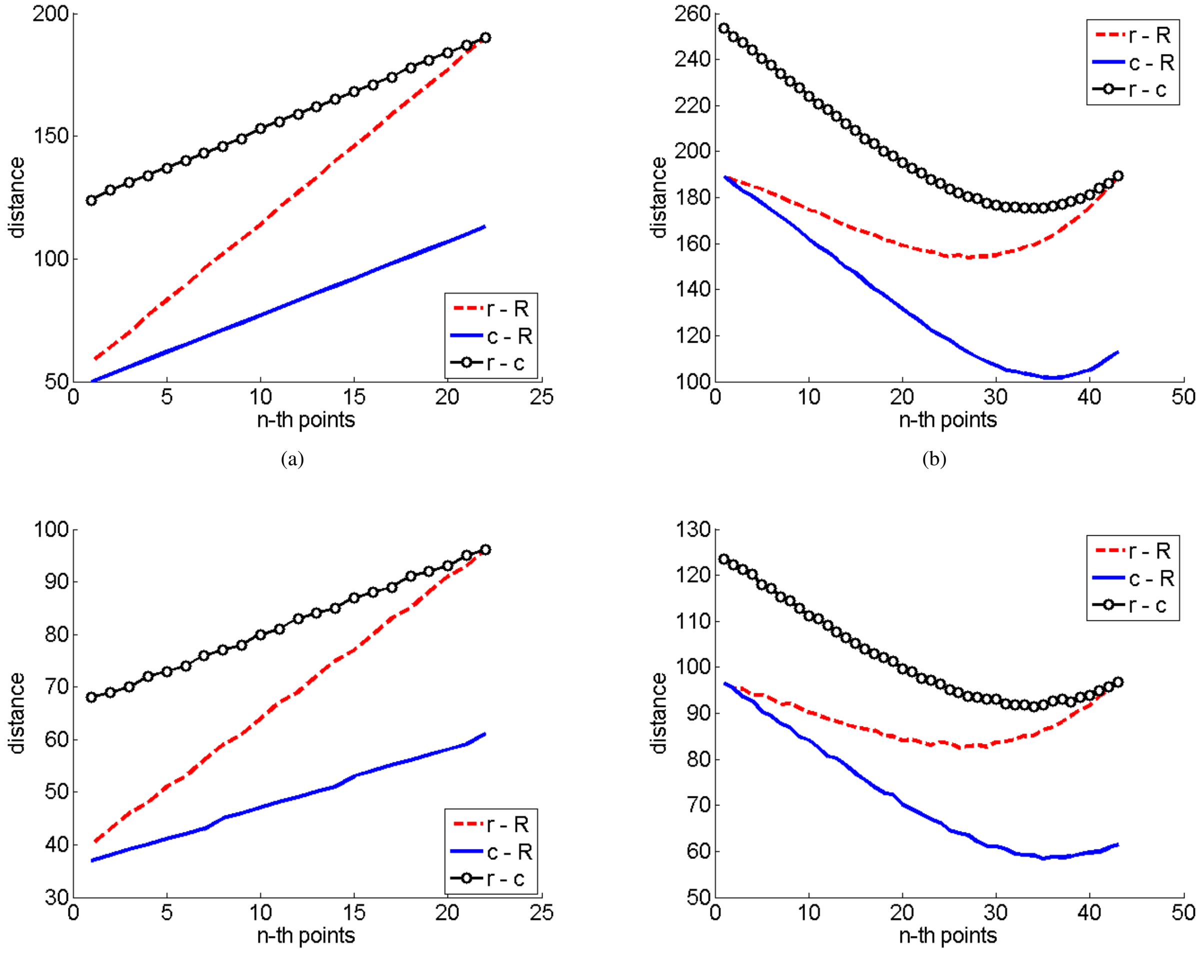
[6] K. Roy and P. Bhattacharya. Collarette area localization and asymmetrical support vector machines for efficient iris recognition. In *Proceedings of the 14th International Conference on Image Analysis and Processing (ICIAP 2007)*, pages 3–8, 2007.

[7] Z. F He, T. N. Tan, Z. A. Sun, and X. C. Qiu. Toward accurate and fast iris segmentation for iris biometrics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(9):1670–1684, 2009.

[8] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 1 edition, 2007.

[9] H. Proena and L. A. Alexandre. UBIRIS: A noisy iris image database. In *Proceedings of the 13th International Conference on Image Analysis and Processing (ICIAP 2005)*, volume LNCS 3617, pages 970–977, 2005.

[10] P. J Phillips, K. W Bowyer, and P. J Flynn. Comments on the CASIA version 1.0 iris data set. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(10):1869–1870, 2007.



(a)

(b)

(c)

(d)

Figure 8. The experimental performance in UBIRIS.v1 Sessao and CASIA.v1 with different iris localization methods and normalization algorithms. (a)(b) UBIRIS.v1 Sessao 1, (c)(d)CASIA.v1

Development of Teacher Competencies in a New Learning Environment in Higher Education

András Benedek
Department of Technical Education
Budapest University of Technology and Economics
Budapest, Hungary
e-mail: benedek.a@eik.bme.hu

György Molnár
Department of Technical Education
Budapest University of Technology and Economics
Budapest, Hungary
e-mail: molnargy@eik.bme.hu

***Abstract*— The development described in this paper aims to form a service-oriented teaching-learning model supporting the modernization of education, which can be the pillar of a gradually established university specific Competence Centre. The goal of the programmes of the Competence Centre conceived last year and being developed is to ensure that university lecturers and teachers, as well as research workers can fulfil the demands of an extending international/European higher education learning environment. The competence development necessary for 21st century university lecturers focuses on such key areas as Lifelong Learning approach and Information and Communications Technology literacy. The authors present the practical realization of the above mentioned development at Budapest University of Technology and Economics following the emergence of a concrete demand. Specifically, the service-oriented organizational and operational frameworks of teaching-learning were established in a way that allows the actual use per organizational unit or respectively person, the success of the participants as well as the satisfaction of those initiating the training to be determined. The main conclusion is that the results of the e-learning-based training programs developed within the framework of the research are a good pedagogical indicator of the transformation process when the further training of teachers in an institution of higher education of significant traditions, an „up-to-date training of trainers" is being developed, and based on the experiences the process is getting institutionalized. The concrete use of the research is the development of the trainers' professional competencies, attitude formation and efficient knowledge transfer. The innovation frameworks necessary for the development of the lecturers' competences were provided by the learning and development environments of the Institute of Applied Pedagogy and Psychology as well as the e-learning material developers of the co-department.**

Keywords- Competence; e-learning support systems; Moodle system; teacher training.

I. INTRODUCTION

The specific, final form of the Technology Transfer Project as a training program was initially proposed as a concept of a training framework within the process of modernizing education, characterized as follows:

It was recommended to develop of an in-house training system relating to the BME (Budapest University of Technology and Economics) Research University Program, with the emphasis on its horizontal elements (human resources development, support for the recruitment of teaching staff, raising standards of educational achievement), which would provide a suitably flexible learning framework for innovatively minded teachers.

The innovational background was to be provided by the Department of Technical Education (MPT) and the Centre for Learning Innovation and Adult Learning of the GTK Institute of Applied *Pedagogy* and Psychology (APPI), as well as by the current "Trainer Training" TAMOP project, while the Institute of Continuing Engineering Education (MTI) would provide the infrastructural organizational background at university-wide level.

The institute's specialised e-learning environment development server also has such software packages as Wimba Creator and Adobe eLearning Suite Extensions, along with electronic learning environments which vary across the training forms and programmes, while providing similar services (Moodle- Modular Object - Oriented Dynamic Learning Environment, Mahara [11], LimeSurvey [12]). Progressive methodological developments (the introduction of on-line forums and online examinations) are realized in the Institute of Applied *Pedagogy* and Psychology framework in conjunction with the operation of the Moodle e-learning system. The experience thus accumulated is important for implementing and broadening the range of opportunities for autonomous learning and for ICT supported teaching applicable in teacher's own practice. The Institute of Applied *Pedagogy* and Psychology is at the forefront of current practice in education innovation and open, distance learning and information and communication technology assisted learning (e-learning), as well as other innovative training employing unconventional methods [1]. In developing a training program which meets these requirements, the internationally successful Web 2.0 model, was found to provide a good fit, as depicted in the jigsaw below. It is obvious that the web 2.0 content management is provided by a special content creation tool (CCT) in addition to the content management systems (CMS). A student management system (SMS) is also applied to help student activity, while a financial-accounting

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-202-8

system (AS) has been formed to manage administration and statistical tasks.

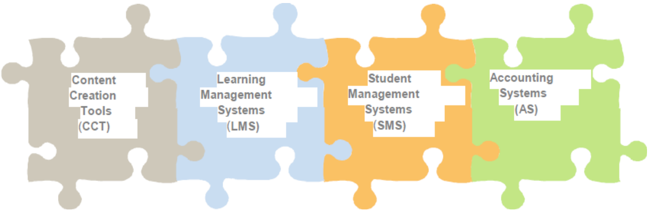

Content Creation Tools (CCT)
Learning Management Systems (LMS)
Student Management Systems (SMS)
Accounting Systems (AS)

Figure 1. Web-Education system in Europe, Source [8]

II. BODY OF PAPER

A. Antecedents

The aim of the development was to establish a service oriented teaching-learning model, furthering the process of the modernization of education, upon which a BME-specific Competence Centre could gradually be built – responding to the needs of departments, institutes and faculties. The objective of the programs offered by the Competence Centre is for the university's teachers and researchers to be able to meet the challenges arising from expansion into the international, European Higher education sector. The training covers the following major fields of competence required to function as university teachers in the 21st Century:

- Lifelong learning
- Teaching in a modular system
- Developing student centred methodologies and materials
- The use of alternative teaching methodologies
- Effective learning, and applying presentation techniques which support individual study
- The use of collaborative and networked study methods
- The application of suitable styles of verbal and non-verbal communication
- Using modern presentation techniques
- Familiarity with methodological opportunities afforded by ICT and incorporating them into the teaching process
- The development and application of new evaluation methods
- The application of ICT based information and knowledge management systems
- The practical application of modern forms of educational organization

In the medium term, the basic system of the in-house continuing vocational training recognized at university-wide level is a flexible 100 hour program, adaptable to the needs of the individual, which current plans indicate is to be reviewed and revised at 5 year intervals. A third of the program hours (30-35) are contact hours of theoretical and practical work in small groups, assisted by an electronic learning environment (Moodle), while the other two thirds of the course (65-70 hours) are to be realised individual preparation and professionally supported consultation. The Moodle [3] system to be used – which is used by more and more university departments – is according to international sources can be applied to "blended learning" type support for students and teachers. Currently, within our institute, with nearly 50 courses and 10000 users we have several years of experience in supporting online training, combined tutoring, online examinations and consultations, operating forum systems and E-portfolio applications.

The integration of the developed e-learning course modules into the Moodle system did not imply either content, or methodology, or technology problems. The Moodle is ready for the management of the course materials, so it is developed in accordance with the methodological structure of e-learning. The course materials were prepared in SCORM format, making possible the integration of the video elements as well.

The service oriented teaching-learning organizational and operational frameworks were designed to allow the monitoring at the level of organizational units and of individuals of course participants' usage levels and achievements, and the degree to which the courses satisfied their needs. The evaluation and financing models of the development take this as their basis. Any given time period – over a whole academic year – can be employed flexibly, and, depending on the subject chosen, its nature and complexity may comprise 5-20 contact hours. The actual achievement of course objectives per academic year, measured against an internal credit system is recognised by the issue of certificates. Course participation alone is not a sufficient criterion – students' progress is also relevant, as is the measurement and evaluation of their acquisition of the target competences, along with effective support for the teacher's self-development.

B. The process of electronic teaching materials development

As an atypical form of teaching, e-learning requires that teaching materials go beyond the traditional digitalised or digitally authored materials, and that they be authored in an e-learning environment. In the course of preparing these electronic materials many factors play a role, which do not arise when creating simple written documents. Electronic teaching materials contain multimedia elements (e.g., spoken narration, animation, simulation tasks, highlighted notes etc.), which are difficult to express in writing alone.

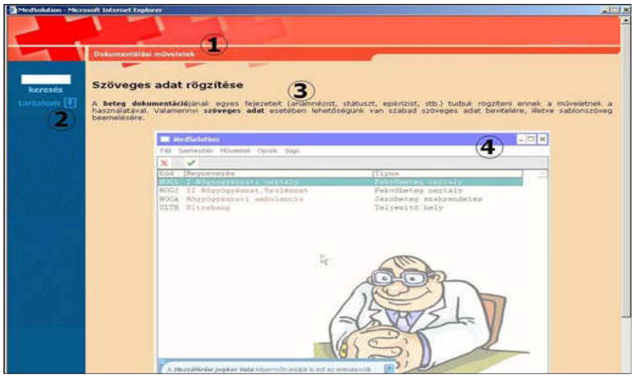
Furthermore, in many cases electronic teaching materials are produced by personnel who are not in fact competent in the details of the subject matter presented. A major problem arising from this is that of how a specialist in a particular field – the content provider – can convey to the personnel who are producing the materials – the materials developers - what should feature in the teaching materials. The creation of electronic teaching materials – due to the factors outlined above – requires an effective quality control process:

Firstly, the materials developers show the *content authors* what possibilities are available in electronic teaching materials, and give them a written description of the development environment. For this, it is necessary to use the appropriate terminology and review the range of available options. Following this they jointly write the so-called script (synopsis) of the teaching materials, on the basis of which the materials are produced. The most important attribute of the script is clarity; that is to say, the material developer should know exactly where the materials will be used, and what will feature in them. The next stage is the process of checking and editing, professionally, in terms of content, and linguistically. It is important that even at this stage the script be precisely written enough that the editors will know, even without having produced electronic materials, what will feature in the materials and how. The professional and linguistic edits are then reviewed by a further editor, who accepts or rejects them as appropriate. Next, the rough materials are produced, which comprise the:

- Written texts,
- Images,
- Studio recordings,
- Video materials and animated sequences required for the teaching materials.

The best solution is for the rough materials, or a large proportion of them, to be prepared by the content authors / script writers themselves, or for them to supervise this preparation. This also greatly aids the work of editors and proofreaders, and there is less chance of the type of mistakes which arise when materials are being developed. During the next stages of the process one of the most notable tasks is tackled, pedagogically speaking, that of curriculum development. In the last stage in the process, the electronic material is subjected to the further editing and testing, to determine how much it conforms to the script.

It can be seen in this process that the key to materials development is an appropriate script. The script has to be in all regards exact and unambiguous, especially as its narrator may in many cases have little idea of what he is saying. He will thus be scarcely able to correct or supplement any unclear or questionable material. The diagram below gives a general overview of electronic teaching materials. This kind of overview is vital, to make the content developer aware of the possibilities and limitations of electronic teaching materials.


MedSolution - Microsoft Internet Explorer

Szöveges adat rögzítése

A **beteg dokumentáció**jának egyes fejezeteit (anamnézist, státuszt, epikrízist, stb.) tudjuk rögzíteni ennek a műveletnek a használatával. Valamennyi **szöveges adat** esetében lehetőségünk van szabad szöveges adat bevitelére, illetve sablonszöveg beemelésére.

Kód	Megnevezés	Típus
NOG1	I. Nőgyógyászati osztály	Fekvőbeteg osztály
NOG2	II. Nőgyógyászat, Szülészet	Fekvőbeteg osztály
NOGA	Nőgyógyászati ambulancia	Járóbeteg szakrendelés
ULTR	Ultrahang	Teljesítő hely

Figure 2. Electronic Study Material, Source: Internal teaching resource

Once the electronic teaching materials have been uploaded the lesson screen appears in a new window, which is composed of several parts:

- Frame, other lesson-related information (1.)
- Notes on content (2.)
- Text (theoretical) information (3.)
- Video window (flash animation, interactive task(s)) (4.).

C. *Real world realisation*

The strategic aims of the program were achieved in several developmental phases. The first phase, in mid-2011 (April-August) the topic areas of the Technology Transfer Project – with the agreement of the participants – were set out in brief training modules (maximum 10 hours). These courses partly consisted of contact hours with full or partial e-learning support (for individual preparatory work and practice). 6 of the 11 modules present and order the general characteristics of innovation, which then link into a module on product development. The technology development projects being pursued at the BME (Budapest University of Technology and Economics) serve as case studies of innovation in the other 4 modules demonstrating how innovation can be supported by knowledge transfer. The modules focus on the application of the results of innovation and the establishment and development of suitable business models for practical application of innovation. The topics of e-learning modules actually developed were:

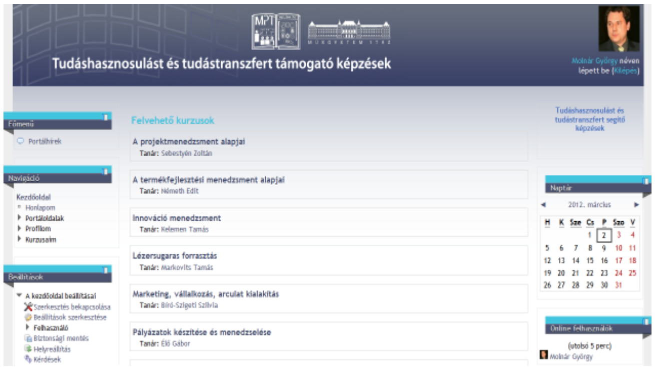
- Fundamentals of Project Management
- Starting businesses, spin-off firms, Raising capital, Economic Analysis of investments
- Syntactic metal foams - metal matrix composites reinforced by special particles
- Laser soldering
- Fundamentals of Product Development Management
- Intellectual Property Rights, industrial copyright, patents
- Preparing and managing tenders/applications
- Innovation management
- Marketing, business, image building

The specific tasks were devised in the following process:

1. Finalizing the task specification and clarifying its financial aspects
2. Selection and recruitment of technical writers/course advisors collaborating on the project
3. Skills training for technical authors
4. Collection of professional materials necessary for E-learning material development
5. Creation of E-learning modes: early summer pilot stage (4 modules)
6. Moodle framework development and downloading of pilot course materials
7. Briefing course advisors on on-line consultations and tests

8. Announcing and initiating pilot courses
9. Uploading the complete course materials onto Moodle and commencing the training – the final test and validation of the teaching materials
10. A brief (6-8 minute) video (in VCAM format) was prepared for each module, functioning as a freestanding lesson, supplementing the written materials, which together make up an individual work (downloadable, executable), which is embedded in the e-learning materials.
11. The supporting framework for the courses was completed (Moodle) [4]
12. . The teaching materials were uploaded onto this, connected to the Google Analytics statistical services, which allows the courses to be monitored. The visual interface: http://visibleexpression.co.uk/mythemes/aardvark_makeover.zip
13. Information relating to the course applicant, and the relevant application form was placed on the electronic learning environment delegated for educational support of the project, which can then be used to help draw up an adult education contract.

An average of 2 A0 sheets worth of “course materials” are prepared per module, and these e-learning study materials, which are suitable for individual study, are accessible to registered users in an inclusive manner on the electronic framework, along with 10-20 related test questions per module. Course participants and teachers alike will have access to Forums for the duration of each module, which are moderated by the technical author and provide opportunities to conduct consultations and to exchange experience. This links in with the course ending, examining/knowledge checking process, allowing for evaluation of the course and statistical analysis. Co-ordinating activities are included in the above processes, as are monitoring and evaluation (quality control), along with administration required for the course (course registration, approval, signing contracts, certificates of achievement, collating statistics). The figure below shows the opening page of the Moodle system which facilitated this e-learning.



Tudáshasznosulást és tudástranszfert támogató képzések

Molnár György néven lépett be (Kilépés)

Felvehető kurzusok

A projektmenedzsment alapjai
Tanár: Sebestyén Zoltán

A termékfejlesztési menedzsment alapjai
Tanár: Németh Edit

Innováció menedzsment
Tanár: Kelemen Tamás

Lézersugaras forrasztás
Tanár: Markovits Tamás

Marketing, vállalkozás, arculat kialakítás
Tanár: Bíró-Szigeti Szilvia

Pályázatok készítése és menedzselése
Tanár: Élő Gábor

Tudáshasznosulást és tudástranszfert segítő képzések

Naptár

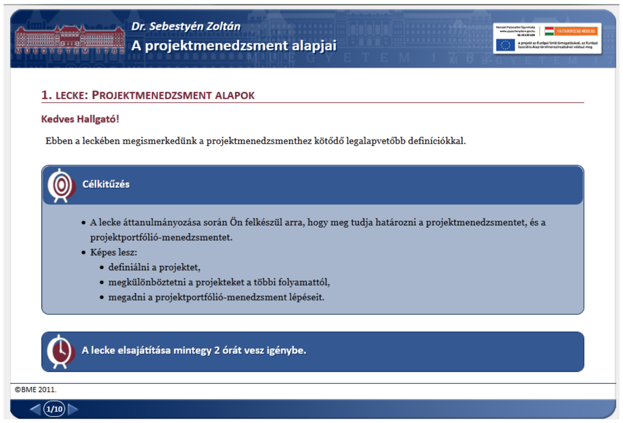
2012. március

H	K	Sze	Cs	P	Szo	V
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

Online felhasználók
(utolsó 5 perc)
Molnár György

Figure 3. Opening page of the Moodle LMS of the Technology Transfer Development, Source: own photo

The study pages of the individual modules can be found under the course category headings, and within them the lessons. The screenshot below shows the contents of one lesson.



Dr. Sebestyén Zoltán
A projektmenedzsment alapjai

1. LECKE: PROJEKTMENEDZSMENT ALAPOK

Kedves Hallgató!

Ebben a leckében megismerkedünk a projektmenedzsmenthez kötődő legalapvetőbb definíciókkal.

Célkitűzés

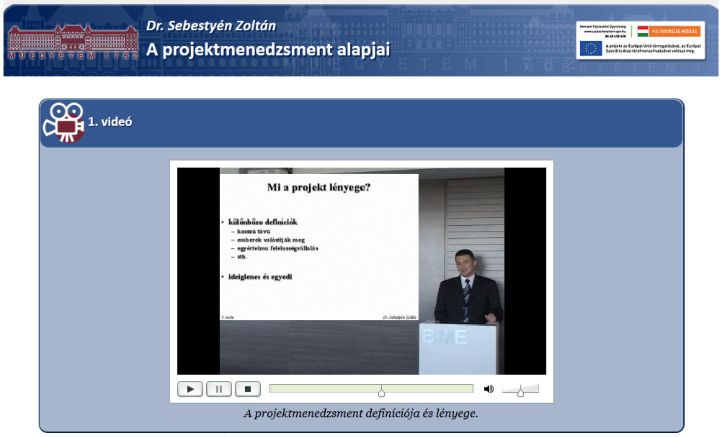
- A lecke áttanulmányozása során Ön felkészül arra, hogy meg tudja határozni a projektmenedzsmentet, és a projektportfólió-menedzsmentet.
- Képes lesz:
  - definiálni a projektet,
  - megkülönböztetni a projekteket a többi folyamattól,
  - megadni a projektportfólió-menedzsment lépéseit.

A lecke elsajátítása mintegy 2 órát vesz igénybe.

©BME 2011.

Figure 4. First lesson of the course of the Moodle system, Source: own photo

The following screenshot shows the video recordings embedded in a lesson, which the student can stop or replay at will.



Dr. Sebestyén Zoltán
A projektmenedzsment alapjai

1. videó

Mi a projekt lényege?

- különböző definíciók
  - hosszú távú
  - emberek valósítják meg
  - egyértelmu felelosségvállalás
  - stb.
- ideiglenes és egyedi

A projektmenedzsment definíciója és lényege.

Figure 5. Opening page of the Moodle LMS of the Department, Source: own photo

III. CONCLUSION

Students who have successfully completed any one of the 9 e-learning-based training modules gain skills which will afford them greater mobility and measurable knowledge, as well as in future years supporting the Lifelong Learning concept. 1466 out of 1965 students successfully completed the selected module. A notable outcome of the programmes was the co-operation between teachers from different faculties and the effort the put in to develop an effective way of conveying knowledge. The participants were able to select the most appropriate training programme for their particular competency set, and to immediately apply the knowledge and practical experience thus gained in their teaching practice. The success of the training programmes of this project is also evinced, from a pedagogical standpoint by the process of transformation in which a higher education institute with long traditions is moulded by the further professional training of its academic staff, and on the principle that experience is systemized suggests that this may lead to this process becoming institutionalised. Beyond allowing teaching materials to be downloaded, the Moodle system also facilitates communication between students and teachers, measuring of their activity levels, the functioning of public forums, and implementation of monitoring and evaluation in an electronic environment. The patterns of usage of the system are shown in the following two diagrams, from November 2011 – February 2012, which clearly shows, for instance, the peak date of online test completion on December 12 2011. It can also be established on the basis of Google analytics statistics, that on December 5 2011, 1023 people visited the courses' study pages to study the online study materials contained therein, preparing for the final exams.

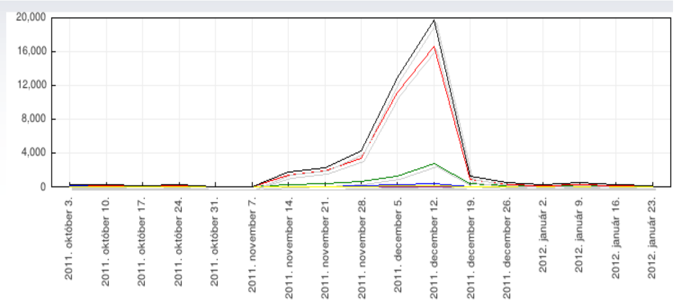
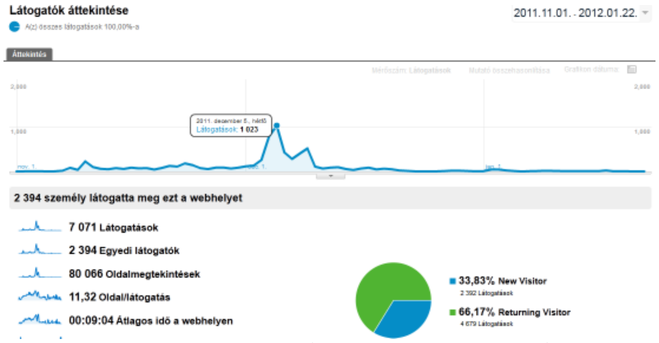


Figure 6. Student and teacher activities of Technology Transfer course, Source: own photo



2 394 személy látogatta meg ezt a webhelyet

7 071 Látogatások

2 394 Egyedi látogatók

80 066 Oldalmegtekintések

11,32 Oldal/látogatás

00:09:04 Átlagos idő a webhelyen

33,83% New Visitor
2 392 Látogatások

66,17% Returning Visitor
4 679 Látogatások

Figure 7. Statistical indicators of TT modules with the aid of Google analytics.

REFERENCES

[1] András Benedek, Multimedia Sensors in Learning by Mobile Communication. In: Alan Tait, András Szűcs (ed.): EDEN Annual Conference: Media Inspirations for Learning. What makes the impact? 9-12 June 2010, Valencia. Book of Abstract, ISBN 978-963-06-9429-2, European Distance and E-learning Network, 2010, pp. 60-65

[2] György Molnár, The requirements and development areas of the ICT aided learning environment. In: Dr. András Benedek (ed.): Digital pedagogy – Typotext Budapest 2008, pp. 225-255

[3] András Benedek and György Molnár, The empirical analysis of a Web 2.0-based learning platform, In: Constantin Paleologu, Constandinos Mavromoustakis, Marius Minea (ed.): ICCGI 2011, The Sixth International Multi-Conference on Computing in the Global Information Technology, Luxembourg, June 19-24, 2011., ISBN: 978-1-61208-008-6, pp. 56-62, retrieved: 05.2012

[4] http://ttp.moodle.appi.bme.hu, retrieved: 03.2012

[5] András Benedek, György Molnár and János Horváth Cz., Moodle-based E-portfolio used in teacher training. (In: Sixth EDEN Research Workshop, User Generated Content Assessment in Learning. Enhancing Transparency and Quality of Peer Production. Emerging Educational Technologies and Digital Assessment Methods, 24-27 October, 2010, Budapest, Hungary. Edited by Morten Flate Paulsen and András Szűcs, Budapest, 2010. ISBN 978-963-87914-4-3, pp. 131-133

[6] András Benedek, György Molnár and János Horváth Cz., Jump over the shadow? From the traditional education to the non-typical one or the experiences of an electronic learning environment. IADIS Multi Conference on Computer Science and Information Systems, Freiburg, Germany, 26-29 July, Proceedings of the IADIS International Conference e Learning, Volume II (Ed. Miguel Baptista Nunes, Maggie McPherson) ISBN:978-972-8939-17-5, pp. 243-246

[7] György Molnár: Flashes or steady light? Or the potentials of developing networked learning, In: Miguel Baptista Nunes, Maggie McPherson (ed.): Proceedings of the IADIS International Conference e Learning, IADIS international conference E-learning 2011, Volume II. Rome, Italy, July 20-23, 2011, ISBN: 978-972-8939-38-0, pp. 405-408

[8] Morten Flate Paulsen, Online Education Systems: Definition of Terms, In: Web-Education Systems in Eu, ZIFF Papiere 118, FernUniversität Hagen, Okt. 2002, pp. 23-28, retrieved: 03.2012

[9] M.A. McPherson and J.M. Nunes, "Critical Issues for e Learning Delivery: what may seem obvious is not always put into practice". JCAL, 24(6), 2008, pp. 433-445, retrieved: 03.2012

[10] Piet Kommers, ICT as explicit factor in the evolution of life-long learning. International journal of continuing engineering education and life-long learning, 20(1/2010), pp. 127-144., retrieved: 03.2012

[11] http://e-portfolio.appi.bme.hu/, retrieved: 04.2012

[12] http://appi.bme.hu/survey/admin/admin.php, 05.2012

Technology Powered Learning at Academic Institutions

Personal Credit Portfolio

Danuse Bauerova
Innovation in Education Institute, Department of Mathematical Methods in Economics
VSB – Technical University of Ostrava
Ostrava, the Czech Republic
danuse.bauerova@vsb.cz

***Abstract*—This article aims to describe the methodology for the realization of particular although partial steps when accepting a new paradigm of learning or other school processes. A part of the paper deals with the verification of partial methodologies of the Model of Technology Powered Learning at school. The student-centric approach and the impact of information and communication technologies on education and university society are applied. Positive digital identity of personality is an outcome of human activity while at the same time it accepts the technology of cloud computing and deep web. A new term – *Personal Credit Portfolio* – has been invented. The processes are implementable for students and for all the members of academic staff, and consequently have an impact on the credit of the whole – University.**

Keywords-Personal Learning Portfolio; Personal Credit Portfolio; Technology Powered Learning; Social Web; Internet-based Education

I. Introduction

Many research programs explored how new technology brings mass collaboration and changes the world. For example, New Paradigm team [1] has conducted several investigations to understand how the Web 2.0 changes the corporation. Collaboration and relationship has been newly shifted to the new paradigm principles.

A. Contextualization and Importance of the Theme

Web 2.0 tools do not only provide different ways of communication which should enhance learning and interaction in the virtual environment. They also offer a real opportunity to create a classroom without walls [2]. Boyd [3] claims that the social aspects of Web 2.0 have great potential for enhancing education, while many authors suggest that Web 2.0 concepts could support lifelong learning communities. The teachers´ role is to encourage eLearners to creative thinking and to stimulate them to be active users (i.e., prosumers) [4].

Learning through ePortfolio is offen offered. Hellen Barret describes Balancing the Two Faces of ePortfolios [5] and separates the process of creation (Portfolio as Workspace) and the process of final display of the product (Portfolio as Showcase). By this approach the process of positive digital identity building is solved only partially. In this article, a proposal of good personality image building by a so called Personal Credit Portfolio (PCP) is newly formulated. For the process of the PCP creation is significant continuous interaction of personalities and technologies.

Links are the new CVs, portfolios aren't just for artists anymore, and experience reigns. The most important skill we'll have in a world where 50% of people see self-employment as more secure than having a full-time job is the ability to go out and get the right knowledge for the right purpose at the right time [6].

The academic exclusivity has for centuries been based on the strictly hierarchical structure of the processes. The evaluation of students, pedagogues and scientific staff is a direct outcome of academic activities. Being closed in is the means of building exclusivity. The role of the so-called gatekeepers is integrated within the system, and the openness and sharing create strong concerns. As it can be generally seen, nowadays numerous methodologies of hierarchic evaluation of academic institutions and their individual members are being introduced.

Hierarchic teams at universities are not very efficient. The open web environment with its flat net structure of the processes with natural activities of end participants offers more. Openness and sharing are not typical of teams within universities and thus the effective connection to the outside world based on the continuous feedback is missing. There is still a tendency to adhere to publishing the results of pre-reviewed scientific research which does not correspond to the speed of development. The natural characteristics of the open web environment concerning the criteria of the evaluation with continuous feedback are missing. Any growth is thus supported with intolerable delay.

The openness of the Internet is a threat. Especially worried are those whose outputs are not shared and naturally used by professionals. It is not always academics who refuse changes; sometimes the same can be said about students. The newly created net connection brings development but also the necessity of continuous work. However, students are not always willing to make an effort because it is the evaluation which is considered the main product of their work not the real personality development. Some students admit that they are studying only to get a degree. Continuous shared

ISBN: 978-1-61208-202-8

collaborative work, for which the activity of each member is a necessity, seems to be a useless complication on their way towards a degree. In this way, the core of the interest lies not in the real personal development but in the achievement of an academic title.

The theme is important and relevant as a fight for survival of a hierarchic structure against the danger of the flattening of the world by net structure is a strong characteristic of the present academic life. New paradigms cause disruption and uncertainty, even calamity, and are nearly always received with coolness, hostility, or worse. Vested interests fight against the change, and leaders of the old are often the last to embrace the new. Consequently, a paradigm shift typically causes a crisis of leadership [7].

The openness of the Internet is a threat for the hierarchic structure and as a consequence it leads to the refusal of new technologies by those whose good reputation (credit) is supported by the hierarchic evaluation. Pundit Andrew Keen issues a long bleat claiming that the world is awash with drivel because it is so easy to propagate ideas on the Internet. Keen pines for the good old days—a time when apparently only smart and credible people had access to printing presses and the airwaves. A healthy society needs gatekeepers, he argues.

Throughout history academic institutions have organized themselves according to strict hierarchical systems of authority. At the time of the Internet it does not seem to be effective to report on the strength of an individual or an academic institution through various hierarchical systems of evaluation. Today, you advance in the world based on your performance, not a piece of paper declaring your expertise in "knowing a little about a lot of things" [6]. The good reputation (credit) of universities, individual students, pedagogues, and scientists could be developed through openness. Sharing outputs brings along a natural evaluation of quality. The flattening of the evaluation processes towards the net structure with active participation of end users is beneficial.

Individuals and institutions reject these changes. No clear examples of such change implementation have been observed in the surrounding environment. There has not been a methodology put into practice yet.

B. *Research question and Objectives*

This article aims to describe the methodology for the realization of particular although partial steps when accepting a new paradigm of learning or other school processes. A part of the paper deals with the verification of partial methodologies of the Model of Technology Powered Learning at school. Is the end-user approach applicable on education and university society? Would a Positive Digital Identity of personality be an outcome of human activity and the technology of cloud computing and deep web at the same time? How can end-users contribute to their personal identity (whether they are students, pedagogues or scientists) and further to identity and prestige development of the whole institution within the process? Are the students able to create systematically their Personal Learning Portfolios and perceive a power of these outputs when creating their digital identity? Could be realized processes generalized?

The aim is to use the results of long term qualitative and quantitative analysis to formulize generally valid methodology. The outputs are monitored when transferring them from Learning Management System usage to freely accessible web applications. The subject of interest a monitoring of implementation effects of Technology Powered Learning is. The research is in process when educating thousands of students of academic institution within more than 10 years. The outputs are generalized in proposed methodology to be transferred to Technology Powered Learning Model.

A new term – *Personal Credit Portfolio* – has been invented.

C. *Structure of the Paper*

In the introduction the context and rationale of using different Web 2.0 tools in teaching and learning are mentioned.

Section II presents the results of examples study. Facts and methodologies of successful world universities are reminded; these universities use an openness and power of Internet more effectively than universities in the Central Europe do. Section is involved to strengthen previous argumentations in favor of research, hypotheses formulation and methodology of innovations implementations to the environment of the Central Europe universities.

Section III involves the results of research being provided at the university to argumentation for the change and proposes the procedures. Here the results of long term research in usage of Learning Management System (LMS) Moodle to support academic education are stated. Methodology of the transition from LMS Moodle to the open access web applications and creation of Personal Learning Networks is described. The way of using the power of social networking when implementing the model of Learning Powered by Technology is shown.

Section IV brings new knowledge and implements a new term. It is linked to previous contributions, which separate two parts of ePortfolio creation – one as the process and second one as the product. The new implemented term Personal Credit Portfolio points out inseparability of both parts. Building of positive identity is presented as continual process. Participants here is not only the man but hidden technology processes as well. Experiences are used accepting the fact that Every work starts with the potential to be a Thought Paper that might be read by millions. I am responsible for building my resume or portfolio since my first day in college. [8].

The final section presents a concluding discussion of the findings of the study and university reality from different angels. The pros and cons of different Web 2.0 tools at universities and the synergy effects of the integration of the selected tools to university processes are also examined. Discussion about the future direction of the Web 2.0 application at the university environment is open.

II. SUCCESSFUL UNIVERSITIES ARE UNIQUE BY THEIR OPENNESS

Universities as well as their academics strive for the best in their field. The danger consists in persuading ourselves that we can get among the successful ones through the currently used methods of evaluation. The winners are those who can accept the fact that there is no need to develop the processes and criteria applicable in the times before the Internet. The challenge of today's world is to overcome the existing paradigm in the most direct way. The focus of the new evaluation methodology lies in openness and sharing, which leads to a natural evaluation based on relevant feedback.

Openness and sharing demonstrably increase the effectiveness of the present processes. Academic processes are no exception. Openness and sharing are what increases the personal and the academic institutions´ credit, and as such they are necessary to be included among the relevant evaluation criteria. The growth of the credit refers to:

- student´s, graduate's and teacher´s personalities but only in case that it is the "learning" not the "testing" which forms the center of the pedagogical processes (personality´s credit is built on continuous learning not on evaluation by testing),
- research and science workers but only if the Internet is used for the open scientific cooperation with shared outputs and for continuous feedback,
- administrative workers but only if the Internet is used for the shared collection and continuous generating of customized data needed for academic processes,
- institutions consisting of students, graduates, pedagogues, and scientific and administrative staff who have relevant knowledge and who use the current power of the Internet as described above.

III. SOCIAL NETWORKING AND THE MODEL OF LEARNING POWERED BY TECHNOLOGY

The present state of eLearning at the Central European universities is based mostly on the pre-internet time processes. Old processes are only performed with new technology usage but they are not powered by technology.

The power of new Internet Web 2.0 is not fully utilized, even though there has been positive feedback from end users/students. The focus of the present approach is a pedagogue not the social networking of learners. So far there are not accepted activities and processes of the Model of learning powered by technology.

A. Out of the classroom, out of the Learning Management Systems

To increase competitiveness of labor force in global knowledge-based economy it is necessary to accept new paradigm of education. The future of learning needs to move out of the box – whether it is the little red school house or the course management system. We need a new way of thinking about teaching and learning that put not just the learner, but learning at the center [8].

B. Methodology of the transition from LMS Moodle to the open access web applications

The key word for all Learning Management Systems (further LMSs) such as Moodle is the management. Students' work managed by the teacher is a clear indication of the hierarchic structure of activities. Within the last ten years, the team of the Innovation in Education Institute at the Technical University Ostrava has conducted large group (tens of thousands of students) investigations to understand how the Learning Management System works. It can be said that LMSs play de facto 3 functions as you can see in Figure 1.

1. Making the content accessible (static presentation of the text or multimedia study materials created by the teacher or borrowed ones).
2. Administration of students´ work such as reporting, evaluation, archiving etc.
3. Management of students' activities, favorably their *guiding* and the overall *communication* management.

Ad 1. The first part – static presentation of *content* continuously decreases its imperfections. Such created materials exceed the LMS, namely to the open access web applications.

- During the first step forward the files (although multimedia ones but still files) are *uploaded to web outside LMSs*. Their characteristic feature is more flexible accessibility, and they can be used for further courses, mostly done by the same teacher.
- Later on the content is to be designed through the open access applications, i. e. directly in web space. Content created in GoogleDocs can serve as the example but also wiki, blogs etc. are used. There need to be said that the present LMSs, e.g., Moodle version 2.x, support the creating of some hybrids, i.e., they accept the outputs of the above mentioned web applications and they can implement them into their management systems.

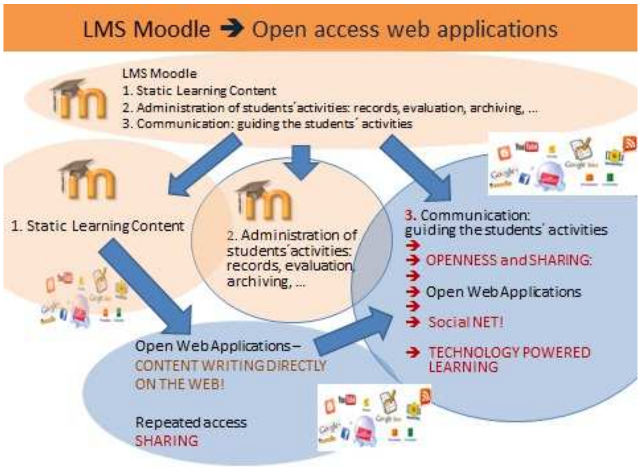

LMS Moodle ➔ Open access web applications
LMS Moodle
1. Static Learning Content
2. Administration of students´activities: records, evaluation, archiving, ...
3. Communication: guiding the students´ activities
1. Static Learning Content
2. Administration of students´activities: records, evaluation, archiving, ...
3. Communication: guiding the students´ activities
➔ OPENNESS and SHARING:
➔ Open Web Applications
➔ Social NET!
➔ TECHNOLOGY POWERED LEARNING
Open Web Applications – CONTENT WRITING DIRECTLY ON THE WEB!
Repeated access SHARING

Figure 1. LMS is transferring benefitting from open web applications.

Ad 2. Gradually, it appears that the right role of LMS lies in the second group of the above mentioned functions, i.e., the *administration* of students ‘work. Reporting, testing, tasks evaluation, and the archiving of the students´ performance can be done in LMS effectively, probably even more effectively than anywhere else. The strengthening of testing and evaluation is not in accordance with the concept of a new lifelong learning model on one hand; however, the current academic environment cannot in fact operate without them. Let us leave them then inside the LMS stating that in this way they are effective, even though their effectiveness for the Model Powered by Technology is more than doubtful.

Ad 3. The creation of the learning content by open access web applications meets with the third part of the above mentioned activities of eLearning in LMSs, i.e., the students‘ work guiding (Figure 1). It is more effective to *communicate* with the students outside LMS, i.e., in other applications designated to it. It is the s*ocial web* which is used. The guidance by a teacher is lowered and the communication among students increases. In other words, the *mutual continuous connection* among learners is much more stimulating in a process of a creation of collective intelligence. Pedagogue is one of the nodes, only possibly distinguished by the coaching aspects. When the network is set up, Model of Learning Powered by Technology begins. Lifelong Learning goes on in a continuous mutual interaction.

IV. Personal Learning Portfolio and Personal Credit Portfolio

The LMSs provided a low level entry on the web. In the new model a course exists outside of the student’s own learning space. A course is one of many hubs of designed and facilitated learning objects and activities, and it links to an established social network of students and professionals exploring the same content and activities. A course becomes a hub of activities facilitated by the expert – the faculty member. But it remains outside of the traditional course space, in students’ own ePortfolios, which transcend the limited space and time that any single course can be expected to provide [8].

Personal Learning Networks are a great way to widen knowledge and learning beyond won [9]. PLNs are created by the individual learners. To extend relevant connection of those who are learning with those of the same interest and knowledge can fulfill specifics of their needs. PLNs provide with an access to significant personalities and experts worldwide. Those form the communities around them, to which others have access. Everyone has opportunity to get the sources and knowledge, which would be inaccessible behind the school walls. Individual learners create their personal learning network as in Figure 2 you can see [10]. Personal Credit Portfolios as social networking products are created continuously. Students are encouraged to make their learning visible. Activities in the course of study subjects are the part of Personal (Learning or Credit, resp.) Portfolios are accessible any time. Accountability and employment of the course participants including teacher is continuously increasing.

As in [5], social networking and personal portfolios creation are used in education:

1. for personal knowledge improvement (Personal *Learning* Portfolio), i. e.
- to store outputs and share experiences,
- to reflect on learning,
- to take feedback for improvement,
2. to showcase achievements (Personal *Credit* Portfolio), i. e.
- to showcase achievements and accomplishments and to facilitate accountability and employment searches.

Both parts act in mutual interaction and *are not dividable*. The more the man would try to distinguish the “working” parts (Learning Portfolio) and „showcases“ (Credit Portfolio) the more work he would have (and never would finish). He would still have to keep in mind unexpected “tricks” of never ending cleaning, setting and structuring the outputs being implemented by the tools of deep web and Cloud Computing. Cleaning is pointless work – we all know that HR officer can always find something what have not been showed in „Credit“ Portfolio by applicant. On the contrary, an applicant makes effort to hide imperfect information.

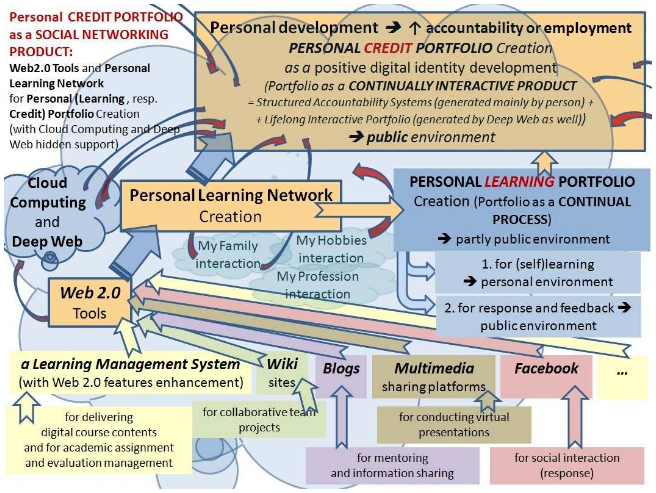

Personal CREDIT PORTFOLIO as a SOCIAL NETWORKING PRODUCT:
Web2.0 Tools and Personal Learning Network for Personal (Learning , resp. Credit) Portfolio Creation (with Cloud Computing and Deep Web hidden support)
Personal development ➔ ↑ accountability or employment
PERSONAL CREDIT PORTFOLIO Creation
as a positive digital identity development
(Portfolio as a CONTINUALLY INTERACTIVE PRODUCT
= Structured Accountability Systems (generated mainly by person) +
+ Lifelong Interactive Portfolio (generated by Deep Web as well))
➔ public environment
Cloud Computing and Deep Web
Personal Learning Network Creation
PERSONAL LEARNING PORTFOLIO Creation (Portfolio as a CONTINUAL PROCESS)
➔ partly public environment
My Family interaction
My Hobbies interaction
My Profession interaction
1. for (self)learning ➔ personal environment
2. for response and feedback ➔ public environment
Web 2.0 Tools
a Learning Management System (with Web 2.0 features enhancement)
Wiki sites
Blogs
Multimedia sharing platforms
Facebook
...
for delivering digital course contents and for academic assignment and evaluation management
for collaborative team projects
for conducting virtual presentations
for mentoring and information sharing
for social interaction (response)

Figure 2. Personal Credit Portfolio [9].

We have to consider the fact that by each click on web we create „*Lifelong Interactive Portfolio*“! In this way hardly controlled *Mush-up* such as reports on Facebook, Blogs, wikis, Twitter, Ning, YouTube, Flickr, Picasa etc. is created. Students (as well as pedagogues and scientists) have to be awarded that *each their trace can be seen by millions of Internet users*! They contribute to their Personal Credit Portfolio by all work they do on the Internet.

A. Literacy to PLN creation

The man has to be equipped with literacy, i.e., how to access the community or to build it, how to find people and sources being trusted. It is more complex process than to sit in the class; it is more independent overreaching curriculum of the subject.

New role is hard for both students and pedagogues. Teacher has to be prepared for his students´ activities, also for the danger resulting from the fact students can go anywhere and speak with anybody to fulfill their ambitions. The pedagogue is able to manage the students to active safely and effectively in a way socially demanded. Active approach of the pedagogue to educational process demands to be equipped with ability to open conversation, which is the base of continuous connection, providing and receiving feedback. Literacy to PLN creation can be acquired step by step and the advices of predecessors can be used as well.

B. *The way from PLN 1.0 to PLN 2.0*

To develop their personality students get ideas how to start with building their own PLNs. There exist many recommendations but all have one thing the same – to proceed by two phases from PLN 1.0 to PLN 2.0.

1. The best way to enter the new world is to become consumer first, i.e., to acquire the skills to work at reader level. The aim of this phase is PLN 1.0. It is the mark overtaken from the term Web 1.0, whose basis is the reading on the Internet (Read Web alias Web 1.0).
2. Then the personality's development towards 2.0 skills is able. Students and pedagogue can go towards creator role. Individual starts to be active. PLE 2.0 is analogical extension of the term Web 2.0 where it is about not only reading but also about active writing (Read-Write Web alias Web 2.0) [11].

Above mentioned two steps are applied gradually to the process of acquiring the knowledge to work with others and others web tools suitable for interaction and reflexing. Among those chronological reports, wiki for web creation, multimedia for active creation of web content, microblogging, today Twitter and social networking, today Facebook, Google+ are. Everything can be shifted to hundreds of web applications but all the time the effectiveness of such activity has to be considered.

C. *Tools integration*

We can integrate Web 2.0 tools into (university) course delivery. One of the possibilities is using widely integrated tools of Google as an environment enabling easier integration of partial tools while logged only ones. In this case it is suitable to create an account at Google and to create own iGoogle. Gradually it can be supplemented by some other selected tool beyond Google group. Entering the new and new fields can be repeated in the model from 1.0 to 2.0.

Blogging. It is good to devote the first cycle "from 1.0 to 2.0" acquiring skills to blog. The real output of it is going to be not only skill but the first part of own PLN as well.

1. To become involved in professional social network according personal interest by choosing some bloggs 5 at most being interesting for the individual as for determined aims. By RSS channel to log for receiving chosen bloggs and to watch them through Google Reader (to make it the part of personal iGoogle).
2. To manage own blog by own inspiration e.g., using Google service Blogger or Edublog or others. To believe in an opportunity to be involved in conversation and commenting bloggs. Most of authors wait for reactions of the readers and are prepared to answer them. There is no reason for fear.

Microblogging. At the same time or with small time lag the two phases' cycle can begin e.g., microblog. First to get known it by reading, later to become active in writing notes and finally to write comments to the others´ notes:

1. To become involved in microblogging by reading Tweets at Twitter. Five well know communities can be chosen to watch their content being produced and shared. In short moment when 140 symbols are read there can be learned a lot. Especially by creating the connections between earlier acquired terms. Besides the professional microblogs we can watch also public personalities such Twitter BarackObama. As a rule big surprise comes how effective such short notes can be.
2. Later to create own Twitter and be active on it. The last thing to do is to comment Twitters of others.

Google+ (Facebook). To start the work on Facebook the same scheme suits from reading (1.0) to writing and active contributing (2.0).

Wiki. Analogically can be proceeded when using other tools, e.g., wiki should not be excluded from any educational community at universities – students "cooperate" on daily basis. Wiki environment can be used advantageously to administrate web space of online course as accessory of Learning Management System. LMS is good to keep for achieving of selected outputs relevant for academic crediting. But to manage activities and cooperation there is opened web space suitable to use. This role wiki fulfills greatly.

Others. Many web application can be used when building own personal learning network. Some of them are unique; others are alternated in few modifications. By Top 100 Tools for Learning being elaborated every year the picture about them can be made.

1. Students write their own notes out of information acquired to their PLN at different spaces on the Internet.
2. They watch the work of their colleagues as for topic and discuss it.
3. By teacher's support they achieve agreement in so far work.
4. They create and publish own material for the community of study subject's online course.
5. They write comments to colleagues' materials.
6. They achieve solutions and results by cooperative work

Students develop their present Personal Network. At the university they go further and further from the closest surroundings to general audience.

1. When entering the university the students are managed to build their own portfolio in such way to be seen by their schoolmates, potential members of working teams and their teachers.
2. Later on their openness and outputs sharing should expand through grades as well as universities.

3. And finally, their portfolios become opened to the general public.

Each step of the way, their audience gets wider:

1. Students start from a small circle of trust,
2. get feedback,
3. incorporate that feedback into their work,
4. and venture into the next larger audience circle.
5. By the time their work reaches the general audience, their portfolios have been vetted and debated by multiple audiences starting from the students´ closest circles of peers and mentors [8].

Next, we consider an exchange of opinions beneficial as for the objects, by which the portfolio concept should be receiving. In the frame of education there is an option to systemize a few portfolios, some of them individual, some connected with universities. Washington State University in its concept Eportfolios for Learning implements three areas:

1. Student Portfolios promoting student's engagement and ownership of learning.
2. Teaching Portfolio offering a faculty a method of assessing, reflecting on, and improving teaching skills.
3. Program Portfolios highlighting the alignment of learning outcomes on multiple scales.

There is another, fourth option to develop above mentioned resulting from Innovation of Education Institute of VŠB – Technical University's experience:

4. Researcher's Portfolios promoting researcher's engagement and ownership of research.

In the research area the support of individuality and institution through Web 2.0 is even more uncommon. Researchers are tired of not ending results reporting by the databases to be filled with data because of insufficient integration and systems´ inability to generate such data to be used by different views and purposes. Practically there is no such a research being interactively supported by portfolios development enabling the author to invite the others to the space to give feedback or collaborate.

V. Conclusion and Future Work

As it can be seen, when going through internet sources, a growing number of institutions see the benefits of new technology and collaborative models and respect Web-enabled transparency as a new force. The university may be going through the biggest change as well. Confirming case studies having been provided showed that it is possible to act similarly also in more conservative environment of Central Europe universities. Methodology of implementation of end-users role in the processes of academic education has been proposed and proved. Respondents of statistically significant size declared significant changes as for their approach to study and its results. They were able to work more by themselves and independently, to find and use sources out of their community and the outputs of the others motivated them to better study etc. Quantitative and qualitative results analysis exposed an increase of significance of users-creators and effectiveness of educational activities compared to currently used educational model under Learning Management Systems. The results become touchable contribution to the graduates´ employment in a shape of their Personal Credit Portfolio.

Academic life prefers idleness and pseudo-safety of a hierarchic structure and non-transparency and thus rejects the openness and sharing. The reason for such behavior is the fear of losing its exclusivity. It is widely known that the changes are brought about by the information and communication technologies. And while these seemingly uncomfortable changes are unwanted, it leads to rather cold acceptance of the corresponding new technologies. Such academic institution is far away from the model Technology Powered Learning. Institutions with this philosophy cannot gain more than a locally-limited significance and thus their importance naturally descends although they try to show the opposite by all the evaluation mechanisms.

The way towards a new conception can be paved by a creation of personal virtual communication networks. Learners (also teachers and researchers) profit from Web 2.0 applications' effort and create their own mutually connected net of their individual learning, Personal Learning Networks (PLNs). Mush-ups of activities and continuous responses through open web applications are the content of their Personal Learning Portfolio.

By continuous activity a man build his Structured Accountability System as showcase achievements for accountability or employment. The tools of Deep Web and Cloud Computing help by facilitating employment searches. Continuously interactive lifelong Personal Credit Portfolio (PCP) is built as an individual expression of Positive Digital Identity.

Personality and its development are powered by technology. The result is the continuous development of personal credit and accountability in the labor market.

This article offers to discussion the experiences resulting from acceptation of collectivism principles in university education and from the practice of supporting the building individual PLNs. Such observations are to disposal when the participants (students and pedagogues) have been managed to go outside their classes, work continuously even when they are not together in the class, to go beyond curriculum of presently studying subject. Concentration of content being prepared in advance has been shifted to the principle of users' generated content as the centre of learning. It appeared the students wanted and was able to find other environments and people with the same interest. They carefully have been finding the effort of realized conversation (using different tools of social net) in the field they want to learn or – by pedagogue's view – which is the aim of subject's curriculum. This all greatly increased their internal motivation to learn.

In real net connection of all participants during the education the literacy of the students has been developed and they have been able to build their own PLNs. Students gained skill to find people and sources in determined context, which can be trusted to, and the ability to connect own net to relevant and inspirational bundles.

The ability to create one's own Personal Learning Network or Personal Learning or Credit Portfolios is very useful for lifelong learning and employability support. A self-motivated approach to one's own active systematic learning creates an environment known as a sustainable development of competencies [12]. The ability to create one's own PLN supports the ability to be adaptable. And to be adaptive is to be perpetually current.

Students are drifted by their own positive digital identity development on Web 2.0 cloud. They explore that learning starts to be the natural omnipresent part of their lives, change their life quality and improve their accountability or employment. Real development of their personalities is an indisputable fact – on Web 2.0 clouds.

A Model of Learning Powered by Technology enables, motivates and inspires all students to leverage the power of technology. Experiences in how to provide personalized and student-centric learning and enable continuous and lifelong learning can be transferred further.

Universities have to find by themselves how to reinvent the rules of business to survive the flat world created by a global Internet as those revolutionary changes have to accept all human activities. To improve university's and its staff's credit (and thus to increase their employability) means:

- Open our activities to the world, actively build own Personal Credit Portfolios.
- Support their aggregation into dynamically kept university credit – University Portfolio.
- Declare the fight against such university's presentation as the statistic websites are.
- Open University Net to communicate with wide public.
- Support the strategy of personal (and institutional as well) positive digital identity development on Web 2.0 cloud.

Innovations being implemented from down to up can help to destroy university bureaucracy and support natural methods of research work. Our Experiences show that by implementation of such principles as openness, peering, sharing and acting globally [1] brings great results not only in business sphere but at universities as well. New principles contribute to university credit's improvement not putting finances into the marketing. The same principle governs one's credit development – student's, pedagogue's or researcher's.

REFERENCES

[1] D. Tapscott, Wikonomics; How Mass Collaboration Changes Everything. Penguin Books Ltd. ISBN 978-1-59184-367-2; 2008.

[2] T. Barlow, Web 2.0: Creating a classroom without walls. Teaching Science, 54(1), 46-48; 2008.

[3] D. Boyd, The significant of social software. In T. N. Burge and J. Schmidt (Eds.), BlogTalks reloaded: social software research & cases (pp. 15-30). 2007. Norderstedt, Germany: Books on Demand.

[4] M. Pankowska, Integrated Information and Computing Systems for Natural, Spatial, and Social Sciences, Claus-Peter Rückemann (ed.) IGI Global, Hershey, New York, 2013 (in printing).

[5] H. Barret, Balancing the Two Faces of ePortfolio. Available at: ttp://eft.educom.pt/index.php/eft/article/viewFile/161/102. Accessed 13.3.2011; 2010.

[6] M. Karnjanaprakorn, Does The Online Education Revolution Mean The Death Of The Diploma? 2012. Available at: http://www.fastcoexist.com/1679315/does-the-online... Accessed 8. 4. 2012.

[7] D. Tapscott and A. Caston, Paradigm Shift: The New Promise of Information Technology (New York: McGraw-Hill, 1993).

[8] Ch. Handley, A. Wilson, N. Peterson, G. Brown and J. Ptaszynski. Out of the Classroom & Into the Boardroom. Higher Ed. Consorcium. 2007. Available at: http://www.microsoft.com/presspass/events/educause/docs/EducauseWhitepaper.pdf. Accessed 31. 12. 2011.

[9] W. Richardson, Five Things You Can Do to Begin Developing Your Personal Learning Network. 2008. Available at: http://theinnovativeeducator.blogspot.com/2008/04/5-things-you-can-do-to-begin-developing.html. Accessed 5. 4. 2011.

[10] D. Bauerová, Web 2.0 and Competitiveness Improvement (Web 2.0 y la Mejora de la Competitivida); 291-296; in: 10.3989/arbor.2011. Extra-3n3160. Vol. 187 - Extra 3 - d iciembre (2011). ISSN: 0210-1963.

[11] D. Bauerová Danuše, Positive Digital Identity Development on Web 2.0 Cloud; 337-348, in: Antonio Méndez-Vilas, Education in a technological world: communicating current and emerging research and technological efforts (pp. 624); available from http://www.formatex.org/ict. Formatex Research Center, 2011. ISBN (13): 978-84-939843-3-5.

[12] S. Schaffert, W. Hilzensauer, On the way towards Personal learning Environments: Seven crucial aspects. ELearning Papers, No 9, July 2008.

Knowledge-Based Visualization of Textual Information Applied in Biomedical Text Mining

Joseph Leone
Dept. of Computer Science and Engineering
University of Connecticut
Storrs, CT 06269-3155 USA
Joseph.2.Leone@uconn.edu

Dong-Guk Shin
Dept. of Computer Science and Engineering
University of Connecticut
Storrs, CT 06269-3155 USA
shin@engr.uconn.edu

***Abstract*—This paper describes a system, called VisuText, which creates visualized diagrams from textual descriptions. This work was motivated by the awareness that if additional contextual knowledge is appropriately utilized, one can develop a visualization system that systematically translates recognized objects and their relationships into a collection of one or more cohesively assembled pictures. VisuText first translates text into a computable representation, called SP form. SP forms are then converted into schematic diagrams by combining words and appropriate small images which themselves are stitched together to form a bigger meaningful picture. VisuText is especially suited for visualizing text that describes processes, particularly, those expressing similar facts and relationships in a large quantity. We find one excellent application area of VisuText is using it as a post-processing step after gene regulatory relationships are extracted through text mining of biomedical literature to pictorially represent discovered gene regulatory relationships for easier understanding by biomedical scientists. We illustrate how VisuText works by creating a pictorial representation of gene regulatory relationships from a set of statements extracted from the biomedical literature.**

Keywords-Text visualization; document visualization; natural language processing; text; semantic processing; dynamic ontology development; collaboration system; information retrieval; search; biomedical literature mining; gene regulatory relationships; cell signalling; picture rendering.

I. INTRODUCTION

The adage “a picture is worth a thousand words” is universally applicable when biomedical scientists summarize gene regulatory relationships from the literature. In the biological literature genomic structures, proteins, and other phenonena are generally described using natural language. The textual descriptions recount of elements that interact in very complex ways and the manner in which the elements interact to express gene regulatory and/or cell signaling relationships. Grasping these complex descriptions when they are presented in text is not an easy task. The problem becomes more difficult when these descriptions are not contained within a single document, but dispersed throughout various documents and need to be combined. The biomedical community, particularly, those working on discovering gene regulatory relationships face this problem more seriously, because each scientist may work on only a small set of genes and yet the community need to understand the big picture of how over 27,000 genes (in human case) work together.

In general, scientists currently read the biomedical literature and manually create schematic diagrams depicting the gene regulatory relationships summarized in the text. Examples include BioCarta [11], KEGG [12], and GenMapp [13]. They also extend existing diagrams when new information is garnered from the literature. The diagrams that they create, being a more adequate medium than language in conceptualizing complex interactions, help researchers quickly comprehend gene regulation relationships. Unfortunately, this process to convert textual information into a schematic diagram is done manually in most cases—an activity that is very laborious and prone to error. The question is whether one can design a system that can automate the process of generating pictorial representation of complex relationships, at least to a substantial degree, if doing that cannot be done entirely automatically.

This paper describes a system, called VisuText and its application in creating pictorial diagrams of gene regulatory relationships from textual descriptions. VisuText has been evolved from one of our earlier system, namely SPS [1, 2, 3], a system that performs phrase search of Web content and uses semantic processing to produce search results of very high quality and relevance.

In the rest of this paper, we describe VisuText in the following way. Section II discusses related works. Section III briefly describes SPS, the precursor of VisuText. Section IV describes VisuText’s architecture. Section V describes Picture Painter, a VisuText component that creates schematic representations from text. Section VI presents an example text and demonstrates how VisuText creates schematic diagrams. Finally, Section VII is the conclusion.

II. RELATED WORK

The previous works in visualizing texts are generally categorized into two groups, analytic ones [4, 5, 6] and artistic ones [7, 8, 9]. The analytic approaches include phrase nets [4], word trees [5], and two-word tag clouds [6]. The artistic approaches include Literary Organism Maps [7], Document Contrast Diagrams [8], and Directed Sentence Diagrams [9]. The artistic ones, generally, have no tie-in between the text and its depiction, and we consider they are

ISBN: 978-1-61208-202-8

remotely related to our work. We omit further discussion of this genre of works.

Phrase nets [4] visualize relationships indicated by a pattern (e.g., as shown many times in Bible, "X begat Y"), between words or phrases. A phrase net displays a graph whose nodes are words (i.e., one node for X and one node for Y) and whose edges indicate that the two words are linked by the user-indicated relation (e.g., “begat”). A high frequency pattern is displayed using a larger font size.

Word trees [5] visualize a user-supplied word or phrase and all the different phrases that follow it. The user-supplied phrases and the follow-up phrases are arranged in a tree-like branching structure.

Two-word tag clouds [6] show the most frequent two-word phrases in a body of text. Each two-word tag is displayed with font size varying by frequency of occurrence of the two-word phrase. Since two-word tag clouds provide more contexts by adding an additional word, this method aims to give a better sense of the text content than a single-word cloud.

The aforementioned approaches are mostly concerned with visualizing text words “in verbatim”, meaning they merely transform texts/phrases into either two- or multi-dimensional representation of expressed words in their exact forms. In contrast, our approach aims at visualizing phrases/sentences after extracting semantic meanings associated with them and use “that understanding” in formulating pictorial counterparts in which the diagrams may contain rephrased words and related words in strategic locations along with contextual images so that the whole picture can provide scientists with the intuition inferred in the adage “a picture is worth a thousand words”. Our approach first captures the text meaning (i.e., the context of the stated phrases/sentences), discourse structure, and discourse thread by using a computable knowledge representation. We then visualize, using pictograms that differ from the text they depict, the meaning of the text and not the text itself. For example, "cell wall" is depicted as an arc, "cell nucleus" as a circle, "interact" and "activate" as arrows, etc. Guided by the discourse structure and thread, the pictograms are combined into a schematic picture that reflects the totality of the stated text meaning.

III. SPS AS GENESIS OF VISUTEXT

Semantic Processing System (SPS) was initially developed to improve the relevancy of web search results. Web search can be divided into two phases: a “look” phase and a “find” phase. In the “look” phase a user presents keywords to a search engine and the search engine returns a set of pages the engine considers relevant to the user. In the “find” phase the user sifts through the search engine results to find the actual relevant/interesting information.

In SPS the "look" phase is performed by the *retrieval* subsystem, which receives a user’s phrase query, increases the quality of the keywords contained in the phrase query, and using a traditional search engine retrieves web pages containing those keywords. The "find" phase is carried out by the *relevance* subsystem, which automates the user cognitive task of sifting through search engine results (i.e., retrieved pages) to find the actual relevant/interesting information. A detailed SPS description can be found in [1, 2, 3].

VisuText is a spinoff of SPS in the sense that we conjectured use of three SPS components, SP Form, NL Parser, and Knowledge Lattice, could form a solid foundation for developing an automated visualization method that can pictorially depict relationships obtained from SPS driven discoveries. In particular, when the SPS discoveries find a large amount of similar, homogenous facts/relationships, we hypothesize that by utilizing additional contextual knowledge, one can develop a visualization system that systematically positions recognized objects and their respective relationships into one or more cohesive pictures.

IV. VISUTEXT ARCHITECTURE

The overview of VisuText architecture is presented in Figure 1. It consists of a GUI, Picture Painter, and three SPS components: SP Form, NL Parser (not shown), and Knowledge Lattice.

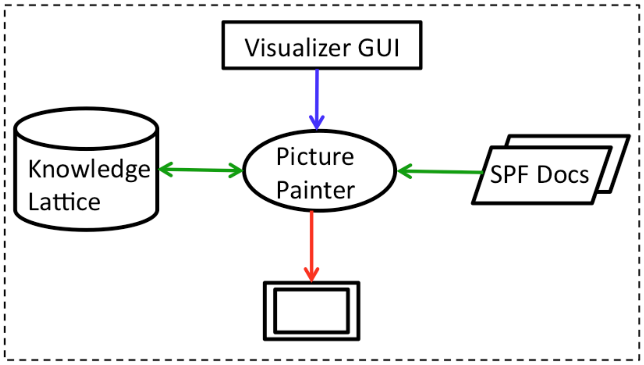


Figure 1. VisuText Architecture

A. SP Form

SP form [1] is the internal knowledge representation formalism used by both SPS and VisuText. A SP form expresses a sentence lexical structure in a computable format. A sentence consists of multiple phrases/clauses. Each phrase/clause is composed of syntactic and semantic elements. Syntactic elements, i.e., subject, verb, object, complement, adverb, etc. are *participants* in the meaning of a clause. Semantic elements, i.e., agent, instrument, affected, etc. are *roles* participants play. Each phrase/clause is encoded in SP form as a triple comprising a role and two participants.

(<role> (<direction1> <participant1>) (<direction2> <participant2>))

The collection of such phrases (i.e., SP forms) constitutes a sentence.

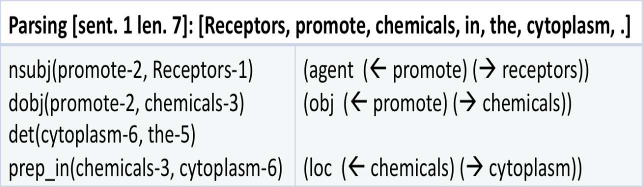
For example,

(agent (← activate) (→ chemicals))
"agent *of* activate *is* chemicals"

The direction symbol → that points away from the role is read as "is", and the direction symbol ← that points to the role is read as "of".

B. *NL Parser: Stanford typed dependencies*

NL Parser implements the Stanford typed dependency (SD) [10] parser. The SD parser represents sentence grammatical relationships as typed dependency relations, i.e., triples of a relation between pairs of words, such as "the subject of *promote* is *receptors*" in the sentence "*Receptors promote chemicals in the cytoplasm*". Each sentence word (except head of sentence) is the dependent of one other word. The dependencies are represented as *relation_name* (<*governor*>, <*dependent*>). All are binary relations: grammatical relation holds between a governor and a dependent.



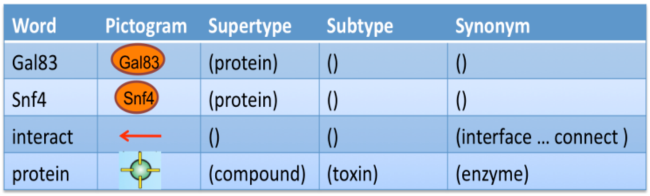
Parsing [sent. 1 len. 7]: [Receptors, promote, chemicals, in, the, cytoplasm, .]	
nsubj(promote-2, Receptors-1)	(agent (← promote) (→ receptors))
dobj(promote-2, chemicals-3)	(obj (← promote) (→ chemicals))
det(cytoplasm-6, the-5)	
prep_in(chemicals-3, cytoplasm-6)	(loc (← chemicals) (→ cytoplasm))

Figure 2. Parser dependency output and SP Form

The representation, as triples of a relation between pairs of words, is well suited for mapping SD parser output to SP forms. Figure 2 shows an SD parse of the sentence "*Receptors promote chemicals in the cytoplasm*". The parse output, i.e., the syntax tree (not shown) and the SD dependencies, is mapped to SP forms.

C. *Knowledge Lattice / Image Element Depictions*

The *Knowledge Lattice* (KL) is a data structure for storing words, their subtype / supertype relationships, their synonyms, and their pictograms. Pictograms are used to compose pictures from text. The subtype / super-type relations comprise the word's hypernyms and hyponyms. Note that the KL stores no word definitions. Included with the data structure is a set of operations for reasoning about the relations between words. The Knowledge Lattice is updated and extended by the *Interactive Learning Component*. A detailed description of the Knowledge Lattice and Interactive Learning Component can be found in [3].



Word	Pictogram	Supertype	Subtype	Synonym
Gal83	Gal83	(protein)	()	()
Snf4	Snf4	(protein)	()	()
interact	←	()	()	(interface ... connect)
protein	⌖	(compound)	(toxin)	(enzyme)

Figure 3. Knowledge Lattice Fragment

Figure 3 shows the computational representation of a Knowledge Lattice fragment. In Figure 3, the word "interact" has an arrow pictogram, no supertype or subtype, but many different synonyms. When composing a picture involving the word "interact" or any of its synonyms, the red arrow is used in the picture's composition. The arrow orientation is determined by the phrase in which the word "interact" occurs.

1) *Knowledge Lattice Operations*

Knowledge Lattice Operations, described in [3], compute word synonyms, hypernyms, and hyponyms. When a word pictogram is missing, the pictogram of the word's synonym or the word's supertype could be used in picture composition.

V. Picture Painter

Picture Painter creates and renders the actual text visualization. Picture Painter interprets *SPS logical form* as a *Picture Description Language* (PDL), creates images from phrases, combines the various images into a whole (i.e., a picture), and finally places the whole into a frame for viewing.

A. *Picture Description Language*

Words are the *SPS logical form* primitives. Words are combined to create an SP form expression, which consists of a role and two participants (see Section IV.A). A collection of SP forms constitutes a sentence. Picture Painter re-interprets SPS logical form as a *Picture Description Language* (PDL).

PDL is treated as a pictorial analogue of SPS logical form. In PDL, words are still primitives; but the words are interpreted as pictograms -- words (and their synonyms) are bound to pictograms in the Knowledge Lattice (see Figure 3). Words (i.e., pictograms) are combined to form images, which are the pictorial analogue of SP forms. A collection of images is combined into a picture.

B. *Picture Composition*

A picture is composed bottom-up by first creating an image (i.e., the pictorial analogue of SP form), and then combining the images.

1) *Image Creation*

Image creation is specified by the following rules.

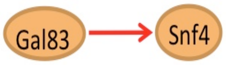
1. Roles determine pictogram orientation.
2. Participants denote pictograms.
3. Links signify the connected participants.
   A link is the common participant that connects two or more roles in multiple SP forms.
4. A connector is the link pictogram. A connector generally has two ends for stitching the participants.

An example illustrates application of these rules.

(agent (← interact) (→ (protein Gal83))
(obj (← interact) (→ (protein Snf4))

The roles are "agent" and "obj". The participants (i.e., pictograms) are "interact", "Gal83", and "Snf4" (see Figure

3). The link is “interact”. The connector is the “interact” pictogram. *Image-creation* aligns the “agent” participant to the connector base and the “obj” participant to the connector top, thus producing the following image:



Note that if “Gal83” or “Snf4” were not bound to pictograms, the participant’s supertype (i.e., “protein”) pictogram would be used.

2) Image Combination

Image combination is specified by the following rules.

1. Links signify the connected images. A link is a common participant that connects two or more images, or an image and a pictogram.
2. Role-of-link determines the alignment/orientation of images, or image and pictogram.

Application of these rules is illustrated by the example: "*Growth factors attach to receptors in the cell membrane*." This sentence’s PDL, separated into the images it produces, is shown below. Rendered images are shown in Figure 4 dashed-rectangle 1.

Image 1:
(type (← factors) (→ growth))
(agent (← attach) (→ factors))
(dest (← attach) (→ receptors))

Image 2:
(type (← membrane) (→ cell))
(loc (← receptors) (→ membrane))

Image 1 has three distinct pictograms: “attach”, “growth factors”, and “receptors” (“type” role dictates that “growth” and “factors” be treated as a single pictogram). Image 2 has only the “membrane” pictogram; “receptors” pictogram is available from image 1. The link that connects the two images is “receptors”, and the role-of-link is “loc” which connects image 1 to the pictogram “membrane”. The role-of-link is “loc” instead of “dest” because in the “loc” phrase the link participant is an “of” participant. The role “type” causes a labeling, which is handled by picture rendering.

Image-creation stitches “growth factors” to “receptors” to create image 1 (i.e., grouping). *Image-combination* stitches “receptors” on “membrane”. The result of image-combination is shown in Figure 4 dashed-rectangle 1.

C. Picture Rendering

When images are created and stitched together, the picture that is formed is placed within a parallelogram-shaped frame for viewing. Picture orientation and alignment of its elements is determined by rendering primitives, type of pictogram, and amount of zoom.

1) Rendering Primitives

During image-combination, as images are created and stitched together, a *rendering expression* is formed. A rendering expression is built from the following primitives.

```
<expression> ::=   <id>
                   | (beside <id> <expression>)
                   | (below  <id> <expression>)
                   | (diag1 <id> expression)
                   | (diag2 <id> expression)
                   | (on <id> <expression>)
        <id> ::=   <pictogram> | <image>
```

The discourse thread guides the expression formation. Pictograms within images (and picture) are linearly arranged/aligned in the relative direction of the discourse thread. The completed rendered expression is used by the rendering system to place/locate the images within the frame. Placement can be *vertical*, *horizontal*, *diagonal*, or *scattered*. A scattered placement results when no discourse thread exists, but the text nonetheless has common elements (e.g., sentences, with common participants, collected from different documents).

2) Pictogram Types

A picture is composed of images, which are in turn composed of pictograms. Pictograms are of two types: mutable and immutable. Immutable pictograms cannot be scaled. All pictograms stored in the KL are immutable and their relative size is constant.

Mutable pictograms instead can be scaled and stretched. These pictograms do not actually exist in the KL, but instead are drawn by the rendering system. Examples of such pictograms are arcs, circles, ovals, lines, rectangles, hexagons, pentagons, diamonds, etc.

Mutable pictograms depict entities that are containers for other entities. For example, an arc could represent a cell wall and an oval could represent the cell itself. Pictograms of entities such as cells must be mutable, because as more elements are placed inside the cell, the extent of the cell (i.e., oval) and the size of the cell wall (i.e., arc) need to increase.

Also, if an entity contains another identical entity, both entities to be distinguished must be of different size. For example, if a cell and cell nucleus are both represented as a circle, the two circles must both be of different size, with the outer circle bigger than the inner circle.

3) Zoom

The picture elements visible within a frame depend on whether a picture is rendered from a long-shot or a close-up (i.e., zoom). For example, a zoom-in of a cell might show only a portion of the cell membrane near the frame edge and a very large cell nucleus, whereas a zoom-out of a cell would show the entire cell membrane within the frame and a tiny cell nucleus.

VI. Example: Picture Composition and Rendering

This section illustrates, via an example, the workings of VisuText as it converts a natural language text into a picture.

The natural language text:

> *"Growth factors attach to receptors in the cell membrane[1]. The receptors promote chemicals in the cytoplasm[2]. The cytoplasm chemicals activate kinases[3]. Kinases activate chemicals that can pass through the wall of the cell nucleus to turn-on transcription factors[4]. Transcription factors turn-on the genes that make the cell divide[5]."*

A. *Natural Language Parsing*

NL Parser translates the text into SP form (numbers correspond to sentence identifiers given in paragraph).

1: a. (type (← factors) (→ growth))
b. (agent (← attach) (→factors))
c. (dest (← attach) (→ receptors))
d. (type (← membrane) (→ cell))
e. (loc (← receptors) (→ membrane))

2: a. (agent (← promote) (→ receptors))
b. (obj (← promote) (→ chemicals))
c. (loc (← chemicals) (→ cytoplasm))

3: a. (type (← chemicals) (→ cytoplasm))
b. (agent (← activate) (→ chemicals))
c. (obj (← activate) (→ kinases))

4: a. (agent (← activate) (→ kinases))
b. (obj (← activate) (→ chemicals))
c. (agent (← pass) (→ chemicals))
d. (affirm (← pass) (→ can))
e. (obj (← pass) (→ wall))
f. (type (← nucleus) (→ cell))
g. (kind (← wall) (→ nucleus))
h. (agent (← turn-on) (→ chemicals))
i. (obj (← turn-on) (→ factors))
j. (type (← factors) (→ transcription))

5: a. (agent (← turn-on) (→ factors))
b. (obj (← turn-on) (→ genes))
c. (type (← factors) (→ transcription))
d. (agent (← make) (→ genes))
e. (agent (← divide) (→ cell))
f. (result (← make) (→ divide))

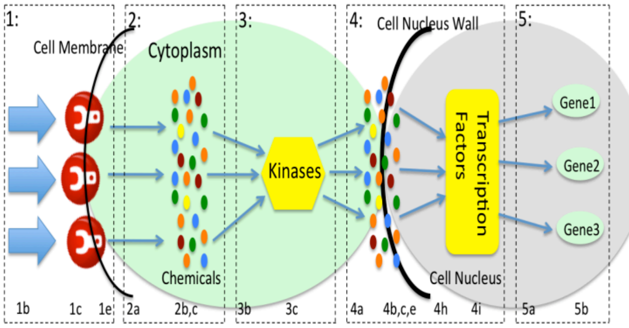
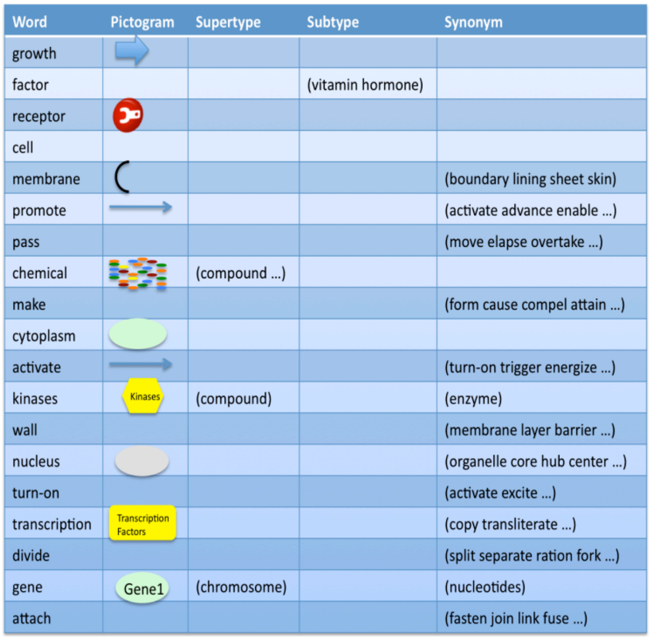


Figure 4. Picture Rendering -- Horizontal

B. *Image Creation, Combination, and Rendering*

Picture Painter interprets the SP forms as PDL. From each sentence, an image is created and combined with images from other sentences. Figure 4 shows the images created. Each dotted box encloses an image that corresponds (via the number) to the PDL clauses (i.e., sentence) from which the image is created. Numbers beneath the pictograms refer to the individual clause from which the pictogram is derived. Figure 5 shows a KL fragment containing pictogram depictions of various words.



Word	Pictogram	Supertype	Subtype	Synonym
growth				
factor			(vitamin hormone)	
receptor				
cell				
membrane				(boundary lining sheet skin)
promote				(activate advance enable ...)
pass				(move elapse overtake ...)
chemical		(compound ...)		
make				(form cause compel attain ...)
cytoplasm				
activate				(turn-on trigger energize ...)
kinases	Kinases	(compound)		(enzyme)
wall				(membrane layer barrier ...)
nucleus				(organelle core hub center ...)
turn-on				(activate excite ...)
transcription	Transcription Factors			(copy transliterate ...)
divide				(split separate ration fork ...)
gene	Gene1	(chromosome)		(nucleotides)
attach				(fasten join link fuse ...)

Figure 5. KL Participant Pictogram Depiction

1) *Image Creation*

a) *Box 1:*

Box 1 says that the "agent" of "attach" is "factors" which are of type "growth", the "dest" of "attach" is "receptors", and the "loc" of "receptors" is "membrane" which is of type "cell". Figure 5 shows the KL pictogram depiction of these participants: "growth factors" (1b) as a thick solid blue arrow, "receptors" (1c) as wrench symbol, and "membrane" (1e) as a black arc. These participants are united according to the rules in Section V.B.1 to create the box 1 image.

b) *Box 2:*

Box 2 says that the "agent" of "promote" is "receptors", the "obj" of "promote" is "chemicals", and that the "loc" of "chemicals" is the "cytoplasm". Pictogram depictions: "promote" (2a) as a thin blue arrow, "chemicals" (2b) as a collection of multi-color ovals, and "cytoplasm" (2c) as a large green oval. These participants are united, according to Section V.B.1 rules, to form the image shown in box 2.

c) *Box 3:*

Pictogram depictions: "activate" (3b) as a single solid blue arrow with a thin body and "kinases" (3c) as a labeled yellow hexagon. Note that if "activate" did not have a

ISBN: 978-1-61208-202-8

pictogram, then the pictogram of its synonym "promote" could be used.

d) Box 4:

Box 4 contains many participants that have already been seen from boxes 1, 2, and 3. The new participants are "pass", "nucleus", "wall", "turn-on", and "transcription". Of these participants only "nucleus" (4i), which is depicted as a grey oval and "transcription" (4ji), which is depicted as a rounded rectangle, have a pictogram. "Pass", which in the context of sentence 4 does not denote a thing but instead describes an occurrence, does not have an associated pictogram. "Wall" and "turn-on" also do not have a pictogram; consequently, the synonyms of their pictograms are used: a black arc for "wall" (4e), and a thin blue arrow for "turn-on" (4h).

e) Box 5:

The new box 5 participant is "gene" (5b) which is depicted as a labeled light green oval.

2) Image Combination

Participants (i.e., pictograms) are united according to the rules in Section V.B.1 to create images (see Figure 4 boxes). Images are stitched together to create a picture.

The combining of images is guided by the discourse thread, which is encoded as a rendering expression (section V.C.1). In this example the discourse thread is "*growth factors – receptors – chemicals – kinases – chemicals – cell nucleus – chemicals – transcription factors – cell divide*".

The rendering expression is *(beside (beside (beside (beside 1b (on 1d 1e)) (beside 2a (on 2b 2c))) (beside 3b 3c)) (beside (beside 4a (on 4b 4e)) (beside 4h 4i)) (beside 5a 5b)).*

3) Picture Rendering

The rendering system uses the rendering expression to place/locate the stitched images within a frame. Placement can be vertical, horizontal, diagonal, or scattered. In Figure 4 placement is horizontal.

VII. Conclusion

We have presented a framework that is designed to carry out a post processing following a text mining step in order to covert the recognized relationships obtained from a text mining into a set of pictorial diagrams. We demonstrated that our automated methodology is well suited for better representing text mining outcomes of gene regulatory relationships buried in the biomedical literature. Using a series of examples we have illustrated that our proposed method can indeed capture textual meanings of stated words using knowledge lattice and can create visual depiction of the key elements of the involved objects at the appropriate conceptual level automatically. In a nutshell, we point out that incorporating this extra layer of visual knowledge into the picture creation is what makes the user's understanding of diagrams far more intuitive than simple narration of multiple related sentences. VisuText is especially suited for visualizing text that describe processes, such as gene regulatory relationships, which consist of various elements that interact with each other or trigger interactions. Currently we are refining the methodology and are experimenting, using large scale text mining of biomedical literature, with a prototype in order to gauge its performance.

References

[1] J. Leone, "A Semantic Processing System (SPS) for Web Search", Ph.D thesis, University of Connecticut, 2011 (under preparation).

[2] J. Leone and D. G. Shin, "SPS: A Web Content Search System Utilizing Semantic Processing," Content 2011: The Third International Conference on Creative Content Technologies, Rome, Italy, September 25-30, 2011.

[3] J. Leone and D. G. Shin, "A Semantic Processing System (SPS) for Biomedical Literature Web Search," Advances in Data Mining 11th Industrial Conference, ICDM 2011, New York, USA, August/September 2011.

[4] F. van Ham, M. Wattenberg, and F. B. Viégas, "Mapping Text with Phrase Nets", Proc. IEEE InfoVis, 2009. http://www.research.ibm.com/visual/papers/phrase-net-rev5.pdf [retrieved: April, 2012]

[5] M. Wattenberg and F. Viégas, "The Word Tree: An Interactive Visual Concordance," Proc. IEEE InfoVis, 2008. http://researchweb.watson.ibm.com/visual/papers/wordtree_final2.pdf [retrieved: April, 2012]

[6] F. Viégas and M. Wattenberg, "Tag Clouds and the Case for Vernacular Visualization", ACM Interactions, XV.4 July/August, 2008. http://www.research.ibm.com/visual/papers/vernacular_visualization.pdf [retrieved: April, 2012]

[7] http://www.itsbeenreal.co.uk/index.php?/wwwords/literary-organism [retrieved: April, 2012]

[8] http://www.neoformix.com/2008/DocumentContrastDiagrams.html [retrieved: April, 2012]

[9] http://www.neoformix.com/2008/DirectedSentenceDiagrams.html [retrieved: April, 2012]

[10] http://www-nlp.stanford.edu/software/stanford-dependencies.shtml [retrieved: April, 2012]

[11] http://www.biocarta.com/Default.aspx [retrieved: April, 2012]

[12] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe, "KEGG for integration and interpretation of large-scale molecular data sets", Nucleic Acids Res. 2012 Jan;40(Database issue):D109-14. Epub 2011 Nov.10.

[13] N. Salomonis, K. Hanspers, A. Zambon, K. Vranizan, S. Lawlor, K. Dahlquist, S. Doniger, J. Stuart, B. Conklin, and A. Pico. GenMAPP 2: new features and resources for pathway analysis. *BMC Bioinformatics, Jun 2007; 8: 217*

Requirements Defect Density Reduction Using Mentoring to Supplement Training

John Terzakis
Intel Corporation
e-mail: john.terzakis@intel.com

***Abstract*— Requirements authors typically receive little formal university training in writing requirements. Yet, they are expected to write requirements that will become the foundation for all future product development. Defects introduced during the requirements phase of a project impact multiple downstream work products and, ultimately, product defect and quality levels. Many companies, including Intel Corporation, have recognized this skills gap and have created requirements training classes to address this issue. While effective in providing the fundamentals of good requirements writing, much of this knowledge can be misapplied or lost without proper mentoring from a requirements Subject Matter Expert (SME). Our experience over the last decade at Intel has found that adding SME peer mentoring improves both the rate and depth of proper application of the training, and improves requirements defect density more than training alone. This paper will present data from a case study demonstrating the issues with training alone and the benefits of combining training with SME mentoring in order to achieve a greater than 75% reduction in requirements defect density.**

Keywords-requirements specification; requirements defects; requirements defect density; training; mentoring.

I. INTRODUCTION

While bachelor degrees exist for a variety of Engineering disciplines, degrees and even undergraduate courses in Requirements Engineering are scarce. Primary requirements authors (those whose primary role is to elicit and write requirements) may have some training. However, secondary authors (those whose primary role is architecture, development, testing, etc.) may have little or no training. As Berenbach, et al, state “Requirements analysts typically need significant training, both classroom and on the job, before they can create high-quality specifications.” [1]. To close this skills gap, many companies have created in-house requirements courses or contracted third-party trainers to teach the basics of well-written requirements. Many are based on the IEEE 830 standard, [2] or the good, practical books published in the field over the last decade [3], [4]. At Intel, in-house requirements courses have been taught to over 13,000 students since 1999. While useful for providing an initial understanding of the issues and challenges of requirements authoring, the knowledge gained through these courses can be misapplied or lost due to the inexperience of authors in writing effective requirements. By pairing with a SME, the authors can be provided with early feedback on the deficiencies of their requirements.

This paper examines the requirements defect density rates for two secondary authors on software projects who attended a requirements writing course and then were mentored on subsequent revisions of their requirements specifications.

II. INITIAL CLASSROOM TRAINING

Both requirements authors attended a training session on requirements writing prior to beginning work on their Software Requirements Specification (SRS). These training sessions focused on the issues with natural language, attributes of well-written requirements, a consistent syntax for requirements and an introduction to Planguage (Planning Language). Issues with natural language in the training included ambiguity, weak words, unbounded lists and grammatical errors. Ten attributes of well-written requirements were shown and explained in detail. A requirements syntax of the form:

[Trigger][Precondition] Actor Action [Object]

was presented in the internal training. Finally, an overview of Tom Gilb’s Planguage [5] was taught, along with exercises to reinforce the concepts.

Following the class on requirements writing, both authors began writing their requirements and submitted early samples for review. These early samples showed requirements defect densities of about 10 and 5 major defects per page respectively. These figures represent the baseline for this paper. While some of the key concepts were applied (a consistent syntax, use of Planguage), other key concepts were not (authors’ continued use of weak words, failure to check requirements for the ten attributes, logic issues, etc.). With this baseline in place, we began mentoring each of the authors. Note that the examples that follow have been slightly modified from their original form to maintain author confidentiality

III. MENTORING

Our mentoring consisted of reviewing the requirements, identifying requirements quality issues and then working with the authors to rewrite the requirements. Here is an initial sample requirement from the first author:

> *The software should have radio style buttons to enable/disable graphics cards.*

Issues with this requirement include use of a weak word (should), design statement (radio style buttons), the use of a slash and vagueness (“graphics cards”). Our mentoring sessions focused on discussing how to correct the issues and rewrite the requirements. The updated requirement became:

> *The software shall display an option to enable or disable graphics cards installed in the PCIe bus.*

By the latter revisions of the SRS, this author was self-reviewing requirements using a checklist provided in the requirements training class. Our reviews of subsequent requirements revealed that they required only minor rewrites and contained far fewer defects.

Initial samples from the second requirements author demonstrated similar issues. Here is a sample:

> *The software needs to provide the ability to wake on a wireless LAN event.*

This requirement was missing a trigger (what causes the software to wake?), lacked an imperative (needs) and is ambiguous (what event?). After mentoring, the rewritten requirement became:

> *When the operating system (OS) is in a sleep state and the software detects a Magic Packet on the wireless network, the software shall wake the OS.*

> *Defined: Magic Packet: A broadcast frame containing anywhere within its payload 6 bytes of 1's (0xFFFF FFFF FFFF) followed by 16 repetitions of the system MAC address.*

This particular author embraced the training to the extent that he would help others to correct their requirements during review meetings.

IV. RESULTS

The requirements defect densities for each author were tracked from an initial version (0.3) of the SRS to a released version (1.0). This process took approximately one year in each case. The results appear in Tables I and II that follow.

Table I: Requirements Defect Density, Author #1

Rev	**# of Defects**	**# of Pages**	**Defects/ Page (DPP)**	**% Change in DPP**
0.3	312	31	10.06	
0.5	209	44	4.75	-53%
0.6	247	60	4.12	-13%
0.7	114	33	3.45	-16%
0.8	45	38	1.18	-66%
1.0	10	45	0.22	-81%
Overall % change in DPP revision 0.3 to 1.0: ***-98%***				

Table II: Requirement Defect Density, Author #2

Rev	**# of Defects**	**# of Pages**	**Defects/ Page (DPP)**	**% Change in DPP**
0.3	275	60	4.58	
0.4	350	78	4.49	-2%
0.5	675	125	5.40	+20%
0.7	421	116	3.63	-33%
0.75	357	119	3.00	-17%
1.0	115	122	0.94	-69%
Overall % change in DPP revision 0.3 to 1.0: ***-79%***				

V. CONCLUSIONS

The requirements defect density data indicates that large reductions can be achieved by combining training with mentoring and that mentoring benefits continue for many months after training. Initial defect density rates following training were high in each case (about 10 and 5 defects per page respectively). By combining requirements SME mentoring with this initial training, defect rates dropped by over 75% in each case. Similar reductions have been observed with other requirements authors when mentoring is combined with training. .

REFERENCES

[1] Berenbach, B., Kazmeier, J., Paulish, D. and Rudorfer, A., *Software & System Requirements Engineering in Practice*, McGraw Hill, March 26, 2009

[2] IEEE Std 830-1998, "IEEE Recommended Practice for Software Requirements Specifications", the Institute of Electrical and Electronics Engineers, Inc., June 25, 1998

[3] Wiegers, K., *Software Requirements, 2nd Edition*, Microsoft Press, March 26, 2003.

[4] Kotonya, G. and Sommerville, I., *Requirements Engineering: Processes and Techniques*, John Wiley & Sons Ltd., August 25, 1998.

[5] Gilb, T., *Competitive Engineering: A Handbook For Systems Engineering, Requirements Engineering, and Software Engineering Using Planguage*, Butterworth-Heinemann, June 25, 2005.

MoBiSiS: An Android-based Application for Sending Stego Image through MMS

Rosziati Ibrahim
Department of Software Engineering
Faculty of Computer Science and Information Technology (FCSIT), Universiti Tun Hussein Onn Malaysia (UTHM),
Parit Raja, Johor, Malaysia
rosziati@uthm.edu.my

Law Chia Kee
Department of Software Engineering
Faculty of Computer Science and Information Technology (FCSIT), Universiti Tun Hussein Onn Malaysia (UTHM),
Parit Raja, Johor, Malaysia
qiqilaw1989@msn.com

***Abstract*— A Steganography algorithm is used to hide data from third party in such a way that people are unable to detect the existence of the hidden message inside the stego image. This algorithm is used to maintain the confidentiality of valuable information, and to protect the data from possible sabotage, theft, or unauthorized viewing. Before mobile services, the stego image is sent via e-mail. The recipients have to be connected to Internet and log into their mailbox to download the stego image. This paper introduces a mobile application named MoBiSiS (Mobile Steganography Imaging System). MoBiSiS improves the capability of steganography algorithm by implementing the steganography algorithm for Android-based application. MoBiSiS is able to send the stego image through the Multimedia Messaging Service (MMS) and the stego image can be retrieved from the device's message inbox to extract the hidden message inside the stego image.**

Keywords-steganography algorithm; secret key; image processing.

I. Introduction

This paper presents an android-based mobile application named MoBiSiS (Mobile Steganography Imaging System). MoBiSiS is a mobile application that is capable of hiding the data inside the image. The image containing the data can then be sent via MMS (Multimedia Messaging Service). MoBiSiS can be used by various users who want to hide data inside an image without revealing the data to other parties. Implementing steganography algorithm [1] in Android-based application makes the usability of steganography increased since mobile is more convenient for user to be brought anywhere and use anywhere. By sending the stego image through MMS, the user is able to get announcement instantly once the stego image received. Therefore, MoBiSiS provides more opportunity for hiding information efficiently.

Steganography is the Greek word for hiding information that invisible to the observer's sense. Steganography is intended to provide secrecy in such a way that others unable to detect the existence of the hidden message. Steganography algorithm helps to hide data and ensures the privacy of the data. This algorithm is used to address digital rights management, conceal secret and protect the confidential information from possible sabotage, theft, or unauthorized viewing.

Steganography algorithm is very important for the purpose of hiding information inside an image. Therefore, the proposed application is being implemented using steganography algorithm to protect the privacy and secrecy of data. The proposed application is an android-based application which allows the user to send or retrieve the hidden data inside the stego image. With a mobile on hand, user can send or retrieve the stego image instantly. The communication media of sending and receiving the steganography image is using the Multimedia Messaging Service (MMS). This application provides an image platform for user to input image, a text box to input the message and allow user to set the key or password of the stego image. Thus, the data is being protected by the key or password.

The rest of the paper is organized as follows. Section 2 reviews the related work and Section 3 presents the details of the implementation of MOBiSiS. Section 4 discusses various results obtained from testing the functionalities of MOBiSiS. The PSNR (Peak signal-to-noise ratio) value of the stego images are also presented and finally, we conclude the paper in Section 5.

II. Related Work

Hiding data is the process of embedding information into digital content without causing perceptual degradation [2]. In data hiding, three famous techniques can be used. They are watermarking, steganography and cryptography. Steganography is defined as covering writing in Greek. It includes any process that deals with data or information within other data. According to Lou *et al.* [3], steganography is hiding the existence of a message by hiding information into various carriers. The major intent is to prevent the detection of hidden information.

Research in steganography technique has been done back in ancient Greek where during that time the ancient Greek practice of tattooing a secret message on the shaved head of a messenger, and letting his hair grow back before sending him through enemy territory where the latency of

ISBN: 978-1-61208-202-8

this communications system was measured in months [4]. The most famous method of traditional steganography technique around 440 B.C. is marking the document with invisible secret ink, like the juice of a lemon to hide information. Another method is to mark selected characters within a document by pinholes and to generate a pattern or signature [4]. However, the majority of the development and use of computerized steganography only occurred in year 2000 [5]. The main advantage of steganography algorithm is because of its simple security mechanism. Because the steganographic message is integrated invisibly and covered inside other harmless sources, it is very difficult to detect the message without knowing the existence and the appropriate encoding scheme [6]. There are several steganography techniques used for hiding data such as batch steganography, permutation stehanography, least significant bits (LSB), bit-plane complexity segmentation (BPCS) and chaos based spread spectrum image steganography (CSSIS).

Research in hiding data inside image using steganography technique have been done by many researchers, for example in [1], [7], [8], [9], [10], [11] and [12]. Warkentin *et al.* [7] proposed an approach to hide data inside the audiovisual files. In their steganography algorithm, to hide data, the secret content has to be hidden in a cover message. El-Emam [8], on the other hand, proposed a steganography algorithm to hide a large amount of data with high security. His steganography algorithm is based on hiding a large amount of data (image, audio, text) file inside a colour bitmap (bmp) image. In his research, the image will be filtered and segmented where bits replacement is used on the appropriate pixels. These pixels are selected randomly rather than sequentially. Chen *et al.* [9] modified a method used in [10] using the side match method. They concentrated on hiding the data in the edge portions of the image. Wu *et al.* [11], on the other hand, used pixel-value differencing by partitioning the original image into non-overlapping blocks of two consecutive pixels.

Rosziati Ibrahim et al. [1] propose a steganography algorithm for hiding secret message inside an image. A bitmap (bmp) image is used to hide the data. Data is then embedded inside the image using the pixels. Then the pixels of stego image can then be accessed back in order to retrieve back the hidden data inside the image. Based on the steganogrpahy algorithm in [1], an android-based application is developed to send the stego image through MMS. This android-based application is known as MoBiSis (MoBile Steganography Imaging System). MoBiSiS used the technology of MMS to send or receive the stego images. MMS is a technology that allows a user of a properly enabled mobile phone to create, send, receive and store messages that include text, images, audio and video clips [13]. Users would be able to benefit from the MMS technology for secretly exchange hidden messages and keys, without arousing suspicion of their existence.

III. MOBISIS IMPLEMENTATION

Based on the algorithm proposed in [12], an android-based application is implemented for the purpose of sending the stego image via MMS. This android-based application is written in open source programming language consisting of Java language, Extensible Markup language (XML) and Apache Ant scripting language.

Figure 1 illustrates the activity diagram that represents the flow of activities for proposed application. Activity diagram is one of the Unified Modeling Language (UML) specifications that describe coordination among activities of the application and its external actor by showing the workflow of application. The purpose of activity diagram is to illustrate possible navigation paths through the interface and connections to other parts of the system functionality. However, an activity diagram is differed from a traditional flowchart as it shows concurrency as well as branches of control.

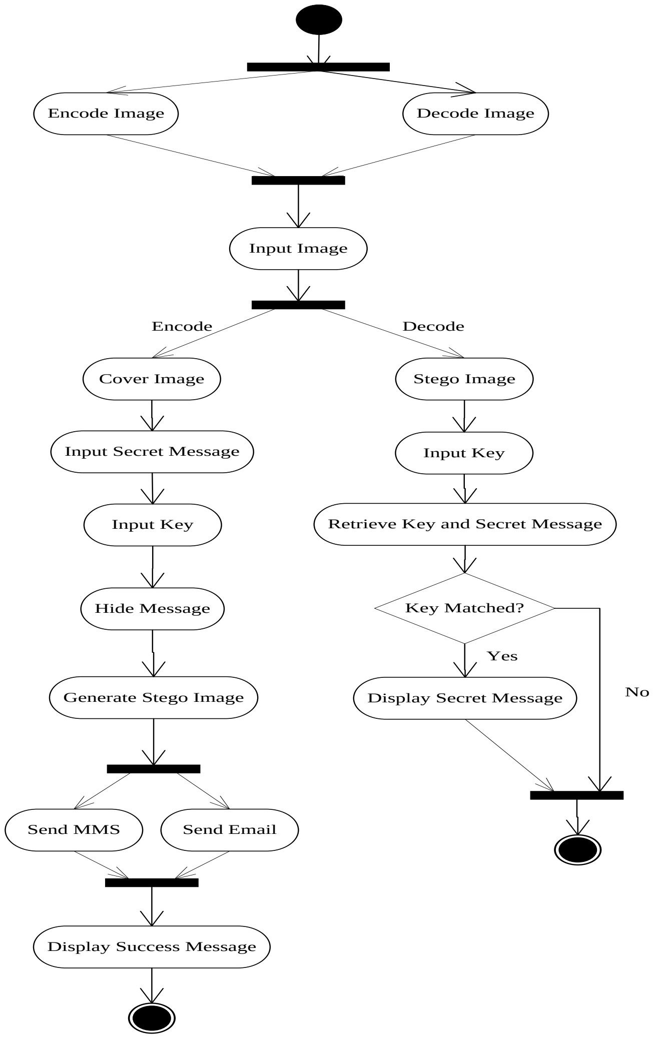


Figure 1. Activity Diagram for MoBiSiS

Based from Figure 1, the application starts with the user selecting to decode message or encode message from menu indicator, which is an android mobile built in function to perform additional functions. A cover image and a secret message are needed before a key (password) is entered where the key is required to allow generation of stego image. The stego image can be sent through MMS or Email after the stego image is generated. The success message will be displayed and the application ends after the stego image is sent. Stego image and the same key, on the other hand, are needed to retrieve the secret message. Secret message can be retrieved only if the key is matched, or else, the process is failed. Finally, the application ends where the stego image has been generated or secret message has been retrieved and displayed on the text box.

The process of embedding and extracting the message is illustrated in Figure 2.

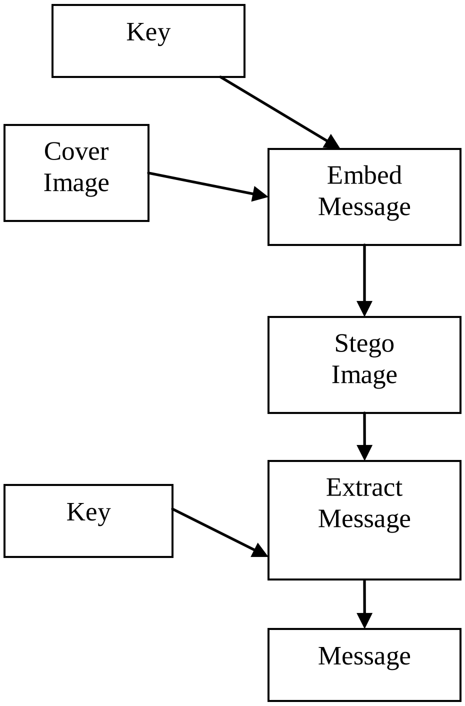


Figure 2. The process of Embedding and Extracting Message

Based from Figure 2, cover image is needed in order to embed the secret message inside the image together with the secret key. Then the message is embedded inside the image. This new image is known as stego image. In order to extract the hidden message inside the stego image, the secret key is needed. Once the correct secret key is provided, the message will be able to retrieve from the stego image. Note that the secret key has to be agreed between the sender and the receiver. If the sender has agreed on a secret key, the sender has to tell the receiver the secret key that has been used for the image. The operational requirements for the application are as follows:

i. Stego image will be generated only after the inputs of cover image, secret message and key (password).
ii. Stego image will be generated in portable network graphic (png) format only.
iii. Secret message will be retrieved only after the input of stego image and the key where the key is matched with the key that has set previously.

MoBiSiS has four functional requirements as stated in Table I. The functional requirements are then used for the functionalities of MOBiSiS.

TABLE I. Functional Requirements of MoBiSiS

i	Generate the Stego Image (Encode)
ii	Retrieve the Secret message from the Stego image (Decode)
iii	Send Steganography Image
iv	Provide additional tools

Based on Table I, the first functional requirement is to generate stego image by the application using the input of cover image, secret message and key from user. The second functional requirement is to retrieve secret message by the application using the input of stego image and the key from user. The third functional requirement is to send the generated stego image through MMS or Email by the application. The forth functional requirement is to enhance the application by providing additional tools for user to interact.

Figure 3 shows the interface of MoBiSiS that allows a user to select the cover image from capture a new photo by mobile camera or from gallery. This page consists of two button which are “From Camera” and “From Gallery” button. The camera function will be switched to on when the “From Camera” button is pressed. On the other hand, the phone’s gallery will be shown when the “From Gallery” button is pressed.

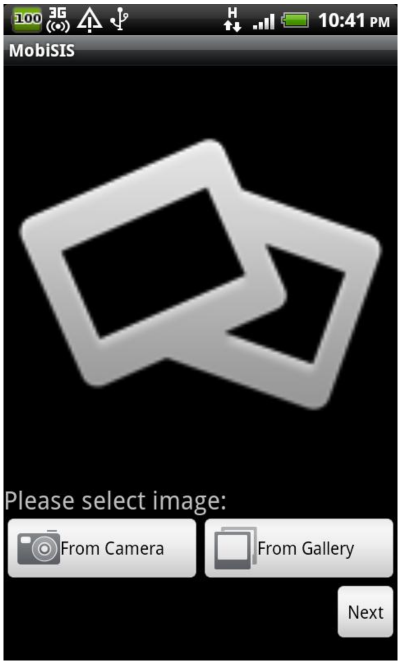


Figure 3. Main Interface of MoBiSiS

Once the image has been choosen (either from camera or from gallery), the secret message can be type and embeded inside the selected image. Figure 4 shows the interface for this process.

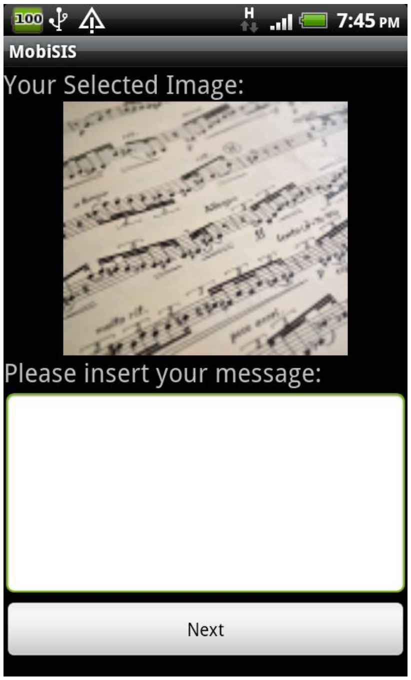


Figure 4. Interface to type the message for embedding

The process flow for MoBiSiS is shown in Figure 1. Once the information has been stored inside the stego image, this stego image can be sent via MMS or Email without exposing the information embeds in the stego image. The hackers would not be able to retrieve information inside the image. The information can only be retrieved from the stego image with the system (MoBiSiS) installed in the mobile and the secret key for the image. Figure 5 shows the option of sending the stego image via MMS or Email.

IV. RESULTS AND DISCUSSION

The functionalities of MoBiSiS are then tested using various images. Figure 6 shows some of the cover images that have been used for testing. These images are used to embed the secret message, send the stego images via MMS and retrieve the secret message**.** Note that the images are used to test the steganography algorithm used in MoBiSiS.

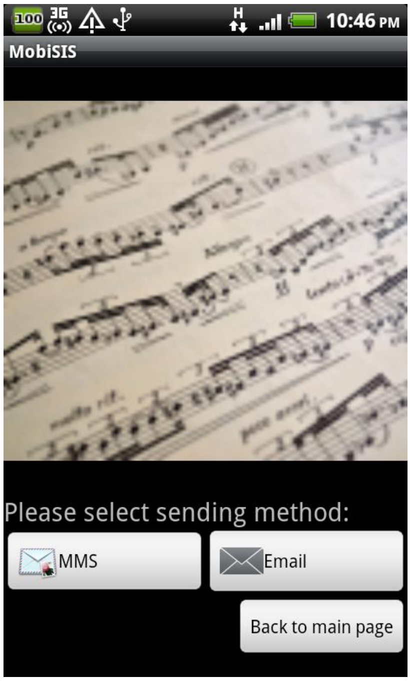


Figure 5. Option to send the stego Image

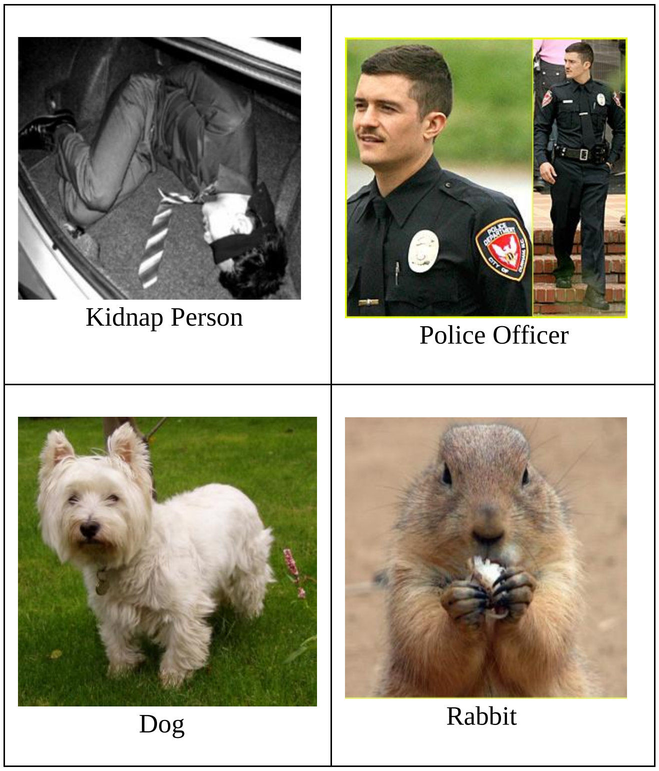


Figure 6. Images that are used for Hiding Data

We also tested the stego image for its PSNR (Peak signal-to-noise ratio). PSNR is a standard measurement used in steganograpy technique in order to test the quality of the stego images. The higher the value of PSNR, the more quality the stego image will have.

If the cover image is *C* of size *M x M* and the stego image is *S* of size *N x N*, then each cover image *C* and stego image *S* will have pixel value *(x, y)* from *0* to *M-1* and *0* to *N-1* respectively. The PSNR is then calculated as follows:

$$PSNR = 10.\log_{10}\left(\frac{MAX^2}{MSE}\right) \qquad (1)$$

where

$$MSE = \frac{1}{MN}\sum_{x=0}^{M-1}\sum_{y=0}^{N-1}(C(x,y) - S(x,y))^2$$

Note that MAX is the maximum possible pixel value of the images. For example, if the pixels are represented using 8 bits per sample, then the MAX value is 255.

If the stego image has a higher PSNR value, then the stego image has better quality image. Table II shows the PSNR value for stego images in Figure 6. The PSNR is calculated using the equation of PSNR in (1).

TABLE II. The PSNR Value of Stego Images

Image	PSNR for Stego Image
Kidnap Person	70.7586
Police Officer	69.0479
Dog	74.6493
Rabbit	72.6493

Based on Table II, the PSNR value shows that the stego images have higher value, which confirms that the quality of the stego image is still high.

The image file format used for MoBiSiS can be in JPEG, GIF, PNG and BMP format which supported by mobile application. However, since MMS is only compatible in JPEG image file format, the generated stego images are in JPEG format. Note that, MMS only support to send image size that less than 30 Kilo Bytes (KB) to maintain its actual size and pixel. Hence, the proposed application compresses the image which larger than 10 KB in order to generate a stego image which will not exceed 30 KB. However, Huffman Encoder assigns shorter codes for characters that appear more often and longer codes for characters that appear less often. Thus, the shorter code improves the capacity of hiding character inside the stego image. To increase as much as characters that can be hidden, zip technique is used to reduce to total size of file and to enhance the security of the file. Table 3 shows the comparison of different image file format and different image size by using MoBiSiS. These JPEG, GIF, PNG and BMP images are used as cover images to encode the zipped file within it.

V. CONCLUSIONS

This paper discusses an android-based application named MoBiSiS (Mobile Steganography Imaging System). MoBiSiS has been developed using the steganography algorithm proposed in [1]. The algorithm used has been tested in term of the quality of the stego image and its PSNR value. The application of steganographic algorithm has been enhanced to mobile application. MoBiSiS can be used by users who want to hide the data inside the image without revealing the data to other parties. MoBiSiS maintains privacy, confidentiality and accuracy of the data.

ACKNOWLEDGMENT

This research is supported under the Fundamental Research Grant Scheme (FRGS) Vot 0738.

REFERENCES

[1] Rosziati Ibrahim and Teoh Suk Kuan (2011), Steganography Algorithm to Hide Secret Message inside an Image, Journal of Computer Technology and Application 2 (2011) 102-108.

[2] Chen M., Memon N., and Wong E.K. (2008). Data Hiding in Document Images. In Nemati H. (Ed.). *Premier Reference Source – Information Security and Ethics: Concepts, Methodologies, Tools and Applications*, Volume 1, Chapter 1.32. New York: Information Science Reference. pp. 438-450.

[3] Lou D.C., Liu J.L., and Tso H.K. (2008). Evolution of Information – Hiding Technology. . In Nemati H. (Ed.). *Premier Reference Source – Information Security and Ethics: Concepts, Methodologies, Tools and Applications,* Volume 1, Chapter 1.32. New York: Information Science Reference. pp. 438-450.

[4] Schneider (2000). *Secrets & Lies*, Indiana:Wiley Publishing.

[5] Cole E. (2003). *Hiding in Plain Sight: Steganography and the Art of Covert Communication.* Indianapolis: Wiley Publishing.

[6] Jahnke T. and Seitz J. (2008). An Introduction in Digital Watermarking Applications, Principles and Problems. In Nemati H (Ed). *Premier Reference Source – Information Security and Ethics: Concepts, Methodologies, Tools and Applications,* Volume 1, Chapter 1.42, New York: Information Science Reference. pp. 554-569.

[7] Warkentin M., Schmidt M.B., and Bekkering E. (2008). *Steganography and Steganalysis.* Premier reference Source – Intellectual Property Protection for Multimedia Information technology, Chapter XIX, pp. 374-380.

[8] El-Emam N.N. (2007). *Hiding a Large Amount of Data with High Security using Steganography Algorithm.* Journal of Computer Science 3 (4), pp. 223-232.

[9] Chen P.Y. and Wu W.E. (2009). *A Modifed Side Match Scheme for Image Steganography.* International Journal of Applied Science & Engineering 2009, 7, 1:53-60.

[10] Chang C.C. and Tseng H.W. (2004). *A Steganographic Method for Digital Image using Side Match.* Pattern Recognition Letters 25 2004, pp. 1431-1437.

[11] Wu P.C. and Tsai W.H. (2003). *A Steganographic Method for Images by Pixel-Value Differencing.* Pattern Recognition Letters 24 (2003), pp. 1613-1626.

[12] Rosziati Ibrahim and Teoh Suk Kuan, (2010). *Steganography Imaging System (SIS): Hiding Secret Message inside an Image,* Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering and Computer Science 2010, WCECS 2010, 20-22 October, 2010, San Francisco, USA, pp. 144-148.

[13] Jain, Y. K., Kumar, R., and Agarwal, P. (2011). Securing data using jpeg image over mobile phone. *Global Journal of Computer Science and Technology*, 11(13), pp. 5-6.

TABLE III. Comparison of different image file format in different image size

IMAGE FILE FORMAT	**FILE SIZE**			**HIDE MESSAGE**	**RETRIEVE MESSAGE**	**RETRIEVE MESSAGE AFTER MMS**	**DISTORTION**
	COVER IMAGE	ZIPPED FILE	STEGO IMAGE				
JPEG	3.06 KB	179 bytes	3 KB	√	√	√	No
JPEG	3.06 KB	223 bytes	2 KB	√	√	√	No
JPEG	3.06 KB	236 bytes	Failed	—	—	—	—
JPEG	9.08 KB	179 bytes	16 KB	√	√	√	No
JPEG	9.08 KB	223 bytes	16 KB	√	√	√	No
JPEG	9.08 KB	236 bytes	Failed	—	—	—	—
GIF	3.08 KB	179 bytes	2 KB	√	√	√	No
GIF	3.08 KB	223 bytes	2 KB	√	√	√	No
GIF	3.08 KB	236 bytes	Failed	—	—	—	—
GIF	9.05 KB	179 bytes	10 KB	√	√	√	No
GIF	9.05 KB	223 bytes	10 KB	√	√	√	No
GIF	9.05 KB	236 bytes	Failed	—	—	—	—
PNG	3.11 KB	179 bytes	Failed	—	—	—	—
PNG	9.08 KB	179 bytes	3 KB	√	√	√	No
PNG	9.08 KB	223 bytes	3 KB	√	√	√	No
PNG	9.08 KB	236 bytes	Failed	—	—	—	—
BMP	3.05 KB	179 bytes	1 KB	√	√	√	No

Development of a Virtual Input Device Using Stereoscopic Computer Vision to Control a Vehicle in a Racing Game

Thiago Ribeiro de Azeredo
Universidade Candido Mendes
Campos dos Goytacazes, Brazil
E-mail: thiagoribeiro@gmail.com

Italo de Oliveira Matias
Universidade Candido Mendes
Campos dos Goytacazes, Brazil
E-mail: italo@ucam-campos.br

Weverson Machado de Oliveira
Universidade Candido Mendes
Campos dos Goytacazes, Brazil
E-mail: weversonmachado@yahoo.com.br

***Abstract*—Nowadays, in the technological world is not hard to see the growing concern about the interaction between users and electronic devices, where the dynamism and the usability have been a decisive factor in the design of new projects and where new developments with ability to increase the experience of the users are in focus. In this light, a system was developed being capable of performing the interaction between the user and the computer using cameras as input devices. It was specifically designed to allow a car in a racing game be controlled only by the user's hand movements without use of markers, where it is possible to turn left, to turn right, accelerate and to brake only moving the hands. For this, haar classifiers, stereoscopic vision and video card programming were used. A racing game was also created to perform tests and validate the proposal. All work was done on Linux environment using C++ with OpenCV , Ogre3D, ODE and CUDA libraries. The system called "virtual wheel" proved satisfactory, having good quality and speed of response, even on a home computer.**

Keywords-stereoscopic vision; human-computer interface; racing game; computer vision.

I. Introduction

The electronics are increasingly present in our day by day and how to interact with them is a crucial factor for their use to be increasingly simple, practical and functional. The desire to facilitate the use of devices has contributed to a significant increase in agility and dynamism to the user, making it easier to perform tasks. Another important factor is the aid digital inclusion of those who, regardless of reason, have difficulty dealing with keyboards, mice and joysticks.

It is easy to see that computing has evolved dramatically in recent years, where the devices have become increasingly faster and smaller operating systems and seeking to deliver more features and flexibility for the user. But usability has always been restricted to the keyboard and mouse that despite undergoing improvements, they had no substitutes. Today, new forms of access are being explored. The new touch screens, and tilt position sensors, voice commands and computer vision are drawing the attention of the industry in several areas such as video games, smartphones, tablets, automobile industry, among others.

Although this theme is not new, the time consumed by the algorithms was an impediment to its use in an application that needs answers almost immediately. Today, some techniques have evolved into the performance, progress has been made in processing power of home computers and video cards have a programmable interface idle low complexity for mathematical computations. These facts combined with the low cost of good quality cameras contributed to the popularity of this line of research, although some are still complicating factors for the techniques of computer vision and therefore compromise the quality of results, such as the background unstable lighting conditions. These benefits can be seen at [1], that presents a technique for fast objects recognizing in images, [2] and [3] shows techniques for obtaining the distance of objects creating disparity maps.

Being a computer environment for experimentation with new forms of interaction, this study aims to create a computer system capable of converting movements performed with the hands free in front of cameras, control actions in a car. Namely: Accelerate, brake, turn left and turn right. This process must occur without markings or devices in the hands of the user, and have to be quick enough to get smooth movements. This system is called a virtual steering wheel.

The computer vision techniques employed are even more interesting is observed that serve as basis for several important tasks, such as aid for the visually impaired locomotion, control of robots and automatons touchless control interface [4] (very useful for environments such as hospitals).

Section 2 will describe the steps to prepare the project, as the images capture , background substract, hands detection, obtaining the hands depth. In the end, the performance analysis. In Section 3 the virtual racing simulator creation as a test environment will be explained, and the proposal used to integrate it into the project will be described. Section 4 will present conclusions and suggestions for future work.

II. Building the "Virtual Wheel"

As previously reported, the paper proposes the development of a system capable of capturing images generated by cameras and, online, treat them and turn them into motion. Specifically in this case was chosen to identify the hand's closed fist and its conversion into motion, controlling the car with four functions: turn left, turn right, accelerate and brake.

All these functions with the degree of sensitivity adjusted, allowing a precise control.

In order to allow this project, the "virtual wheel" was created using the OpenCV library [5] (Open Source Computer Vision) and CUDA [6]. OpenCV is cross platform and consists of more than 2500 already optimized algorithms for capturing, creating, editing, processing and obtaining information from images. Since CUDA is a library developed by *NVIDIA* company that allows the development of algorithms for mathematical calculation can be run on video cards.

A. Image Capture

To make capturing the images you must first make a selection of cameras. This factor is vital to the process due to several variables that relate to this device, such as level of distortion of the lens opening angle lens, focusing system, quality shutter, shutter speed, among others.

For this work the camera chosen was the ps3 eye (Fig. 1). Created by *Sony* for the play station 3, was born with the goal of combining quality with performance. This camera has a lens opening 56 degrees, a microprocessor capable of sending images without compression, allowing for better utilization and capacity to generate 60 frames per second capture frames of 640x480 pixels and 120 frames per second at 320x240 pixels. Its speed was the primary factor for the choice, because the response time of the system must be transparent to the user. Thus, two cameras were purchased to be used in conjunction with this system.

Figure 1. Image of Sony PS3 Eye

The image size used in this study was to 640x480 pixels, in order to obtain large amounts of information to work.

At each step an image of each camera is captured and converted into two other images in grayscale. One with the average color and the other containing only the red channel. Each image will have a utility in the following steps.

B. Background Substract

The subtraction of the background consists of making the algorithm learn what was already on the scene and what is new. In this case, this process is done so that the whole environment is eliminated and only the user is recognized. This is done to eliminate irrelevant information for the process.

The process of background subtraction used is quite simple. Using only the red channel of the camera, the user should avoid being exposed during the first 100 frames, at which time the catch is made. The average of the images from each camera is generated separately, taking then the average color of each pixel of the same to the left and right image. These images are used as average of comparison, where the color of each pixel of a new image is compared in the same coordinate with the image of the average so that if the variance between the color exceeding a predetermined threshold, this new color is considered something new on the scene. So is created a mask for the image.

C. Hands Detections

Object detection systems use feature detection algorithms, there are several techniques that can be employed for this task, but the detector created by Vioja-Jones has been widely used recently [1]. The Viola-Jones is able to detect objects with precision, high accuracy rate, low false positive rate and low computational cost. The algorithm consists of 3 parts: The first is the representation of the image in a feature space based on the Haar filter. The second is the assembly of a classifier based on Boosting able to select the most relevant features. The third part concerns the cascade combine classifiers to ensure good performance and processing speed (Fig. 2).

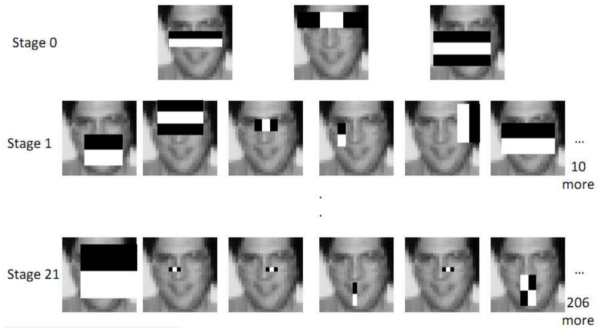


Figure 2. Example identifying characteristics using the algorithm of Viola Jones removing [1]

Both the speed of recognition as the quality of the final result of the Haar like are affected by factors such as:

- Number of images used in training;
- Images size;
- Number of features used for recognition;

So these parameters must be adjusted to achieve a good quality without losing performance. This work requires the effort to make several attempts to reach the goal, which in this case occurred six times.

The training process starts with getting the images. In many areas of science you can get them in stock photos public which are widely used in research projects. This is useful because whith it is possible for example, compare the quality of work using the same images. But there was no success in finding images containing fists stock photos, thereby obtaining such images was made from photographs.

There was developed a small program with the function of capturing a sequence of frames of the cameras and converting each individual image files. During this process, the program recorded 250 frames from each camera, then completing 500 paintings. This little program was used twice to take pictures of the user on different days, then generating 1000 photographs. Of these 1000, 306 were randomly selected to be part of the training process, of which 578 examples were generated for left and right hands, some samples were discarded because of poor quality. Fig. 3 examples a photo used and Fig. 4 shows some hands selected.

In the quest for improved quality of recognition, the equalization of each image in the training and recognition was used.

These hands were then placed for training over 94 background images. These images were selected to ensure a diversity of environments and consist of rooms, textures and people. It is worth noting that all photographs of persons used in this step became an issue in order to remove his hands. Among these images, 29 were obtained using the camera Nikon Coolpix S4000, 5 were obtained from the camera ps3 eye, while the other 59 were obtained from images free databases.

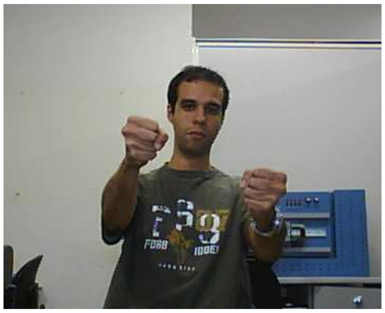


Figure 3. Example of photography used for training

In generating the final version, the training process took about 8 hours, where 42 iterations were performed for a total of seven features were identified to obtain a good quality results.

After obtaining the location of both hands on the image, the center point of each is identified and used for the discovery of the angle which they form, as showed on Fig. 5.

This angle is then adjusted to suit the sensitivity of the system and then is used to rotate the wheel.

To turn possible to visually evaluate the behavior of the system, an ellipse was displayed on the screen with the second function of rotating the angle generated.

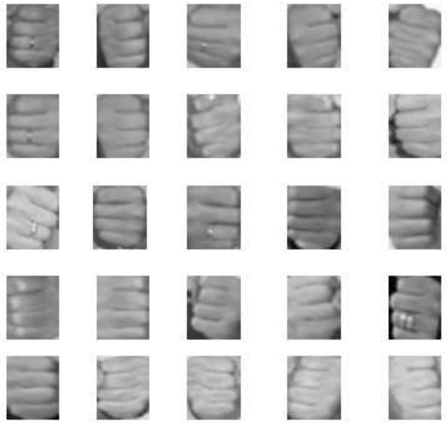


Figure 4. Example of hands identified

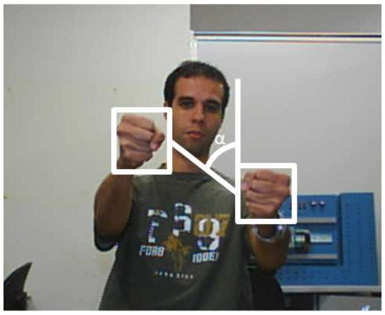


Figure 5. Proof of identification of the angle between the hands

D. Obtain the deph of hands

The term comes from stereoscopic two Greek words that represent "vision" and "solid" and the origin of this knowledge area can be rescued at least the year 1838.

In general, the depth information is obtained from a system for processing stereoscopic (Fig. 6) three main problems: calibration correlation and reconstruction. In the calibration seeks to determine the parameters which describe the acquisition system used. The correspondence problem is to determine which element in the captured image from a point of view corresponds to a given element in the captured image from another point of view. In turn, by rebuilding seeks to retrieve information from depth based on parameters obtained in the calibration step and pairs of corresponding points obtained in the step of matching.

For this study, we used the algorithm proposed by [3] that is capable of generating a disparity map quality and extremely fast using a GPU.

The GPU (Graphics Processing Unit) is responsible for all calculations to generate graphics on electronic devices, especially three-dimensional environments. Initially created to offload the CPU, allowing streamline tasks such as texture mapping and rendering polygons, received more functionality over time, such as rotation and translation of vertices and

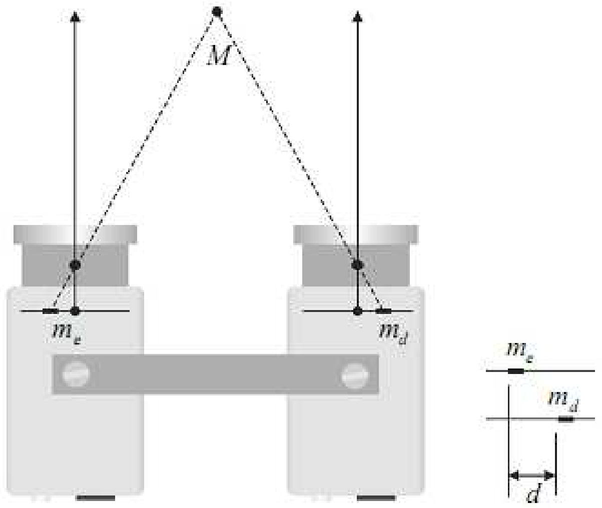


Figure 6. Simplified model of stereoscopic vision([2])

now support lighting calculations and processing of vertices, and direct manipulation of pixels. These features are designed to grow with the gaming industry, not only computers but also videogames, working with similar architecture.

As you can see in Fig. 7, the architecture of the central processing unit (CPU) includes communication with many other structures and has a throughput small compared with the GPU. The difference becomes large when compared to the speed of memory access.

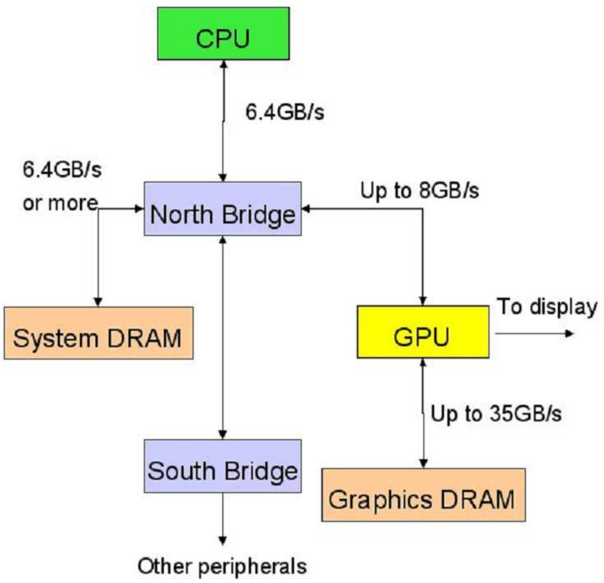


Figure 7. Defining the outer structure of the CPU and GPU capabilities with maximum data transfer. [7]

The Table 1 compares two of the most modern and powerful devices found in Brazil today, and hence its high price. You can clearly see that the *NVIDIA* GTX580 is actually better prepared for the scenario of mathematical calculations that the processor Intel core i7 3930k.

Table I
COMPARISON OF CALCULATION POWER AND PRICE BETWEEN CPU AND GPU. INFORMATION FROM [8], [9], [10], [11], [12]

	Intel Core i7-3930K	*NVIDIA* GTX580
Kernels	6 (12 threads)	512
GFLOPS	105	1.581
Memory Speed	21 GB/s	192.4 GB/s
Cache	12 MB	768 KB
Frequency	3.2 Ghz	772 Mhz
Price (Brazilian Real)	R$ 2.399	R$ 1.899

Knowing that calculations images are nothing more than multidimensional arrays with numeric values and that the procedures for rotation, translation, transformation and lighting are just mathematical equations is easy to understand as it was thought the creation of the CUDA (Compute Unified Device Architecture). This, in turn, was created in 2007 by *NVIDIA* and became a powerful platform to perform mathematical calculations in parallel directly to the video card, using high-level language. Unavailable until now, where it was necessary to program in assembly. Process that began around 1998.

The technique of [3] was compared with the literature to assess the quality of their results as seen in Fig. 8. This technique was then used in the captured images to provide the map of disparities. With prior knowledge of the location of the player's hands, together with the removal of the background it was possible to determine the distance of the hands using the average of the colors within the bounding box, each hand. This value is then framed between -1 and 1, where numbers greater then 0 means speed increases and smaller then 0 means speed decreases. The variation occurs in relation to its proximity to either 1 or -1. Thus, the farther the hands are the camera, the lower the speed and the closer, the greater the speed.

E. Performance

At the end of development, the system was tested and it was possible to obtain the execution time. At this point, each loop step took an average of 128ms to run.

Table 2 shows the time of each step separately and in the order they are executed. Looking for increase the speed of execution, an analysis was made in the steps used in order to locate those who consume reasonable time and that can run in parallel with others, in this way could make this time was absorbed without affecting the system structure.

It was identified that the detection step of hands would be a good candidate for this change. This was then ported to a

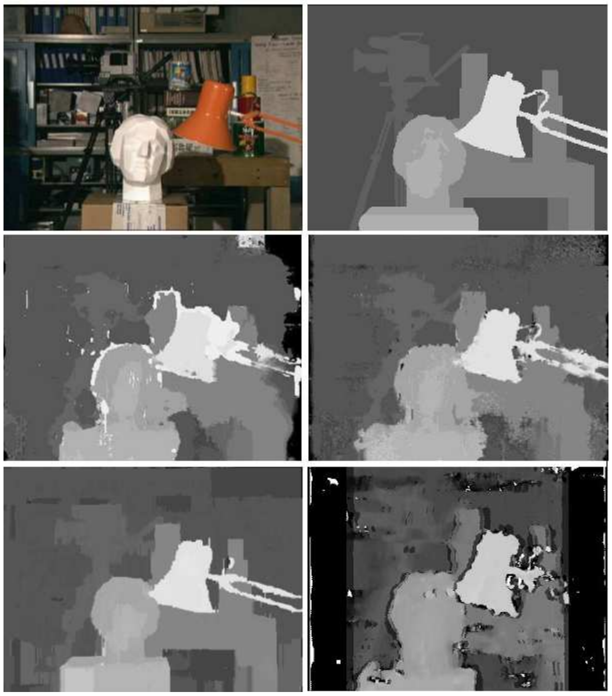


Figure 8. Comparison of disparity maps. a) original image b) expected result; c) result from[13]; d) result from [14]; e) result from [15] f) algorithm from[3] used on this work.

Table II
TIME SPENT BY EACH PROCEDURE

Step name	Time(ms)
Get new frames from cameras	40
Images preprocess	10
Hands detection	37
Disparity calculation	40
Move calculation	1

thread that runs in parallel whereas the main runs to obtain the disparity map, and then its time taken up by the main process steps.

Thus has a gain of 37ms, lowering the average time of each iteration to 91ms on average, generating 11 frames per second. This change has backpack fluidity for the result, but it was observed that there is another step where the same strategy could be applied, the image capture cameras. The capture process alone consumes 40ms, since each camera needs 20ms to return the image. With this further modification of a gain medium in 16ms, then lowering the average of the iteration to 75ms, which represents 13.3 frames per second. Compared to the initial time of 128ms with 7.8 frames per second, was obtained a gain of 41% and a degree of fluidity pleasant.

III. THE GAME

To evaluate the quality of results generated by the virtual wheel, a computational experiment was developed. The virtual wheel was integrated with the virtual racing car created by [16], [17], where its gameplay could be checked in a subjective way, after the user experience should be taken into account. The Fig. 9 examples the experiment views.

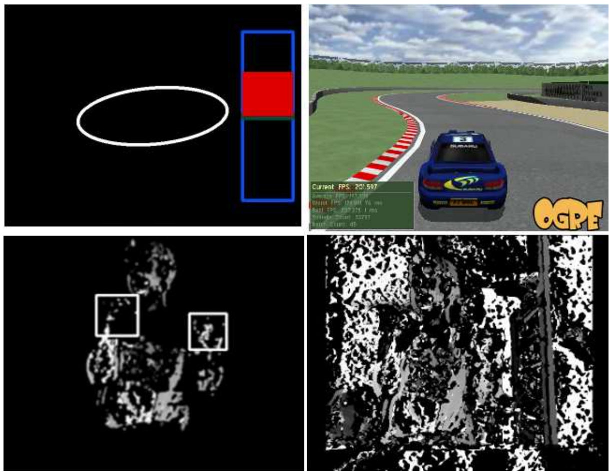


Figure 9. Representation of four views (a, b, c, d). a) represents the steering angle on the ellipse (inverted relative to the image) and acceleration based on the distance of the hands for the cameras at the bar; b) presents the response of the car in the game to the command; c) Represents the disparity map with background removal and hands found; d) Represents the disparity map of the moment.

The virtual scenario used here was developed in 2005, in Windows, in C + +, using Ogre3D [18] and ODE [19] engines. Since the virtual wheel was developed in Linux and libraries Ogre3D and ODE suffered several updates over the past six years, so it was necessary to translate the project to linux environment and update it to use the latest versions of libs applied. In this previous work, the whole project was used except the autopilot, not be relevant to the project scope.

As the project was running architected to allow the inclusion of new controls for the vehicle, a new control was implemented to integrate with the virtual wheel. This was developed using the network protocol UDP (User Data Protocol) which fits better in the proposal to be extremely fast due to the failure to implement controls and safeguards in existing TCP (Transmission Control Protocol). The use of network protocol for integration also allows new environments to integrate virtual driving future.

On average, 82% of the cycles was the detection of two hands, which during the game guarantees a quality experience satisfactory. To maintain control of the vehicle frame in which at least one hand is not located in the last value of the detection is sent as well as the last value of the acceleration and braking. As in a common environment lighting is not constant, small variations occur in the map of disparities any time by changing the values of acceleration, but are not observed during the game.

IV. Conclusion and Future Work

The algorithms used to develop this work demonstrate adequate for this task. The viola-jones algorithm requires an effort to match the quality and performance using trial and error. The algorithm [3] proved to be extremely fast to generate the disparity map, although its quality is not fitted with other techniques from the literature, it is enough to reach the proposed objective. The results in uncontrolled environments demonstrate that the algorithm is quite robust. It is also possible to treat the result to minimize noise, improving the final quality.

The 13 frames per second achieved were sufficient to pass the feeling of immediate response to the player, allowing a good gameplay and spending confidence.

The atmosphere can be adapted to new features such as control of robot arm, moving the mouse, among others.

References

[1] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '01)*, vol. 1, pp. 511–518, 2001.

[2] M. E. Stivanello, E. S. Leal, N. PALLUAT, and M. R. STEMMER, "Desenvolvimento de uma biblioteca para sistemas de visao estereoscopica para robotica movel," *VIII Conferencia Internacional de Aplicacoes Industriais*, 2008.

[3] D. Gallup, J. Frahm, and J. Stan, "Real-time local stereo using cuda," *NVIDIA Research*.

[4] J. P. Wachs, H. I. Stern, Y. Edan, M. Gillam, J. Handler, C. Feied, and M. Smith, "A gesture-based tool for sterile browsing of radiology images," *Journal of the American Medical Informatics Association*, vol. 15, pp. 321–323, 2008.

[5] (2012, Fev.) Opencv. [Online]. Available: http://opencv.willowgarage.com/

[6] (2012, Fev.) Cuda. [Online]. Available: http://www.nvidia.com/cuda

[7] A. Cuno and J. R. M. Vianna. (2011, Dec.) UFRJ. [Online]. Available: http://www.lcg.ufrj.br/Cursos/GPUProg/gpuintro

[8] (2011, Dec.). [Online]. Available: http://www.waz.com.br/produtos/101319

[9] (2011, Dec.) Techpowerup. [Online]. Available: http://www.techpowerup.com

[10] (2011, Dec.) Intel. [Online]. Available: http://ark.intel.com/products/63697/Intel-Core-i7-3930K-Processor

[11] (2011, Dec.) Hexus. [Online]. Available: http://hexus.net/tech/reviews/graphics/29509-gigabyte-geforce-gtx-550-ti-oc-graphics-card-review/

[12] (2011, Dec.). [Online]. Available: http://www.waz.com.br/produtos/101385

[13] R. Yang and M. Pollefeys, "Multi-resolution real-time stereo on commodity graphics hardware," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '03)*, vol. 1, p. 211, 2003.

[14] R. Yang, M. Pollefeys, and S. Li, "Improved real-time stereo on commodity graphics hardware," *Computer Vision and Pattern Recognition Workshop*, 2004.

[15] J. C. Kim, K. M. Lee, B. T. Choi, and S. U. Lee, "A dense stereo matching using two-pass dynamic programming with generalized ground control points," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 2, pp. 1075–1082, 2005.

[16] T. R. Azeredo, P. S. VIEIRA, A. A. S. NETO, and I. O. MATIAS, *Inteligencia Artificial aplicada a tomada de decisao em jogos eletronicos.*, Bacharel em Ciencia da Computacao, Universidade Candido Mendes Monografia, Aug. 2006.

[17] M. R. Junior, R. C. B. P. Gomes, S. F. P. P. Judice, and I. O. Matias, *Simulacao computacional aplicada a jogos eletronicos*, Bacharel em Ciencia da Computacao, Universidade Candido Mendes Monografia, Aug. 2006.

[18] (2012, Fev.) Ogre3d. [Online]. Available: http://www.ogre3d.org

[19] (2012, Fev.) Ode. [Online]. Available: http://www.ode.org

Automated Infarction Core Delineation

Comparison between using Cerebral Blood Volume and Perfusion Blood Volume Maps

Petr Maule, Jana Klečková
Department of Computer Science and Engineering
University of West Bohemia
Pilsen, Czech Republic
e-mail: pmaule@kiv.zcu.cz, kleckova@kiv.zcu.cz

Vladimír Rohan, Radek Tupý
Department of Neurology
The University Hospital in Pilsen
Pilsen, Czech Republic
e-mail: rohan@fnplzen.cz, tupyr@fnplzen.cz

***Abstract* — This article is focused on development of a tool supporting physicians with an appropriate treatment decisions at patients with acute ischemic stroke. The automated tools for infarction core area delineation could provide important information about the volume of the infarction core. This article describes such automated method results used on both cerebral and perfusion blood volume computed tomography maps compared with manual infarction core delineations made by two physicians.**

Keywords - ischemic stroke; infarction core delineation; perfusion blood volume maps; computed tomography

I. Introduction

This paper deals with acute ischemic strokes, which are the third leading cause of death and the first leading cause of disability in population over 60 years old. Patients undergo several types of computed tomography (CT) examinations and based on the results appropriate treatment follows. Possible treatment is a thrombolytical treatment, which can not be indicated if the patient exceeds certain level of the volume of the infarction core. Studies like [1] and [2] deal with finding of the best threshold value for the infarction core. The largest study [1] at 130 patients found threshold at 2 ml/100g using cerebral blood volume maps (CBV) provided by Perfusion Computed Tomography (CTP) examination. The threshold can vary from patient to patient and the threshold is also dependent on the used method.

Thrombolytical treatment infarction core volume level limitation must be evaluated from the whole brain but the CTP examination is often limited in the covered area. Different methods are to be used for the whole brain infarction core volume evaluation.

Several studies described process of construction of so called perfusion blood volume maps (PBV) [3], which are constructed from non contrast computed tomography examination (NCCT) and computed tomography angiography (CTA) [4]. It expresses also the blood volume level as CBV but the provided information is not the same. CBV maps are computed from a series of images observing spreading of the contrast material while the PBV maps are constructed by adjusted subtraction of two values – densities with and without contrast material, which is also depended on the quantity of the contrast material in the time of the CTA data acquisition [3]. Another difference can be seen also in the slice thickness of the used CTA and NCCT examinations. The NCCT slice thickenss is often about 5 mm while the CTA slice thickness can be 1 mm. This difference increases partial volume impact in the final PBV maps.

This article shows comparision of automated infarction core delineation method using CBV maps and PBV maps. First, we describe the used material and its adjusting for our use. Next, the method itself is introduced and results are presented. Discussion summarizes our findings and proposes future steps.

II. Material

In cooperation with the University Hospital in Pilsen we had an access to 24 anonymized examinations from 12 patients with a supratentorial stroke. 12 examinations were CBV maps and 12 PBV maps. Both examination types are with the whole brain coverage acquired on dual-source CT (Somatom Definition, Siemens Healthcare, Forchheim, Germany) and PBV maps were constructed using commercial software Siemens syngo Neuro PBV. We also had available manually delineated best opinion prediction of infarction core provided by one radiologist and one neurologist experienced in CT evaluation. CBV and PBV maps including the manual delineations were mutually registered. PBV and CBV maps were available in DICOM format and after the mutual registration they had dimensions 512x512 with 44 images per examination with used units *ml/l*.

We can refer the two physicians as Ph1 and Ph2. Ph2 in one patient's examination did not mark any area meaning the opinion that there is no infarction core at all. The average mutual correspondence between the findings of the two physicians expressed by Matthews correlation coefficient is 62.09% for CBV maps and 56.90% for PBV maps.

III. Segmentation Adjustement

The CBV examinations were already segmented by the instrument and the bones and cerebral ventricles were removed. The PBV examinations have already segmented bones and cerebral ventricles but probably because of the partial volume effect the steep values changes persist on the two different tissue types borders. Both examination types contain a rests from non-ideal segmentation at the bottom part under the skull base, which would significantly influence the following methods' results. Because of this reason we performed segmentation adjustement step, which selects one image as a divider and for further processing are used only images from the top to the divider image. All images below are ignored. We used following technique for finding the divider image.

A. Divider Image Detection

If we refer one examination image as Im_i where $i=0$ corresponds to the top of the examination (in the sense of the top of the head) and i_{max} to the last image (in the sense of the bottom of the head), then, for every image, we can calculate following equation:

$$C_i = C_i^{IN} / C_i^{OUT}. \quad (1)$$

C_i^{IN} means count of voxels of the image i, which values are above zero, C_i^{OUT} means count of voxels of the image i, which values are below or equal to zero (except voxels with value -1024 representing the outer space).

One of the image becomes the divider image according to the C_i value. For this purposes, we start comparing C_i values from the bottom image with $i=max$ towards the upper image with index $i=0$. The first image with index i, which satisfies condition $C_i<Th$ becomes the divider image. We found optimal Th value for CBV examination *2.0* and for the PBV examination *1.0*.

This segmentation adjustment step (example in Fig. 1) is just auxiliary for following method comparison to remove posterior fossa from further processing. We evaluated that this segmentation adjustment step is successful in 95.8% (23) cases from all 24 examinations in the sense of providing enough area by the segmentation adjustment step to be possible to delineate enough area to cover 100% of manual infarction core tracking. The only one unsuccessful case reduces the area but it is still possible to find 96.3% of the manually marked infarction core volume.

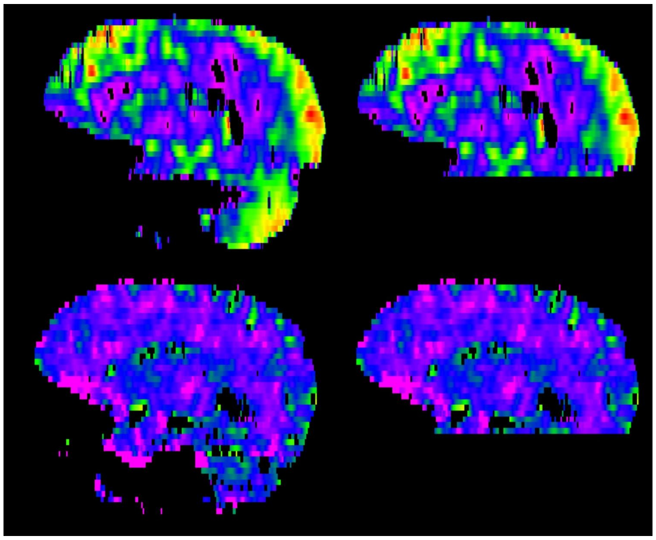


Figure 1. CBV (upper row) and PBV (bottom row) segmentation adjustment step, original examination on the left and adjusted examination on the right using sagittal views. Displayed values range is from 0 ml/l to 150 ml/l using a common color-scale.

IV. Automated Delineation Method

Details of the image processing method used in study [1] are not presented enough to reproduce it by own prototype software. PBV and CBV maps, which we have available, are also provided by commercial software and so that we tried to use simple thresholding for processing it. We found a need for the examination preprocessing and a need for focusing to area of the infarction core to avoid false-positive voxels, which would be included into the whole volume of the infarction core.

We developed a prototype software, which processes CBV and PBV examinations. The automated infarction core delineation is based on examination's preprocessing, and following thresholding and it is focused to delineate only the infarction core area and thus to reach higher specificity.

A. Image Preprocessing

We used preprocessing examination in a form of edge preserving smoothing as mentioned in [1]. We used Curvature Anisotropic Diffusion smoothing with the usually used parameters defined by [5]:

Time step:	*0.0625*
Conductance:	*3.0*
Iterations:	*5*

B. Thresholding

After the preprocessing step we tried to find the best threshold corresponding to the highest specificity. We used thresholds from *0 ml/l* to *21 ml/l*. The value *21 ml/l* corresponds to the found threshold value from [1], which is *2 ml/100g*. The used increment was *1 ml/l*. The thresholding process is simple in CBV examinations - all values from 0 to threshold are marked as infarction core but in PBV examinations we faced to high amount of negative values, which belongs to the imperfect bone and cerebral ventricles segmentation but also to supposed infarction core areas. In order to avoid marking such voxels we reduced them by lower limitation of thresholding to arbitrarily used value *-50*.

C. Infarction Core Area Selection

We faced to too many false-positives voxels and thus we grouped all adjacent voxels, after the thresholding step, into groups and we discarded all groups instead of the largest one, which we believe to be the one corresponding to the infarction core. The disadvantage of this step is a possibility that the largest group does not correspond to the infarction core while the correct group was discarded since it is smaller than the largest group.

D. Match Evaluation

We have two patterns of how the automated findings should ideally look like. We compare our automated findings according to the patterns separately. Firstly, for the CBV examinations, and secondly, for the PBV examinations. Let us call the match between CBV automated findings and findings of Ph1 as CBV-Ph1, similarly CBV-Ph2 and also PBV-Ph1 and PBV-Ph2.

For each match we can evaluate 4 voxel counts. TP (true positive) - increased when both voxels were marked as infarction core, by physician and also by automated method, TN (true negative)– both voxels were marked as non-infarction core, FP (false positive) – the automated method marked voxel as infarction core while the physician marked the same voxel as non-infarction core and FN (false negative) – automated method marked voxel as non-infarction while physician as infarction core.

Tables I and II present sensitivities and specificities as an average values from all 12 patients excluding those with incorrect match, which count is presented in IM column.

E. Incorrect Match

We call incorrect match the case when the automated method findings and the physician's findings have marked no common infarction core voxel, it means TP = 0 while FP > 0.

V. Results

Results of the described method are presented in Table I for CBV examinations and in Table II for PBV examinations. Fig. 2 demonstrates using of different threshold values on PBV examination. The specificities seem to be high but lets consider that all examinations have 512x512x44=11534336 voxels and for example the average infarction core area from manual marking of the Ph1 contains 28603 voxels. If the automated method would find all voxels marked by physician but it would mark the 3 times larger area, the specificity would be 99.5%.

We can see that for CBV examinations we can obtain very high specificity, almost 100% (thresholds from *0* to *7 ml/l*) but for the PBV examinations the maximum specificity was found only 99.24378%, which is not satisfying.

VI. Discussion

We believe that the low specificity in PBV images is caused mostly by partial volume effect, which is caused by

TABLE I. CBV Results

Threshold [ml/l]	*Sensit. Ph1 [%]*	*Specif. Ph1 [%]*	*IM[a] Ph1*	*Sensit. Ph2 [%]*	*Specif. Ph2 [%]*	*IM[a] Ph2*
0	2.79	99.98695	8	2.10	99.98542	8
1	21.77	99.97208	6	19.39	99.97134	6
2	19.52	99.99688	4	17.38	99.99704	4
3	26.08	99.99348	3	23.28	99.99424	3
4	35.12	99.98405	3	30.11	99.98456	3
5	41.86	99.97453	3	37.67	99.97327	4
6	43.51	99.96923	2	38.91	99.96935	3
7	52.46	99.94637	3	46.45	99.94888	4
8	57.97	99.90980	3	51.71	99.91527	4
9	61.95	99.88400	3	55.85	99.89189	4
10	65.77	99.85239	3	60.28	99.86398	4
11	75.79	99.33449	3	67.38	99.35707	3
12	73.22	99.11246	2	65.10	99.13822	2
13	78.39	98.82117	1	70.50	98.84521	2
14	83.34	98.42550	1	75.58	98.43674	2
15	86.17	98.03474	1	79.07	98.02557	2
16	88.78	97.58111	1	81.92	97.55204	2
17	91.13	97.03577	0	83.22	96.99077	1
18	92.57	96.54330	0	85.66	96.48031	1
19	93.74	96.03487	0	87.75	95.95049	1
20	94.72	95.51506	0	89.48	95.41960	1
21	95.54	94.93930	0	90.92	94.81682	1

a. Count of incorrect matches from total 12 patients

TABLE II. PBV Results

Threshold [ml/l]	*Sensit. Ph1 [%]*	*Specif. Ph1 [%]*	*IM[a] Ph1*	*Sensit. Ph2 [%]*	*Specif. Ph2 [%]*	*IM[a] Ph2*
0	62.28	99.20641	4	58.97	99.24378	4
1	67.33	99.07244	4	63.35	99.11330	4
2	69.19	98.96140	3	64.01	99.00619	3
3	72.12	98.82390	3	66.53	98.87054	3
4	74.46	98.70119	3	69.10	98.75185	3
5	76.84	98.55058	3	71.73	98.60509	3
6	78.93	98.38614	3	74.08	98.44412	3
7	75.47	98.40800	2	72.15	98.46359	2
8	77.62	98.31482	1	74.71	98.32093	2
9	79.42	98.16158	1	76.89	98.16444	2
10	81.41	97.99354	1	79.36	97.99234	2
11	82.98	97.81751	1	81.37	97.81185	2
12	78.36	97.68276	0	77.15	97.67533	1
13	80.55	97.50674	0	80.08	97.49548	1
14	81.90	97.33071	0	82.00	97.31522	1
15	83.17	97.13386	0	83.70	97.11169	1
16	84.39	96.94731	0	85.24	96.92112	1
17	85.68	96.76089	0	87.12	96.72960	1
18	86.87	96.56792	0	88.41	96.52946	1
19	87.84	96.36111	0	89.65	96.31516	1
20	89.63	95.80393	1	90.92	95.71346	2
21	90.38	95.58968	1	91.79	95.49068	2

a. Count of incorrect matches from total 12 patients different slice thickness of source images and also by the imperfect segmentation of the bones and cerebral ventricles. Because of this reason the PBV examinations contain steeper changes of values and also high amount of negative values especially at two different environment borders including the infarction core. The edge preserving smoothing at least with the used settings is not strong enough to make the infarction core distinguishable by used thresholding.

We also believe that the use of different kind of filters like meaning can be useful for the PBV examinations despite of the loss of details and in combination with a local neighborhood features better results could be obtained.

VII. Conclusion

We presented simple combination of edge preserving smoothing with selecting the largest continuous area, which is considered to be infarction core. Using the thresholding technique we evaluated correspondence between automated method and manual infarction core delineations provided by 2 physicians. We can see that while the same method can in the case of CBV maps provide almost 100% specificity, it is almost unusable in the same form using the PBV maps. In discussion we mentioned our opinion of the low PBV specificity and we proposed our ideas how to improve results on PBV maps.

Acknowledgment

The work presented in this paper is supported by The Czech Science Foundation project 106/09/0740.

References

[1] Wintermark M., et al., “Perfusion-CT Assessment of Infarct Core and Penumbra: Receiver Operating Characteristic Curve Analysis in 130 Patients Suspected of Acute Hemispheric Stroke”, Stroke 2006;37;979-985.

[2] Murphy B.D., et al., "Identification of penumbra and infarct in acute ischemic stroke using computed tomography perfusion-derived blood flow and blood volume measurements". Stroke 2006;37:1771–77.

[3] Srinivasan A., Goyal M., Azri F. A., Cheemum L., "State-of-the-Art Imaging of Acute Stroke", RadioGraphics 2006; 26:S75-S95.

[4] Hamberg L.M., Hunter G.J., Kierstead D, Lo E.H., Gilberto González R, Wolf G.L., "Measurement of cerebral blood volume with subtraction three-dimensional functional CT", AJNR Am J Neuroradiol. 1996; 17: 1861–1869.

[5] Ibanez L., Schroeder W., Ng L., Cates J., “The ITK Software Guide”, Kitware, Inc. ISBN 1-930934-10-6, 2003. http://www.itk.org/ItkSoftwareGuide.pdf [retrieved: April, 2012]

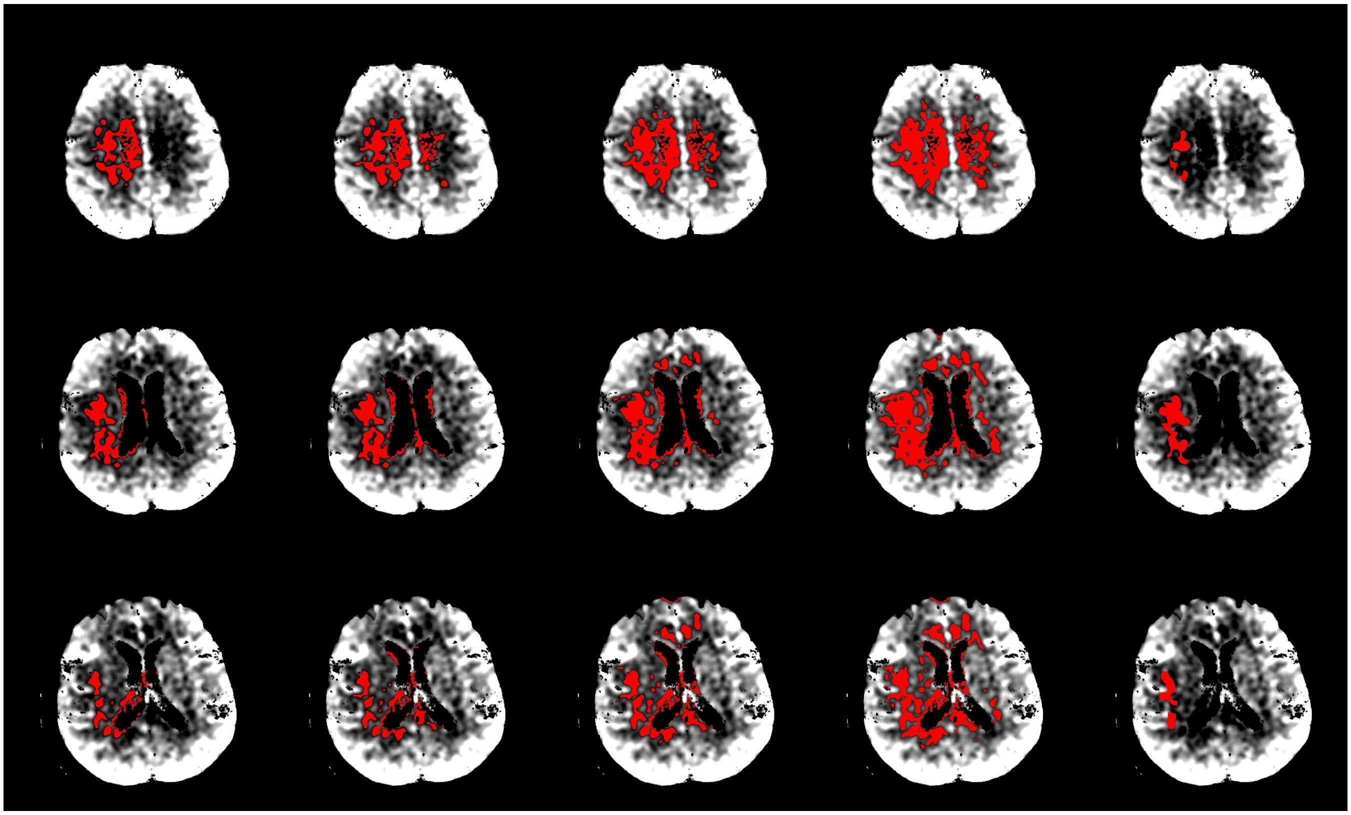


Figure 2. Example of infarction core detections (red color) on PBV examination using edge preserving smoothing. Columns from left correspond to thresholds 0, 2, 7, 12 ml/l and the last column expresses manual tracking by Physician 1. Rows correspond to different locations in the examination. The used lower limit for infarction core -50 ml/l was used.

An Application of the Checkstand Antitheft System by Means of the Image Recognition Technique

Sung Min Hong, Dong Ryoung Seo, Dong Sin Kim, Do Won Hyun and Ju Wook Jang
Department of electronic engineering
Sogang University
Seoul, Korea
e-mail: feverbypassion@hotmail.com, drs@monet1.sogang.ac.kr, freeman@sogang.ac.kr, snatcher@monet1.sogang.ac.kr , jjang@sogang.ac.kr

***Abstract* - This study aims to develop an application to recognize any missing barcode scan despite normal scanning motion by means of the video image processing algorithm in order to prevent an omission of product barcode scanning, intentional or mistaken, by a checkstand. The algorithm that searches for a barcode from a moving object in a video is divided into two steps: the step of extracting the moving object from a video; the step of extracting a barcode of the moving object. The process of extracting moving objects consists of blocking the video frame images, calculating motion vectors of each block, and clustering vectors of similar direction and size in the order. The process of extracting a barcode from a moving object includes the step of changing the color of the moving object image to that of the HSL color model, selecting areas, part of whose block is colorless, for the portion to be scanned for the barcode, and recognizing the barcode when the portion consists of a certain number of blocks in the order. The application has been developed by means of this algorithm for Microsoft Windows. If no barcode scan output is found even though the scan motion was normally implemented, this application will check if the barcode exists and prevents barcode omission from taking place at checkstands of a discount outlet.**

Keywords-Using motion vector; Clustering; Image recongnization; Application implementation.

I. INTRODUCTION

According to the managers of major discount stores, the financial loss due to intentional or mistaken barcode scan omission by a cashier is as much as 2% of the total sales, but there is no system that can effectively supervise the process. In a major discount outlet in Korea, for example, the sales in 2009 reached some 10 trillion (HI Investment & securities co. ltd Research Center, October, 2009), which indicates the loss of more than 200 billion won. In the end, this type of loss results in increase of retail prices and damage to a wide range of consumers. Thus, the need for a system to completely prevent barcode scan omission from happening has been consistently emphasized.

This study, therefore, suggests and develops an algorithm to find a barcode on a moving object in a video with relatively small amount of operation. First of all, adopted was the moving object extracting method suggested previously to find a moving object [1]. This is a method to block each frame in an image and find the motion vectors with the central point of the block as the center. In consideration of the fact that a moving object has motion vectors of the similar direction and size, the moving object is detected by clustering vectors of the similar direction and size. The use of this technique reduces the operation compared to that of finding vectors by means of characteristics or colors. Then the color of the extracted object is changed and the barcode is checked. As the barcode is searched for, not on the entire frame, but on the area of the moving object, this further contributes to the reduction of the operation. This study embodied and verified this application by means of the algorithm above in Microsoft Windows System.

Section 2 explains the way of extracting moving objects from a video, section 3 the way of extracting the barcode from the extracted moving object, section 4 the results of the actual representation, and section 5 concludes the study and presents issues for the future study.

II. SEPARATION OF A MOVING OBJECT FROM BACKGROUND

This section suggests a swift and effective algorithm to detect a moving object. The suggested algorithm block each frame of an image, and then extracts motion vectors with the central point of the block as the center. Extracting motion vectors for each block reduces the operation more than in extracting for all pixels or characteristics. Figure 1 shows the example of extracting motion vectors for each block.

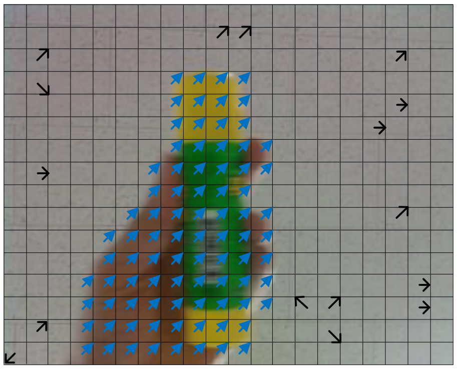


Figure 1. Image blocking and motion vector extracting for each block

For the motion vectors for the extracted block, a 3-dimensional histogram is prepared. The X, Y coordinates represents the directions of motion vectors while the Z coordinates represents the sizes of them. Figure 2 shows the 3D histogram mapping of the extracted motion vectors.

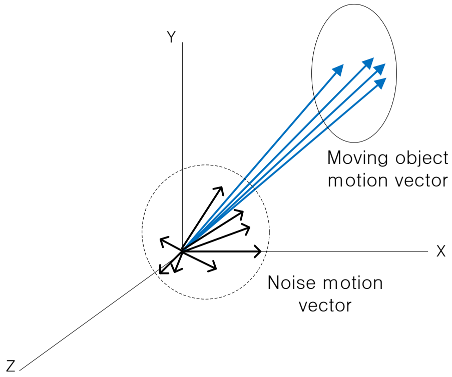


Figure 2. 3D histogram of motion vectors extracted for each block

As in Figure 2, motion vectors of the block of the moving object have similar directions and sizes. Noise vectors generated due to lighting or shadow may have different elements from those of the moving object vectors. After the vector clustering, remains the block of vectors of similar direction and size while removed are other blocks. Then only the moving object is extracted. However, there is a possibility that other blocks than the moving object may have similar directions and sizes with the moving object block. Thus, only when the block has more than a certain number of similar vectors, it is recognized as a moving object, and others are removed as noise. Figure 3 shows the result of extracting the moving object in the way above.

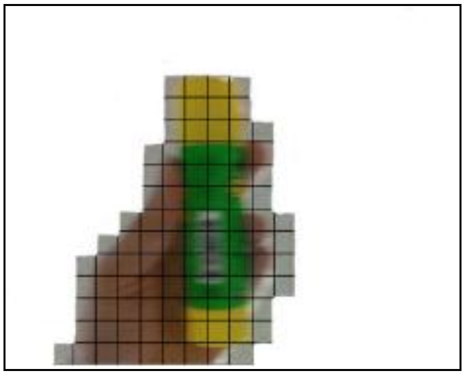


Figure 3. Extraction of blocks with the moving object's vectors and then of the moving object

Figure 4 is the flowchart of the algorithm to extract the moving object by means of motion vectors after the image blocking.

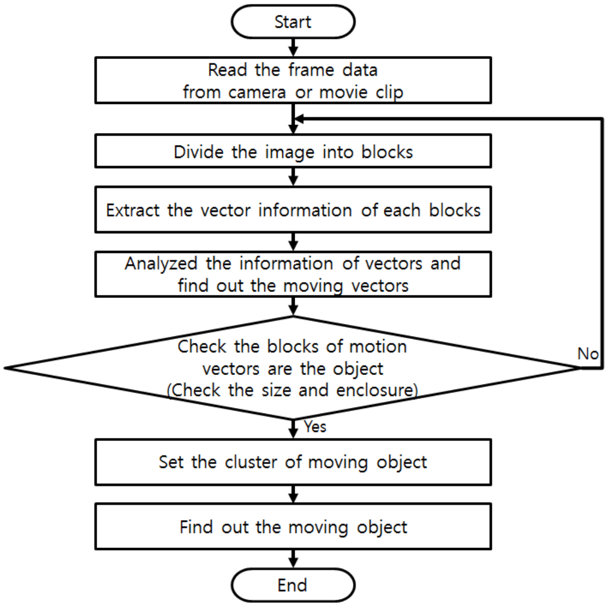

Start
Read the frame data from camera or movie clip
Divide the image into blocks
Extract the vector information of each blocks
Analyzed the information of vectors and find out the moving vectors
Check the blocks of motion vectors are the object (Check the size and enclosure)
No
Yes
Set the cluster of moving object
Find out the moving object
End

Figure 4. The moving object detecting algorithm by means of the motion vectors of the blocked image

This technique is appropriate for an embedded system as it adopts the blocking and clustering of motion vectors and thus reduces the amount of operation more than existing techniques.

III. EXTRACTING BARCODE FROM OBJECT IMAGE

This chapter explains the way of extracting the barcode area from the image of the moving object extracted in Chapter 2. Since the barcode scanning is implemented only in the area of the extracted moving object, the amount of operation may be reduced more than in scanning over the entire image.

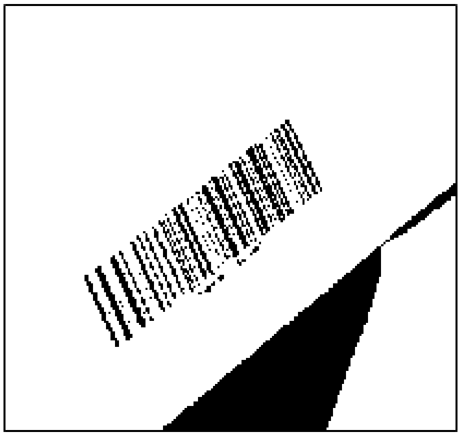


Figure 5. The image of changing a frame of the video with HSL color elements

A barcode commonly consists of the white background and black bards. Thus, areas in the extracted moving object where a certain amount of white and black parts exist will be selected for barcode scanning. To this end, the image area of the moving object that consists of RGB color elements is reformed with HSL(Hue, Saturation, Lightness) color elements. Figure 5 shows that a frame of the video has been changed with HSL color elements.

When the saturation value of pixels in the image block of the moving object is 0.3 or less, it is judged as colorless, and when the colorless part is more than 70% of the block, this block is selected for barcode searching. This procedure is to remove color blocks among moving blocks. In addition, the black and white parts of the image blocks of each moving object are distinguished, the ratio of black and white is calculated, and any part whose percentage is 10% or less is excluded. This procedure is to exclude blocks that only consist of no other achromatic color than white and black parts. Although this method may effectively detect the block that contains the barcode, a part of the image that seems similar to the barcode can be detected. Thus, calculated is the complexity of pixels in the block. When the value of difference of luminance is lower than a certain degree, this is regarded as monochrome. This is possible since the blocks extracted earlier are colorless. Difference of luminance may not be the only factor to be considered if colors exist. Based on the process above, finally the areas for barcode searching are decided.

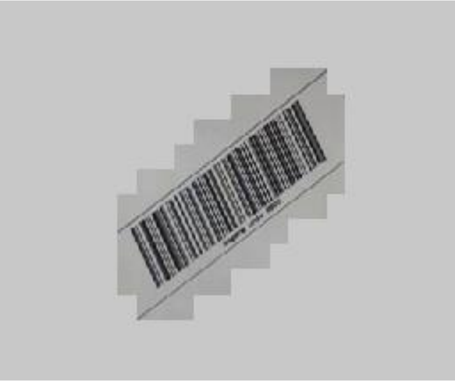


Figure 6. A barcode image to expand each block outward

Once the areas for barcode searching are decided, the shape of the block is checked to see if it is a quadrangle like a barcode, and then each block is expanded. The blocks are expanded to prevent a part of the barcode image from being removed or deleted in the process of blocking and to make the barcode image clearer. Figure 6 shows the expanded image of each block around the barcode.

Figure 7 shows the flow chart of the process of finding the barcode through color conversion.

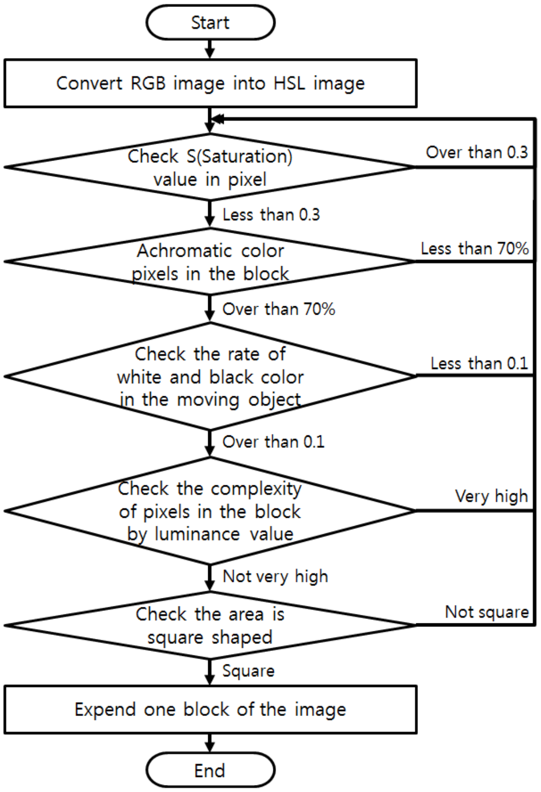

Start
Convert RGB image into HSL image
Check S(Saturation) value in pixel
Over than 0.3
Less than 0.3
Achromatic color pixels in the block
Less than 70%
Over than 70%
Check the rate of white and black color in the moving object
Less than 0.1
Over than 0.1
Check the complexity of pixels in the block by luminance value
Very high
Not very high
Check the area is square shaped
Not square
Square
Expend one block of the image
End

Figure 7. Flowchart of Extracting Barcode from Object Image

IV. IMPLEMENTATION AND RESULTS

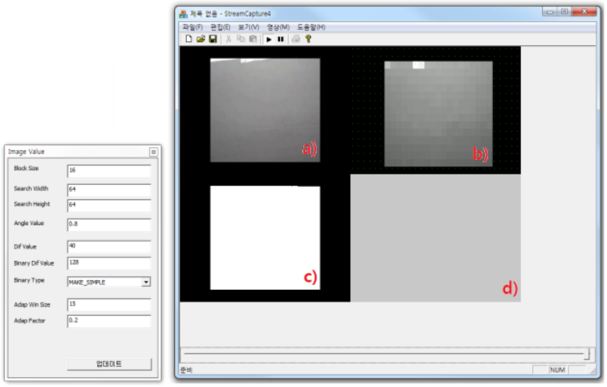
The following are the environments and software adopted to develop the application of this study:

- OS : Windows 7 Professional Edition 32bit
- Compiler : Visual Studio 2010
- Programming Language : C/C++

A. Application View

1) Main Window

Upon executing the application, the following screen in Figure 8 is displayed:


제목 없음 - StreamCapture4
파일(F) 편집(E) 보기(V) 영상(M) 도움말(H)
a)
b)
c)
d)
Image Value
업데이트
준비
NUM

Block Size	16
Search Width	64
Search Height	64
Angle Value	0.8
Dif Value	40
Binary Dif Value	128
Binary Type	MAKE_SIMPLE
Adap Win Size	15
Adap Factor	0.2

Figure 8. Main window of application

The small window to the left is an option window with functions to enhance the accuracy in barcode detection. The options of the program may be reset in this window. The big window to the right displays the following from the left: a)the original video; b)blocking and extracted motion vectors; c)black & white image of the area for barcode searching; and d)the detected barcode image.

2) Option Window

The application provides the following option window as in Figure 9:

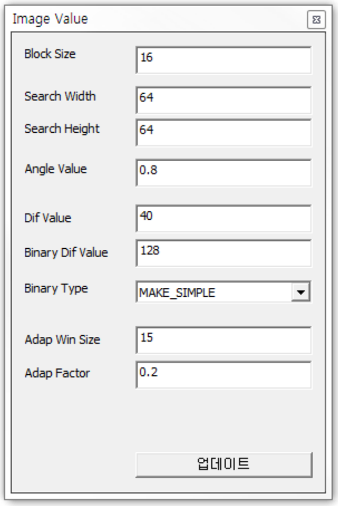


Image Value

Block Size	16
Search Width	64
Search Height	64
Angle Value	0.8
Dif Value	40
Binary Dif Value	128
Binary Type	MAKE_SIMPLE
Adap Win Size	15
Adap Factor	0.2

Figure 9. Option Window

The currently available options are as follows, and those that are not stated are currently not used.

- Block Size: the number of pixels for one block. If you input '16,' a 16x16 sized block is produced.
- Search Width: the width of the searching window to enhance the efficiency of block searching(unit: pixel)
- Search Height: the height of the searching window to enhance the efficiency of block searching(unit: pixel)
- Angle Value: the range of toleration to check if the motion vectors are in the same direction(0 ~ 1)
- Dif Value: The value of color difference to check if it is the same block(0 ~ 255)

B. Implementation Issues

1) Implementation issues of seperation of a moving object from background

The solutions to the implementation issues of separation of a moving object are as follows:

a) Detection of moving blocks is time consuming

- Improve the efficiency by limiting the processing area
- Parameterize the range of search in the program options

b) Saturation occurs when the surface of the moving object is wide and uniform in color

- Determination of the blocks of similar colors by prioritizing the proximity search
- Utilize the process of filling in the interior of folia

c) The boundary between the moving object and the background is irregular due to blocking

- Ignore the unclear boundariy as the goal is not in extracting the perfect imagery of moving objects

2) Implementation issues of extracting barcode from object image

a) Non-barcode area of wide achromatic color is erroneously extracted

- Discard the extracted area if the ratio of black and white is less than 10%

b) Retangular and linear images, visually close to barcode, are errornesoul extracted

- Identify barcode by calculating the complexity of pixels in the block
- Treat the pixels as of uniform color if the difference of lunimance is below certain level

C. Results

The application worked quite appropriately, and the barcode recognition of a moving object was outstanding.

Figure 10 shows the entire process of extracting the barcode from the camera image:

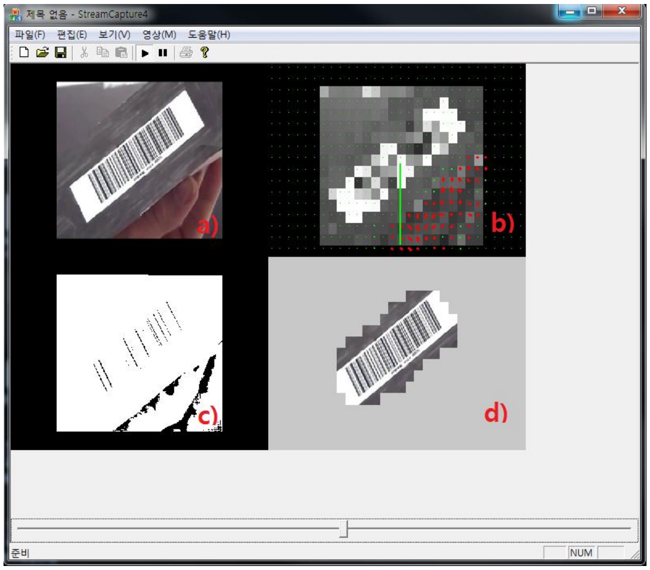

제목 없음 - StreamCapture4
파일(F) 편집(E) 보기(V) 영상(M) 도움말(H)
a)
b)
c)
d)
준비
NUM

Figure 10. Extraction of the barcode from the object with the black background

Figure 10 shows the process of extracting the barcode from the object with the black background. a) the original video image taken by a camera; b) the detection based on the blocking and motion vectors; c) detection of the area for barcode searching based on the colors and shadowing; and d) the final result of locating the barcode.

This process was successfully implemented even when the object was white or in various colors.

Figure 11 shows the result of barcode extraction from an object with the white background. Figure 12 shows the result of extracting the barcode from an object in various colors.

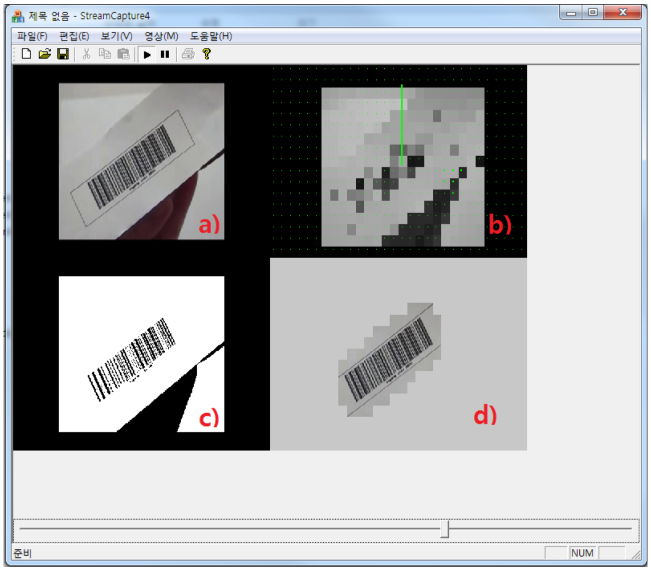

제목 없음 - StreamCapture4
파일(F) 편집(E) 보기(V) 영상(M) 도움말(H)
a)
b)
c)
d)
준비
NUM

Figure 11. Barcode extraction from an object with the white background

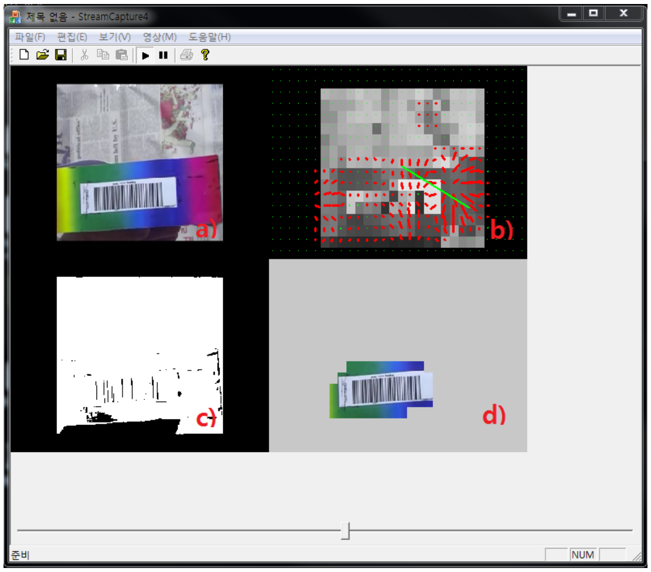

제목 없음 - StreamCapture4
파일(F) 편집(E) 보기(V) 영상(M) 도움말(H)
a)
b)
c)
d)
준비
NUM

Figure 12. Barcode extraction from an object in various colors

V. CONCLUSION AND FUTURE WORK

This study aims to develop an application to detect any missing barcode scan despite the normal scanning process by means of the video processing algorithm to prevent any intentional or mistaken barcode scan omission from happening at a checkstand.

To extract a moving object, each frame of the image is blocked, and the motion vectors are extracted with the central point of each block as the center. Among the extracted motions vectors, those of the similar direction and size are clustered and extracted. The image of this set of block is the moving object. The image of the extracted moving object is changed to a HSL color model, and the ratio of black and white and size are checked to extract the barcode. The actual application that adopts the algorithm above has been developed and tested with various video clips.

The primary factor affecting the success rate of barcode extraction was the optical focusing of the barcode area in the video images. There were cases where barcode was not properly extracted when the barcode area in the video images was out of focus, whereas the video images of which barcode area was in focus were properly processed by the algorithm. The failures occurred because the shadows in the out of focus video images were not sufficiently clear enough for the barcode searching algorithm to properly extract the barcode based on the colors alone.

It is planned to comparatively analyze the rate and accuracy of extracting an object and a barcode in comparison with existing algorithm in the future.

ACKNOWLEDGMENT

This work (Grants No. 000428980111) was supported by Business for Cooperative R&D between Industry, Academy, and Research Institute funded Korea Small and Medium Business Administration in 2010

REFERENCES

[1] Hyun-Sung Kim, Ju Wook Jang, “Using motion vector clustering for separation of a moving object from background,” in proc. Conf. Institute of Embedded Engineering of Korea, vol 1. November, 2011

[2] In-kwon Kim, Jae-bok Song, “Moving Object Detection based Motion Vector and Color Information,” in proc. CASS, 5pages, 2007.

[3] Ji-man Kim, Dae-jin Kim, “Detection of Moving Object from Moving Camera,” Workshop on Image Processing and Image understanding 2010, 4pages, 2010.

[4] G. D. Borshukov. G. Bozdagi, Y. Altunbasak, and A. M. Tekalp, "Motion segmentation by multistage affine classification", IEEE Transaction on Image Processing, 1997, Vol. 6, issue. 11, pp. 1591-1594

[5] Y. Weiss, "Smmothness in layers: Motion segmentation using nonparametric on homographics", Proceeding IEEE Conference on Computer Vision Pattern recognition, 1997.

Extending RTOS Functionalities: an Approach for Embedded Heterogeneous Multi-Core Systems

Shuichi Oikawa Gaku Nakagawa Naoto Ogawa Shougo Saito
Department of Computer Science
University of Tsukuba
Tsukuba, Ibaraki, Japan
{*shuioikawa,gakutarou,onaoto0707,shougosaitoh*}*@gmail.com*

***Abstract*—This paper proposes an approach to extend real-time operating system (RTOS) architecture for embedded heterogeneous muti-core processors, which consist of processors with different processing power and functionalities. The architecture splits the RTOS kernel into the two components, the proxy kernel (PK) and user-level kernel (UK). The PK runs on a less powerful core, and delegate its functions to the UK that runs on a powerful core as a user process. The experiment results running micro benchmark programs show that a communication cost between the UK and its user process is negligible and that there are cases where UK outperforms the monolithic kernel. These results confirm that the proposed approach is practically useful.**

Keywords- Real-Time Operating Systems; Heterogeneous Multi-Core Systems; Embedded Systems.

I. INTRODUCTION

As embedded devices, such as mobile smart phones, tablets, Internet TV sets, and so on, require more functions to respond to consumers' needs, their processors have been becoming more powerful. Since it is important for embedded processors to maintain their power consumption as low as possible, they cannot simply make their clock frequencies higher to increase their performance; thus, they nowadays consist of multiple processor cores and provide symmetric multi-processor (SMP) environments.

Some processors even go further and include different kinds of processor cores. The Texas Instruments OMAP4 processor [1] and the Renesas Electronics R-Mobile processor [2] are such examples. The OMAP4 processor consists of dual ARM Cortex-A9 cores and dual Cortex-M3 cores, and the R-Mobile processor consists of a Cortex-A9 core and a Renesas SH core. The OMAP4 processor incorporates Cortex-M3 cores, which are designed as microcontrollers and much smaller than but incompatible with A9 cores, to offload multimedia processing and to achieve faster real-time response. The R-Mobile processor incorporates an SH core also to offload multimedia processing. Therefore, incorporating more smaller cores to offload the specific types of processing can be a trend for future embedded processors.

Systems software, especially the operating system (OS) kernel, is, however, rather slow to respond to such an architectural change. While there have been researches conducted to support muti-core systems [3], [4], [5], they targeted server class systems with highly functional processors and their approaches do not fit into embedded processors. Since there has been no support for such embedded heterogeneous muti-core processors, only approach currently available is to execute independent OS kernels on different processors. A typical configuration for the OMAP4 processor is to execute the Linux SMP kernel on the dual Cortex-A9 and to execute an real-time OS (RTOS) on the less powerful Cortex-M3. The problems of such an architecture are 1) Linux and RTOS run independently with few cooperations between them and 2) only static functions can be provided by the software executed on the RTOS.

This paper proposes an extensible RTOS architecture for embedded heterogeneous muti-core processors. The architecture splits the RTOS kernel, which runs on a less powerful core, such as Cortex-M3 for OMAP4, and delegate its functions to the user-level kernel (UK) that runs on a powerful core, such as Cortex-A9 for OMAP4. The kernel on a less powerful core is called a proxy kernel (PK) since the global decisions are made by the UK and it works as a proxy of the UK. Figure 1 shows the overall architectures of the existing and proposed systems on OMAP4. In the figure, A9 and M3 stand for Cortex-A9 and Cortex-M3 cores of OMAP4, respectively. The architecture addresses the problems of the existing systems that consist of the independent OS kernels by making the PK closely coupled with Linux and controlled flexibly via the UK.

The rest of this paper is organized as follows. Section II describes the related work. Section III present the proposed system architecture. Section IV describes the current status and shows experiment results. Finally, Section V concludes this paper and describes our future work.

II. RELATED WORK

Recent researches conducted to support muti- or many-core systems all take basically the same approach. They consider a single system as a distributed system in order to amortize different characteristics. Multikernel [3] and Corey [4] target symmetrical processor systems with non-uniform memory access (NUMA) characteristic. By considering such

ISBN: 978-1-61208-202-8

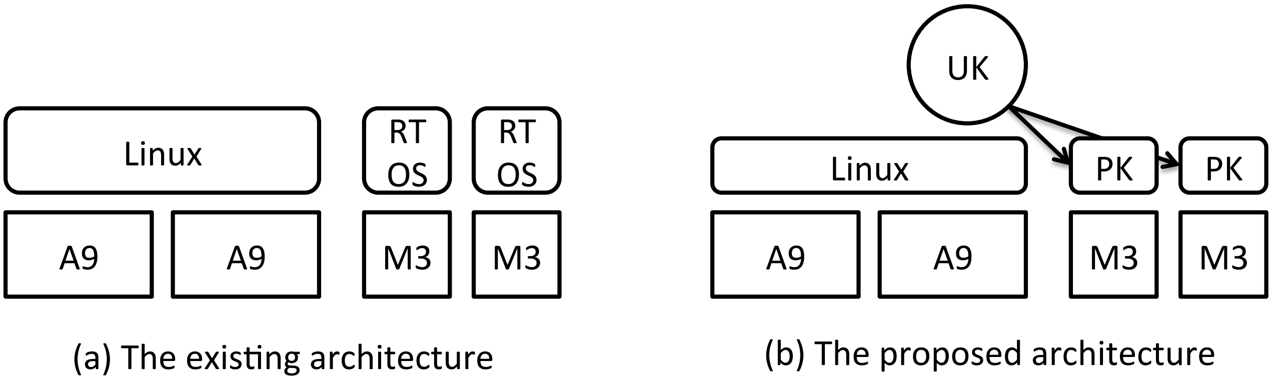


Figure 1. Overall architectures of the existing and proposed systems.

systems as distributed systems, they can partition systems into clusters of processors with the same characteristic, and can hide NUMA characteristic. On the other hand, the architecture proposed in this paper targets heterogeneous multi-core systems, of which processors have different functionalities.

Helios [5] targets heterogeneous multiprocessor systems and supports different kinds of processors by employing a satellite kernel, which is a microkernel [6]. Each satellite kernel on a different processor exports the same API, so that programs can be executed on a processor that fit the programs' requirements. Although the architecture proposed in this paper targets heterogeneous multi-core systems, which is similar to the target of Helios, the kernels on different kinds of processors are not the same. It aims to make programs on a less powerful core closely coupled with Linux on a powerful core by complementing the functionalities of the small kernel on a less powerful core.

III. Proposed System Architecture

This section describes the proposed RTOS architecture and its components in detail. As depicted in Figure 1, the proposed architecture consists of three major components, the PK (Proxy Kernel), the UK (User-Level Kernel), and Linux. The PK and the UK constitutes the RTOS kernel. The PK is executed on a less powerful core, and its functionality is supported by the UK that is executed on a powerful core as a user process of Linux.

A. PK: Proxy Kernel

The PK is a simplified RTOS kernel. It is a standalone kernel; thus, it consists of basic RTOS components, such as interrupt and exception handlers, a scheduler, and synchronization mechanisms. It works with the UK so that its functions are complemented by the UK. It does not perform dynamic resource management except for task scheduling. It simply picks up the highest priority task and dispatch it. It processes interrupts, and unblocks tasks when needed. When a fixed task set is executed on it, it works the same way as an ordinary RTOS.

The PK works in different ways from an ordinary RTOS when dynamic management features are involved. It outsources such features to the UK, so that it can keep itself as simple as possible while its functionality can be extensible. When a task on the PK invokes a system call the PK itself cannot handle, the system call is transferred to and processed by the UK on behalf of the PK. Some exceptions are processed in the same way. By outsourcing the functions to the UK, the OS functionalities provided for tasks on the PK become flexible and extensible. For example, file access and networking features can be easily provided through the UK since it is executed on Linux. Moreover, the PK outsources its memory management to the UK since it is unnecessary to execute a fixed task set. It is possible because the physical memory of the PK is mapped into the UK's address space. The task management, especially the creation and deletion of tasks, uses the memory management functions. Thus, creating a new task is a function of the UK, and the PK simply dispatches it when it becomes ready to run. When a task exits, such an event is transferred to the UK, and the memory used by the existed task is reclaimed by the UK.

B. UK: User-Level Kernel

The UK processes the requests issued by the PK's tasks on behalf of the PK as described above. The UK is executed as a user process of Linux; thus, it can utilize the full functionalities of Linux. It can access files on Linux's file systems and operate on networks just as Linux's user processes can. It can provide the PK with such Linux's functionalities without increasing the complexity of the PK, and can easily introduce dynamic features to the PK.

The other benefit for the UK to be a user process of Linux is that it is free from maintaining its own execution environment, and can focus on managing the execution environments of the PK. Such delegation of functions makes the implementation of each components as simple as possible.

Finally, there must be a communication means between the PK and the UK. As a user process of Linux, the UK cannot directly communicate with the PK without the support from Linux. There are two ways for the UK to communicate with the PK. One is through the read and write system calls, and the other is through the shared memory. The first method is simpler while it involves overheads of using system calls. The latter is faster while it complicates the interactions between the UK and the PK. Both methods need to be considered for the better implementation.

IV. Current Status and Experiment Results

This section describes the current status, experiment results, and a performance improvement based on profiling. The PK and UK were implemented based on the XV6 operating system [7], which is a reimplementation of UNIX V6 [8]. The current implementation is based on an Intel IA-32 multi-core processor because we could not obtain an OMAP4 based system when we started the work. It statically considers some cores as powerful ones and some as less powerful ones. Therefore, all experiments described below were performed on a PC-AT compatible system, which is equipped with an Intel Core i7-920 2.66GHz CPU. The hyper threading and power management features were disabled to perform all benchmarks. We used the Scientific Linux 6.1 x86_64, which is based on the Linux kernel 2.6.32, to execute the UK. While the Linux kernel executes in the 64-bit mode, the UK is a 32-bit program. The original XV6, which is used for comparisons, executes in the 32-bit mode.

A. Current Status

The implementation consists of 3 parts, the PK, the UK, and the linux device driver to interact with the PK, as described in Section III. The PK consists of total 1534 lines, which are 1448 lines of the C program and 86 lines of the assembly program. The linux device driver consists of total 830 lines, which are 723 lines of the C program and 107 lines of the assembly program. As far as the UK is concerned, 12 files, mostly for device drivers, were deleted, 5 files were added, and 13 files were modified from the original XV6. The total number of lines of the added files are 494 lines.

The current implementation is stable enough to perform micro benchmark programs as follows.

B. Micro Benchmarks

We first executed several micro benchmark programs on the UK and also on the original XV6 in order to investigate the performance penalty to realize the proposed architecture. We chose 4 programs, getpid, pipe, fork, and fork+exec to measure the functions without and with dynamic resource management. The getpid program invokes the getpid system call to find the cost of calling the kernel. The pipe program makes 2 processes communicate with each other through 2 pipes, so that it can measure the cost of context switches between them. The fork program creates a copy of the current process, and the fork+exec program executes a different program in a newly created process. These programs can measure the process management costs.

Table I
MICRO BENCHMARK RESULTS [IN μSEC]

Benchmark	UK	UK (mmap buf)	Original XV6
getpid	1.06	0.68	0.20
pipe	10.47	8.98	2.77
fork	23.80	22.05	82.63
fork+exec	49.99	48.17	168.56

Table I shows the results of executing micro benchmark programs. In the table, UK uses the linux device driver to communicate with the PK. On the other hand, UK (mmap buf) directly communicates with the PK through a buffer that is mapped in the UK and the PK's address spaces; thus, it does not use the linux device driver to communicate with the PK. The original XV6 was executed directly on a system.

The results from the getpid and pipe programs show the overheads incurred by splitting the execution of the UK and its user processes on different processors. Since the getpid program only invokes the getpid system call, the difference between the results of UK and XV6 is the communication cost between them; thus, the communication cost is 0.86 μsec for UK and 0.48 μsec for UK (mmap buf). Direct communication through the mapped buffer without the linux device reduces the communication cost by 44%. While the overall cost for UK (mmap buf) to invoke the getpid system call increases as much as 3.4 times more than XV6, the cost increases by only 0.48 μsec, which is negligible in the total computing time of applications and other more complicated system calls. Moreover, such simple system calls as getpid, which obtains the state of in-kernel resources but does not manipulate them, can be optimized by embedding them within the PK. In this case, the communication costs are eliminated, and the costs to invoke such simple system calls becomes the same as XV6.

The results from the fork and fork+exec programs show that UK performs better than XV6. This is an advantage of the proposed architecture. The control flows to process the fork and exec system calls currently remain the same for UK and XV6 since no optimization, such as batching multiple system calls and parallelizing the execution of the kernel and user processes, has been applied to UK. We consider the differences arise due to the effects of cache and TLB. There is no need for the UK to flush TLB of the processor it is running since actual process manipulation is done on another processor user processes are running. The PK runs on the processor that runs user processes. It is however extremely small; thus, the effect to it is

Table II
PROFILING RESULTS (TOP 5)

Function Name	%
memmove	35.48
jpkwaitevent	11.93
memset	9.18
freevm	6.92
getcallerpcs	3.78

negligible. Moreover, having the UK and user processes run on different processors increases the chances for them to remain on cache. Therefore, the UK architecture can decrease the number of cache and TLB flushes and increase the performance.

C. Profiling Results and Improvement

In order to further improve the performance of UK, we analyzed the hot spots in the UK. We used OProfile, which is a system wide profiler for Linux systems. Profiling was taken while running the fork and fork+exec programs. Table II shows the results of profiling. The table only shows the top 5 function names where the most of CPU cycles were consumed. These top 5 functions consumes 67.29% of CPU cycles in total.

The function that most consumes CPU cycles is memmove, which copies data from one place to the other. The second one is jpkwaitevent, which busy waits the completion of the user process side processing; thus, it does nothing. The third one is memset, which sets a memory region to a specified value. The two string functions, memmove and memset, consumes 44.66% of CPU cycles in total; thus, their performance should impact the overall performance.

Intel Core-i7 supports SSE4, which is a SIMD unit that has 8 128-bit long registers. Since a SIMD unit similar to Intel SSE is also available for the ARM architecture as NEON, we decided to utilize it to accelerate memory operations. A single SSE instruction can move 128-bit (16-byte) data between a SSE register and memory. We utilized a SIMD instruction to accelerate memmove and memset. By using the SIMD versions of them, the performances of the fork and fork+exec benchmark programs were improved 27% and 13%, respectively.

Figure 2 summarizes the results from the performed micro benchmark programs.

V. SUMMARY AND FUTURE WORK

This paper proposed an extensible ROTS architecture for embedded heterogeneous muti-core processors, which consist of processors with different processing power and functionalities. The architecture splits the RTOS kernel into the two components, the PK and UK. The PK runs on a less powerful core, and delegate its functions to the UK that runs on a powerful core as a user process. The experiment results running micro benchmark programs show that a communication cost between the UK and its user process is negligible and that there are cases where UK outperforms the monolithic kernel. We now obtained an OMAP4 based evaluation board [9], and are currently porting the proposed architecture on it.

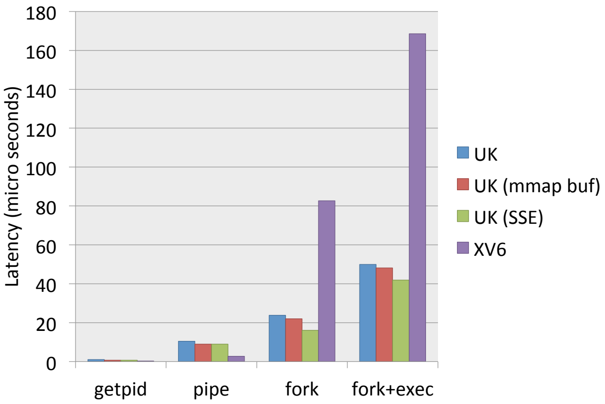


Figure 2. Summary of Micro Benchmark Results

REFERENCES

[1] D. Witt. OMAP4430 Architecture and Development. Hot Chips Symposium, August 2009.

[2] M. Ito, et. al. SH-Mobile G1: A Single-Chip Application and Dual-mode Baseband Processor. Hot Chips Symposium, October 2006.

[3] A. Baumann, P. Barham, P.-E. Dagand, T. Harris, R. Isaacs, S. Peter, T. Roscoe, A. Schüpbach, and A. Singhania. The Multikernel: a New OS Architecture for Scalable Multicore Systems. In *Proceedings of the 22nd ACM Symposium on Operating System Principles*, pp. 29-44, October 2009.

[4] S. Boyd-Wickizer, H. Chen, R. Chen, Y. Mao, F. Kaashoek, R. Morris, A. Pesterev, L. Stein, M. Wu, Y. Dai, Y. Zhang, and Z. Zhang. Corey: an Operating System for Many Cores. In *Proceedings of the 8th USENIX Conference on Operating Systems Design and Implementation*, pp. 43-57, December 2008.

[5] E. B. Nightingale, O. Hodson, R. McIlroy, C. Hawblitzel, and G. Hunt. Helios: Heterogeneous Multiprocessing with Satellite Kernels. In *Proceedings of the 22nd ACM Symposium on Operating System Principles*, pp. 221-234, October 2009.

[6] D. Golub, R. Dean, A. Forin, and R. Rashid. Unix as an Application Program. In *Proceeding of the USENIX Summer Conference*, pp. 87-95, June 1990.

[7] Xv6, a Simple Unix-like Teaching Operating System. http://pdos.csail.mit.edu/6.828/xv6/

[8] J. Lion. Lion's Commentary on UNIX V6.

[9] Pandaboard. http://pandaboard.org/ (as of 3 April 2012).

The Impact of a Requirements Specification on Software Defects and Quality Indicators

John Terzakis
Intel Corporation
e-mail: john.terzakis@intel.com

Abstract— **Capturing requirements in a written, centralized specification is a recognized product development best practice. But, how does this document impact software defect rates and quality? Is there any correlation between a well-written, properly reviewed requirements specification and software defect levels and other quality indicators? This paper will present data from an Intel case study illustrating the "before and after" scenario for a requirements specification. In the former, a minimal set of requirements were scattered across various documents for a first generation (older) product. In the latter, requirements were written and reviewed in a single requirements document for a second generation (newer) product. Software defect rates, feature commit vs. delivery, requirements volatility, and defect closure rates all improved dramatically even with the increased complexity of the newer product.**

Keywords-requirements specification; requirements defects; reviews; software defects; software quality.

I. INTRODUCTION

This case study involves two generations of a software product at Intel. The first generation product was developed without a requirements specification (e.g., Product Requirements Document or Software Requirements Specification). The requirements that existed were scattered across a variety of design documents, emails and web sites. There was no centralized source or repository for these requirements. The second generation product was developed based on a requirements specification. A standardized template was used along with a requirements management tool. Architecture specifications, design documents and test cases were developed from this specification. The requirements were rigorously reviewed by both technical content experts and a requirements Subject Matter Expert (SME). The second generation software product was more complex than the first in that the software had to run with, and implement functionality for, a next generation Intel processor. In addition, it had to combine code bases with a similar product from another business group.

II. DEFECT POTENTIAL COMPARISON

In general, there are many factors that impact the number and severity of software defects including: maturity of the team (development and validation), number of new features, complexity of the new features, test coverage and stability of the code base at the start of the project. In comparing the two software development efforts, the teams were of about equal size and maturity and their development methodology was the same (waterfall). The validation teams were also of similar size and maturity. There was some overlap of personnel between projects. As for a comparison of the two products, the newer product had more features, those features were more complex, the underlying hardware went through an architectural change, test coverage increased and the starting code base was less stable (due to the code merge from the other business group). Given all of these factors, the defect potential [1] should be higher for the second generation product than the first.

The most notable difference for the second generation product was the requirements specification. What impact would it have on overall software defect levels, quality, features delivered, number of change requests and defect closure rates?

III. REQUIREMENTS AND REVIEWS IN BOTH PROJECTS

Requirements for the first generation product were spread across documents, emails and web sites. That loose collection of requirements captured only about half of the initially intended product functionality. Reviews were held for those requirements that existed.

For the second generation product, the primary requirements author used a requirements management tool (RMT) to enter the requirements. The requirements were organized logically using key product features as section headers. The RMT had the capability to export to a document format. Reviews were based on this document.

We were the requirements SMEs assigned to work with the author to review and provide feedback on requirements quality. Initial requirements defect levels were high as this was the first set of requirements written by the author. However, with mentoring, peer reviews and stakeholder reviews, the requirements defect density for the requirements specification was reduced from about 4.75 defects per page in an initial revision to about 1.18 defects per page a later revision, a reduction of about 75%. The requirements specification became the basis for all architecture, design and test documents that followed.

IV. ACTUAL VALIDATION RESULTS

The following data presents a comparison of software defects, requirements volatility, feature variance and defect closure efficiency between the first generation ("Gen 1") and second generation ("Gen 2") software products.

ISBN: 978-1-61208-202-8

Table I shows the total number of defects by type per product at the end of validation testing. Overall, the second generation product had about 50% fewer defects.

Table I: Total Number of SW Defects

Defect Type	Gen 1	Gen 2	Delta
Critical	21	3	-86%
High	137	69	-50%
Medium	111	62	-44%
Low	24	6	-75%
Totals:	293	140	-52%

Table II shows the requirements volatility per product at key milestones during development. Some requirements volatility is due to scope creep (requests for new features) but most of it is due to changes needed due to missing, incomplete or incorrect requirements. At release, the second generation product had almost half the volatility of the first generation.

Table II: Requirements Volatility at Major Milestones

Milestone	Gen 1	Gen 2	Delta
Alpha	0.4	0.4	0%
Beta	1.2	0.7	-42%
Release	1.7	0.9	-47%

$$\text{Volatility} = \frac{\text{\# of added+changed+deleted requirements}}{\text{Total \# of requirements}}$$

Table III shows the feature variance per product at key milestones during development. This metric shows how well the features delivered in final product matched what was committed by the team to be delivered. The second generation product was able to deliver many more features than the first generation product at release.

Table III: Feature Variance at Major Milestones

Milestone	Gen 1	Gen 2	Delta
Alpha	0.05	0.15	+300%
Beta	0.15	0.25	+167%
Release	0.15	0.35	+233%

$$\text{Feature Variance} = \frac{\text{(Current - Planned Features)}}{\text{Planned Features}}$$

Finally, software defect closure efficiency (cumulative SW defects closed / cumulative SW defects submitted) at the end of validation testing improved from about 69% in the first generation product to about 87% in the second generation product, an improvement of over 25%. Note that a higher percentage indicates that defects are being closed more rapidly. This means the development and validation teams are spending less time identifying, researching and correcting software defects.

V. CONCLUSIONS

A number of factors could have had some impact in reducing the number of software defects from the first to the second generation product. They include applying lessons learned from the first development to the second, augmented developer experience and maturity, improved code review practices and more rigorous unit testing prior to the start of validation. No doubt these factors had some influence on improving software defect levels. However, given the increased complexity of the second generation product, they should have had a minimal effect on total software defect density levels. The key software quality indicators showed a dramatic improvement in the second generation product. Some other factor was playing a dominant role in these improvements.

Clearly, a well-written, properly reviewed requirements specification was the major contributing factor to these improvements in software defects and other quality indicators on the second generation product. This set of requirements had a positive influence on the total number of software defects (down 50%), requirements volatility (down 50%), feature variance (improved 2x) and software defect closure rates (improved by 25%). A third generation product is currently in development. Results from that project will be analyzed in a future paper.

REFERENCES

[1] Jones, Capers, *Software Quality: Analysis and Guidelines for Success,* International Thomson Computer Press, June 14, 2000.

ISBN: 978-1-61208-202-8

Height Map Viewer

Tool not only for the Laser Engraved Sample Exploration

Jana Hájková
Department of Computer Science and Engineering
University of West Bohemia
Pilsen, Czech Republic
e-mail: hajkovaj@kiv.zcu.cz

***Abstract*— The paper describes a complex software tool designed for exploring height maps. The tool facilitates tasks like viewing the height maps in 2D and 3D perspectives, measuring their dimensions, computing various statistics for them and, most importantly, mutually comparing several of them for similarity. The tool originates from a research in the area of modeling and simulation of the laser engraving process and is used for exploring surface laser engravings digitized by a confocal microscope. It may, however, be used with any general height maps described in the input format, which is designed to be easy to create and human readable. The main aim of this paper is to introduce the useful tool to the experts.**

Keywords-height map; visualization; exploration; statistics; comparison.

I. Introduction

The effort to develop a height map viewer was part of a larger project dealing with modeling and simulation of the laser engraving process. Because the results of physical laser engraving are very sensitive to various physical parameters of the used material as well as the settings of the laser device, a simulation model has been developed at the New Technology Centre at the University of West Bohemia in Pilsen for predicting the engraving results in order to save both time and material costs. The predictions are made based on data obtained from physical samples engraved so far and various statistics computed from them. The physical samples vary both in used material and laser settings. They are scanned using a confocal microscope and saved in the form of a height map. However, this is not sufficient. In order to use these height maps with the simulation model, they must be explored in more detail to discover various dependencies, similarities or differences, and to reject those samples that are significantly affected by any adverse effects not related with the laser engraving process, such as damages in the used material. Because the huge amount of data and the high costs of human resources, a high level of automation is desirable, especially in comparing several samples for similarity.

As no suitable tool was found among the existing ones, it was decided to develop a new one that fits required purposes the best. The description of this tool is the main aim of this paper. Even though some its functions are closely associated with laser-engraved samples (especially the pit/area detection algorithms), most its functions are general enough to use it for detailed explorations and comparisons of any height maps represented in the corresponding input format, which is described in Section III.A. The functions of the height map viewer are described in Section IV in more detail.

Laser engraving is a technique frequently used in various branches of industry and science. During the laser engraving process, the surface of a material is exposed to the activity of a laser beam, which affects and modifies it. As the electromagnetic radiation of the laser beam strikes the surface, the material starts to heat and, if the laser intensity is high enough, the rising temperature evocates material ablation. More detailed description of the ablation process can be found in [1][2].

During our research, two different types of engravings were processed: single-point engravings and motion engravings. Examples of both can be seen in Figure 1. For each physical sample, a scenario containing a set of key parameters was prepared (number of laser pulses, laser beam intensity, motion speed, and the like). Then, a corresponding physical sample was engraved (BLS-100 Nd:YAG solid-material, lamp-pumped laser with the wavelength of 1064nm was used) and scanned using a confocal microscope (for described samples Olympus LEXT OLS3100 [11] was used).

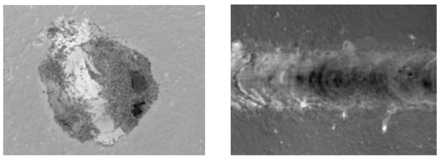


Figure 1. Single-point and motion engravings (used material: cermet).

A 3D view on a sample exported from the confocal microscope can be seen in Figure 2 (the original figure was white on black, so this is a negative). The confocal microscope also provides other data formats on export, such as 2D top views (both in color and in gray scale) and, more importantly, heights description in the form of CSV file. The scanning of the sample surface is provided in a discrete way. Each sample is measured in many parallel cross-sections with a pre-defined step distance. The height coordinates (one for each cross-section) are saved in plain text CSV file. Input and output file formats are the main content of Section III.

The application and its main functions are described in Section II and Section IV. Section V concludes the paper.

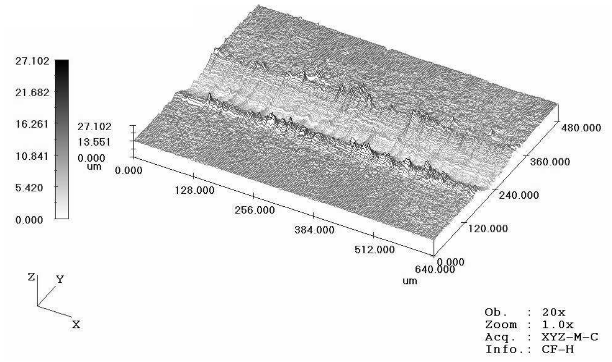


Figure 2. Sample 3D view exported from the confocal microscope (negative image).

II. Program Description

The Height Map Viewer is a desktop application licensed under CC-BY-NC-S. Downloads are available under [9].

The application works in two modes that can be switched between: map exploration mode and map comparison mode. The map exploring mode can be used to visualize maps in 2D and 3D perspectives. Moreover, it offers tools for performing measurements, exploring cross-sections, as well as some functions specific for laser-engraved samples, such as automatic detection of heat-affected area [7]. It also allows exporting the sample or its parts in several formats. The map comparison mode contains tools that can be used to mutually compare two samples for similarity. The samples can be overlapped and the difference visualized. Also, various statistics related to the difference can be computed.

III. Inputs and Outputs

The input and output formats of data are mostly text-based in order to allow preparing and pre-processing the input data and post-processing the output data using the many powerful text utilities and to allow preparing a batch script for it.

A. *Height Map Input Format*

The application uses an input format that some confocal microscope can produce directly and, if not the case, that the output of most confocal microscopes can be easily converted into. This is important in order to save time on data pre-processing when measuring physical samples. The format is relatively simple - it consists of several headlines containing metadata followed by an array of floating point values representing the height coordinates.

Table I summarizes the list of possible headlines that can be included in the input file. The headlines can be used in any order. However, it is important that the DataLine describing the array of height coordinates is used before the coordinates themselves appear. The first column of the table shows the format of each individual line, the second one explains its meaning and shows other acceptable values if there are some. The color of rows in Table I specifies, whether or not the line has to be included in the file, (mandatory lines are highlighted with gray color). Besides settings described in Table I, the headlines of the file can also contain other information such as calibration data, various file names or comments. Such lines are ignored during the file load process.

TABLE I. Headlines that can be used in the input file.

Line Format	***Explanation***
Data type = Height	indicates a height map

Xcv = 0.625
Ycv = 0.625
	step size in the X and Y direction (sampling raster size); this value is used for computing the real size of the sample. If no value is defined, step = 1 is used

Xunit = um
Yunit = um
Zunit = um
	real units for all three axes (um = µm)

Along x-axis	indicates the direction of sample measuring (x-axis: each data line represents one column of the sample; y-axis: sample is represented row by row)
DataLine, Pos = 0,Pos = 1, ... , Pos = 300	line placed before the height coordinates data, the number of values indicates the number of rows (x-axis) or columns (y-axis) measured in the sample
0.0,2.6499, 2.6556,2.6631...	floating point values representing the height coordinates of the sample surface in the particular measuring points; separated with comma, the first value of the row represents the real distance of the row (column) from the borderline of the sample

To gain a better imagination of how the data look like, the following example shows a description of a small (five rows by ten columns) sample. The measurement was performed along to the `x-axis`; it means that lines of the height map file represent columns of the sample (Figure 3a). Than the direction of saved values (headline `Along`) from `x-axis` to `y-axis` was changed. In such case, the file would represent a sample with ten rows and five columns. In both figures, each value is depicted as the square of the gray scale adequate to the height value (Figure 3b).

```
Data type = Height
Xcv = 0.5
Ycv = 0.5
Xunit = um
Yunit = um
Zunit = um
Along x-axis
DataLine,Pos = 1,Pos = 2,Pos = 3,Pos = 4,Pos = 5
0.00,0.0,0.0,0.59,1.26,1.62
0.50,0.0,0.92,2.0,2.82,3.27
1.00,0.59,2.0,3.27,4.31,4.92
1.50,1.26,2.82,4.31,5.63,6.53
2.00,1.62,3.27,4.92,6.53,7.99
2.50,1.62,3.27,4.92,6.53,7.99
3.00,1.26,2.82,4.31,5.63,6.53
3.50,0.59,2.0,3.27,4.31,4.92
4.00,0.0,0.92,2.0,2.82,3.27
4.50,0.0,0.0,0.59,1.26,1.62
```

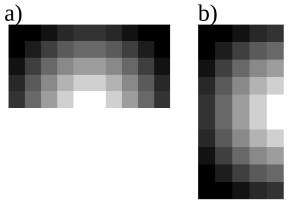


Figure 3. Visualization of the sample, where the value for `Along` as: (a) `x-axis`; (b) `y-axis` was used.

B. Outputs and Exports

During the exploration of a height map, some results typically need to be saved for future use. Thus, the application allows saving parts of the height map, their visualizations or cross-section plots. The possible outputs and their formats are described in the following sections.

1) Sample Surface

The whole sample can be saved in two different ways. The first one saves the whole sample or its part bordered by a selection rectangle in the text format described in Section A. It can be used, e.g., to cut off border areas of the height map which should not be processed or to crop the height map after rotating it. If the whole height map is selected, no height map values are lost in comparison with the original surface. The saved file can be reloaded and processed in the same way as the original one.

The second way saves a grayscale representation of the current view on the height map. The user can choose from several image formats – BMP, JPG, and PNG. The current view is always saved with all the tools (selection, cross-section lines, ruler, etc.) that are actually shown. This is useful when measures or statistical data are desirable to view along with the height map itself (as shown in Figure 4), which is typical for images intended to be published.

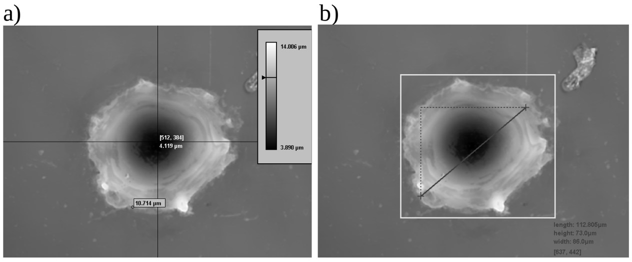


Figure 4. Images saved with (a) cross-sections and height scale; (b) ruler and selection rectangle.

2) Cross-Sections

When working with height maps, it is sometimes useful to save separate cross-sections for further processing. Because the 2D visualization uses them as its fundamental part, the possibility of exporting them is a native part of the height map exploration. The cross-sections can be saved both as images and as sets of values. The set of values is exported as a simple CSV file. This allows further processing of the data, for example in a spreadsheet processor. The file contains the described direction (x – vertical cross-section; y – horizontal cross-section; line – cross-section along the line segment) followed by the actual height coordinates, each on its own line. The export as an image enables its further comparison, presentation or retainment. Several parameters of the image export (background color, plot height scale, font size, and the like) can be set. An example of a simple cross-section plot can be seen in Figure 5. It is possible to export vertical or horizontal cross-sections from the whole sample or just from the area bordered with the selection rectangle. It is also possible to draw an arbitrary line and to get the surface cross-section along that line. For the line rasterization, the DDA algorithm [13] is used.

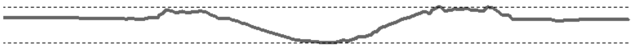


Figure 5. Exported simple cross-section plot.

IV. IMPORTANT FUNCTIONS

The important and useful functions of the viewer are described in the following sections. Besides the principal function of visualizing the data and couple of specialized functions for exploring them, there are also general functions not described in detail. Especially functions for opening and saving the data (the format alternatives were mentioned in Section III.B), selecting them, rotating them, and the like belong into this group.

A. Height Map Visualization

During the height map visualization, it is necessary to map the three-dimensional object (surface of the sample) onto a two-dimensional plane (screen or paper). There are two possible approaches that can be successfully combined, namely the three orthogonal 2D views on the object along with its 3D visualization. This combination is also often used by many 3D modeling programs.

1) 2D Visualization and Control

The first approach of the 3D object visualization is the usage of three orthogonal views on a 2D plane. As shown in Figure 6, there is one top view together with two cross-section plots.

The top view is represented by a grayscale image, where each point of the surface is represented in gray shade corresponding to its height coordinate. There are also two orthogonal lines in the top view. They indicate the position of the cross section plots shown left and below the top view. By moving these lines, it is possible to explore the exact shape of the surface. To explore the height map in a more precise way, a ruler and a heightscale can also be used. Their usage is described in Section B.

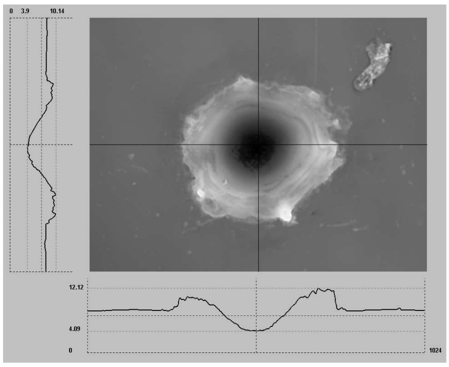
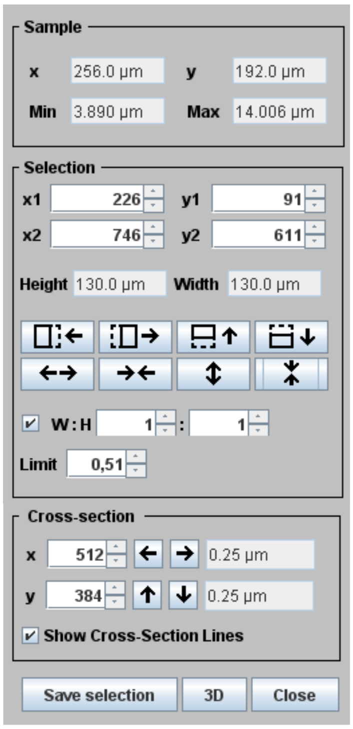


Figure 6. The three views of the 2D visualization.

For the control of 2D visualization it is possible to use only the mouse clicking and dragging and also a specialized control panel that is placed on the right side of the main window. In the main (top) view on the sample selection rectangle can be shown or hidden (only by one single click in the view), its size and position can be changed (if the user drags the border by mouse). If cross-section lines are shown, they can be moved (also by drag and drop) and the cross-section relief of the moved direction is modified immediately. Selection and cross-section lines positions can be adjusted also in vertical and horizontal views.

To enable better and exact height map exploration, ruler and heightscale can be used. Both tools are activated on the viewer initiation; they can be activated and deactivated as needed (in menu or by shortcuts). Their usage is described in Section B.

The control panel (shown in Figure 7) in the right part of the window is divided into three parts. The first one (`Sample`) informs about the height map size (according to values `Xcv` and `Ycv` defined in the input file – see Table I) and the minimal and maximal heights on the surface. The second part (`Selection`) enables exact setting of selection rectangle (its each border), its motion or changing of its height and width. For each operation a special button can be used. It is also possible to set the exact proportion of height and width of the selection area. If it is activated, the selection proportions are fixed during each manipulation. In this part of control panel also the real size of selected area is computed. Value named as `Limit` is a threshold computed automatically from the sample that is used during the automatic detection of the heat-affected area described in Section IV.C. The last part of the control panel (`Cross-section`) works with cross-section lines, enables their exact motion (also according to their real position) and enables to show and hide them. Control panel contains also several individual buttons for the selected area saving, 3D visualization initializing and closing the whole viewer.



Sample

x 256.0 µm
y 192.0 µm
Min 3.890 µm
Max 14.006 µm

Selection

x1 226
y1 91
x2 746
y2 611
Height 130.0 µm
Width 130.0 µm

☑ W:H 1 : 1

Limit 0,51

Cross-section

x 512 0.25 µm
y 384 0.25 µm

☑ Show Cross-Section Lines

Figure 7. The control panel on the right side of the viewer window.

2) 3D Visualization and Control

If viewing the height map as a whole is desired, 3D visualization can be used (see Figure 8). The 3D plot is very realistic, but exploring the height map in detail or performing precise measurements is difficult using this view, because the control of the object space orientation is more complicated. There are two modes of visualization that can be switched between - a smooth surface (see Figure 8) and a wire model (see Figure 9). In both modes, the surface plot can be moved, scaled, and rotated. Both modes use the height map visualization module [8].

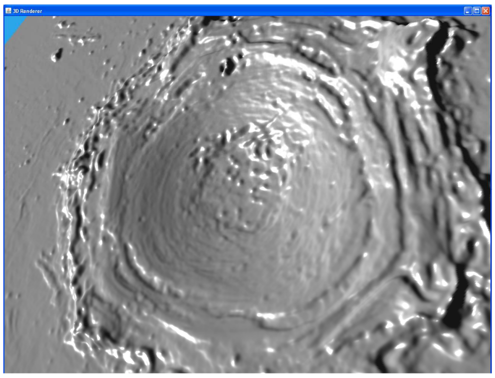


Figure 8. 3D height map visualization of a smooth surface.

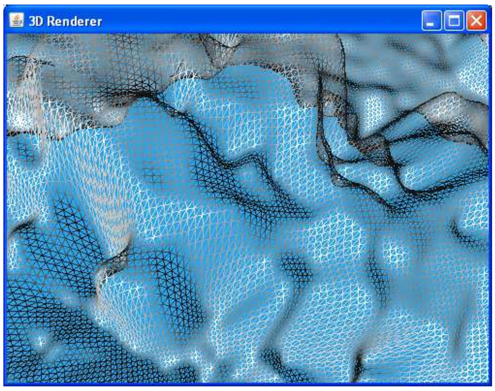


Figure 9. Wire model of the 3D surface with a help shown in the top left corner.

B. Admeasurements and Volume Computations

The height map viewer provides detailed surface exploration using several measurement tools. The first one is called ruler. If activated, the coordinates of the current cursor position are shown in the right bottom corner of the central view. To measure a distance, two points on the surface have to be selected. This can be done by the clicking on the surface using the right mouse button. The selected points are visualized on the surface as small crosslines. When the second point is selected, their distance is computed and the result is shown in the right bottom corner of the central view. A similar function can be used in both cross-section views. In this case, the distance of heights of the two points can be measured as well. An example of using the ruler can be seen in Figure 10 (the top image shows the central view and the bottom image the horizontal cross-section view). The selected points can be discarded by another click using the right mouse button. The ruler control points in each particular view are independent of each other.

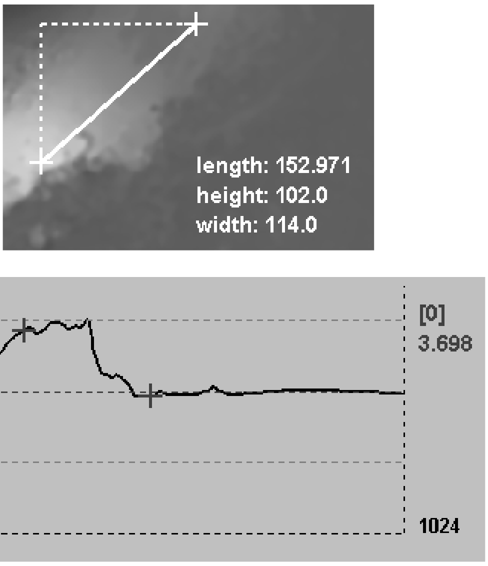


Figure 10. Ruler function - an example of results.

The second tool is the heightscale. If activated, the grayscale from white (maximum) to black (minimum) is shown in the right top corner (see Figure 11). The height of the current cursor position is marked on the height scale. When the cursor moves outside the central view, the height of the intersection of the cross-section lines is marked instead. The heightscale tool shows also the height of the current cursor position (as shown in Figure 12) and description of the position of cross-section lines crossing. It is activated in all the three views simultaneously using the heightscale tool.

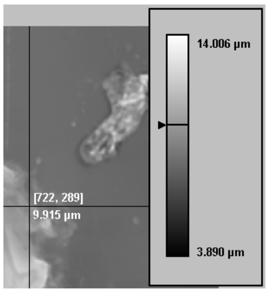


Figure 11. Heightscale function – an example of results.

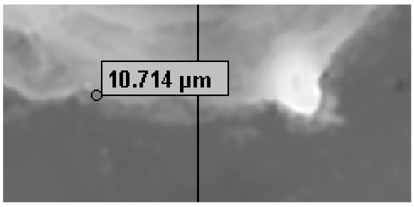
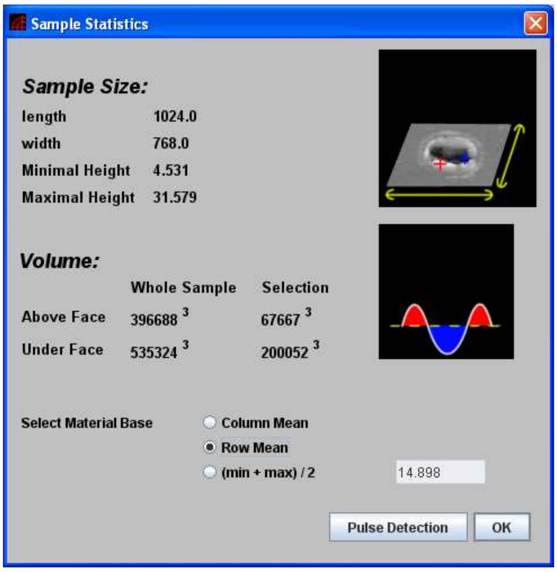


Figure 12. Height of the current position.

The third tool that can help in exploring the sample is statistics calculator. It summarizes the information about the sample (e.g., its size or minimum and maximum values) and computes the volume of retained/ablated material. The volume above and below the basic material level is computed either for the whole surface or for the surface of the currently selected area. It is also possible to provide an automatic detection of the heat-affected area directly from the statistics dialog. Because the computed volumes depend on the material basic level setting, the user can choose from three different modes of computing it (also shown in Figure 13).



Sample Statistics

Sample Size:

length	1024.0
width	768.0
Minimal Height	4.531
Maximal Height	31.579

Volume:

	Whole Sample	Selection
Above Face	396688 ³	67667 ³
Under Face	535324 ³	200052 ³

Select Material Base
○ Column Mean
● Row Mean
○ (min + max) / 2
14.898

Figure 13. Statistics function - an example of results.

C. Automatic Engraved Area Detection

These functions may not be useful for different types of height maps, but they can significantly speed up the exploration of laser engravings, where the material is modified by the laser beam in a single area. The algorithms used for these methods are described in [3][6]. The detection methods may be used for all height maps, but because they are designed for a specific type of height maps, the advisability for different height map types is disputable. All detection methods are available from the Pulse Detection menu. The user can choose from the following methods: statistical, spiral, or clipping. The result of these methods is a selection sized and positioned according to the result of the detection. It is also possible to run all the methods and to compare their results. This approach was used in [7], for instance, and is shown in Figure 14. There are also several other functions that can be used during data preprocessing and that work with semi-results of some of the detection methods.

D. Height Map Comparisons

The second mode of the Height Map Viewer serves for comparing a pair of height maps. To visualize the differences, the 2D visualization mode is used. Both samples should be opened at the same time and overlapped into the optimal and fitting position. The overlapping can be done automatically or by the user (the second sample can be moved by holding the Ctrl key).

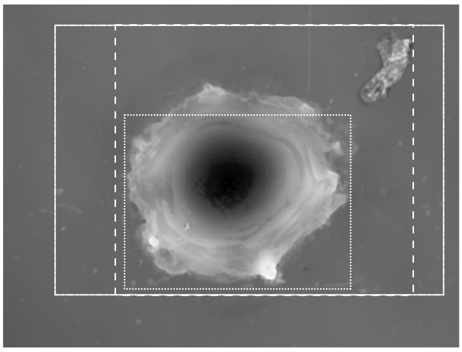


Figure 14. Comparison of results of all the detecting methods.

Finally, the samples can be visualized in three different ways - as both samples overlapped over each other (can be seen in Figure 16) or as the difference image in either linear or logarithmic scale [10] (Figure 15).

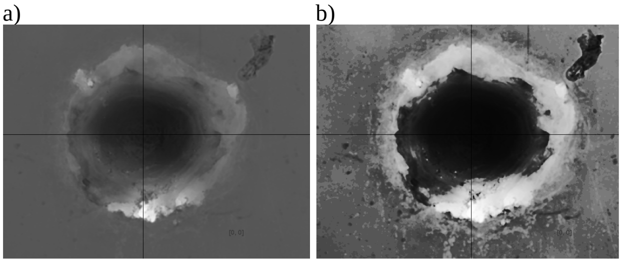
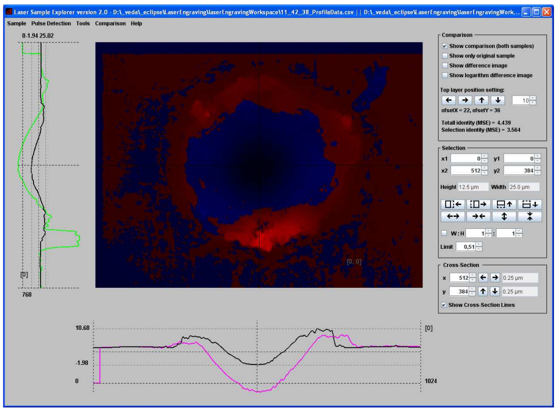


Figure 15. Comparison of two height maps: (a) difference image in linear scale; (b) difference image in logarithmic scale.

For either case, all the functions for exploring 2D visualization described above can be used and their results are adapted to the comparison mode. That means, if both height maps are shown together, the heights for both of them are shown and it is possible to measure differences between them. An screenshot of the comparison mode can be seen in Figure 16.



Laser Sample Explorer version 2.0 - D:_veda_eclipse\LaserEngraving\laserEngravingWorkspace\11_42_38_ProfileData.csv | | D:_veda_eclipse\LaserEngraving\laserEngravingWork...

Comparison

☑ Show comparison (both samples)
☐ Show only original sample
☐ Show difference image
☐ Show logarithm difference image

Top layer position setting: 10

ofsetX = 22, ofsetY = 36

Totall identity (MSE) = 4.439
Selection identity (MSE) = 3.564

Selection

x1 0 y1 0
x2 512 y2 384
Height 12.5 µm Width 25.0 µm
☐ W:H 1 : 1
Limit 0,51

Cross-Section

x 512 0.25 µm
y 384 0.25 µm
☑ Show Cross-Section Lines

Figure 16. Visualization of compared height maps.

V. Conclusion and Future Work

The Height Map Viewer has been created to simplify exploring of height map surfaces and to enable to process height maps with a user friendly tool. It offers a broad range of functions for data processing and its export. The tool is used, among others, by the team of experts from the New Technology Center at the University of West Bohemia in Pilsen performing research in the area of lasers and helps them to explore laser-engraved samples measured by a confocal microscope.

The tool can be downloaded from [9]. However, thanks to the universal text-based input format, it can be also used for different purposes, where height maps need to be explored in detail. The functionality of the tool can be further extended in dependence on actual demands.

Acknowledgement

I acknowledge Ing. Petr Zelenka for language help and proofreading of this paper.

References

[1] S. I., Anisimov, Vaporization of metal absorbing laser radiation. Soviet Physics JETP 1968, Vol. 27, pp. 182.

[2] N. B. Dahotre and S. P. Harimkar, Laser Fabrication and Machining of Materials, Springer, New York 2008, USA.

[3] Official Java website [online] http://www.java.com/en/ [cit. 2011-10-05]

[4] Java3D official website [online] http://java3d.java.net/ [cit. 2011-10-05]

[5] J. Hájková, LASER SIMULATION - Methods of Pulse Detection in Laser Simulation in Proceedings of the Third International Conference on Software and Data Technologies ICSOFT 2008, Porto, INSTICC, 2008, pp. 186-191.

[6] J. Hájková, Data Processing for Simulation of Laser Beam Impact – Statistical Method for the Heat-Affected Area Detection in Proceedings of the First International Conference on Computational Intelligence, Modelling and Simulation CSSim 2009, Brno, Czech Republic, 2009, pp. 69 – 74.

[7] J. Hájková, Laser Engraving Modelling – Comparison of Methods for the Heat-Affected Area Detection in Proceedings of the 10th International Conference APLIMAT 2011, STU Bratislava, 2011, pp. 1191-1199.

[8] JHeightMap Module [online] http://www.kiv.zcu.cz/~hajkovaj/sw/JHeightMap/index.html [cit. 2011-10-03]

[9] Height Map Viewer [online] http://www.kiv.zcu.cz/~hajkovaj/sw/HeightMapViewer/index.html [cit. 2011-10-04]

[10] Logarithm Operator [online] http://homepages.inf.ed.ac.uk/rbf/HIPR2/pixlog.htm [cit. 2011-03-13]

[11] Olympus LEXT OLS3100 [online], [cit. 2012-04-18]. http://www.olympus.co.uk/microscopy/26_LEXT_OLS3100.htm

[12] W. M. Steen, Laser Material Processing. Springer-Verlag, New York Berlin Heidelberg, 1991.

[13] A. Watt, 3D Computer Graphics, 3rd edition, Addison-Wesley, 2000.

Variability Identification by Selective Targeting of Significant Nodes

Anilloy Frank
Institute of Technical Informatics,
Technische Universitt
Inffeldgasse 16, 8010 Graz, Austria
Email: anilloy.frank@student.tugraz.at

Eugen Brenner
Institute of Technical Informatics,
Technische Universitt
Inffeldgasse 16, 8010 Graz, Austria
Email: brenner@tugraz.at

***Abstract*—The automotive industry is characterized by numerous product variants, often driven by embedded software. With ever increasing complexity of embedded software, the electrical/electronic models in automotive applications are getting enormously unmanageable. Significant concepts for modeling and management of variability in the software architecture are under development. Models are hugely hierarchical in nature with numerous composite components deeply embedded within projects comprising of Simulink models, implementations in legacy C, and other formats. Hence, it is often necessary to define a mechanism to identify reusable components from these that are embedded deep within. The proposed approach is selectively targeting the component-feature model (CF) instead of an inclusive search to improve the identification. We explore the components and their features from a predefined component node list and the features node vector respectively. It addresses the issues to identify commonality in identification, specification and realization of variants within a product development. Since the approach does not depend on the depth of the components or on its order, it serves well with all the scenarios, thereby exhibiting a generic nature. The results obtained are faster and more accurate compared to other methods.**

***Keywords*-Design Tools; Embedded Systems; Feature Extraction; Software Reusability; Variability Management.**

I. Introduction

Embedded systems are microcontroller-based systems built into technical equipment mainly designed for a dedicated purpose, where communication with the outside world occurs via sensors and actuators [1]. Although this definition implies that embedded systems are used as isolated units, there is also a trend to construct distributed pervasive systems by connecting several embedded devices, as noted by Tanenbaum and van Steen [2].

The current development trend in automotive software is to map software components on networked Electronic Control Units (ECU), which includes the shift from an ECU based approach to a function based approach. Also, according to data presented by Ebert and Jones, up to 70 electronic units are used in a car containing embedded software consisting of more than 100 million lines of object code, which is mainly responsible for the value creation of the car.

Variants of embedded software functions are vital in customizing for different regions (Europe, Asia, etc.), to meet regulations of the respective regions. Also different sensors / actuators, different device drivers, and distribution of functionality on different ECUs necessitate variants. Managing variability involves extremely complex and challenging tasks, which must be supported by effective methods, techniques, and tools [3].

Ebert and Jones present recent data about embedded software in [4], stating that the volume of embedded software is increasing between 10 and 20 percent per year as a consequence of the increasing automation of devices and their application in real world scenarios.

The proposed strategy is to introduce a variability identification layer. It intends to facilitate a reusable software solution. We start by analyzing the textual representation of the model structure. Based on this we form a concept to extract an element list to facilitate the identification of variability. Both implementation and evaluation of the proposed strategy is based on a technically advanced adaptation of a formal mathematical model, which is beyond the scope of this paper.

II. Software reuse

In the 1960s, reuse of software started with subroutines, followed by modules in the 1970s and objects in the 1980s. About 1990 components appeared, followed by services at about 2000. Currently, Software Product Lines (SPL) are state of the art in the reuse of software.

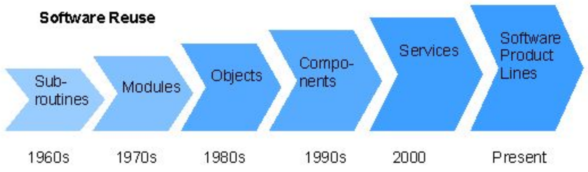


Figure 1. Software reuse history.

Figure 1 shows a short history of the usage of reuse in software development. The key idea of Product Lines is very old; it is based on Henry Ford's mass customization to provide a effective way for cheap individual cars. Today, many different approaches exist to the implementation of Software Product Lines.

A SPL is a set of software-intensive systems that share a common set of features for satisfying a particular market segment's needs. SPL can reduce development costs, shorten time-to-market, and improve product quality by reusing core assets for project-specific customizations [3][5].

Despite of all the hype, there is a lack of an overall reasoning about variability management.

The SPL approach promotes the generation of specific products from a set of core assets, domains in which products have well defined communalities and variation points[6].

Although variability management is recognized as an important issue for the success of SPLs, there are not many solutions available [7]. However, there are currently no commonly accepted approaches that deal with variability holistically at architectural level [8].

III. Variability management

One of the fundamental activity in Software Product Line Engineering (SPLE) is Variability management (VM). Throughout the SPL life cycle, VM explicitly represents variations of software artifacts, managing dependencies among variants and supporting their instantiations [3].

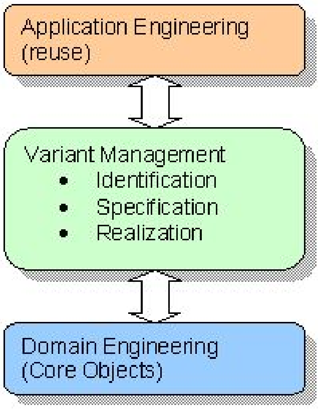


Figure 2. Variability management in product lines.

To enable reuse on a large scale, SPLE identifies and manages commonalities and variations across a set of system artifacts such as requirements, architectures, code components, and test cases. As seen in the Product Line Hall of Fame [9], many companies have adopted this development approach.

SPLE as depicted in Figure 2 can be categorized into domain engineering and application engineering [10][11]. Domain engineering involves design, analysis and implementation of core objects, whereas application engineering is reusing these objects for product development.

Activities on the variant management process involves variability identification, variability specification and variability realization [12].

- The Variability Identification Process will incorporate feature extraction and feature modeling.
- The Variability Specification Process is to derive a pattern.
- The Variability Realization Process is a mechanism to allow variability.

IV. Software Architecture

Figure 3 depicts a layered software architecture that is considered in the proposed architecture. It shows a comparison of distributed systems and platform with the proposed layered architecture and the feasibility of mapping the corresponding artifacts and responsibilities for each layer.

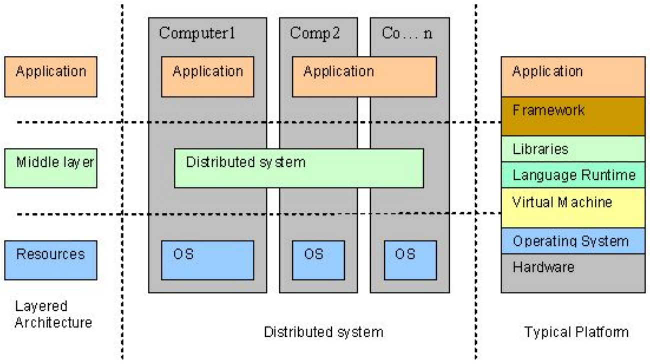


Figure 3. Comparison of architecture, system, and platform.

The definition of software architecture given in the ISO/IEC 42010 IEEE Std 1471-2000: "The fundamental organization of a system embodied in its components, their relationships to each other, and to the environment, and the principles guiding its design and evolution [13]."

In the middle illustrates a distributed system. Tanenbaum and van Steen define distributed systems as "A distributed system is a collection of independent computers that appears to its users as a single coherent system [2]."

Similarly to the right side is depicted a typical platform as specified by Atkinson and Kühner in Model Driven Architectures (MDA) [14].

V. Specification of the Cases

To enable identification of variability for software components in a distributed system within the automotive domain [15][16], we enlist the specifications below:

- *Specification of components by compatibility*
  The product is tested using software functions of a certain variant and version. These products may exhibit compatibility issues between functional blocks, whilst

using later version of the function may fail to perform as expected.

- *Extract, identify, and specify features*
  To enable parallel development, it is necessary to be able to extract features, and to identify and specify the functional blocks in the repository based on architecture and functionality.
- *Usability and prevention of inconsistencies*
  A process that tracks usability and prevents inconsistencies due to deprecate variants and versions in the repository is required.
- *Testing mechanism for validations*
  A testing mechanism for validations in order to maintain high quality for components and its variants has to be established.
- *Mechanism for simplified assistance*
  The developer has to be assisted by a process to intelligently determine whether a functional block or its variant should exist in the data backbone to avoid redesign of existing functions, thereby improving productivity.

VI. Proposed approach for Variability Identification

Models confirming to numerous tools like ESCAPE®, EAST-ADL®, UML® tools, SysML® specifications and AUTOSAR® were considered. Although this concept is not limited to automotive domain alone.

A. Project analysis

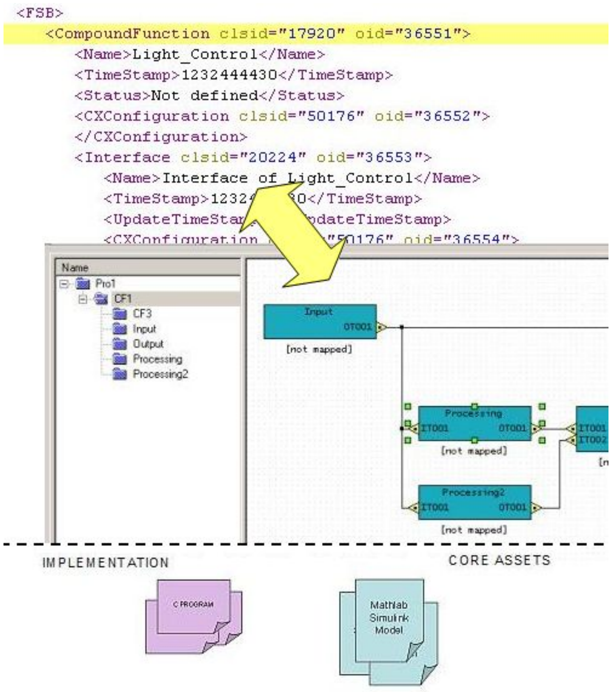
An analysis of the models exhibits a common architecture. Figure 4 depicts the textual representation that underlies several graphical model. The textual representation usually is given in XML, which strictly validates to a schema. A heterogeneous modeling environment may consist of numerous design tools, each with its own unique schema, to offer integrity and avoid inconsistencies. Developed projects have to be strictly validated to the schemas of these tools.

A closure examination of the nodes in the textual representation of models depicted in Figure 5 reveals some interesting information. The nodes outlined in rectangles provide important information regarding the identity, specification, physical attributes, etc. of a component, but are insignificant from the perspective of variant.

B. Concept and approach

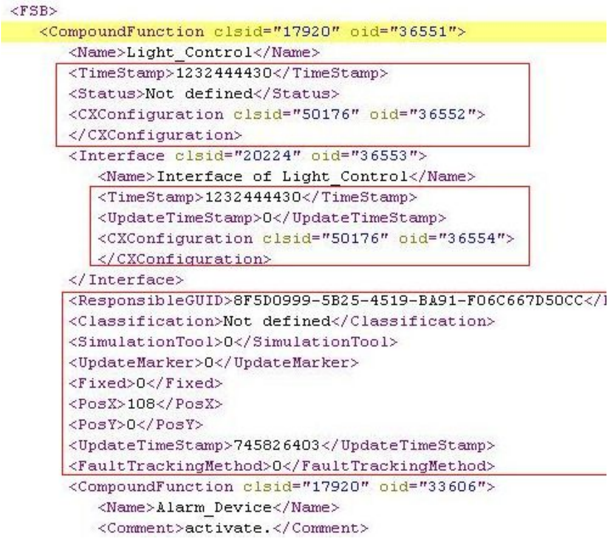
The basic concept to identify variability is depicted in Figure 6.

The left side is a set of projects that have software components hierarchically embedded. These projects validate to the corresponding schemas. The middle layer is an identification layer with three functional blocks. A set of component lists is derived from the node list in the schema. Similarly a feature vector is derived from it that corresponds to components.



```
<FSB>
   <CompoundFunction clsid="17920" oid="36551">
      <Name>Light_Control</Name>
      <TimeStamp>1232444430</TimeStamp>
      <Status>Not defined</Status>
      <CXConfiguration clsid="50176" oid="36552">
      </CXConfiguration>
      <Interface clsid="20224" oid="36553">
         <Name>Interface of Light_Control</Name>
         <TimeStamp>1232    0</TimeStamp>
         <UpdateTimeStam     pdateTimeStamp>
         <CXConfiguration      "50176" oid="36554">
```

Figure 4. Mapping textual and graphical representations.



```
<FSB>
   <CompoundFunction clsid="17920" oid="36551">
      <Name>Light_Control</Name>
      <TimeStamp>1232444430</TimeStamp>
      <Status>Not defined</Status>
      <CXConfiguration clsid="50176" oid="36552">
      </CXConfiguration>
      <Interface clsid="20224" oid="36553">
         <Name>Interface of Light_Control</Name>
         <TimeStamp>1232444430</TimeStamp>
         <UpdateTimeStamp>0</UpdateTimeStamp>
         <CXConfiguration clsid="50176" oid="36554">
         </CXConfiguration>
      </Interface>
      <ResponsibleGUID>8F5D0999-5B25-4519-BA91-F06C667D50CC</
      <Classification>Not defined</Classification>
      <SimulationTool>0</SimulationTool>
      <UpdateMarker>0</UpdateMarker>
      <Fixed>0</Fixed>
      <PosX>108</PosX>
      <PosY>0</PosY>
      <UpdateTimeStamp>745826403</UpdateTimeStamp>
      <FaultTrackingMethod>0</FaultTrackingMethod>
      <CompoundFunction clsid="17920" oid="33606">
         <Name>Alarm_Device</Name>
         <Comment>activate.</Comment>
```

Figure 5. XML Nodes that are not significant for variability.

The second block is a customized parser that generates a relevant lexicon from the set of software components within a project. The third block is a set of rules (viz., mandatory, optional, exclude) to govern the identification of variability.

The basic concept can be extended to obtain a working model for the identification of variants. The work flow is depicted in Figure 7. The top layer here represents the domain or core assets. The middle layer is a semi-automatic identification layer for variants. A component list and a

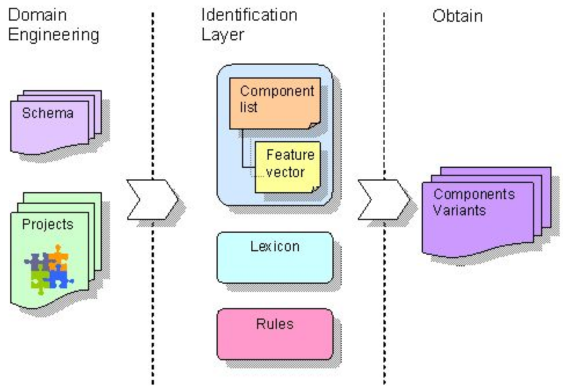


Figure 6. Basic Concept.

feature vector is derived manually from the schema of the project; a collection of elements that represent components and their descriptive features that significantly contribute to the identification of the component's variant.

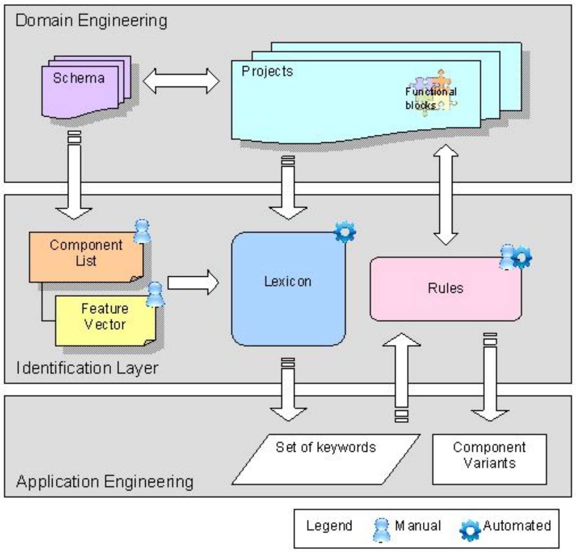


Figure 7. Work flow of the identification process.

The workflow can be further extended to adapt a heterogeneous environment which consist of projects developed using several modeling and simulation tools. The identification layer is separated into two parts. Numerous component lists and feature vectors can be derived for each distinct schema as depicted in Figure 8, whereas a common lexicon and common rules govern the identification process.

C. Evaluation

A prototype of the architecture presented here has been implemented. These case studies targeted the design of model-based software components firstly in an industrial use case where the project model was developed using the design tool ESCAPE® [17], and secondly in a case study targeting the execution of specific paradigms based on the naming convention of AUTOSAR® [18].

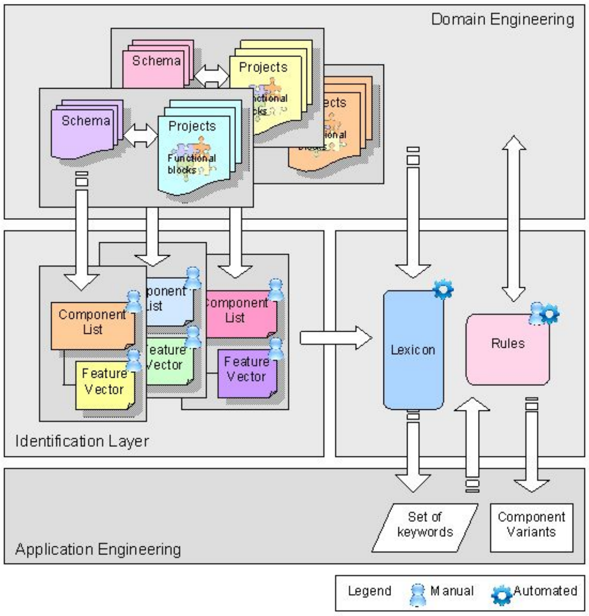
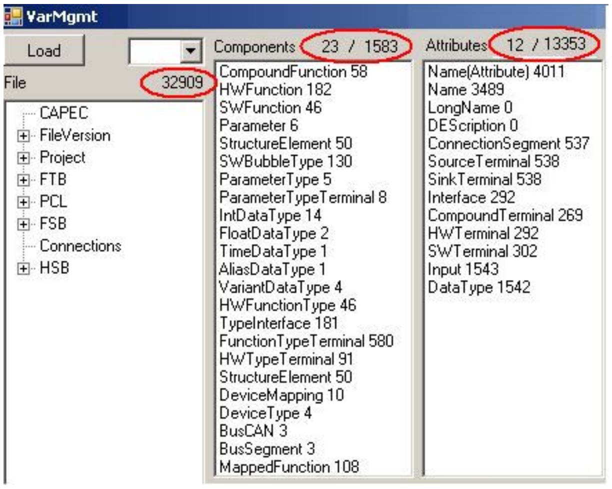


Figure 8. Work flow of the identification process for heterogeneous systems.

The specific project data set depicted in Figure 9, which was used to verify the implementation, consisted of a total of 32909 elements. Of these elements a total of 1583 elements signify components, these were categorized into 23 categories when enlisted in the component list. A total of 13353 elements signified features that were assigned into 12 categories.



File 32909

- CAPEC
- FileVersion
- Project
- FTB
- PCL
- FSB
- Connections
- HSB

Components 23 / 1583

CompoundFunction 58
HWFunction 182
SWFunction 46
Parameter 6
StructureElement 50
SWBubbleType 130
ParameterType 5
ParameterTypeTerminal 8
IntDataType 14
FloatDataType 2
TimeDataType 1
AliasDataType 1
VariantDataType 4
HWFunctionType 46
TypeInterface 181
FunctionTypeTerminal 580
HWTypeTerminal 91
StructureElement 50
DeviceMapping 10
DeviceType 4
BusCAN 3
BusSegment 3
MappedFunction 108

Attributes 12 / 13353

Name(Attribute) 4011
Name 3489
LongName 0
DEScription 0
ConnectionSegment 537
SourceTerminal 538
SinkTerminal 538
Interface 292
CompoundTerminal 269
HWTerminal 292
SWTerminal 302
Input 1543
DataType 1542

Figure 9. Dataset summary of project using ESCAPE design tool.

Three different approaches were adopted to evaluate and determine the performance with respect to matches and time.

- **Evaluation using a single element specification set** The first experiment was conducted on a single element specification set. A group of ten sets formed the input to determine the result set in both comprehensive (global) search and selective search as illustrated in Figure 10. The notion of comprehensive search is used, when scanning all occurrences of the specification set within projects, irrespective of whether they are components or features of those components. This can return a result set that contains false matches.

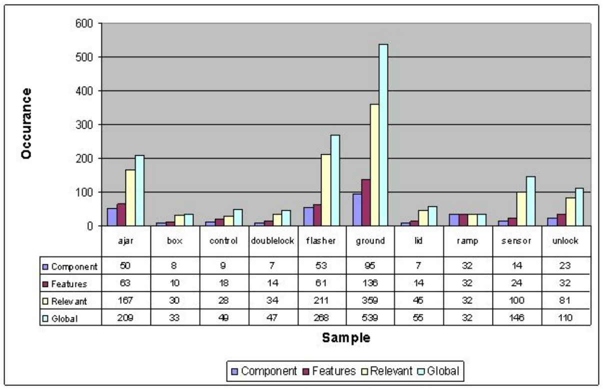


Figure 10. Occurrence graph for a single element specification set.

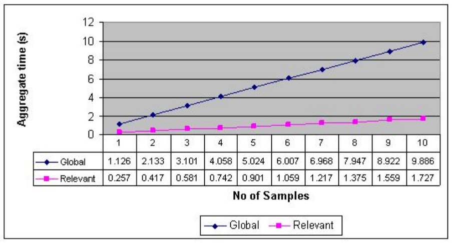


Figure 11. Time graph for a single element specification set.

The pattern of the results displayed similar behavior.

Observations

  - The comprehensive search yields a result set that contains every occurrence of the specification set, even if these nodes do not characterize a component.
  - The nodes representing components yield a result set which is somewhat realistic, though these do not epitomize the complete set desired. This is often observed when the component nodes do not match, but their features collectively match the specification set.
  - These nodes along with the feature set yield a more elaborate result set. A match contained by any node in a set of features would result in representing the component to which it belongs.

Figure 11 depicts the time taken to obtain the specification set illustrated in Figure 10. The time graph depicts the aggregate time required for global and selective search for a set of ten specification sets.

Observations

  - It is evident from these figures that the time required for comprehensive search exceeds the selective search - which is the method proposed in this article - by almost a factor of 5; this may be a dominant factor for large specification sets.

- **Evaluation using multiple element specification set** The second experiment was conducted using one up to seven element specification sets as a group illustrated in Figure 12.

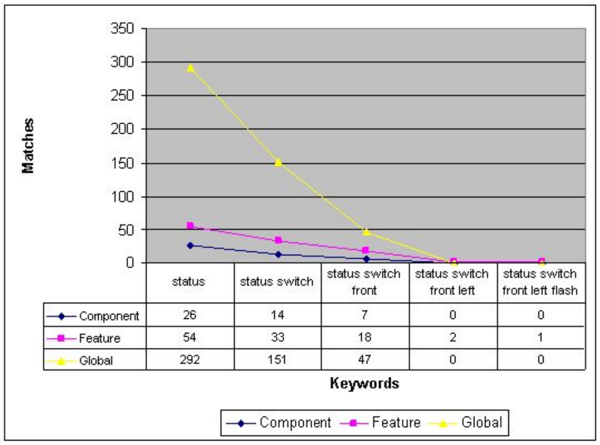


Figure 12. Occurrence graph for multiple element specification sets.

Observations

  - The comprehensive search often yielded large result sets, as it searches in individual nodes that are treated as atomic.
  - The exhibited behavior is similar to the varying size of the specification set. As observed in Figure 12, the selective component-feature search result set demonstrates a value when the size of specification set exceeds 3, because in this case the matches take place across the boundary of the feature within the component. On the other hand the other methods return null result set as the search is only within the boundary of the element.
  - For any given size of specification set, the selective component-feature search returns a much smaller result set and is more precise.
  - Convergence is optimal with a specification set of size 3. If the size of the specification is too large the result may be null for both methods as shown in Figure 12.

- **Evaluation using different starting points for elements in specification sets**

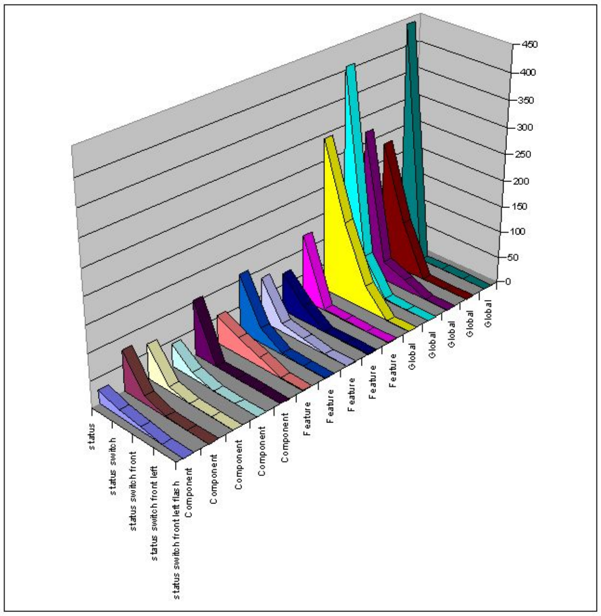


Figure 13. Occurrence graph for different starting points.

The third experiment was conducted searching for elements within specification sets using different starting points. Figure 13 depicts the result sets in comprehensive search and selective search.

To determine the effect of different starting points, a multiple-element specification set was used, where the orders of the elements were changed to obtain five sets. The result set for this exhibits the same pattern as the two experiments above.

VII. Conclusion

Managing variants is of utmost importance in today's large software bases as they reflect legal constraints, marketing decisions, and development cycles. As these software bases often grew from different sources and were developed by different teams using different tools it is in many cases very complicated if not nearly impossible to find artefacts that might be variants, both for historical reasons as for development purposes.

Searching algorithms have to reflect both the capability to match keywords and to reflect the structure that characterizes a component. Our proposed method is capable of both aspects and therefore helps the developer to find matches even in large and heterogeneous databases. In addition to that not only the required time for the search is a lot shorter, but also accuracy of the retrieved set of candidates is highly improved.

The developed prototype is itself independent of a specific tool as it works on textual descriptions that typically are available in XML.

References

[1] Ebert, C. and Salecker, J.; *Guest editors' introduction: Embedded software technologies and trends*. Software, IEEE, Vol 26(3): pp. 14-18, 2009

[2] Tanenbaum, A.S. and van Steen, M.; *Distributed Systems: Principles and Paradigms (2nd Edition)*. Prentice Hall, 2006

[3] Clements, P. and Northrop, L.; *Software Product Lines: Practices and Patterns*, Addison-Wesley, 2007

[4] Ebert, C. and Jones, C.; *Embedded software: Facts, figures, and future*. Computer, IEEE Vol 42(4): pp. 42-52, 2009

[5] Gomaa, H. and Webber, D.L.; *Modeling Adaptive and Evolvable Software Product Lines Using the Variation Point Model*. The Proceedings of the 37th Hawaii international Conference on System Sciences, 2004

[6] Oliveira, E., Gimenes, I., and Maldonado, J.; *A variability management process for software product lines*. CASCON 2005, The conference of the Centre for Advanced Studies on Collaborative research: pp. 225 - 241

[7] Heymans, P. and Trigaux, J.; *Software product line: state of the art*. Technical report for PLENTY project, Institut d'Informatique FUNDP, Namur, 2003

[8] Galster, M. and Avgeriou, P.; *Handling variability in software architecture: Problem and implications*. WICSA 2011, Ninth Working IEEE/IFIP Confernce on Software Architecture: pp. 171-180

[9] PRODUCT LINE HALL OF FAME; *"http://splc.net/fame.html"*. retrieved: 04,2012

[10] Bachmann, F. and Clements, P. C.; *Variability in Software Product Lines*, Technical Report -CMU/SEI-2005-TR-012, 2005.

[11] Bosch, J.; *Design and Use of Software Architectures: Adopting and Evolving a Product-Line Approach*, Addison-Wesley, 2000

[12] Burgareli, L.A., Selma, Melnikoff, S.S., and Mauricio Ferreira, G. V.; *A Variation Mechanism Based on Adaptive Object Model for Software Product Line of Brazilian Satellite Launcher*, ECBS-EERC 2009, First IEEE Eastern European Conference on the Engineering of Computer Based Systems: pp. 24-31

[13] IEEE; *Iso/iec standard for systems and software engineering - Rrecommended practice for architectural description of software-intensive systems*. Technical report, IEEE, 2000

[14] Atkinson, C. and Kühne, T.; *A generalized notion of platforms for model-driven development*. Model-Driven Software Development, Springer-Verlag, Berlin: pp. 119–136, 2005

[15] Frank, A.A. and Brenner, E.; *Model-based Variability Management for Complex Embedded Networks*. ICCGI 2010, The Fifth International Multi-conference on Computing in the Global Information Technology: pp. 305-309

[16] Frank, A.A. and Brenner, E.; *Strategy for modeling variability in configurable software*. PDES 2010, The 10th IFAC workshop on Programmable Devices and Embedded Systems

[17] ESCAPE; *"http://www.gigatronik2.de/index.php?seite=escape_produktinfos_de&navigation=3019&root=192&kanal=html"*. retrieved: 04,2012

[18] AUTOSAR; *"http://www.autosar.org/download/conferencedocs/03_AUTOSAR_Tutorial.pdf"*. retrieved: 04,2012

Impacts of a Whole Person eAssessment on Students' Learning Performance and Faculty Development

Koichi Nakajima
Faculty of Economics
Tezukayama University
7-1-1 Tezukayama, Nara-city, Nara, Japan
koichi2@tezukayama-u.ac.jp

***Abstract*— We have developed an open education community based on our homegrown instructor-centric eTeaching system called TIES since 1996. Its mission is to share educational content and pedagogical knowledge via the interuniversity collaboration. We currently host 83 universities in Japan and abroad with about 1,300 instructors and 70,000 students as users, and have more than 39,000 sharable materials. TIES has an eAssessment system that assists an instructor to evaluate learning outcomes and levels of attainment of her students from a wide spectrum of their academic as well as non-academic efforts and performance. The purpose of the system is to encourage students to self-review their intellectual growth, reflect on their personal attributes, and understand their strengths and limitations. In this paper we elaborate impacts of this system on students' learning performance from faculty development perspectives. We also report preliminary results of the new questionnaire that approximates students' learning preferences, and analyze if such preferences can be correlated with the specific assessment attributes in the TIES eAssessment.**

Keywords-TIES; eTeaching; eAssessment; faculty development; learning styles

I. HOW TO EVALUATE STUDENT LEARNING?

We all know that assessment is the most important issue with students and it defines their learning behavior in higher education. Unfortunately, we observe a well-known problem of "surface learning" or memorization-only learning with little retention or use of knowledge after passing course. Thus we may have to conclude that our conventional assessment method of grading students' achievement via tests and quizzes may not be enough to motivate and direct their learning toward "deep learning" [1].

In this paper, we first illustrate our eAssessment system that complements a standard method of marking students' academic performance by encouraging students to recognize and develop their personal attributes and social skills. Second, we report preliminary results of the questionnaire developed to identify and approximate a learner's learning preferences through his preferences for teaching styles. The questionnaire attempts to identify learning preferences of a student in four criteria: Logic, Planning, Emotion, and Creativity (LPEC in short). Third, we draw implications based on the data obtained from courses, and elaborate possible impacts of the new assessment approach on our pedagogical thinking and faculty development (FD in short).

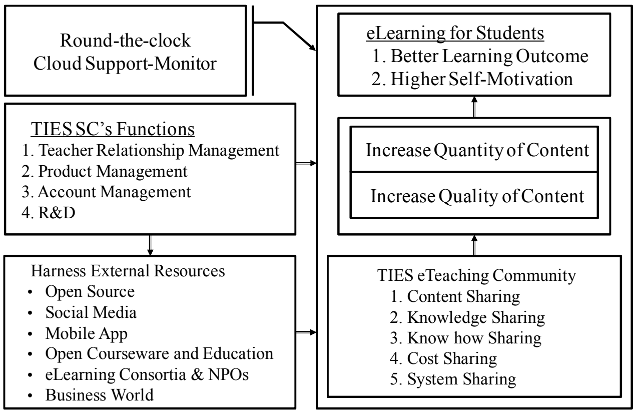
II. A BRIEF REVIEW OF TIES SYSTEM

We have developed an instructor-centric "eTeaching" system called TIES since 1996 at Tezukayama University. Its goal is to help motivate and direct instructors to use IT effectively and happily, so that students in turn can get motivated and self-directed to improve learning by engaging in face-to-face, online or blended courses more happily and willingly.

TIES community has started as a grass-roots initiative among a few instructors, and developed the concept of eTeaching based on the three principles of (1) interuniversity collaboration, (2) content and knowledge sharing, and (3) contribution to society.

Fig. 1 summarizes the concept of eTeaching, where eLearning is considered to be a subset of the system inside the TIES eTeaching community supported by TIES Support Center (TIES SC in short) and interuniversity membership.

TIES SC has helped us to develop the instructor-centric teaching-learning culture, and enabled our community to grow steadily. We are currently hosting 83 universities mostly in Japan, with about 70,000 student users, and almost 1,300 faculties. The educational materials created by instructors are sharable, and amount to about 40,000 as shown in Table I.



Round-the-clock
Cloud Support-Monitor

eLearning for Students
1. Better Learning Outcome
2. Higher Self-Motivation

TIES SC's Functions
1. Teacher Relationship Management
2. Product Management
3. Account Management
4. R&D

Increase Quantity of Content

Increase Quality of Content

Harness External Resources
- Open Source
- Social Media
- Mobile App
- Open Courseware and Education
- eLearning Consortia & NPOs
- Business World

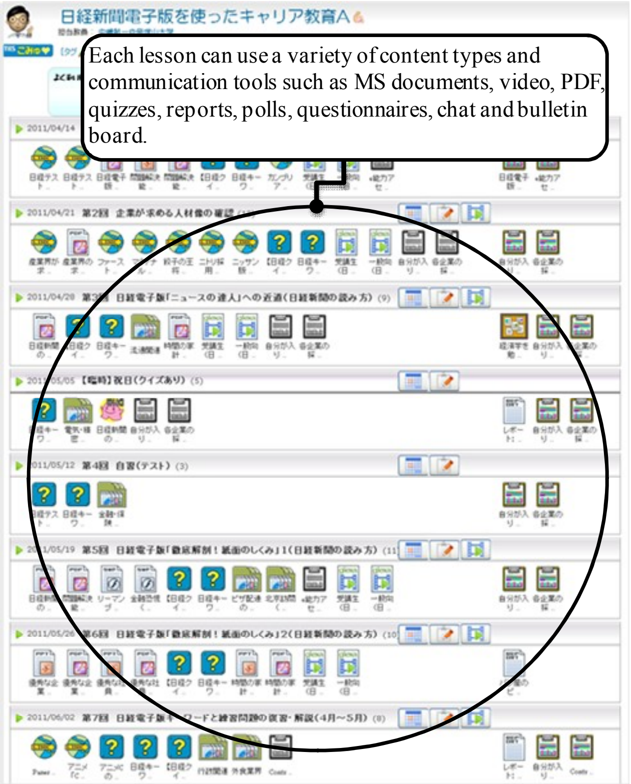
TIES eTeaching Community
1. Content Sharing
2. Knowledge Sharing
3. Know how Sharing
4. Cost Sharing
5. System Sharing

Figure 1. The Concept of eTeaching

TABLE I. THE RECENT GROWTH OF TIES COMMUNITY

Year	2006	2007	2008	2009	2010	2011Feb.
Institutional Users	51	66	73	74	78	83
Instructors	320	801	907	1,021	1,099	1,271
Students	15,099	32,935	46,667	51,783	60,065	70,359
Lectures	548	817	1,053	1,345	1,582	1,810
Video Lectures	660	1,879	3,212	6,181	8,470	11,040
Sharable Content	9,861	15,429	20,801	27,052	33,258	39,417
Lectures Open to the Public	134	186	228	254	258	267

Fig. 2 shows a snapshot of TIES unique interface, where an instructor can have a bird's-eye view of her syllabus, and flexibly create each lesson by selecting and arranging icons according to her own instructional design. She can use all the basic eLearning functionalities such as report and quiz systems, video editor, mobile learning, as well as a Web conference and lecture recording system, ePortfolio and eAssessment systems. Its design also reflects the cultural preference of the Japanese students' love of "cuteness" [2].



Each lesson can use a variety of content types and communication tools such as MS documents, video, PDF, quizzes, reports, polls, questionnaires, chat and bulletin board.

Figure 2. TIES User Interface

III. TIES ASSESSMENT SYSTEM AND FD

In Japan we face the fundamental challenge of how to motivate students to learn more willingly and effectively, and also make them engage in learning more proactively.

The development of TIES eAssessment system is a part of our efforts to solve the issue. The eAssessment project started in 2008 with a premise that our conventional grading system may be overly focused on academic skills, and that a new assessment system is needed to augment the current system by evaluating a student more as a whole-person [3].

The goal of this new assessment is to encourage students to understand their personal strengths and weaknesses, and to reflect on their social skills and self-review their intellectual growth. After two years of intensive discussion involved by every department at Tezukayama University, we have laid out three basic evaluation criteria as follows: (1) academic attributes, (2) personal qualities, and (3) social skills.

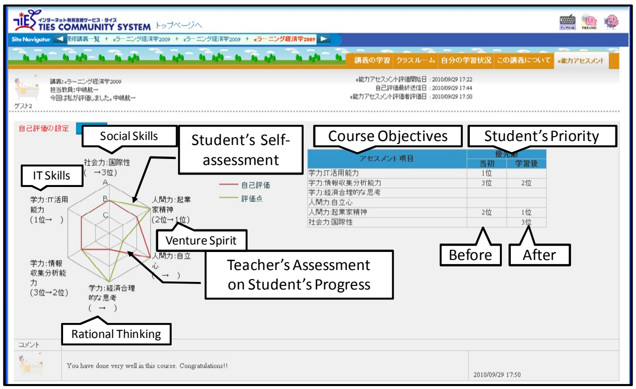
Based on the three criteria, we have identified basic attributes and skills necessary for all students to acquire as well as those specific to their academic majors. For example, academic attributes include abilities such as problem finding and solving, logical and critical thinking. Personal qualities include business manners, aptitude of empathy, and venture spirit, among others. Social skills cover abilities to negotiate and communicate with others as well as the capacity to cope with stress, for example. Instructors are then advised to specify in the syllabus which skills and attributes are related to the course objectives for students to learn.

The eAssessment system is made of a four-step process. In the first step, at the beginning of the course, each student is asked to rank listed attributes and skills according to his or her priority of importance. The second step takes place at the end of the semester, where the student again evaluates the criteria in order of importance to see if there is any change of order after taking the course. The third process is for each student to self-evaluate his progress of attainment on each criterion according to (1) significantly acquired more than before (select A), (2) acquired more than before (select B), and (3) unchanged (select C). That is, if a student thinks he has acquired the required attributes and skills significantly more (more or unchanged) after taking the course than before the course, he selects A (B, C) in the system, respectively. In the final step, the instructor evaluates the progress of the student same way by observing the difference of the student's attainment of each criterion before the course and after the course.

Fig. 3 illustrates the final outcome, and it appears in the student's ePortfolio with a summary and radar chart. Table II is an example of the class data that help an instructor to grasp how students changed their attribute priorities before and after the course. We often observe that students tend to mark A's to the highest order of attributes. Table II can also help us to identify a student who tends to select C's regardless of his order of importance, and receive poor final grades often due to his low self-esteem and lack of self-confidence.

Evaluating students from a wide spectrum of academic attributes, personality development and social skills, thus, has many implications on FD. First, each instructor has to pay more attention to each student as an individual with vast potentials and abilities, and recognize him or her as “a whole person”. Next, it requires an instructor to be aware of how her course is related to the educational mission of the university, and of its relevance to her department curriculum, when she prepares her syllabus. Furthermore, she needs to focus more sharply on the objectives of her teaching context and related assignments in terms of the attributes that she expects her students to learn and acquire. Last but not least, she has to assume additional responsibility for her students’ personal development as well as their academic performance.

Though it requires more time and duty for an instructor to work on, this eAssessment system assists her to establish a close relationship with her students by understanding them better. We believe that this understanding of a student as a whole person will empower an instructor to influence and motivate her students to learn more effectively for better learning outcome.



講義:eラーニング経済学2009
担当教員:中嶋航一
今回は私が評価しました。中嶋航一

e能力アセスメント評価開始日 : 2010/09/29 17:22
自己評価最終送信日 : 2010/09/29 17:44
e能力アセスメント評価者評価日 : 2010/09/29 17:50

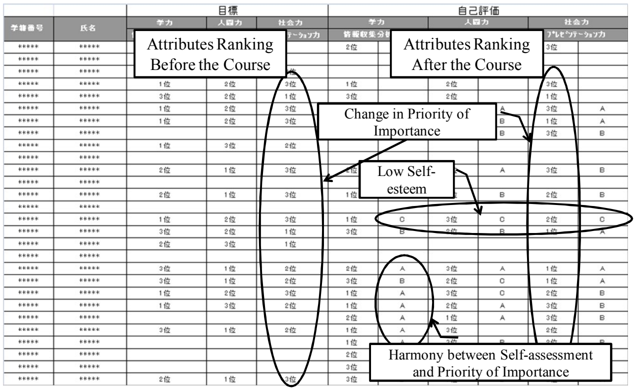
アセスメント項目	優先順 当初	 学習後
学力:IT活用能力	1位	
学力:情報収集分析能力	3位	2位
学力:経済合理的な思考		
人間力:自立心		
人間力:起業家精神	2位	1位
社会力:国際性		3位

You have done very well in this course. Congratulations!!

2010/09/29 17:50

Figure 3. TIES eAssessment

TABLE II. DATA FROM EASSESSMENT


Attributes Ranking Before the Course
Attributes Ranking After the Course
Change in Priority of Importance
Low Self-esteem
Harmony between Self-assessment and Priority of Importance

		目標			自己評価					
学籍番号	氏名	学力	人間力	社会力	学力		人間力		社会力	
		[illegible]	[illegible]	[illegible]テーションカ	情報収集分[illegible]		[illegible]		プレゼンテーションカ	
*****	*****	[illegible]	[illegible]	[illegible]	2位	[illegible]	[illegible]	[illegible]	3位	
*****	*****	[illegible]	[illegible]			[illegible]	[illegible]	[illegible]		
*****	*****	[illegible]	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]		
*****	*****	1位	2位	3位	1位		2位		3位	
*****	*****	3位	2位	1位	3位		2位		1位	
*****	*****	1位	2位	3位	[illegible]	[illegible]	[illegible]	A	3位	A
*****	*****	1位	2位	3位	[illegible]	[illegible]	[illegible]	B	1位	A
*****	*****				[illegible]	[illegible]	[illegible]	B	3位	B
*****	*****	1位	3位	2位						
*****	*****									
*****	*****	2位	1位	3位	2位	[illegible]	[illegible]	A	3位	B
*****	*****									
*****	*****	2位	1位	3位	1位	[illegible]	[illegible]	B	2位	B
*****	*****									
*****	*****	1位	2位	3位	1位	C	3位	C	2位	C
*****	*****	3位	2位	1位	3位	B	2位	B	1位	A
*****	*****	2位	3位	1位						
*****	*****									
*****	*****	3位	1位	2位	2位	A	3位	A	1位	A
*****	*****	3位	1位	2位	3位	B	2位	C	1位	A
*****	*****	1位	2位	3位	1位	A	3位	C	2位	B
*****	*****	1位	3位	2位	1位	A	2位	A	3位	B
*****	*****				2位	A	1位	A	3位	B
*****	*****	3位	1位	2位	1位	A	3位		2位	
*****	*****				1位	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
*****	*****				2位	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
*****	*****				3位	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
*****	*****	2位	1位	3位	3位	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

IV. LPEC QUESTIONNAIR AND FD

In order to complement the eAssessment, we have also investigated the potential impact of different teaching styles on students’ learning motivation and performance. As a result, we have developed a questionnaire called LPEC to assess students’ learning performance by identifying their preferences for teaching and class management styles, and tested its validity since 2005. The questionnaire consists of two sets of questions, asking students what kind of class management and teaching styles they like and dislike. Each student selects 8 out of a first set of 24 styles of class management and teaching styles that he likes, and another 8 out of a second set of 24 styles that he dislikes.

Then each one of those 8 selected answers per set of question is classified to four criteria of logic (L), plan (P), emotion (E) and creativity (C). Finally, they are combined to yield the average class distribution of LPEC preferences.

Questions to identify a student’s preference for L are like “teaching style based on logic, fact and evidence”, while “teaching style emphasizing creativity and new knowledge, a big picture and holistic approach” are categorized as a preference for C. Similarly, keywords such as “a step-by-step learning”, “concrete and procedural”, “teaching with a clear answer” are considered to belong to P, while keywords like “group work”, “role playing”, “student empowerment” are considered to show students’ preference for E.

In this research area, the seminal work has been done by Felder and others [4][5]. They categorize students into four main learning styles as (1) active vs. reflective learners, (2) sensing vs. intuitive learners, (3) visual vs. verbal learners, and (4) sequential vs. global learners. In order to identify students’ learning categories, they have created an Index of Learning Styles Questionnaire (ILSQ in short). ILSQ is made of 44 binary questions, asking a student to answer the questions like “I find it easier” with (a) to learn facts, or (b) to learn concepts [6].

In addition to ILSQ, Glynn et al. propose the Science Motivation Questionnaire (SMQ in short) to use five factors that may influence a student’s learning performance. Those five factors are (1) intrinsic motivation and personal relevance, (2) self-efficacy and assessment anxiety, (3) self-determination, (4) career motivation, and (5) grade motivation [7].

Our LPEC differs from ILSQ or SMQ first that the LPEC is trying to identify a student’s learning preferences by asking his preferences for teaching and class management styles. This approach is based on our casual observation that students in Japan seem to have stronger opinions on our teaching styles and class management rather than their own learning styles. This may be due to the fact that most of the students in Japan have to adapt their learning styles to their instructors’ teaching styles.

Second, we also observe that students seem to know their likes and dislikes better than whether they are sensory or intuitive, and that students do not necessarily understand their motivation as assumed by ILSQ or SMQ. Third, since students often feel lazy to answer many questions in a questionnaire, we have avoided asking them complicated or confusing questions that lose their interests. Last but not least, some of our questions in LPEC reflect Japanese cultural values that may not be covered by ILSQ or SMQ.

While the use of our questionnaire to identify student preferences is valid or not calls for more research, we like to present some data and attempt to interpret preliminary results. Fig. 4 shows the fixed-point observation of the LPEC of the 2011 course called eLearning Economics, which teaches a wide range of topics from economics, finance, and IT.

As for the eAssessment, we have selected six attributes for the course: they are abilities to (1) find appropriate questions, (2) solve questions, (3) collect and analyze data, (4) take actions, (5) apply rational thinking of economics, and (6) pay attention to the global business trends.

The data is collected at the beginning of the course (Apr.15), at the midterm (Oct.14) and at the end of the course (Jan.27). The sample sizes of students answering the questionnaires are 37, 32, and 32, respectively. Most of them are sophomore and junior students.

Though this course is offered as one-year course, we announced at the end of the first half of the course before the summer vacation that we would change the teaching style radically from one-way teaching by an instructor to team learning and students' engagement and empowerment.

Fig. 4 shows a marked shift of students' preference from P to E in the second half of the course. The P type of teaching is a traditional teaching method of deduction, while E is characterized by the keywords like team work, friendly class atmosphere, communication and collaboration.

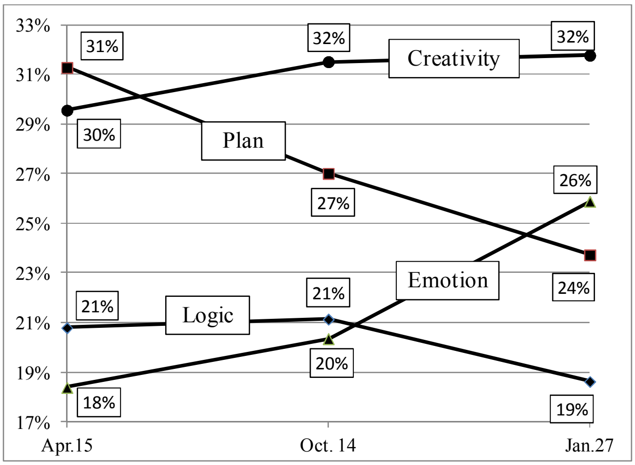


Figure 4. LPEC of eLearning Economics

In addition to the shift of the students' preferences from P to E, data of self-assessment results like Table II provided by students of the class clearly suggest that they are more confident of their learning performance, and that they think they have learned many of the assessment objectives more than they started the course. And unlike a case highlighted in Table II, there was no student marking all C's in this class.

Then, we made more direct question asking the students which teaching style of the semester, first or second, they preferred. The response is that 76% of the students preferred the second semester while the rest liked the first semester. Comments from students indicate that they think they learned more deeply and acquired assessment objectives better in the second semester than the first. Many of them used the phrase like "my learning style harmonizes better with the teaching style of the second semester than that of the first". Thus, we conclude that the class distribution of LPEC can be significantly influenced if an instructor can prepare an appropriate instructional design to align with students' tacit learning preferences.

This result sharply contrasts with the class distribution of LPEC obtained from Development Economics taught in the same year as seen in Fig. 5, where the class size is 28. We did not change the original teaching style of the class, and Fig. 5 suggests that students did not change their learning preferences, either. Thus, the LPEC distribution is fairly stable throughout a year unless the instructor deliberately changes the "rule of the game".

Next, if such a conspicuous shift of the LPEC distribution seen in Fig. 4 frequently occurs or not, we have checked all the available LPEC data of the past eLearning Economics, and summarized them in Table III, where S is a class size.

Unlike the case of 2010 in Fig. 4, it is clear from Table III that none of the LPEC numbers taken from the past four years of the eLearning Economics changed abruptly. Likewise, we have checked the available past data of LPEC of Development Economics and found that they are fairly stable as seen in Table IV.

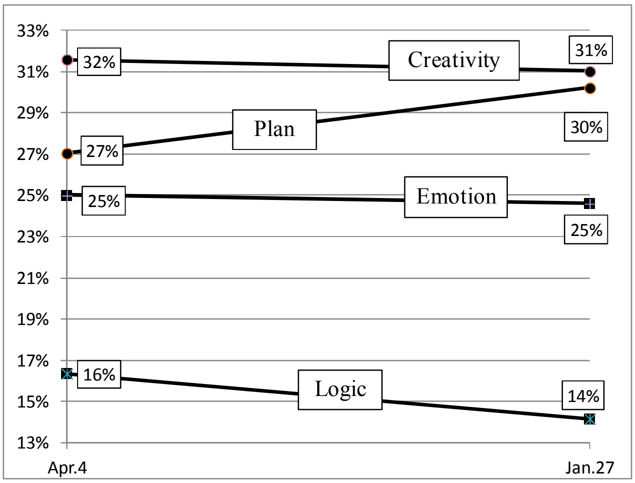


Figure 5. LPEC of Development Economics

TABLE III. LPEC OF ELEARNING ECONOMICS: 2006 - 2009

	2006		2007	
	2006 Apr.14	2007Jan.12	2007Apr.20	2008Jan.25
L	19%	23%	21%	25%
P	29%	25%	27%	28%
E	18%	20%	22%	21%
C	34%	32%	30%	26%
S	115	60	45	26

	2008		2009	
	2008Apr.18	2008Jul.25	2009Apr.10	2010Jan.29
L	22%	22%	21%	23%
P	27%	29%	29%	27%
E	20%	20%	20%	23%
C	31%	29%	29%	28%
S	44	40	40	33

TABLE IV. LPEC OF DEVELOPMENT ECONOMICS: 2007-2009

	2007		2008		2009
	2007 Apr.13	2007 Jul.13	2008 Apr.8	2008 Jul.22	2009 Apr.10
L	21%	21%	19%	18%	18%
P	23%	21%	29%	26%	27%
E	24%	29%	21%	24%	22%
C	32%	29%	31%	33%	32%
S	22	19	27	23	36

With a caveat that the LPEC questionnaire may not be valid, these preliminary LPEC course results appear to suggest the followings: (1) students have tacit preferences for teaching style of the class, (2) these preferences are fairly stable and robust regardless of course characteristics, topics, and content, but (3) appropriate instructional design may be able to alter these preferences considerably.

An immediate implication of these results to FD is that an instructor can accommodate students' learning preferences and their assessment priority by adapting her teaching and class management styles to students' preferences. Without this learning-teaching alignment, both students and an instructor may get frustrated with each other, and students' learning may deteriorate as time passes.

More importantly, if an instructor can design her lecture style appropriately based on the LPEC data, she can manage her class more easily and expect better learning outcomes of students. For example, if the instructor wants her students to acquire logical thinking as one of the assessment objectives, she can use L type of teaching style to change and direct students' preferences more toward logic oriented content and context.

V. CONCLUSION

In Japan we face the urgent issue of implementing a more comprehensive evaluation management system, and creating a spontaneous and self-disciplined learning culture among students.

To solve the problem, we have developed a whole-person approach to assess students from unconventional metrics of academic attributes, personal qualities, and social skills.

To complement the eAssessment, we have also done research on relationship between learning styles of students and teaching styles of instructors, and have developed the questionnaire called LPEC to approximate students' learning preferences via their preferences for class management and teaching styles. This questionnaire is intended to augment the eAssessment system by assisting an instructor to align her teaching style with students' assessment priority and preferred styles of learning.

We have presented preliminary results based on the data obtained from two courses, indicating some usefulness of the approach. However, eAssessment objectives are neither defined precisely to stand a rigorous scrutiny, nor applied unexceptionally. Likewise, the LPEC questionnaire may not truly reflect unobservable nature and preference of students. Furthermore, we need to know how to teach students those attributes in practice. That is, how can we teach a student, say, entrepreneurial spirit, which is not observable? And how can we be sure that the student indeed acquires such a spirit after the course?

One way to approach the problem is to use the data from the student's self-evaluation of the entrepreneurial spirit as a dependent variable, and test its correlation with the LPEC distribution data to find out which style of teaching has a statistically significant coefficient. That is, given that the student's self-evaluation is correct, we can identify which teaching styles influence the student's success of acquiring the qualitative concept of entrepreneurial spirit.

Nonetheless, it seems safe to assume that students appear to have preferred styles of class teaching, and that they seem to be fairly stable, maybe due to their past learning practice. However, our small experiment suggests that a change of instructional method and goal can drastically change their preferences for better learning outcomes. Thus, while we admit that the eAssessment with LPEC questionnaire is only an approximation of the student's unobservable abilities and traits, we conclude that further research is worth pursuing.

REFERENCES

[1] F. Marton and R. Saljo, "On Qualitative Differences in Learning: Outcome and Process," British Journal of Educational Psychology, 46, 1976, pp. 4-11.

[2] K. Nakajima, Mobile Internet Technology for A New Style of Learning and Teaching, The 2010 International Conference on e-Learning, e-Business, Enterprise Information Systems, & e-Government, EEE 2010, Proceedings, CSREA Press, 2010, pp. 16-20.

[3] K. Nakajima, Innovation of TIES: eAssessment, Mobile Learning, and Digital Publishing, Tezukayama Journal of Business and Economics, vol.21, March 2008, pp. 187-197, in Japanese.

[4] R.M. Felder and L.K. Silverman, "Learning and Teaching Styles in Engineering Education," Engineering Education, 78 (7), 1988, pp. 674-681. The same paper with a 2002 preface is at http://www4.ncsu.edu/unity/lockers/users/f/felder/public/.

[5] R.M. Felder, and B.A. Soloman, LEARNING STYLES AND STRATEGIES, http://www4.ncsu.edu/unity/lockers/users/f/felder/public/ILSdir/styles.htm.

[6] B.A. Soloman, and R.M. Felder, Index of Learning Styles Questionnaire, http://www.engr.ncsu.edu/learningstyles/ilsweb.html.

[7] S. M. Glynn, G. Taasoobshirazi, and P. Brickman, Science Motivation Questionnaire: Construct Validation With Nonscience Majors, Journal of Research in Science Teaching, 46(2), 2009, pp. 127-146.

Using Git to Manage Capstone Software Projects

An Empirical Research Report

Zhiguang Xu
Department of Math and Computer Science
Valdosta State University
Valdosta, GA, USA
zxu@valdosta.edu

***Abstract*—Distributed software project development has become a reality not only in industry but also in computer science classes nowadays – students and teachers have to leverage time, talent, and resources collaboratively wherever they reside, especially when everyone is working on his/her own schedule, from his/her convenient location, and using various programming systems. In this paper, we will present an empirical study of how *Git*, "a free & open source, distributed version control system", is used in an undergraduate Computer Science (CS) capstone class to facilitate team collaboration for the students and to ease the project assessment and grading tasks for the teachers. Other Git-related aspects such as preventing plagiarization, hosting online public/private project repositories, and improving the student-teacher interactivity during lecture sessions, are also discussed. Despite of the relatively bumpy and steep learning curve in the beginning of the semester, all four groups of students in the capstone class described in this paper benefitted tremendously from Git, which reduced the burdens of version control and group management on their shoulders, increased the collective productivity of their groups, and helped them in completing their substantial software projects successfully. This paper is concluded with a vision on expanding and standardizing the adoption of Git in other Computer Science classes in the future.**

Keywords - Distributed Student Software Project Management; Distributed Version Control System: Git; Computing and Information Sciences Education.

I. INTRODUCTION

CS 4900, *Senior Seminar*, is a project-driven course designed to provide senior capstone experiences for graduating Computer Science majors at Valdosta State University (VSU). In fall of 2011, twelve students in this class formed four groups to write full-fledged *Ruby on Rails*-based Web server applications that were accessible not only from regular Web browsers but also from *Android* mobile clients that they developed.

In Section II, we present reasons why a *Distrusted Version Control System (DVCS)* is very much needed in CS 4900 and what features that it ought to have. Then, in Section III, we will provide a literature survey of popular DVCSs (Git being one of them) under the umbrella context of Collaborative Development Tools. This is followed by Section IV, an in-depth review of how a Git system is setup, configured, and used in CS 4900. Then, in Section V, three workflows with Git are presented to show what kind of services Git (configured in the way as described in Section IV) provides to students and teachers to increase the overall productivity of the whole class. Finally, this paper concludes with Section VI, a vision on and future works planned for expanding and standardizing the adoption of Git in a wider range of Computer Science programming classes.

II. BACKGROUND

In this section, we will provide the *pedagogical motivations* of incorporating DVCS into CS 4900 in fall 2011. Many issues discussed here are also believed to be common concerns that many students and teachers in a CS programming course would be likely to share.

A. Student's Perspective

One of the central challenges for the students in managing their software project development is handling the *update* process among multiple distributed team members without sacrificing or introducing undue overhead. It is such a process that is too time-consuming, error-prone, and chaotic to be done either manually or using some generic Web content management tools such as Google Docs. What they truly need is an automatic version control system that has the following features –

- *Easy branching and merging*. First and foremost, every group member has a complete "sandbox" of the project. Creating branches for fixing bugs, experimenting different designs, or developing new features is easy, cheap, and fast. When the time comes to merge work outputs from multiple group members back together, even multiple times, the job is done in a snap.
- *Platform Neutral*. When multiple students work in a group nowadays like in CS 4900, it is very likely that their computers are running different kinds/versions of operating systems, mainly Windows, Mac OS X, and Linux. Therefore, they need a version control system that works seamlessly across them.
- *Distributed architecture*. Each group member can work on his/her part of the project and commit the

work output locally without the dependency and/or distraction of an always-on Internet connection. In the age of always-on, broadband Internet connections, we forget that sometimes we do not have access to a network [1]. This was truly a concern when literally every student in CS 4900 was working on his/her laptop; but, unfortunately, the Wi-Fi signal on VSU campus was not ideal all the time. (It was the case even on the day when students did their final project demos.)

B. *Teacher's Perspective*

As what you will see in Section V later, DVCSs satisfy all students' needs above. In fact, they have garnered significant attention in developer communities [2] while attracting relatively little in CS education. Exposing students to and familiarizing them with such an important aspect of the software development process was *the primary motivation* that drove me to include it in CS 4900. In addition, the following items were also behind the adoption of a DVCS for student projects in such a capstone class. They are elaborated in section V.

- *Preventing Plagiarization and "Free-Riders"*.
- Being unobtrusive to undergraduate level students, both conceptually and mechanically.
- *Fitting an Educational Setting*. It should not require a significant commitment of administrative, technical, and financial resources to be successful in an educational institution.

C. *Social Context*

In contrast to faculty members who belong mostly to the "baby boomers" and "X generations", college students sitting in our Computer Science classrooms today are also known as the "*Generation-Yers*" [5], who embrace mobile phones and cloud-computing based social networks as part of their daily lives. The latter is of particular importance to the subject of this paper because it includes not only general social networking sites such as *Facebook* and Web content management sites such as *Google Docs*, but also "social coding" sites such as *Github* – the primary online source code repository hosting site used by the student projects in CS4900. It is the students' digitized cultural background that makes it such a natural and smooth process to transit from the manual, tedious, and error-prone way of managing software projects to a DVCS.

III. DVCS AND GIT

Lanubile et al. did a comprehensive survey on collaboration tools for global software engineering in [2], which include Trackers, Build Tools, Modelers, Knowledge Centers, Communication Tools, Web 2.0 Apps, and of this paper's most interest, version control systems. Subversion is a popular version control system. But, it adopts a traditional centralized architecture, which does not fit well for the educational setting for reasons as described in Section II and in [2]. Git, Mercurial, and Darcs are distributed systems that operate in a peer-to-peer manner, where each local clone of the project is a full-fledged repository with complete history and full revision tracking capabilities, not depending on network access or a central server.

Although there are technical differences between these DVCSs and the decision of choosing Git in CS 4900 was quite of my personal preference, there were a few legitimate factors that reinforced my decision: first, Git is the built-in version control mechanism of Rails, the platform students used to build their Web server applications in this class; second, *Eclipse*, the Integrated Development Environment (IDE) students used for both Rails and Android programming has a Git plug-in that makes version control a natural step in their project developing cycle; and third, Github [3], the most popular online Git repository hosting site, offers educational accounts to host not only public but also private repositories for free, which is greatly convenient for authenticated accesses to both individual and group projects in CS4900.

There are some drawbacks of using Git that one needs to put under consideration. First, Windows support is still lagging behind. You simply cannot use Git from a normal command prompt. Second, there is a long and rough learning curve for students before they feel comfortable using Git.

Next, Section IV discusses how Git is used in CS 4900 from the mechanical view that focuses on various components in such a distrusted system; then Section V covers it from the "Service" view, i.e., workflows that demonstrate how the students and teacher can use and take advantages of Git.

IV. GETTING GIT TO WORK

A. *Local Git Repositories*

For each project, either individual or group, each student has a local Git repository (see the `.git/` directory in Figure 1). The *working tree* is student's current view into the repository [6]. After making changes to the files on the working tree, through the *staging area*, he/she can *commit* the changes to one of the working sets, known as *branches*, in his/her local repository and store a log message/comment explaining what the change did. (The use of such logs and comments will be more covered later in Section V.)

Each student could have as many local branches as he/she wants and *checkout* anyone at any time to start/ continue to work on it. Among these local branches, one is of special importance – The *master* branch serves as the "interface" branch to other group members and the teacher. It always stores the most current version of the project, which gets *pushed* to other students in the same group for sharing the collaborating purposes or to the teacher to be graded. When a newer version of the project from some other group members, or when a graded version of the project from the teacher, becomes available, it gets *pulled* in onto the master branch.

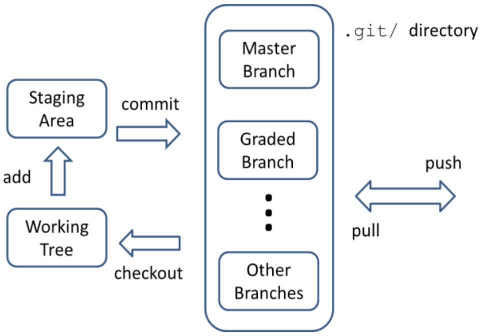


Figure 1. Student's Local Repo per Project

Teacher's project-based local repository looks structurally similar to Students', except that its branches hold graded code turned in by students, i.e., one branch per student (see Figure 2).

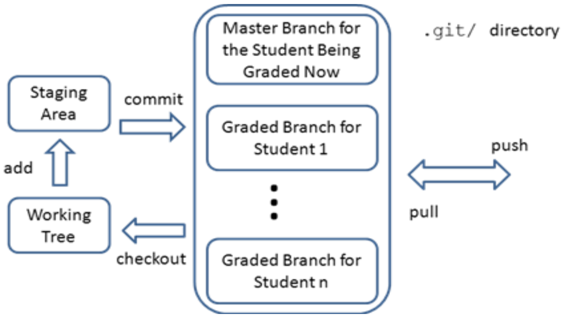
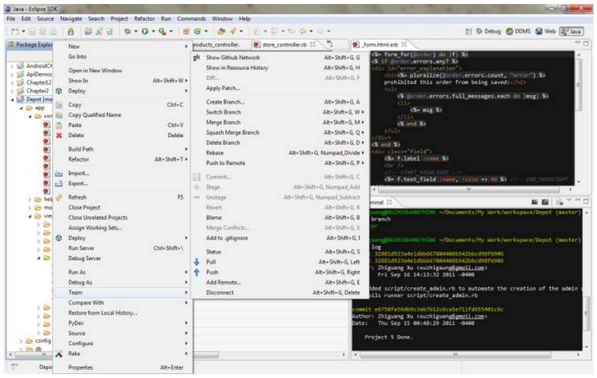

.git/ directory
Master Branch for the Student Being Graded Now
Graded Branch for Student 1
Graded Branch for Student n
Staging Area
commit
add
Working Tree
checkout
push
pull

Figure 2. Teacher's Local Repo per Project

The Git push and pull operations described above are actually performed to and from online Web-based hosting service at *Github* (see Section III.B) and might incur conflicts handling and branch merging [6].

Git plug-in for Eclipse makes it very easy for all Git operations to be conducted from within Eclipse either through GUI items or more conveniently in an embedded shell (see Figure 3).



```
<%= form_for(@order) do |f| %>
<% if @order.errors.any? %>
<div id="error_explanation">
    <h2><%= pluralize(@order.errors.count, "error") %>
    prohibited this order from being saved:</h2>
    <ul>
        <% @order.errors.full_messages.each do |msg| %>
        <li>
            <%= msg %>
        </li>
        <% end %>
    </ul>
</div>
<% end %>
<div class="field">
    <%= f.label :name %>
    <br />
    <!-- START_HIGHLIGHT -->
    <%= f.text_field :name, :size => 40 %> <!-- END_HIGHLIGHT -
```

```
uang@XUZHIGUANGTHINK ~/Documents/My Work/workspace/Depot (master)
 branch
er

uang@XUZHIGUANGTHINK ~/Documents/My Work/workspace/Depot (master)
 log
: 32881d923a4e1dbb667804408b542bbcd98f6906
: 32881d923a4e1dbb667804408b542bbcd98f6906
: Zhiguang Xu <xuzhigaung@gmail.com>
  Fri Sep 16 14:12:32 2011 -0400

dded script/create_admin.rb to automate the creation of the admin
ils runner script/create_admin.rb

commit e6758fa56db9c3ab7b12c6ca5e711fd655401c6c
Author: Zhiguang Xu <xuzhiguang@gmail.com>
Date:   Thu Sep 15 00:48:29 2011 -0400

    Project 5 Done.
```

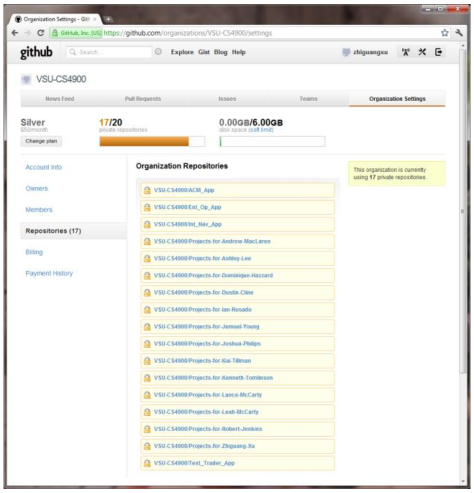
Figure 3. A snapshot of Git in Eclipse

B. *Online Public and Private Repositories on Github*

Github was chosen as the online Git repository hosting site for CS 4900 due to the following reasons:

- Free public AND private repositories, thanks to Gitbub's educational program, that allow students and teacher to access their projects from anywhere at any time
- Secure source code backup in the Cloud (True story – one student's laptop crashed in the middle of the semester and it was his backups on Github that saved his project)
- Clean and fast submission and grading of projects, especially when their sizes go beyond megabytes
- Rich tools for administrating student groups, visualizing students' contributions to their group projects, archiving projects for future course assessments, and much more

We created an *Organization* "VSU-CS4900" on Github that has 13 members (12 students and 1 teacher) and 17 private repositories (12 for individual projects, 4 for group projects, and 1 for the teacher, see Figure 4).



Silver
$50/month

17/20
private repositories

0.00GB/6.00GB
disk space (soft limit)

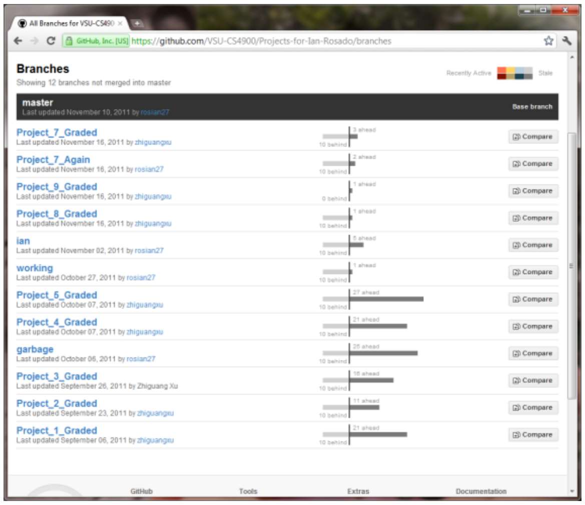
Organization Repositories

This organization is currently using **17** private repositories.

VSU-CS4900/ACM_App
VSU-CS4900/Ent_Op_App
VSU-CS4900/Int_Nav_App
VSU-CS4900/Projects-for-Andrew-MacLaren
VSU-CS4900/Projects-for-Ashley-Lee
VSU-CS4900/Projects-for-Dominique-Hazzard
VSU-CS4900/Projects-for-Dustin-Cline
VSU-CS4900/Projects-for-Ian-Rosado
VSU-CS4900/Projects-for-Jemuel-Young
VSU-CS4900/Projects-for-Joshua-Philips
VSU-CS4900/Projects-for-Kai-Tillman
VSU-CS4900/Projects-for-Kenneth-Tomlinson
VSU-CS4900/Projects-for-Lance-McCarty
VSU-CS4900/Projects-for-Leah-McCarty
VSU-CS4900/Projects-for-Robert-Jenkins
VSU-CS4900/Projects-for-Zhiguang-Xu
VSU-CS4900/Text_Trader_App

Figure 4. Private Repositories on Github

Each individual repository has two owners – a student (e.g., Ian) and the teacher – who both have full privileges, and a number of branches. The *master* branch always stores the most up-to-date version of the current individual project (e.g., #5) that Ian is working on. One the due date, the teacher will pull the project on the master branch and grade it. Once the grading is done, the graded project is pushed up to the *Project_5_Graded* branch on Ian's individual repo for him to review. There might be other branches in his individual repo that Ian creates for himself (see Figure 5).



https://github.com/VSU-CS4900/Projects-for-Ian-Rosado/branches

Branches
Showing 12 branches not merged into master

master
Last updated November 10, 2011 by rosian27

Project_7_Graded
Last updated November 16, 2011 by zhiguangxu
3 ahead, 10 behind

Project_7_Again
Last updated November 16, 2011 by rosian27
2 ahead, 10 behind

Project_9_Graded
Last updated November 16, 2011 by zhiguangxu
1 ahead, 0 behind

Project_8_Graded
Last updated November 16, 2011 by zhiguangxu
1 ahead, 10 behind

ian
Last updated November 02, 2011 by rosian27
5 ahead, 10 behind

working
Last updated October 27, 2011 by rosian27
1 ahead, 10 behind

Project_5_Graded
Last updated October 07, 2011 by zhiguangxu
27 ahead, 10 behind

Project_4_Graded
Last updated October 07, 2011 by zhiguangxu
21 ahead, 10 behind

garbage
Last updated October 06, 2011 by rosian27
25 ahead, 10 behind

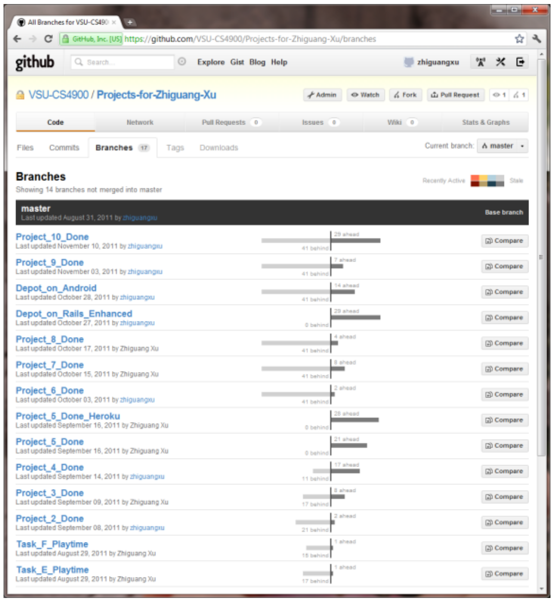
Project_3_Graded
Last updated September 26, 2011 by Zhiguang Xu
16 ahead, 10 behind

Project_2_Graded
Last updated September 23, 2011 by zhiguangxu
11 ahead, 10 behind

Project_1_Graded
Last updated September 06, 2011 by zhiguangxu
21 ahead, 10 behind

Figure 5. A Student's Individual Repo on Github

Teacher's repo has only one owner (the teacher) and a number of branches. The master branch as usual serves as the interface branch and the rest branches store the solutions to the student projects and example projects for class lectures (see Figure 6).



https://github.com/VSU-CS4900/Projects-for-Zhiguang-Xu/branches

VSU-CS4900 / Projects-for-Zhiguang-Xu

Branches
Showing 14 branches not merged into master

master
Last updated August 31, 2011 by zhiguangxu

Project_10_Done
Last updated November 10, 2011 by zhiguangxu
29 ahead, 41 behind

Project_9_Done
Last updated November 03, 2011 by zhiguangxu
7 ahead, 41 behind

Depot_on_Android
Last updated October 28, 2011 by zhiguangxu
14 ahead, 41 behind

Depot_on_Rails_Enhanced
Last updated October 27, 2011 by zhiguangxu
29 ahead, 0 behind

Project_8_Done
Last updated October 17, 2011 by Zhiguang Xu
4 ahead, 41 behind

Project_7_Done
Last updated October 15, 2011 by Zhiguang Xu
8 ahead, 41 behind

Project_6_Done
Last updated October 03, 2011 by zhiguangxu
2 ahead, 41 behind

Project_5_Done_Heroku
Last updated September 16, 2011 by Zhiguang Xu
28 ahead, 0 behind

Project_5_Done
Last updated September 16, 2011 by Zhiguang Xu
21 ahead, 0 behind

Project_4_Done
Last updated September 14, 2011 by zhiguangxu
17 ahead, 11 behind

Project_3_Done
Last updated September 09, 2011 by Zhiguang Xu
8 ahead, 17 behind

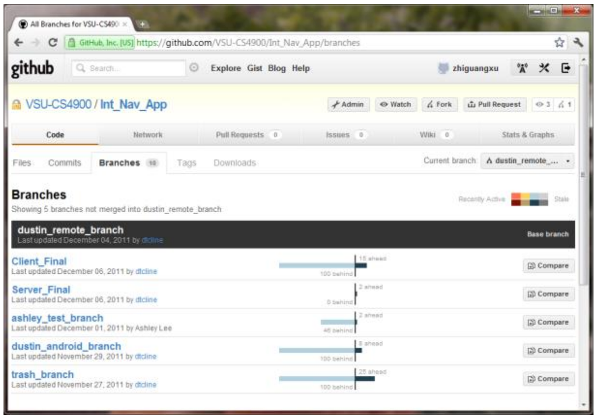
Project_2_Done
Last updated September 08, 2011 by zhiguangxu
2 ahead, 21 behind

Task_F_Playtime
Last updated August 29, 2011 by Zhiguang Xu
1 ahead, 15 behind

Task_E_Playtime
Last updated August 29, 2011 by Zhiguang Xu
1 ahead, 17 behind

Figure 6. Teacher's Repo on Github

Each group repo has four owners (three students in the group and the teacher) and a number of branches, at least two of which store the final version of their client side code and server-side code respectively (see Figure 7).



https://github.com/VSU-CS4900/Int_Nav_App/branches

VSU-CS4900 / Int_Nav_App

Branches
Showing 5 branches not merged into dustin_remote_branch

dustin_remote_branch
Last updated December 04, 2011 by dtcline

Client_Final
Last updated December 06, 2011 by dtcline
15 ahead, 100 behind

Server_Final
Last updated December 06, 2011 by dtcline
2 ahead, 0 behind

ashley_test_branch
Last updated December 01, 2011 by Ashley Lee
2 ahead, 46 behind

dustin_android_branch
Last updated November 29, 2011 by dtcline
8 ahead, 100 behind

trash_branch
Last updated November 27, 2011 by dtcline
25 ahead, 100 behind

Figure 7. A Group Repo on Github

V. THE WORKFLOWS WITH GIT

This section presents three workflows with Git to show what kind of services Git (configured in the way as described in Section IV) provides to the students and the teacher to increase the overall productivity of the whole class. In the end, you will also find discussions on a challenging issue that we have encountered and how we addressed it.

- Developing, Submitting, and Deploying Projects. In a group of three students A, B, and C, student A is the "group leader" (see Figure 8). As the project progresses, each student is able to push and pull his/her newest work output to and from his/her remote branch in their group repo on Github. Only A has the privilege to pull code from everyone's remote branch, merge them, and push the result to the master branch, which consequently stores the most current version of the project for everyone to pull so as to be code-synchronized.

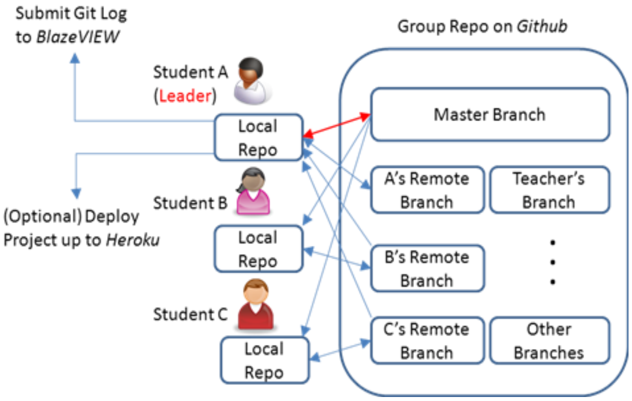


Figure 8. Student Workflow

When the project is finished, the group leader will submit the Git log file to BlazeVIEW, a Blackboard based online course management system at VSU, and optionally deploy the server side of the code to

Heroku [7], a cloud based, Rails friendly application platform.

- Grading Projects. When grading a project, the teacher pulls the code from the master branch in the project's Github repo, builds it locally, and runs it. But more importantly, the teacher heavily relies on the revision logs to see each group member's contributions to the final project (see Figure 9). These logs also provide an audit trail for determining if students followed the incremental process, which will be demonstrated by a logical and coherent sequence of commit messages that indicate a methodical progression toward the end goal. At each commit, students must stop and describe their work in a commit comment, which forces reflective pauses and helps promote an intentional attitude toward their work [4]. The graded project is then pushed up to the Teacher's branch on Github for students to review.

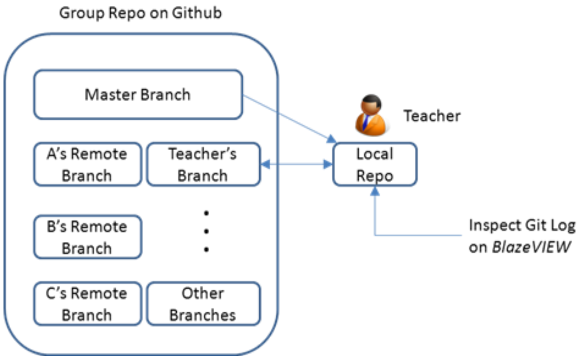


Figure 9. Teacher Workflow

In fact, the flexibility that Git extends in terms of setting up local and online repositories greatly helped how students' projects were graded in CS 4900. In addition to the semester-long projects as mentioned in the introduction section of this paper which constituted the major component in CS 4900, there were around ten "practice" projects that were designed to get students technically ready for their "big deals" (Note, Ruby and Rails and Android programming were new to most of the students in this class), and they had to accomplish these "practice" projects individually. On the other hand, a related issue that concerns lots of CS teachers (me included) is how to assess individual student's performance in group projects. Obviously, the best way to detect cheating in individual projects and free-riding in group projects is to have a version control system that comes with a rich and sane logging history that records each and every commit of intermediate work output along the evolvement of the project, based on which students can justify their progresses towards and contributions to the final product.

- Discussing Example Code in Class Lectures. For better student-teacher interactions and more efficient use of the class lecture time, Git makes it very easy for the teacher to checkout a new branch and elaborate critical code step by step to students in class (see steps 1 and 3 in Figure 10) and skip non-essential parts by checking out the commits that conclude them (see steps 2 and 4 in Figure 10) and move on.

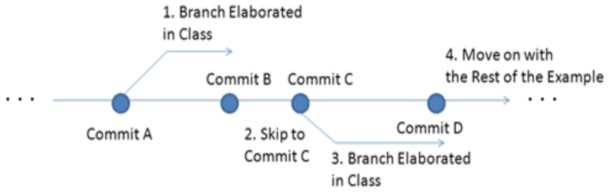

1. Branch Elaborated in Class
4. Move on with the Rest of the Example
Commit B
Commit C
Commit A
2. Skip to Commit C
Commit D
3. Branch Elaborated in Class
. . .
. . .

Figure 10. Discussing Examples in Class Lecture

- Challenging Issues. In addition to the bumpy road in the beginning of the semester mainly to get familiar with Git, inevitably, there were a few issues students encountered that held them from moving on with their projects but fortunately found solutions to [8]. For instance, although Git worked perfectly with Rails for the development of their servers, it gave students hard time merging work outputs from multiple group members on the Android client side into one new version by generating all sorts of conflicts. They found that the .gitignore file was their friend which allowed them to specify which files they did not want Git to track, specifically for Android, the ones in the bin/ and gen/ folders, for they will be automatically generated during the build process anyways.

VI. Conclusions and Future Work

Our initial experience with Git and DVCS in general has been very positive. We have seen senior students in the Capstone class voluntarily and comfortably use Git as the distributed version control system for their projects. Git gives them unique opportunities and exposures to collaborative and real-world practices that are prevalent in today's distributed software development community. As the students gain experience and competitive skills with the version control system that will be integrated into CS 4900, such skills scale with them, enabling them to collaborate with their peers, contribute to open source software projects, and eventually transfer their new knowledge to the work environment [2]. It also streamlines my work as a teacher in terms of grading student projects and giving lectures.

Future work includes expanding the adoption of Git in a wider range of Computer Science classes that emphasize students' programming skills. In particular, we are also going to enrich the process of composing Git commit comments [9] to help keeping them from getting too general, vague, and/or uninformative.

REFERENCES

[1] F. Lanubile, C. Ebert, R. Prikladnicki, and A. Vizcaíno, “Collaboration Tools for Global Software Engineering,” IEEE Software, March/April 2010, pp. 52-55.

[2] B. de Alwis and J. Sillito, “Why are software projects moving from centralized to decentralized version control systems?” CHASE ’09: Proceedings of the 2009 ICSE Workshop on Cooperative and Human Aspects on Software Engineering, IEEE Computer Society, Washington, DC, USA, 2009, pp. 36–39.

[3] http://www.github.com, retrieved: June, 2012.

[4] D. Rocco and W. Lloyd, “Distributed Version Control in the Classroom,” ACM SIGCSE’11, Dallas, Texas, USA, March, 2011, pp. 637–641.

[5] G. Thiruvathukal, K. Laufer, and D. Dennis, “Moving Academic Department Functions to Social Networks and Clouds: Initial Experience,” IEEE Computing in Science and Engineering, September/October, 2011, pp. 84–89.

[6] T. Swicegood, “Pragmatic Version Control Using Git,” Publisher: Pragmatic Bookshelf, ISBN: 1-934356-15-8.

[7] http://www.heroku.com, retrieved: June, 2012.

[8] C. Bird, P. Rigby, E. Barr, D. Hamilton, D. German, and P. Devanbu, “The Promises and Perils of Mining Git,” MSR ’09 Proceedings of the 2009 6th IEEE International Working Conference on Mining Software Repositories, Washington, DC, USA, 2009.

[9] M. D’Ambros, “Commit 2.0: Enriching Commit Comments with Visualization,” ICSE ’10, Cape Town, South Africa, May, 2010.

An Overview of the Student-to-Student Chinese Language Learning Lab

Lilia Cai-Hurteau
Educational Theory and Practice
University at Albany, SUNY
Albany, NY, USA, 12222
liliacai@hotmail.com

Peter Shea
Educational Theory and Practice
University at Albany, SUNY
Albany, NY, USA, 12222
Pshea@albany.edu

***Abstract*— This paper represents a report of mid-term results of a United States Department of Education funded International Research and Studies Program project: The Student to Student Chinese Language Lab. This paper gives a summary of the Lab, a brief assessment of related theory and research, measures of learning, as well as a critical evaluation of other existing resources. We conclude that the lab provides critical additional benefits beyond similar, existing resources.**

Keywords - Language; Chinese; Online Technology; Language Lab.

I. Introduction

Our Lab is a novel Mandarin Chinese language and cultural resource for primary and secondary schools that offers curricular structure, dynamic multimedia content, opportunities for teacher and student input, and mechanisms for interaction. The Lab also offers the distinctive motivational factor of learning from age-matched peers who speak the target language as a first language. This student-to-student feature creates a powerful learning nexus that creates a personal and meaningful context critical for second language acquisition. The Lab also provides an 'indirect communication' approach that allows teachers to bypass typical difficulties associated with 'direct communication,' such as time differences, cultural differences regarding value and expectations, inappropriate student input and a tendency to stray from the curriculum. At the same time, the Lab is associated with opportunities for direct teacher-monitored student interaction. Furthermore, the project pioneers a new learning model that utilizes modern communications technology to promote mutual understanding and cooperation, and can potentially be replicated at different levels and with different languages.

This paper is organized as follows; first we review the intended audience of the lab, then we discuss the overarching goals of the project, related research, metrics of learning and a critical comparison of similar efforts to support Chinese language learning with online resources.

II. Intended audience

The Student to Student Chinese Language Lab is designed for and made available to all Internet-connected schools offering Mandarin Chinese Language classes. Extensive review of curriculum in k-12 US schools provided alignment between common content and materials available through the Lab. Some Lab sections may be used by teachers teaching about contemporary life in China. Colleges, universities, adult and community learning institutions, after-school programs and summer education programs teaching Beginner and Intermediate Mandarin Chinese or Chinese culture may also find the Lab useful. The Lab will also be available for teacher training, research purposes and as a resource for informal learning across many contexts. The project is managed by the Research Foundation of the State University of New York at Albany, and hosted by OneWorld Classrooms, a non-profit organization that promotes curriculum-based international exchanges.

III. Goal of the project

The primary goal is to develop a total of 60 curriculum-based Lab modules, each centering on a cultural theme and grammatical element; containing a primary vocabulary list; audio and video featuring students from China speaking in Mandarin Chinese and sharing different aspects of their lives and cultures as well as video of students writing the Chinese characters on the vocabulary list.

Each module also includes vocabulary flashcards featuring student artwork. Some modules include a related PowerPoint featuring photos of students and video of students introducing themselves in Chinese and English. Another main objective is to develop an attractive, dynamic and intuitive Flash-based multimedia interface to present Lab modules to Mandarin Chinese students.

A corresponding goal of the project is to develop an electronic format through which all teachers who use the Lab may complete Lab evaluations, submit feedback, suggest future modules that match their curricula, and interact with other teachers.

IV. Related Research

The Student to Student Chinese Language Lab (Lab) model addresses needs highlighted in contemporary foreign language acquisition research in a number of ways. The Lab promotes cultural and linguistic competence, simulates socially mediated activity, utilizes multiple modalities, offers multiple ability tasks, and represents a platform for authentic language use. Each of these affordances is discussed in more detail below.

The Lab helps students to develop cultural and linguistic competence. Cultural competence, the ability to interact effectively with people of different cultures [1], and linguistic competence, the ability to speak and understand language in the context of its culture [2], are indisputably integral to foreign language learning. By presenting skits focusing on specific cultural themes, Chinese students introduce cultural knowledge into Lab modules. Students using this content in the US will discover cultural similarities and differences, which, in turn, will stimulate higher order thinking and positive attitudinal responses that benefit language acquisition.

The Lab also simulates socially mediated activity. According to sociocultural constructivism [3], learning is a socially mediated activity and peer interaction is central to knowledge construction. Through active interaction “language learning activity mirrors genuine human communication, the chief locus for understanding the world and self” [4].

In addition, the Lab lowers the learner’s affective filter and utilizes multiple modalities. The Lab simulates a total immersion environment, introducing students in China who model natural conversation as “comprehensible input” [5]. The Lab’s intrinsic interest, challenge and reward improve learners’ motivational level and lower the anxiety of learning a foreign language, thus lowering students’ “affective filter” [5]. The engagement of multiple modalities (image, audio, video, etc.) is also a highly positive contributing factor for the language learning process [6].

The Lab is also an example of Computer-Assisted Language Learning (CALL) that offers multiple ability tasks. [7], in addressing socio-collaborative language learning, terms “multiple ability tasks” as tasks that drive conceptual work and are intrinsically interesting and rewarding; allow different students to make different contributions; use multimedia; involve sight sound and touch; and are challenging.

Finally, the Lab is a platform for authentic language. Computer-mediated communication leads us to rethink the authentic, the authorial, and ultimately, the communicative itself [8]. “Proper” textbook language was artificially modeled to ensure “correct” if obviously staged input, with the result that language learning materials provided highly contrived pattern practice, yielding unnatural and boring dialogues [9]. The Lab features K-12 students in China speaking in naturalistic contexts, offering linguistic input that is up-to-date and true-to-life.

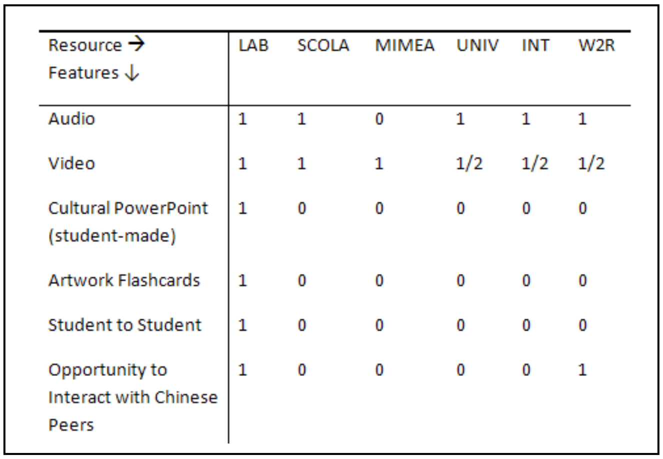
V. Metrics of Learning

With regard to metrics of outcomes of learning, our plan is to continue to develop related assessment and skill testing tools, including: a game and quiz generator that incorporates module content, and an interactive Lab component through which students may record their own voices, save original audio files, and compare their spoken Mandarin with that of their native-speaking peers. Other objectives include a Lab search function that allows teachers and students to locate Lab content that matches their criteria by language proficiency level, cultural theme, grammatical element and vocabulary. We are also seeking to disseminate information about availability of the Lab broadly through various related newsletters, listservs, publications, associations, conferences and networks, and, upon completion of the Lab in Year 3, marketing materials that will be mailed to schools. Finally we plan to assess the impact of the Lab on K-12 student learning of Chinese language and culture and to disseminate results of this initiative to the language teaching and learning community.

VI. Critical Comparison

A table comparing well known existing resources, e.g. SCOLA, Multimedia Interactive Modules for Education and Assessment (MIMEA), University Online Multimedia Chinese Courses (UNIV), General Internet Chinese Resources (INT), common Web 2.0 resources (W2R) and the Lab (LAB) by the features is provided below. In the table, 1 indicates YES, ½ indicates IN SOME CASES and 0 indicates NO. As the table indicates, the Lab is the only resource that offers all 15 features; the next closest (UNIV) offers an aggregate of seven.

Table I. Comparison of the Language Lab and Similar Resources



Resource → Features ↓	LAB	SCOLA	MIMEA	UNIV	INT	W2R
Audio	1	1	0	1	1	1
Video	1	1	1	1/2	1/2	1/2
Cultural PowerPoint (student-made)	1	0	0	0	0	0
Artwork Flashcards	1	0	0	0	0	0
Student to Student	1	0	0	0	0	0
Opportunity to Interact with Chinese Peers	1	0	0	0	0	1

A. Account of Related Materials

The following existing resources, offering hypermedia units for foreign language instruction and assessment, are in some ways similar to the Student to Student Chinese Language Lab.

SCOLA [10] receives and retransmits television programs from around the world in various languages and offers them to schools on a subscription basis. Its Insta-Lessons are hypermedia modules that include a video and an audio clip of a news broadcast segment. Students watch or listen to clips while viewing a transcript, translation, quiz or vocabulary window.

Multimedia Interactive Modules for Education and Assessment (MIMEA) [11] offers a series of language modules in Arabic, Chinese, Korean, Russian, German and Vietnamese. The modules center on video clips that show native speakers and non-native speakers interacting in natural, unscripted situations. Each module also features cultural notes, a quiz and ideas for classroom activities.

University Online Multimedia Chinese Courses (UNIV), such as the California State University's Conversational Mandarin Chinese Online [12], Western Kentucky University's Audio Tutorials of Basic Chinese [13], the University of Oxford's Chinese Multimedia Course [14], Connecticut College's Chinese Video Exercises [15], the Massachusetts Institute of Technology's Open Courseware Learning Chinese [16], and others like them, offer free online theme-based Beginner or Intermediate Level Chinese courses with series of audio and/or video clips and practice activities and grammar explanations.

General Internet Chinese Resources (INT), such as Zap Chinese [17] Language Guide BBC Real Chinese, and others like them, offer lessons in Chinese grammar, vocabulary and the writing system; units on Chinese characters with videos showing how the characters are written, grammar and vocabulary units organized by theme (Zap), audio clips and video (BBC only). 'Web 2.0' Resources (W2R), such as podcasts - like ChinesePod blogs and Popup Chinese, and social networking Websites – like Live Mocha, Mango Languages, and Ning, offer podcast audio and in some cases, opportunities to communicate and interact via the Internet with people who speak Chinese as a first language.

B. *Critical Commentary*

SCOLA's Insta-Lessons and MIMEA's interactive modules are similar to the Student to Student Chinese Language Lab in that they feature video and/or audio, transcripts and assessment. They differ in that they do not feature students who are native speakers or encourage peer-to-peer interaction and connection. SCOLA's videos feature professional newscasters and MIMEA's dialogues cater to college students and adults. Online University courses, while free of charge and very valuable for university Beginning and Intermediate Level Chinese students, are not designed for use in K-12 classrooms and do not feature K-12 students from China in the audio and video sections. The other Internet resources are typically word and sentence-based 'audio-lingual' foreign language teaching resources. With the exception of BBC Real Chinese (which only offers only ten short units), the units do not have dialogues; instead, theme-based sentences are written in Chinese characters and/or pinyin and translated into English. Audio is typically the only interactive feature. Some of these also have outside advertising or require payment for advanced lessons. While podcasts, blogs and social networking Web sites offer independent learning opportunities, a high level of student input and opportunities to communicate via the Internet with people who speak Chinese as a first language, they are typically not designed for K-12 classroom use and therefore do not follow national standards or K-12 Chinese curricula. Likewise, they do not encourage focused and structured learning and do not permit teachers to monitor student communication for appropriateness. While they are a valuable resource for language classes, the do not typically follow scaffolded activities and sequential learning patterns. In some cases, advanced or complete resources are only available for a fee.

The Language Lab can be accessed (for free!) at [19] and [20].

ACKNOWLEDGMENT

This paper was sponsored in part by the US Department of Education International Research and Studies Program. Additional support was provided by the Office of the Provost, State University of New York.

REFERENCES

[1] J. Diller and J. Moule Cultural Competence: A Primer for Educators. Florence, KY: Wadsworth, 2005.

[2] D. Hymes Language in Culture and Society. New York, NY: Harper & Row, 1964.

[3] L.D. Vygotsky, Mind and Society: The Development of Higher Mental Processes. Cambridge, MA: Harvard University Press, 1978.

[4] C. Meskill, C.Computer as tools for sociocollaborative language learning. In CALL: Media, Design and Applications K. Cameron (Ed.) (pp. 141-162). The Netherlands: Taylor & Francis, 1999, pp. 141-162.

[5] S.D. Krashen, The Input Hypothesis: Issues and Implications. United Kingdom: Longman Group.

[6] C. Meskill, "Listening Skills Development Through Multimedia,". Journal of Educational Multimedia and Hypermedia. vol. 5,pp. 179-20.

[7] E. Cohen, Designing Groupwork: Strategies for the Heterogeneous Classroom. New York: Teacher's College Press, 1994

[8] C. Kramsch, F. A'Ness andW.S.E. Lam,"Authenticity and Authorship in the Computer-Mediated Acquisition of L2 Literacy,". Language Learning and Technology, vol. 4, pp. 78-104.

[9] H. Lotherington,Authentic language in digital environments. In CALL research perspectives J. L. Egbert & G. M. Petrie Eds.. New York: Lawrence Erlbaum Associates, 2005, pp. 109-127.

[10] SCOLA (http://www.scola.org) [retrieved: April, 2012]

[11] MIMEA (http://mimea.clear.msu.edu/) [retrieved: April, 2012]

[12] California State University's Conversational Mandarin Chinese Online (http://www.csulb.edu/~txie/ccol/content.htm) [retrieved: April, 2012]

[13] Western Kentucky University's Audio Tutorials of Basic Chinese (http://www.wku.edu/~yuanh/AudioChinese/) [retrieved: April, 2012]

[14] University of Oxford's Chinese Multimedia Course (http://www.ctcfl.ox.ac.uk/Chinese/lessons.htm) [retrieved: April, 2012]

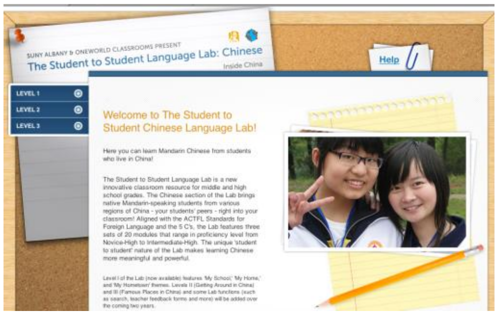
[15] Connecticut College's Chinese Video Exercises (http://www.conncoll.edu/academics/departments/chinese/mhu/videos2/index/index.html) [retrieved: April, 2012]

[16] Massachusetts Institute of Technology's Open Courseware Learning Chinese (http://ocw.mit.edu/OcwWeb/Foreign-Languages-and-Literatures/21F-101Spring-2006/CourseHome/index.htm) [retrieved: April, 2012]

[17] ZAP Chinese (http://www.zapchinese.com/) [retrieved: April, 2012]

[18] http://www.oneworldclassrooms.org/Lab/chinese/app/webroot/index.php/lessons/welcome [retrieved: April, 2012]

[19] http://bit.ly/x0JY9q [retrieved: April, 2012]



SUNY ALBANY & ONEWORLD CLASSROOMS PRESENT
The Student to Student Language Lab: Chinese
Inside China

Help

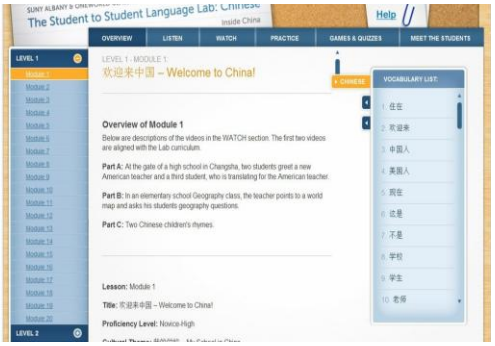
LEVEL 1
LEVEL 2
LEVEL 3

Welcome to The Student to Student Chinese Language Lab!

Here you can learn Mandarin Chinese from students who live in China!

The Student to Student Language Lab is a new innovative classroom resource for middle and high school grades. The Chinese section of the Lab brings native Mandarin-speaking students from various regions of China - your students' peers - right into your classroom! Aligned with the ACTFL Standards for Foreign Language and the 5 C's, the Lab features three sets of 20 modules that range in proficiency level from Novice-High to Intermediate-High. The unique 'student to student' nature of the Lab makes learning Chinese more meaningful and powerful.

Level I of the Lab (now available) features 'My School,' 'My Home,' and 'My Hometown' themes. Levels II (Getting Around in China) and III (Famous Places in China) and some Lab functions (such as search, teacher feedback forms and more) will be added over the coming two years.



SUNY ALBANY & ONEWORLD
The Student to Student Language Lab: Chinese
Inside China

Help

OVERVIEW | LISTEN | WATCH | PRACTICE | GAMES & QUIZZES | MEET THE STUDENTS

LEVEL 1
Module 1
Module 2
Module 3
Module 4
Module 5
Module 6
Module 7
Module 8
Module 9
Module 10
Module 11
Module 12
Module 13
Module 14
Module 15
Module 16
Module 17
Module 18
Module 19
Module 20
LEVEL 2

LEVEL 1 - MODULE 1:
欢迎来中国 – Welcome to China!

Overview of Module 1

Below are descriptions of the videos in the WATCH section. The first two videos are aligned with the Lab curriculum.

Part A: At the gate of a high school in Changsha, two students greet a new American teacher and a third student, who is translating for the American teacher.

Part B: In an elementary school Geography class, the teacher points to a world map and asks his students geography questions.

Part C: Two Chinese children's rhymes.

Lesson: Module 1

Title: 欢迎来中国 – Welcome to China!

Proficiency Level: Novice-High

▸ CHINESE

VOCABULARY LIST:

1. 住在
2. 欢迎来
3. 中国人
4. 美国人
5. 现在
6. 这是
7. 不是
8. 学校
9. 学生
10. 老师

Figure 1. Language Lab Interface

Figure 2. Module View of the Lab

Static Task Allocation Algorithms in Mesh Networks: An Experimentation System and Analysis of Properties

Piotr Franz, Leszek Koszalka
Dept. of Systems and Computer Networks
Wroclaw University of Technology
Wroclaw, Poland
e-mail: leszek.koszalka@pwr.wroc.pl

Iwona Pozniak-Koszalka, Andrzej Kasprzak
Dept. of Systems and Computer Networks
Wroclaw University of Technology
Wroclaw, Poland
e-mail: iwona.pozniak-koszalka@pwr.wroc.pl

***Abstract*—The paper concerns the static task allocation problem in mesh structured system. Three allocation algorithms have been evaluated, including well-known First Fit and Stack Based Algorithm, and newly created by authors the Current Job Based First Fit algorithm. The evaluation of their properties and a comparison of their efficiencies have been done on the basis of simulation experiments. The reported investigations have been made with a designed experimentation system coded in C# language with use of .NET Framework for Windows platform. The discussion of results confirms that the created algorithm seems to be promising.**

Keywords - mesh structure; task allocation algorithm; experimentation system

I. INTRODUCTION

Nowadays, modern computer systems are often created by connecting many processing units into one big structure, in order to solve complex problem more efficient. The performance of such structures depends not only on computing power of single processing units, but also on efficiency of algorithms, which are responsible for allocating tasks in structure and those which are responsible to pick certain tasks from queue of ready for execution tasks. Problems of scheduling (task selection) and allocation are important in terms of reducing cost of computing (saving both time and resources) [1].

In the field of solving allocation problem with efficient algorithm still new ideas are proposed on basis on such approaches as Best Fit or Adaptive Scan or First Fit (see, e.g., [2], [3]) as well as algorithms based on evolutionary concepts (see, e.g., [4]).

The aim of this paper is to examine the three implemented allocation algorithms. The two well-known algorithms, FF (First Fit) algorithm and SBA (Stack Based Algorithm) [3], [5], are considered. We designed the third one, called CJB FF (Current Job Based First Fit), which was initially presented in [6].

The static allocation problem [7] considered in this paper, assumes the two-dimensional mesh topology with closed queue of ready tasks (during allocation process no new tasks are added to the queue/system). We assume that tasks from the queue may be picked for allocation using FIFO or SJF scheme [1].

For the purposes of this paper, the experimentation system was designed and implemented. The system allows multi-aspect comparison of the considered algorithms.

The rest of the paper is organized as follows: Section II contains the used nomenclature. In Section III, the three allocation algorithms are briefly described. Section IV contains description of the experimentation system. In Section V, results of investigations are presented and the obtained results of two complex experiments are discussed. Finally, in Section VI, the concluding remarks are stated.

II. PROBLEM STATEMENT

In order to formulate the task allocation problem considered in this paper, the basic definitions and ideas need to be described.

Mesh is a set of nodes (processors) connected in orderly fashion. The typical, full mesh *M (w, h)* is a rectangular two-dimensional matrix of sizes *w* and *h,* where *w* stands for width and *h* stands for height.

Nodes in a mesh are marked as *(i, j),* where *i* stands for a column and *j* for a row in mesh structure.

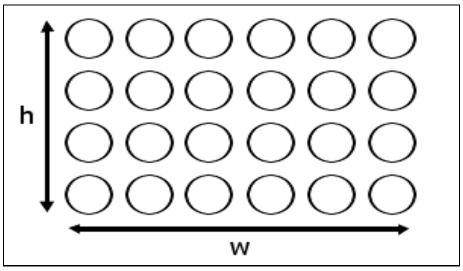


Figure 1. An example of the MESH structure M(6, 4).

Submesh S_M *(i, j, w, h)* is a rectangular set of (*w* x *h) nodes* that belong to a mesh *M (w, h)*. The node *(i, j)* is the foothold of submesh S_M in mesh *M.*

Free submesh is a submesh in which every node is free, i.e. it is not occupied with previously allocated task.

Busy submesh is a submesh in which at least one node is already assigned to execute a task.

Task J (w, h, t) is a rectangular form with known sizes *w* and *h* and execution time *t*. The tasks wait in a queue to be allocated within a mesh. The queue can be a simple FIFO structure or can be sorted (ascending or descending due to execution time of needed nodes number). To allocate each task, the free sub-mesh with a specified size is needed.

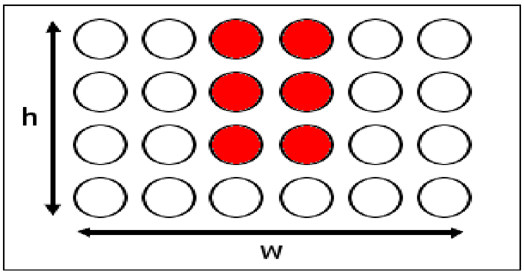


Figure 2. An example: MESH M(6, 4) and a submesh $S_M(3, 1, 2, 3)$.

Expected relative task's width p_w is a ratio of expected task width to mesh size.

Expected relative task's height p_h, similar to p_w, is a ratio of expected task width to mesh size.

Expected relative task's size p is a ratio of expected task size to mesh size (when expected values of task width and task height are equal, then $p = p_w = p_h$).

Allocation problem consists in picking and allocating on a mesh all queued tasks in a way that gives the best results in respect to the introduced quality indicators of allocation efficiency.

Quality indicators. In this paper, the following indicators of efficiency (the indices of performance) are introduced and considered:

The average allocation time t_A (1) needed for algorithm to allocate the task, measured in real time units.

$$t_A = \frac{1}{n}\sum t_{alloc}(i) \qquad (1)$$

where: t_{alloc} (i) – time needed to allocate i-th task, n – total number of task in the system.

The total time T_A. The time needed for computing all tasks, measured in 'abstract' time units (so called mesh ticks). One tick passes when allocation algorithm is not able to allocate new task due to lack of free submeshes.

The average latency L_A (2). This is the average time which task needs to wait in a queue until being allocated.

$$L_A = \frac{1}{n}\sum L_i \qquad (2)$$

where: L_i – latency of i-th task, n – total number of tasks in the system.

The fragmentation f_A. This is the ratio (3) of the total number of free nodes to the total number of nodes in mesh during algorithm's work (excluding the biggest free submesh).

$$f_A = \frac{w \cdot h - P - \sum_i^n w_i \cdot h_i}{w \cdot h - P} \qquad (3)$$

where: w and h – sizes of mesh, P – number of nodes in the biggest free submesh, w_i and h_i – sizes of i-th task.

III. ALGORITHMS

A. First Fit Algorithm (FF)

The First Fit algorithm, is described in details in [2]. The algorithm was implemented as follows:

Step 1. Start searching a given mesh from the node (0, 0) for every single task.
Step 2. Search nodes row by row until free one is found.
Step 3. Check whether a free submesh (containing found free node as a foothold) matching a given task size may be found. If not, go to Step 2.
Step 4. Allocate the task. The matching free submesh becomes busy.
Step 5. End algorithm.

B. Stack Based Algorithm (SBA)

The detailed description of this algorithm can be found in [3]. The main idea of this algorithm consists in finding a base submesh for task, reducing the search space and avoiding unnecessary searches. The algorithm works as follows:

Step 1. For a given task create prohibited area (task if allocated in this area would stick out of mesh).
Step 2. Create coverage areas (respectively if task is going to be allocated in those areas, it will overlap on a busy submesh).
Step 3. Create base areas by spatial subtraction of prohibited and coverage area.
Step 4. Check if exists base area, in which task can be allocated. If yes allocate the task and end algorithm.
Step 5. Rotate the task by 90 degrees and go to Step 1.

C. Current Job Based First Fit Algorithm (CJBFF)

The created algorithm may be treated an improvement of First Fit algorithm. The main idea is to speed up the process of searching free nodes in the mesh structure by omitting already busy nodes belonging to discovered task. The algorithm works as follows:

Step 1. For a given task create a prohibited area (task if allocated in this area would stick out of mesh). Consider only nodes non-belonging to this area.
Step 2.. Start from the node (0, 0).
Step 3. Check whether the node is busy**.** If yes, go to Step 7.
Step 4. Check whether a free submesh (containing found free node as a foothold) matching a given task size may be found. If not, go to Step 7.
Step 5. Allocate the task. The matching free submesh becomes busy.
Step 6. End algorithm**.**
Step 7. Move to the node, next to the last busy node of the encountered task (in the same row). If the task's right edge adjacent to the mesh edge, then move to the next row. Go to Step 3.

IV. EXPERIMENTATION SYSTEM

In order to make simulation of the performance of the considered algorithms, an experimentation system was designed and implemented. The core of the system is simulator with block-scheme shown in Fig. 3.

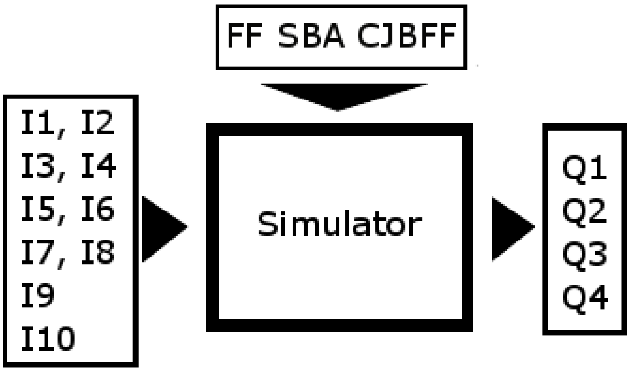


Figure 3. Model of the simulator.

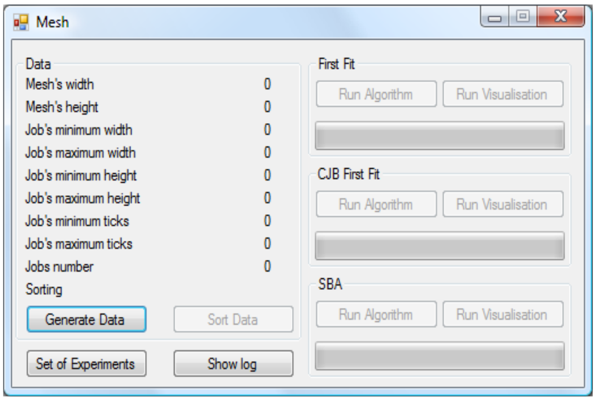
Input parameters of the simulator are:

- I1, I2 – width and height of mesh structure,
- I3, I4 – minimum and maximum width of tasks,
- I5, I6 – minimum and maximum height of a task,
- I7, I8 – minimum and maximum time of a task,
- I9 – number of tasks,
- I10 – sorting type.

Output parameters of the simulator are:

- Q1 – average allocation time,
- Q2 – total computing time,
- Q3 – average latency,
- Q4 – fragmentation.

The system has been implemented using .NET Framework with C# language (it is working well on MS Windows platform with .NET packages). The system possesses the implemented GUI (shown in Fig. 4).

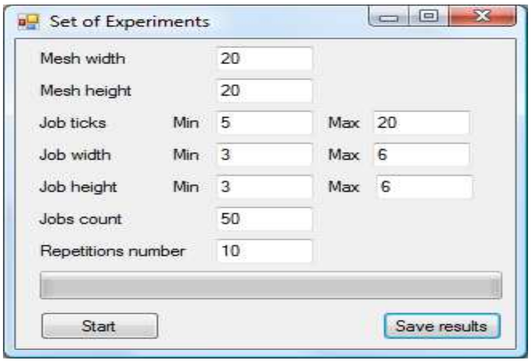


Data

Mesh's width	0
Mesh's height	0
Job's minimum width	0
Job's maximum width	0
Job's minimum height	0
Job's maximum height	0
Job's minimum ticks	0
Job's maximum ticks	0
Jobs number	0
Sorting	

Figure 4. Main window of simulator.

For convenience, the system has implemented function of automatic repetition of the experiment (certain amount of times) for each algorithm with the same input parameters (shown in Fig. 5).



Mesh width		20		
Mesh height		20		
Job ticks	Min	5	Max	20
Job width	Min	3	Max	6
Job height	Min	3	Max	6
Jobs count		50		
Repetitions number		10		

Figure 5. Experiment design window.

V. INVESTIGATION

The aim of the investigations was to compare efficiency of FF, SBA and CBJFF in the same environment. Three efficiency measures were taken into consideration:

t_A – the average allocation time (1),

L_A – the average latency (2),

f_A – the fragmentation (3).

Furthermore, in each experiment the impact of queue sorting on the received latency was examined.

A. *Experiment 1. Increasing number of tasks*

In the first experiment, the set of tasks (queue) was changed in series of experiments - increasing significantly with slightly growing meshes. Experiment design (combination of input values) is shown in Table 1.

TABLE I. INPUTS IN EXPERIMENT 1

Number of Tasks	*Relative Task's Size [%]*	*Mesh width*	*Mesh height*
60	22.5	20	20
140	15.0	30	30
240	11.3	40	40
380	9.0	50	50
540	7.5	60	60
730	6.4	70	70
840	5.6	80	80

Other inputs were taken as follows:

- min – max width of task: 3-6,
- min – max height of task: 3-6,
- min – max execution time of task: 5-20,
- sorting: unsorted, ascending, and descending.

The obtained results are shown in Figs. 6-8.

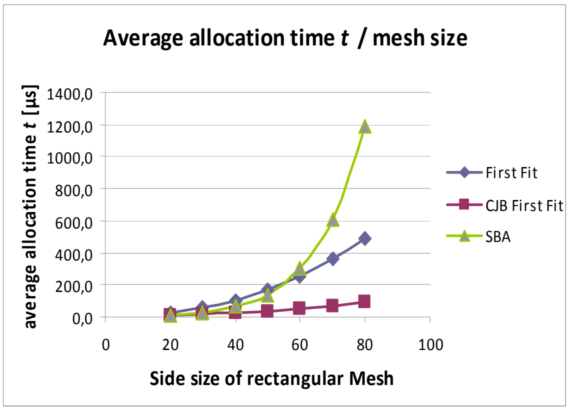


Figure 6. The average allocation time - Experiment 1.

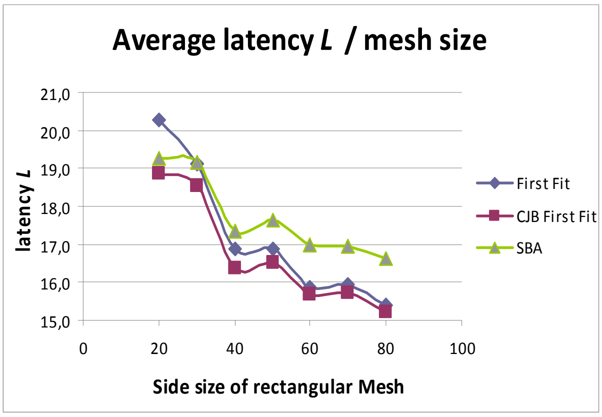


Figure 7. The average latency - Experiment 1.

The created CJBFF was characterized by the best allocation time, significantly lower than the other compared algorithms (Fig. 6). What is more it guaranteed the lowest latency (inversely proportional to mesh size); however for big mesh structures the difference between the CJBFF and FF starts to fade (Fig. 7). The obtained low latencies were possibly the result of low fragmentation maintained by CJBFF algorithm, especially in comparison to SBA (Fig. 8).

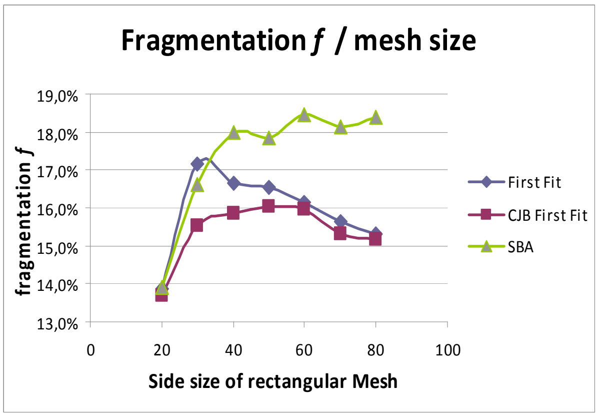


Figure 8. The fragmentation - Experiment 1.

The impact of the chosen queue's sorting type on average latency (in CJBFF) is shown in Fig. 9.

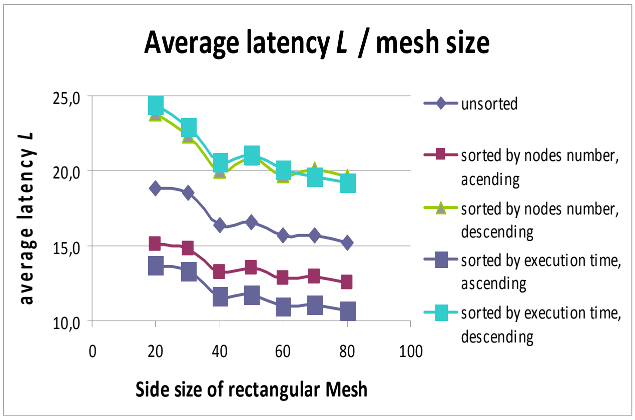


Figure 9. Average latency depending on queue sorting - Experiment 1.

For sorting the tasks in queue by execution time in ascending way, over 25% decrease of latency was obtained comparing to the case when no sorting was used. For descending sorting a remarkable increase of latency was noticed.

B. *Experiment 2. Increasing mesh size.*

In the second complex experiment the mesh size and the task generation parameters were chosen in such a way that the expected relative task's size *p* was always constant and equal 15% for increasing mesh size. Experiment design (combination of input values) is shown in Tab. II.

TABLE II. INPUT S IN EXPERIMENT 2

Task width	*Task height*	*Mesh width*	*Mesh height*
2	5	20	20
3	7	30	30
3	9	40	40
4	12	50	50
5	14	60	60
6	16	70	70
6	18	80	80

Other inputs were as follows:

- number of tasks: 134,
- min – max execution time of task: 5-20,
- sorting: unsorted.

The obtained results are shown in Figs. 10-12.

It may be observed that, in this experiment, the CJBFF was not the fastest among the considered allocation algorithms. In this case the SBA algorithm was slightly faster for larger mesh structures (see Fig. 10). However, once again the created algorithm proved to guarantee the smallest latency from all tested algorithms, as it can be observed in Fig. 11.

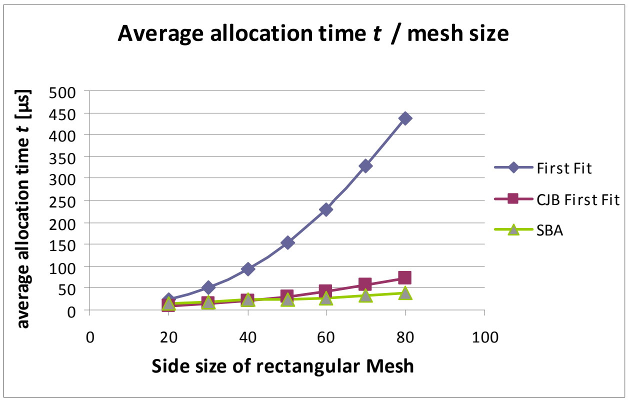


Figure 10. The average allocation time - Experiment 2.

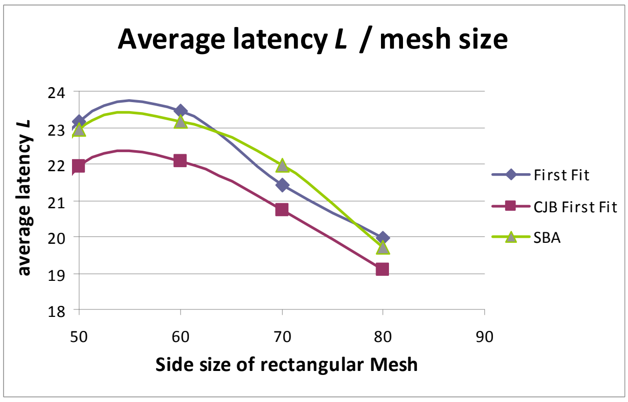


Figure 11. The average latency - Experiment 2.

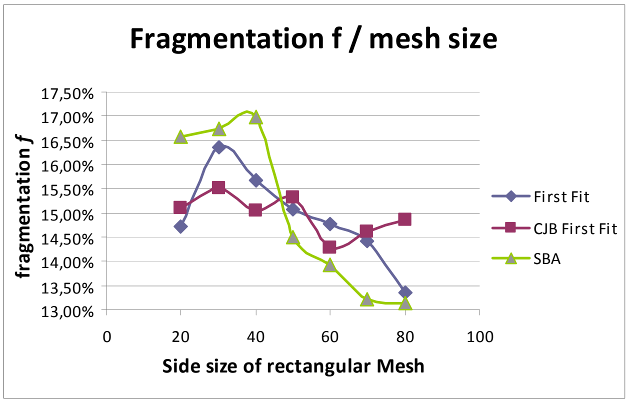


Figure 12. The fragmentation - Experiment 2.

Considering the fragmentation (Fig. 12) it can be seen that the CJBFF algorithm performed as the weakest algorithm; however, only for large meshes. Moreover, it may be observed that the variance of results obtained by all algorithms is rather small and it is not larger than 4%.

The impact of the chosen queue's sorting type (in CJBFF) on average latency is shown in Fig 13.

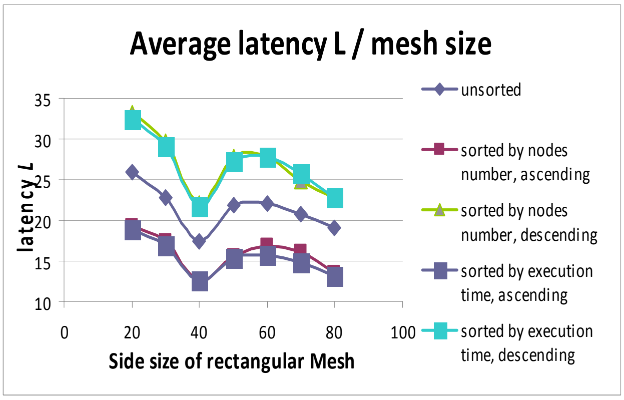


Figure 13. Average latency depending on queue sorting - Experiment 2.

Again, the best results were obtained when used sorting by execution time in ascending order and the worst when sorting in descending order.

VI. CONCLUSION AND FUTURE WORK

The analysis of the results of complex experiments confirms that the designed and implemented CJBFF allocation algorithm is easy to implement and fast in many cases. This algorithm can be recommended to use by designers of multi-processor systems with mesh structures [8], for which the most important factor is the latency of newly added tasks.

Moreover, a big advantage of CJBFF is that with increasing size of a mesh, the time needed for task allocation increases only slightly when comparing to FF and SBA. However, for larger mesh structures the CBJFF has the tendency to fragment the mesh in bigger scale than two other considered algorithms.

To additionally decrease of the latency of tasks (which means improving the allocation process) it may be desirable to apply sorting of task's queue. It is worth to be noticed that ascending sorting by execution times resulted even in a 20% decrease of latency, when comparing to results for unsorted queues.

The further development of the presented in this paper experimentation system will focus on implementing other allocation algorithms, e.g., algorithms based on evolutionary ideas [5].

Moreover, we plan preparing new modules of the system to ensure designing multistage experiments [7] in automatic way and store the results of experiments in problem-oriented data base.

ACKNOWLEDGMENT

We would like to thanks Mr. M. Halaczkiewicz, student of Electronics Faculty, Wroclaw University of Technology for remarkable help in preparing the programmed modules in the experimentation system.

REFERENCES

[1] A. S. Tanenbaum, Modern Operating Systems, 2nd edition, Prentice Hall, 2001.

[2] Y. Zhu, “Efficient Processor Allocation Strategies for Mesh-Connected Parallel Computers”, J. Parallel & Distr. Computing, vol. 16, 1992, pp. 328-337.

[3] B.S. Yoo and C. Das, “A Fast and Efficient Processor Allocation Scheme for Mesh-Connected Multicomputers”, IEEE Transactions on Computers, vol 51, No. 1, 2002.

[4] W. Kmiecik, M. Wójcikowski, L.,Koszałka, and A. Kasprzak, ” Task Allocation in Mesh Connected Processors with Local Seach Meta-heuristic Algorithms”, Lecture Notes in Artificial Intelligence , vol. 5559, Springer, 2010, pp. 215-224.

[5] L. Koszalka, M. Kubiak, and I. Pozniak-Koszalka, “Allocation Algorithm for Mesh-Structured Networks”, Proc. of 5th ICN, IEEE Comp. Society Press, 2006, pp. 24-29.

[6] M. Halaczkiewicz, “Implementation of Static Task Allocation Algorithms in Mesh Networks”, M.Sc. project, Faculty of Electronics, Wroclaw University of Technology, 2009 /in Polish/.

[7] L. Koszalka, D. Lisowski, and I. Pozniak-Koszalka, “Comparison of Allocation Algorithms for Mesh- Networks with Multistage Experiments”, Lecture Notes in Computer Science, vol. 3984, Springer, 2006, pp. 58-67.

[8] D. Zydek, H. Selvaraj, L. Koszalka, and I. Pozniak-Koszalka, “Evaluation scheme for NoC-based CMP with integrated processor management system”, International Journal of Electronics and Telecommunications, vol. 56, no. 2, 2010, pp. 157-167.

ISBN: 978-1-61208-202-8

Does Chaos Exist in Social Network Components? Role of Evolutionary Dynamics and Tool

Hameed Al-Qaheri
Department of Quantitative
Methods and Information Systems,
Kuwait University, Safat, Kuwait
alqaheri@cba.edu.kw

Soumya Banerjee
Department of Computer Science,
Birla Institute of Technology,
Mesra, India
dr.soumya@ieee.org

Goldina Ghosh
Department of Computer Science,
Birla Institute of Technology,
Mesra, India
goldinag@gmail.com

***Abstract*-Chaos and chaotic optimization is a global trend of optimization, customized for several engineering applications. This paper explores the behavior of chaos under social networking sites. Recalling chaotic dynamic characteristics emerging evolutionary network, the *News Feed* behavior of *Facebook* has been modeled for optimization. The research challenge is to address the emphasized presence of chaotic behavior under the social network paradigm. Subsequently, modeling could be collaborated with a class of *evolutionary network*, e.g., *vaccination network*. The outcome of investigating such non linear attribute could be of emerging relevance to demonstrate how far this will affect social network participants to be influenced over a specific social inference. In order to address the results obtained through evolutionary dynamics with clarity, the same problem has been tested through *Genetic Algorithm* components for different visualization and open access to this new research frontier using contemporary evolutionary algorithms.**

Keywords - Social Networking; Evolutionary Dynamics; Genetic Algorithms; Chaos.

I. INTRODUCTION

Chaos [1] is a general nonlinear phenomenon that sustains in the linear system. Chaos exhibits certain main characteristics such as *quasi-randomness, ergodicity and sensitive dependence* on initial conditions. Furthermore, ergodicity could be considered as an effective mechanism to avoid trapping into local minima in the searching process. As such, chaos could be considered as a novel and potential optimization tool of interest [12]. Since the inception of Web 2.0, different forms of social network have been envisaged and raised the relevance of chaos and chaotic optimization. A social network is basically represented as a graph, with individual persons represented as vertices, the relationships between pairs of individuals as edges, and the strengths of the relationships represented as the weights on edges (for the purpose of finding the shortest weighted distance, we treat lower-weight edges as stronger relationships). It has been suggested that there is amalgamation of social chaos model and evolutionary network noticed in its vaccination strategic behavior [10]**.** It is general practice that in any system of the evolutionary design, the intelligence emerges from a chaotic balance between individuality and sociality. The chaotic balances are the characteristic features of the complex system. For a given energy or cost function, by following chaotic *ergodic* orbits, a chaotic dynamic system may converge towards a targeted global optimum.

Considering the broad hybrid evolutionary and chaotic behavior of social media services, (which are the direct or indirect function of user's participation via feeding of content and tagging) this paper explores the possibility of the existence of chaos in the functionality of *Facebook* social components, such as posts and news feed, which has the possibility of demonstrating non linear scales and being sensitive to initial condition. Subsequently, the paper also discusses the relevance of optimized function to distinguish the chaos on these components. This will help to achieve an optimized social container both from the users and social network service provider's space and time complexity perspective.

The remaining part of the paper is organized as follows. Section II introduces elementary background material, the problem statement and the relevance of evolutionary approach of visualizing chaos within social network components. In Section III, we present a news feed modeling technique coined from *Facebook*. Section IV presents the proposed algorithm of chaos modeling. In Section V, the implementation and results are presented, followed by the conclusion in Section VI.

II. BACKGROUND, PROBLEM STATEMENT AND RELEVANCE OF EVOLUTIONARY DYNAMICS

Facebook introduced the News Feed feature on September 5, 2006. This feature gradually becomes clearer to the user, as he/she posts items to personal profile or wall. These items may be visible to friends or acquaintances connected to the *Facebook* connection. Both culturally, and technically, privacy of such news feed propagation has been questioned and the corresponding impact of the user after those news feed has also been scrutinized with a significant survey [3]. Such investigations also reveal the factors that are related to news, used on *Facebook*, and also demonstrate the other contemporary factors such as gender participation, emotional stability and so on [4]. News formation and distribution for various social networking sites could be analyzed by Exponential Random Graph Models (ERGMs), which show the importance of link and their non linear value of centrality [5]. In particular, it has also been observed that *Facebook* has several self-organized features like status

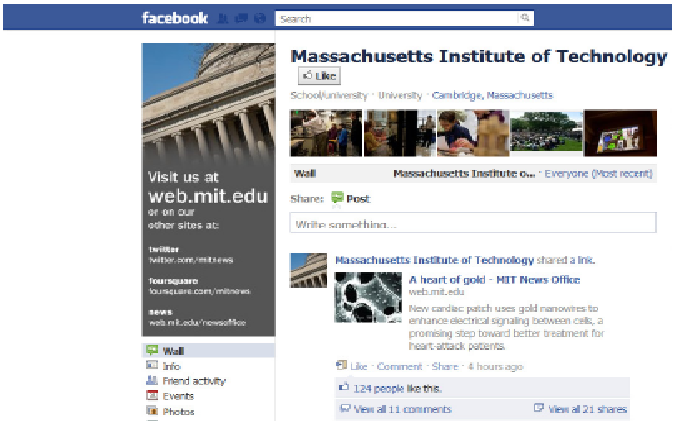
ISBN: 978-1-61208-202-8

updates, wall posts, pokes, tagging and commenting, subscription services, games and applications (developed by external agencies). Among these user-oriented features, the majority of them are news feed or post related activities. Thus, *Facebook* becomes a well defined dynamic brand, referred to and maintained by its users [6]. The content feed subscription in *Facebook* follows dynamic distribution of content to all the subscribers, who are having common membership of the same service provider or of same interest. The most current content of two users can be shared when they meet to extend and optimize network coverage with lower cost of service. This could be another reason to be concerned with in social networking sites, especially when accessed through mobile devices [7]**.** Envisaging different news feed and content generation options, *Facebook* keeps the option open by forming similar opinion of clusters, which can be derived from the user's activities. Recent research emphasizes the model of opinion formation from social network [8], but the research challenge is to keep the contents and feed free from final chaotic behavior. Present *Facebook* structure has the option to subscribe or non-subscribe to certain content page or posts from a particular subscriber, but their trend of linearity and broadcast will still remain ambiguous.

Hence, the concept of optimization has been proposed in social networking sites, such as *Facebook*, from different perspectives. Developers are always thriving to ensure that their sites and apps are publishing stories that make the lead profitable and popular, which has pioneered the concept of "News Feed Optimization" (NFO) in *Facebook*.

A. *Background on Evolutionary Dynamics*

Evolutionary dynamics is an emerging concept, evolved from the realm of game theory. Evolutionary game dynamics is conceived from a set of deterministic differential equations capable of addressing infinitely large and well-mixed populations [2]. In a well-mixed population, the chance of two individuals interacts is equally alike. Some recent approaches consider stochastic evolutionary dynamics in populations of finite size. Evolutionary game dynamics are also influenced by population structure [14]. Analogically, a well-mixed population typically opposes evolution of cooperation, but a structured population can promote it. The whole idea is to determine the strategy of two present strategies deployed by several sets of common or different players, then based on their interaction, the session calculates certain payoff. The payoff actually gives the effectiveness of any such propagated and linked approaches among those players. The trend of the evolutionary dynamics has been characterized through diversified mathematical treatments, and the utility of such treatments are of typical importance on social networking structure. The individuals finally try to converge and optimize certain deterministic measures at the end of a social session. We explore *Fermi's equation* [10] under this paradigm. Interested readers are advised to go through the references mentioned to get a broader understanding on this emerging concept. Figure 1 shows a diagrammatic view of a *Facebook* screen shot, where the users of the particular community share or refer to a particular new item. The view will enable the implantation followed in the model.



Massachusetts Institute of Technology

School/university · University · Cambridge, Massachusetts

Visit us at web.mit.edu or on our other sites at:

twitter twitter.com/mitnews

foursquare foursquare.com/mitnews

news web.mit.edu/newsoffice

Wall Massachusetts Institute o... · Everyone (Most recent)

Massachusetts Institute of Technology shared a link.

A heart of gold - MIT News Office
web.mit.edu
New cardiac patch uses gold nanowires to enhance electrical signaling between cells, a promising step toward better treatment for heart-attack patients.

124 people like this.

View all 11 comments

View all 21 shares

Figure 1. Passive Mode of Information Sharing on *Facebook Screen shot*

B. *General Mathematical Model*

There is a generic mathematical model of the spread of messages and it follows a complex socio-psychological process. An adequate modeling of this process requires both a correct description of the underlying social networks, along which messages propagate, and a quantitative formulation of various behavioral mechanisms that motivate individuals to participate in the spread of any messages. The formal model comprises of a set of plausible rules. Generally, irrespective of any social network, we consider a population consisting of N individuals with respect to the propagation of messages, that is subdivided into different classes of participants. Inspired by Maki and Thompson [15], we can also assume that any messages propagate by direct contact of the participants with others in the population. However, these contacts can only take place along the links of an undirected social interaction network G = (V, E) where V and E denote the vertices and the edges of the network, respectively. The negotiation of contacts between the message originator and the rest of the population are governed by the different set of rules to define the originator, follower, and neutral. Assuming that a node j has k links, g can be considered as a stochastic variable which has the following binomial distribution-

$$\Pi(g,t) = \binom{k}{g}\theta(k,t)^{g}(1-\theta(k,t))^{k-g} \qquad (1)$$

where, θ(k, t) is the probability at time t that an edge transmitting from a neutral node with k links points to the other effected nodes.

III. News Feed Modeling Perspective from *Facebook*

During Facebook's developer conference, f8, on April 22, 2010, *EdgeRank*, a formula which determines the likelihood of an object appears in a *News Feed*, was introduced [11]. The formula is discussed below. However, two important *Facebook* concepts need to be defined prior to introducing the formula, and these are as follows-

(1) An Object is any item, such as a post, a status update or a change to a profile picture that appears in *News Feed*.

(2) An Edge is an interaction with an object, such as a comment or a tag, by another user.

Having defined these concepts, the *New Feed Optimization (NFO)* or the *Edge Rank* formula can be expressed mathematically as follows [11]:

New Feed Optimization (*EdgeRank) =*

$$\sum_{edges\,\mathrm{e}} \mathrm{u_e\ w_e} d_e \tag{1A}$$

where, $\mathbf{u_e}$ is the affinity score between the viewing user and edge creator (a user gets higher affinity score if he/she sends more messages to friends or check their profiles often), $\mathbf{w_e}$ is the weight for this edge type (create, comment, like, tag, etc.) and $\mathbf{d_e}$ is time decay factor indicating how long ago the edge was created (the older an *Edge* is, the less important it becomes). The score returned by the formula indicates *Edge rank* of the object, the higher the value of the *Edge rank score* the more likely the desired object is to appear in the user's feed. It is also worth pointing out that the act of creating an object is considered an Edge, which is what allows objects to show up in user's friends' feeds before anyone has interacted with them.

A. How Chaos exists in Facebook: Functional Validation

Inspired by the evolutionary dynamics of social network of new propagation, where we consider the analogy of process of *Page* diffusion via *Facebook's* News Feed, the diffusion followed leads to chaotic behavior while feeds towards optimization. *EdgeRank* is similar in semantics to a fan page structure. As such, analytically, diffusion of pages occurs when a user becomes fan of the page and finally their friends of friends become fans of the same page as well. Therefore, the question is raised as how to accomplish optimization in terms of feed option. We model diffusion approximation for large populations of size N under *Facebook.*

Let m/N be the fraction of targeted individuals, who are interacting in that propagation of news network of *Facebook*. *Facebook participants* follow each others broadcast based on the group presence, which preferentially copies others with higher influential impact friend or friends of friends. In each round, a randomly chosen individual *i* selects another random individual *j* as cascaded friend, and compares his/her position of influence to that of the actual pivotal friend. Individual *i* adopts the strategy of individual *j* with the probability of feeding the particular post given by the Fermi function [10].

$$\phi(s_i \leftarrow s_j) = f(p_j - p_i) = \frac{1}{1 + \exp[-\beta(p_j - p_i)} \tag{2}$$

where, β represents the intensity of selection of probability of most acclaimed post. The population can change, only if individuals *i* and *j* have different strategies. Hence, subsequently, the probability that the number of news fed individuals increases from *m* to *m + 1* (denoted T^+_m) and the probability that the number decreases from *m* to *m – 1;* this statistics will influence *news feed* and thus the linear behavior of page rank will not sustain any more. We model the affect of chaos, for some emergency and popular pages under these typical phenomena of vaccination network. It has been observed that, for some popular pages, more than 90% of the fans can be part of a single group of people who are all somehow connected to one another. We solicit our prepositions through an example taken from "*Facebook of August 21, 2008, 71,090 of 96,922 fans (73.3%) of the NastiaLiukin (an American Olympic gymnast) Page were in one connected cluster"* [9]. The chain data feed existing in this page and several parameters could also exist analogous to our validation of vaccination network.

Based on the cited instance, we prepare a shadow data set as detailed in Section V.

IV. Proposed Algorithm

The paper explores an elementary, yet evolutionary, optimized approach to minimize the chaos associated with the social network site. The ambiguity of posts and other shareable activities is also partially addressed. The model contains certain variables in the algorithm. The algorithm describes the flow of message among different friends and even to the friends of friends. This flow of message leads to chaos, when propagated to friends of friends. The flow of messages is restricted among the friends, only if there is a decrease in chaos, but the rate of message flow remains the same. Different variables and their corresponding semantics as used in the algorithm are detailed below.

TABLE I. Facebook News and wall Variables with semantics post

Category	Semantics
Message	**existing 'message received' record of the friends**
u message	**new messages with different point value sent by a person to his friends**
Link	**presence or absence of friend of a friend**
Nm	**message , but no representation of chaos**
rec1	**summation of all the values of the maximized function**

rec 2	**summation of all the values of the minimized function**
Rank list	**The summation of age, *Facebook* age and activity count denoting the active user’s rate. A scale from 1 to 100**
News feed	**The different types of newsfeeds with a specific value ranging from 1 to N**

Table I summarizes the different parameters associated with variables, pertain to the semantic post of the user and it also solicit the same to devise the proposed algorithm.

```
Begin
Initialization of values to the variables link
 present or not present
Link present =1
Link Not present=0
Values for variables like Age, FaceBook
       age, Activity count and message received
       are generated on the user’s access report
Generating Rank List
Rank list = ∑ Age, FaceBook Age, Activity count
Generating different valued Newsfeed
 Newsfeed =1, 2… N
Delivering Newsfeed without any restriction
         for i=1 to a do
                 for j= 1 to c do
                     rec (i) = message (i) + umessage(j)
                 end
        end
  Delivering Newsfeeds with a restriction
        for m=1 to c do
          if link(m) = = 0 then
                for i = 1 to a do
                        for j= 1 to c do
                           ch(m)=message(i) +
                                      umessage(j)
                       end
                  end
          else
           ch(m) = message(i)
        end

    Separating Chaos and Newsfeed
         for i = 1 to a do
                 for m = 1 to e do
                    nm(i) = rec(i) – ch(m)
                   end
          end
     Generating Maximization and Minimization value of
      the Function
     $f(x)=x^2 - 10 \cos(2x) + 10$
         value of cos (x) lies between $-1<= \cos(x) <=1$
                 while cos (x)<= -1 do
                       $\cos(2x) = -1 = \cos 180^0$
                   $2x = 180^0$
                   $x = 0.5$
                   end while
                   while cos (x)>= 1 do
                            $\cos(2x) = 1 = \cos 0$
                            $2x = 0$
                            $x = 0$
                   end while
                if x == 0.5 then
                        $f(x)= (-x^2 + 10 \cos(2x) – 10)$
                 end if
                if x == 0 then

           $f(x)= x^2 - 10 \cos(2x) + 10)$
          end if
         Maximizing Newsfeed and Minimizing Chaos
            mf= maxfun(x)
            mif= minfun(x)
         for i = 1 to a do
                rec1(i)= rec(i) + mf
         end for
         for m = 1 to e do
                  rec2(i) = rec(i) + mif
         end for
  Plotting the graph
  Transfer of newsfeeds to all
  Tracking of chaotic behavior due to transfer of
   newsfeed
  Maximization of message flow
  Minimization of chaos
 End Begin
```

V. Implementation and Results: Preparation of Data Set for the Proposed Model

A list of 40 different friends of a specific person from *Facebook* is used as sample**.** These individual friends have their own identity, nature of behavior, and activities. Some of the friends in the list are also interrelated with a few other friends in the list. This signifies the community of that person’s *common friend circle*. The messages transmitted by the person to all his friends at some point in time are also cited.

The attributes of Table II are operationally defined as follows:

Friend List- The friends are denoted by F1 to F40.

Age- It is the age of the individual friends. Based on the age the nature of activity of an individual is also reflected.

***Facebook* Age**- It denotes that since how many years they are connected to this Social Network site.

Activity count- It includes the information like message sent, photo uploaded or *Facebook’s* Wall post. Any such activity done is kept as a count in the record set.

Link present/Not present- This denotes the idea of friend of a friend. Link present=1 and Not present=0

Received Messages- This is a count, which is based on the number of messages, accepted by the friends from the person, denoting the depth of the connectivity. Based on all these attributes we generate a *Rank list* to judge the liveliness of a member. This is the summation of age, *Facebook* age, and activity count denoting the active user's rate. A scale from 1 to 100 is provided to plot the rank.

TABLE II. List of Friends and their Associated Attributes

FRIEND LIST	PARAMETERS				
	Age	Facebook Age (in years)	Activity count	Link Present=1 Not present=0	Received Messages
F1	20	5	10	1	2
F2	18	0	6	1	4
F3	30	5	12	0	4
F4	19	1	7	0	2
F5	28	5	15	1	3
F6	28	2	10	1	1
F7	33	5	25	1	1
F8	24	5	25	0	3
F9	40	4	15	0	4
F10	28	1	1	0	5
F11	36	4	1	1	1
F12	41	5	10	0	3
F13	15	1	0	0	3
F14	45	5	5	1	1
F15	20	2	10	1	2
F16	18	1	11	1	1
F17	22	3	7	0	4
F18	32	3	10	0	2
F19	30	4	9	1	3
F20	48	5	15	0	1
F21	50	4	15	1	4
F22	28	5	15	0	2
F23	22	1	7	0	4
F24	42	3	15	0	4
F25	50	6	9	0	4
F26	22	5	10	0	3
F27	20	4	6	1	1
F28	50	5	17	1	2
F29	48	5	17	1	5
F30	40	3	15	0	1
F31	29	6	15	1	5
F32	42	5	1	0	3
F33	50	4	6	1	5
F34	31	1	17	0	1
F35	20	4	17	1	3
F36	30	1	4	1	1
F37	18	1	4	1	5
F38	47	4	9	0	1
F39	33	5	7	1	4
F40	50	4	17	1	5

The data set shown in Table II has been considered, for the proposed validation using **MATLAB and C++.** In Figure 2, the plot represents the message transferred to all friends along with friend of friends and it leads to chaotic behavior. The different valued messages that are transferred via the *newsfeed* are plotted along the x-axis. Whenever a new message is received by the friend its *"received message"* record in the database is updated. The friend's rate of activity is judged on the basis of the rank. This is plotted on the y-axis. The lines in graph denote that any valued message that is sent is received, in each range of ranks in between 10 to 80. The arrow marks denotes the flow of information in every direction, i.e. is, first to the friends and then to the 'N' number of links connected to each friend.

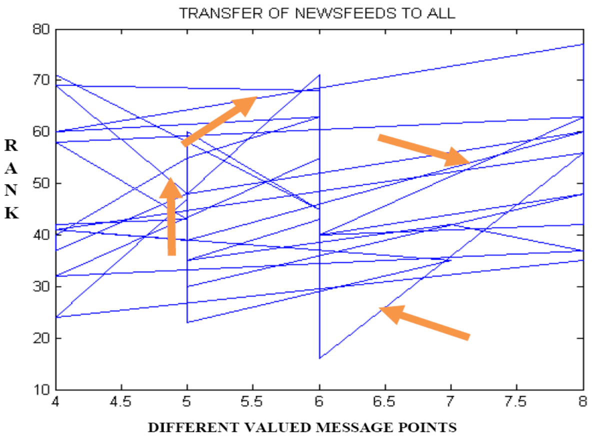


Figure 2. Random Message passing lead to chaos

As it becomes difficult to identify the message sent and the chaos that is generated, a filter is used to distinguish only the chaos, where conditions *for link = = 1***,** *i.e. presence of immediate link* is checked to identify *friends of friends*. Based on this condition, the graph is plotted in Figure 3. This figure demonstrates the different message values along x-axis after updating it in the "received message" section to only those who are linked. The y-axis is the rank. The multiple parallel straight lines in the graph denote that any valued messages that are received each time by all the linked friends create a meshed structure.

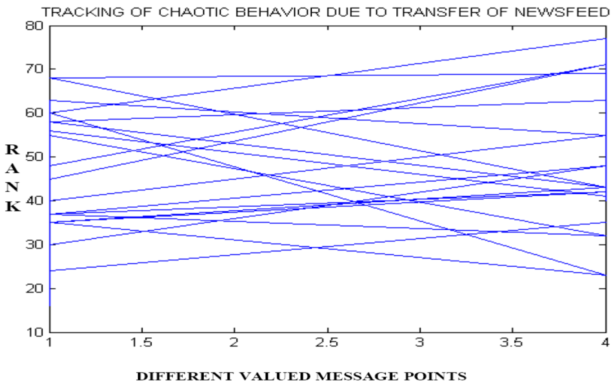


Figure 3. Chaos Present in messages

Hence, the concept of minimizing the chaos arises and maximizing the message transfer takes place. Without affecting the message transfer, we try to reduce the chaos and maximize the message flow, (number of news feed points) which implies no transferring to the friends of friends. By maximizing, we mean transferring all different valued messages to all different ranked users, who do not have any link. This is achieved by maximizing the message chaos function:

$$f_1(x)=\sum_{i=1}^{n}\left[x_i^2-10\cos(2\Pi x_i)+10\right] \quad (3)$$

where, 0 < (x) <0.5 and 0 < f(x) <19.75. The graphical representation is given in Figure 4, where the x-axis gives the representation for all the different messages uploaded for only the unlinked friends with respect to the y-axis that represents the rank. In this graph it is observed that the multiple straight lines are much distinct. The directions are indicated by the arrows.

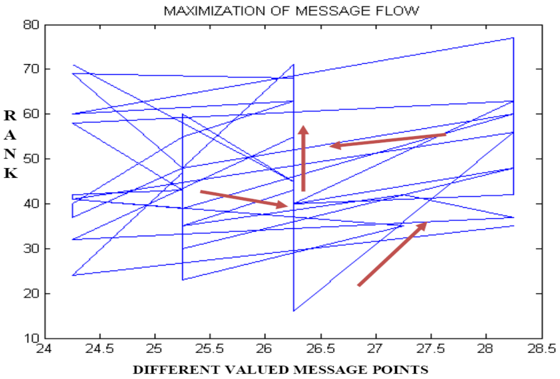


Figure 4. Maximizing the message flow out of chaos

Similarly, after minimizing the function application of the result for the chaotic control is shown in Figure 5. Here the graph denotes the presence of chaos in much lesser number of locations.

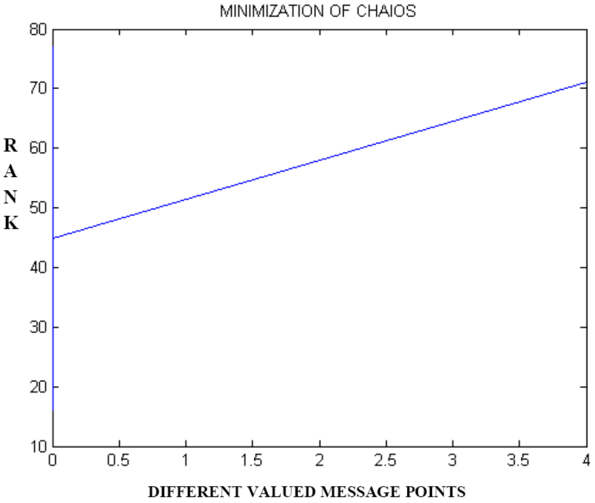


Figure 5. Minimizing the Number of recipients and chaos

A. Observation of news feed using Genetic Algorithm (GA) for similar objective function

To compare the efficacy and implication of the present proposed module for the same optimization problem concerned with *Facebook* news feed and post, the idea of the natural selection process that drives the biological evolution is applied through *Genetic Algorithm (GA)*. MATLAB 7.0.4, toolbox has been utilized on the current population set. Genetic Algorithm considers potentially huge *search spaces* and navigates them looking for optimal solutions. If we are intending to search for a solution to a problem, we look for the best among possible solutions. The space for all possible solutions is called “Search Space”. “Looking for a solution” signifies looking for extremes (either maximum or minimum) in the search space. *Fitness function* quantifies the optimality of a solution. Fitness function is derived from objective function and then used in the successive genetic operations. After obtaining the suitable values from the fitness function, we go to the next step of Genetic Algorithm which is selection procedure. In this procedure, an individual called ‘parent’ that combines the population at the next generation. Cross Rule helps in forming children in the next generation and hence obtaining only the optimized solutions.

Here, by applying equation (3), the maximization function is used to represent the maximum amount of message transfer only to the friends without link. After maximizing the function, a graph is plotted as shown in Figure 6. X-axis in the graph represents the range of ranks of the friends who receive the message and y-axis represents the maximum number of the messages with different values that are transferred (measured by the fitness value).

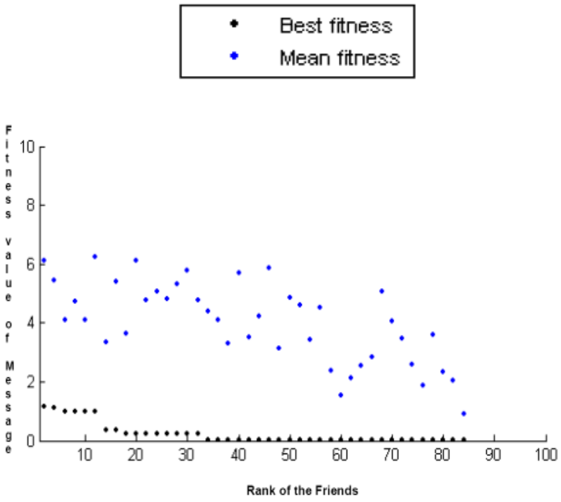


Figure 6. Rank of the Friends with fitness

The blue dots in the scattered graph represent the best fitness value and the black dots represent the mean fitness value. In Figure 7, the average distance between the messages transferred among each individual is shown, where x-axis is the range of ranks of the friends, who receive the message and y-axis is the average points calculated based on the fitness value of the different valued messages that are sent.

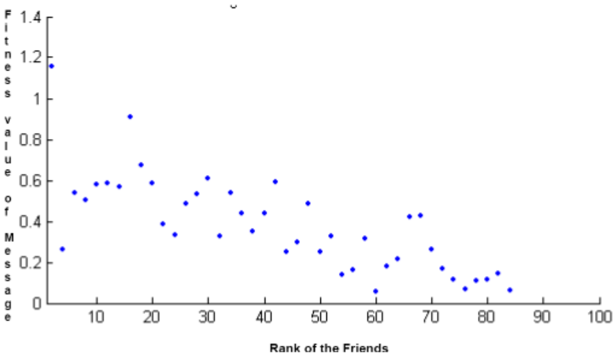


Figure 7. Distance between the messages with rank of friends

In Figure 8, the graph plotted represents the range, *Best*, *Worst* and *Mean*, of the fitness function value of each group of rank in each generation due to the transfer of different valued messages.

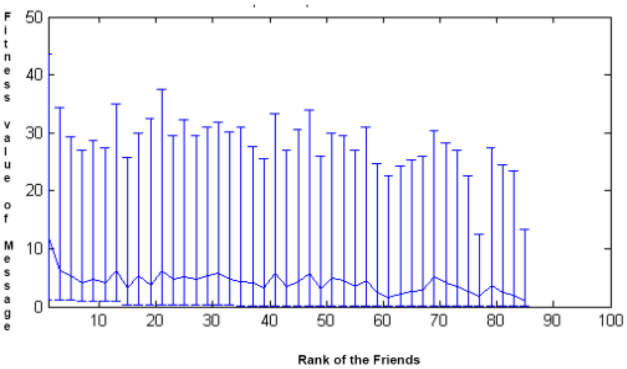


Figure 8. Diversified Fitness Function

Similar to the previous proposed module, the minimization function is used for reducing the chaos, but the message transfer is kept as is with regards to the friends without links. Hence, after minimizing the function, a graph is plotted as shown in Figure 9. X-axis in this graph represents the range of ranks of the friends who receive the message and y-axis as the minimized chaos and maximum number of the messages with different values that are transferred (measured by the fitness value).

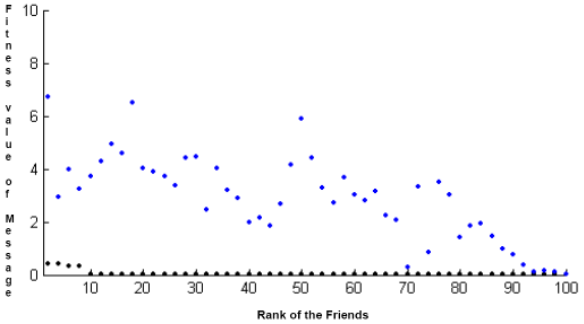


Figure 9. Minimization of chaos

In Figures 10 and 11, the average distance between the messages transferred among each individuals and the minimized fitness value generated are plotted in the graphs that represent Best, Worst and Mean of the fitness function value of each group of rank in each generation due to the transfer of different valued messages respectively.

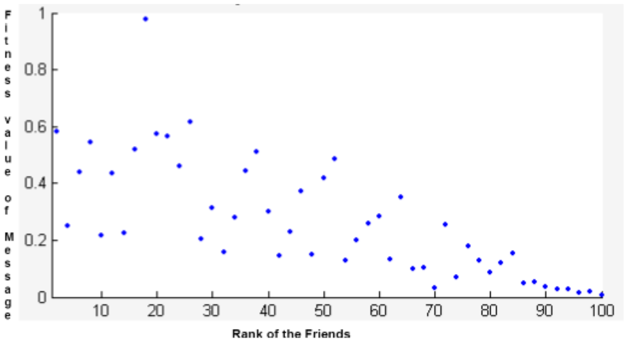


Figure 10. Average Distances of Messages

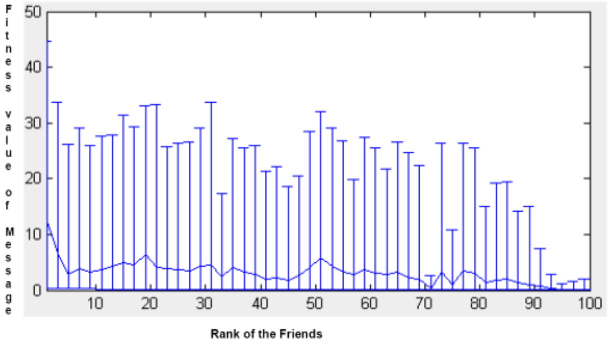


Figure 11. Diversified Fitness values for different Propagated Messages

B. Implication on the Results

Careful observation of two sets of graphs generated for observing chaos, rank of friends and link of message propagation using different evolutionary methodologies reveal certain interesting observations. Firstly, set of results generated on the basis of Fermi’s equation model and its associated evolutionary dynamics concept (equation 2, section 3.1), demonstrates more clarity on the chaos optimization as shown in Figure 4. The counterpart evolutionary tool through GA produces more scattered and distributed optimization index. Moreover, the choice of fitness also requires certain heuristics across all diversified message values and rank of friends. Hence, the first set of evolutionary dynamics paradigm could be considered a better choice for social networking sites to develop software plug-ins in the form of optimization function.

C. Major Observation on Figure.6-Figure.11

Figures 6 to 11 represent the alteration in tendency of message propagation of friends depending on the value of message and distribution strategy. Figure 10 specifically, gives an estimation of distance of message transferred to the individuals. The work discusses about rank list of the shared group messages among the members, reflected in Figure 8.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-202-8

From Figure 9 to Figure 11, the status of the plots has been revisited and significant minimization of chaos, while incorporating Genetic Algorithm has been observed. This is only possible, if distributions of objective function of messages have been optimally configured. The optimization could assist the social network message distribution and privacy scheme in a more validated processes and more such optimal function could enhance design space flexibility of social network.

VI. Conclusion

This paper investigates the presence of chaotic behavior under Facebook's news and post events. The paper also suggests an optimized solution from the perspective of *evolutionary network, e.g., vaccination network*. Real life *Facebook* instances have been envisaged and different observational points of chaotic behavior were studied with an evolutionary optimization function. Several research directions have been evolved out of the present initiatives: particularly, the extraction of *Facebook* data and its associated crawling, or graph visualization which could be made simpler. We came across an unbiased sample of Facebook users by crawling its social graph using certain popular algorithms e.g., *Metropolis-Hasting random walk (MHRW) and a re-weighted random walk (RWRW)* [13]. Similarly, other evolutionary heuristics like *Particle Swarm Optimization (PSO)* or *Differential Evolution (DE)*, could be tested on the same data set of *Facebook* to quantify the chaotic behaviour. The present study can also facilitate *Facebook* application developers to intricate certain smarter and optimized mechanism for wall and message post events.

References

[1] T. Y. Li and J. A. Yorke, Period three implies chaos, Amer. Math. Monthly, 82 , 1975, pp. 985-992.

[2] Martin A. Nowak, Corina E. Tarnita, TiborAntal, Evolutionary dynamics in structured populations Phil Trans R Soc B, 365, 2010, pp,19-30.

[3] Christopher M. Hoadley, HengXu, Joey J. Lee, Mary Beth Rosson, Privacy as information access and illusory control: The case of the *FaceBook* News Feed privacy outcry in Electronic Commerce Research and Applications Elsevier, Vol. 9, 2010, pp. 50–60.

[4] Carroll J. Glynn, Michael E. Huge, Lindsay H. Hoffman All the news that's fit to post: A profile of news use on social networking sites, Computers in Human Behavior, 2011, Elsevier, doi:10.1016/j.chb.2011.08.017.

[5] Sandra Gonzalez-Bailon Opening the black box of link formation: Social factors underlying the structure of the web, Elsevier Social Networks Vol. 31 2009, pp. 271–280.

[6] Patterson A, Social-networkers of the world, unite and take over: A meta-introspective perspective on the *FaceBook* brand, J Bus Res, 2011, Elsevier, doi:10.1016/j.jbusres.2011.02.032.

[7] Stratis Ioannidis, AugustinChaintreau and Laurent Massouli´e, Optimal and Scalable Distribution of Content Updates over a Mobile Social Network, in proceedings of IEEE INFOCOM, 2009, pp.1422-1430.

[8] Gerardo Iñiguez, Rafael A. Barrio, JánosKertész, Kimmo K. Kaski, Modeling opinion formation driven communities in social networks, Computer Physics Communications, Elsevier, Vol.182, 2011, pp.1866–1869.

[9] Eric Sun, ItamarRosenn, Cameron A. Marlow, Thomas M. Lento, Gesundheit! Modelling Contagion through *FaceBook* News Feed, Association for the Advancement of Artificial Intelligence, 2009, www.aaai.org.

[10] Feng Fu, Daniel I. Rosenbloom, Long Wang and Martin A. Nowak, Imitation dynamics of vaccination behavior on social network, *Proc. R. Soc. B* 2011, 278, pp. 42-49 first published online , 28 July 2010.

[11]. http://www.facebook.com/f8 (Facebook Developers Conference September 22, 2011)

[12] Wei Gong, Shoubin Wang. Chaos Ant Colony Optimization and Application.2009 Fourth International Conference on Internet Computing for Science and Engineering. DOI 10.1109/ICICSE.2009.38, pp. 301-303.

[13] M. Gjoka, M. Kurant, C. Butts, and A. Markopoulou. Walking in facebook: a case study of unbiased sampling of OSNs. In Proceedings of the 29th conference on Information communications, pp.2498. 2506, IEEE Press, 2010.

[14] Taylor C, Fudenberg D, Sasaki A, Nowak MA, Evolutionary game dynamics in finite populations. B Math Biol 66: 2004,pp. 1621–1644.

[15] D.P. Maki, Mathematical Models and Applications, with Emphasis on Social, Life, and Management Sciences, Prentice-Hall, Englewood Cliffs, NJ, 1973.

[16] S. Boccaleti, V. Latora, Y. Moreno, D.-U. Hwang, M. Chavez, Complex networks: structure and dynamics, Phys. Rep. 424 , 2006.

Multi-modal Optimization using a Simple Artificial Immune Algorithm

Tad Gonsalves
Dept. of Information & Communication Sciences,
Sophia University,
Tokyo, Japan
E-mail: tad-gonsal@sophia.jp

Yu Aiso
Dept. of Information & Communication Sciences,
Sophia University,
Tokyo, Japan
E-mail: yu.aiso.0920@gmail.com

***Abstract*— Evolutionary Algorithms have an inherent parallelism that should enable them to locate several optima of a multi-modal function. However, in practice they are found to converge to a single (global) optimum. This has led to the research in the design of highly specialized evolutionary algorithms to obtain the maximum number of global and local optima of multi-modal functions. However, this is an over-kill, since in most cases the management needs no more than a handful of optima to make decisions. We demonstrate that the ordinary CLONALG algorithm, without any special modification to handle multi-modal optimization, is powerful enough to obtain several global and local optima to support the decision-making process.**

Keywords - evolutionary computation; multi-modal optimization; artificial immune algorithm;

I. INTRODUCTION

In real-world optimization problems, sometimes we are not satisfied with only one optimal solution. The demand for multiple solutions is more prominent when there exist several near optimal answers to a problem. Evolutionary Algorithms (EA) are widely used for function optimization. The EAs have an inherent parallelism that should enable them to locate several optima of a multimodal function. However, in practice they are found to converge to a single (global) optimum. The Genetic Algorithm (GA), in particular, is found to converge to a single solution when attempting to optimize a multimodal function [10].

The inability of EAs to handle Multi Modal Optimization (MMO) has led to an extensive research in the design of new algorithms. Extended EAs are devised to locate all the global optima and as many local optima as possible. However, at a practical level, the effort and the cost of designing such high-caliber and computationally expensive EAs do not seem to be justified, since the management cannot possibly refer to ***all*** the global and the local optima in making important managerial decisions. Knowledge of ***a handful of optima*** in the multi-modal problem is sufficient to make quick and speedy decisions.

In this paper, we demonstrate that an implementation of the simple CLONALG algorithm meets this end. This algorithm need not be stretched to locate all the optima of a multi-modal function. It has an inherent mechanism to locate some of the global and local optima, which presents a sufficiently comprehensive scenario to aid the managerial decision making process.

The Artificial Immune System (AIS) algorithm is inspired by the biological immune system [5,21]. The biological immune system is made up of primarily two types of cells - B cells which are produced in the bone marrow and T cells which are produced in the thymus. The pathogens like bacteria and viruses invading the body are called antigens. Both the antigen and the receptors on the surface of the B cells have three-dimensional structures. The affinity between the structure of the receptors and that of the antigen is a measure of the complementarities between the two. When an antigen invades the body, the immune system generates antibodies to diminish the antigen. Initially, the invaded antigen is recognized by a few of the B cells with high affinity for the antigen. Stimulated by the helper T cells, these high affinity B cells proliferate by cloning. This process is called clonal selection principle. The new cloned cells undergo a high rate of somatic mutations called hyper-mutation. The mutations undergone by the clones are inversely proportional to their affinity to the antigen. The highest affinity antibodies experience the lowest mutation rates, whereas the lowest affinity antibodies have the highest mutation rates. The high affinity B cells and their clones proliferate and differentiate into plasma cells. Finally, the plasma cells generate a large number of antibodies to neutralize and eliminate the antigens.

After the cloning and hyper-mutation stage, the immune system must return to its normal condition, eliminating the extra cells. However, some cells remain circulating throughout the body as memory cells. When the immune system is later attacked by the same type of antigen (or a similar one), these memory cells are activated, presenting a better and more efficient secondary response.

Among the various mechanisms in the biological immune system that are explored as AISs, negative selection [12], immune network model [6] and clonal selection [7] are the most discussed models. The CLONALG algorithm based on the above clonal selection principle is also use in optimization [7-8]. In this study, we show that the CLONALG algorithm routinely locates several global and local optima of a multi-modal function.

This paper is organized as follows: Section II presents a review of the literature on multi-modal optimization functions. Section III explains the CLONALG algorithm in detail and introduces the multi-modal test functions. The experimental results are presented in section IV. The paper ends with a brief conclusion in section V.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-202-8

II. LITERATURE REVIEWED

EAs are either extended or hybridized with other optimization techniques to solve the MMO problems. In addition, new algorithms are also designed. This section presents a review of the algorithms found in literature.

A. Extended Algorithms

The standard Genetic Algorithm (GA) is extended towards multi-modal function optimization by introducing a niche-preserving technique [10]. The technique deals with finding and preserving multiple stable niches of the solution space possibly around multiple solutions so as to prevent convergence to a single solution [11]. Niching methods maintain diversity in the population and permit the EA to find many optima in parallel. Clearing [18], Crowding [9,16], Clustering[19], Sharing [3, 11], Restricted Selection [13], Species [19] and Conserving [14] are some of the notable niching techniques employed to extend EAs to find solutions to the multi-modal optimization problems.

The Particle Swarm Optimization (PSO) is a simple, but efficient algorithm based on the swarm intelligence metaphor. NichePSO is an extended form of PSO designed to handle multi-modal optimization. The Species-based PSO (SPSO) implements proximity-based speciation and creates turbulent regions around the already found solutions to prevent unnecessary function evaluations [2]. The Bottleneck Assigned Binary Any System is inspired by the traffic organization in real ants under crowded conditions [24].

B. Hybrid Algorithms

Memetic algorithm which introduces local search techniques in PSO is developed by Wang et al. [23]. The PSO disperses the solutions in diverse sub-regions, where an adaptive local search takes place to locate the optima. The Niche Hybrid Algorithm (NHGA) is a hybrid form of the Nelder-Mead’s Simplex Method and GA and is used in the multi-modal optimization of vehicle suspension system [1]. An agent-based hybrid algorithm is found in [15] and a Differential Evolution hybrid algorithm is found in [20].

C. New Algorithms

The Artificial Immune Network Algorithm for multi-modal optimization (opt-aiNet) is a novel algorithm [12]. One of the salient features of this algorithm is the increase in population at every iteration. The increasing population is an indication that the problem has many local optima which the algorithm finds efficiently. Estimation of distribution algorithms (EDAs) are a new set of algorithms used in MMO. Unlike most EAs, EDAs do not make use of variation operators (e.g., crossover and/or mutation) in the combination step. Instead, EDAs generate the offspring population at each iteration by learning and subsequent simulation of a joint probability distribution for the individuals selected [17]. The ensemble of niching algorithms (ENA) approach uses several niching methods in parallel in order to preserve diversity of the populations and to benefit from the best method [22]. Other miscellaneous evolutionary approaches, including the Multi-objective Optimization (MO) algorithms are found in the comprehensive survey on MMO by Das et al. [4].

III. AIS FOR MMO ALGORITHMS

In this section, we define the objective functions and the AIS clonal selection algorithm that is used in finding a set of global and local optima in these MMO benchmark test functions.

A. AIS Algorithm

The AIS clonal algorithm described below consists of the following steps:

Generation of antibody population

A population consisting of N antibodies (Abs) is randomly generated. Each antibody represents a feasible solution to the optimization problem. The Abs in our application are represented as binary bit strings or as real numbers.

Objective function evaluation

The multimodal test functions (Equations 1 ~ 4) are evaluated numerically for each of the antibodies.

Affinity Calculation

The affinity (or the fitness) of each individual antibody is evaluated based on the value of the objective functions mentioned in this sub-section.

Clone Selection

A certain percentage of the antibodies with greater affinities are selected from the population. These are then cloned to produce additional antibodies.

Affinity Proportional Mutation

The clones produced in the above step are subjected to mutations in proportion to their affinities.

Memory Renewal

The antibodies with relatively lower affinities (i.e., with higher values of the objective function) are eliminated. The selected clones are introduced into the antibody population as the immune memory cells. The above steps are iterated for M number of cycles. The Ab with the highest affinity (maximum values) found in all the iterations is the optimal solution.

B. Test Functions

We have chosen the following benchmark multi-modal maximization functions to demonstrate the applicability of the simple CLONALG algorithm: Rastrigin (1), Schaffer (2), Multi-function (3) and Roots function (4).

$$f(x,y) = 10 * n + \sum_{i=1}^{n} (x_i^2 - 10 * \cos(2\pi x_i)) \qquad (1)$$

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-202-8

$$-5.12 \leq x_i \leq 5.12, i = 1, 2, \ldots, n.$$

$$f(x,y) = 0.5 + \frac{\sin^2\left(\sqrt{x^2 + y^2} - 0.5\right)}{\left(1 + 0.001(x^2 + y^2)\right)} \quad (2)$$

$$(x,y) \in [-10, 10]$$

$$f(x,y) = x * \sin(4\pi x) - y * \sin(4\pi y + \pi) + 1 \quad (3)$$

$$(x,y) \in [-1, 2]$$

$$f(z) = \frac{1}{1 + |z^6 - 1|} \quad (4)$$

$$z \in \mathbf{C}，\ z = x + iy，\ (x,y) \in [-2，\ 2]$$

IV. Experimental Results

The different parameters of the algorithms used to produce the experimental results are shown in Table 1. The table shows a modest number of populations evolving through a modest number of generations (iterations). Since our aim is to demonstrate that the CLONALG algorithm produces a good number of optima under ordinary performance conditions, we have tried not to vary the (default) parameters used in conventional test runs.

TABLE I. Execution parameters of algorithms

Algorithm	Parameter		
	Population (N)	Total generations (Ngen)	Mutation Probability
PSO	40	200	-
GA	100	200	0.01
AIS (binary)	100	200	0.01
AIS (real)	100	200	0.01

The test results are plotted as 3D graphs (Figure 1 to Figure 4). In our experiments, PSO is coded as real-valued, while GA is coded as binary. However, in the case of the CLONALG algorithm, binary as well as real-valued algorithm is implemented. Both the implementations of CLONALG find a good deal of the global and local optima of the multi-modal test functions as explained below.

a) Rastrigin function: PSO and GA converge to a single global optimum (shown as black dots) (Figure 1 a, b). As to which optimum these algorithms converge to, depends on the randomly generated initial populations. But the binary as well as the real version of the CLONALG algorithm locate at least one of the global maxima and a couple of the local maxima (Figure 1 c, d).

b) Schaffer's function: All the particles in PSO rapidly converge to the central ridge of the function (Figure 2 a), while the GA is spreadout onto a few global optima. Some of the solutions are cearly sub-optimal. But the binary as well as the real version of the CLONALG algorithm locate at least 2~3 central peaks of the function (Figure 2 c, d).

c) Multi-function: PSO easily converges to the global optimum (Figure 3 a), while the GA surrounds the global optimum and is spreadout in the search space (Figure 3 b).

d) Roots function: PSO locates two of the six peaks (Figure 4 a). On the other hand, GA locates just one of the peaks and is rather spread out (Figure 4 b). Both the versions of CLONALG successfully locate all the six peaks of the Roots function (Figure 4 c, d). In all the experiments, the simple CLONALG algorithm finds a number of global and local optima without using the MMO techniques.

V. Conclusion

In spite of their inherent parallelism, Evolutionary Algorithms are found to converge to a single global optimum when they attempt to optimize multi-modal functions. Extensive research has been done to design EAs to locate all the global optima and as many local optima as possible. However, the effort and the cost of designing such high-caliber and computationally expensive EAs do not seem to be justified, since the management cannot possibly refer to ***all*** the global and the local optima in making important managerial decisions. Knowledge of a handful of optima in the multi-modal problem is sufficient to make quick and speedy decisions. We demonstrated through a series of multi-modal test functions that an implementation of the simple CLONALG algorithm, without any special modification toward multi-modal optimization, meets this end.

References

[1] Alugongo, A.A. ;Lange, J.M., Optimization of multimodal models in mechanical design by a Niche Hybrid Genetic Algorithm, AFRICON 2009, pp.1- 6

[2] H. Cho, D. Kim, F. Olivera, S. D. Guikema, Enhanced speciation in particle swarm optimization for multi-modal problems, European Journal of Operational Research, (2011), 213(1):15-23

[3] A.D. Cioppa, C. De Stefano, A. Marcelli, Where are the niches? Dynamic fitness sharing, IEEE Transactions on Evolutionary Computation 11 (2007) 453-465.

[4] S. Das, S. Maitya, B-Y. Qub, P.N. Suganthanb, Real-parameter evolutionary multimodal optimization — A survey of the state-of-the-art, Swarm and Evolutionary Computation,1(2)71-88, (2011)

[5] de Castro L.N., Von Zuben, F.J.: Artificial immune systems: Part II—A survey of application. State Univ. Campinas, Campinas, Brazil, Tech. Rep. RT DCA 02/0065 (2000)

[6] de Castro L.N., Von Zuben, F.J.: aiNet: An artificial immune network for data analysis. In: Data Mining: A Heuristic Approach, H.A. Abbass, R.A. Sarker, and C.S. Newton (eds). Idea Group Publishing, USA, pp. 231--259 (2001)

[7] de Castro L.N., Von Zuben, F.J.: Learning and optimization using the clonal selection principle. IEEE Trans. Evol. Comput., 6(3): 239-251 (2002)

[8] de Castro L.N., Timmis, J.: An artificial immune network for multimodal function optimization. In: Proc. IEEE Congress on Evolutionary Computation, vol. 1, 699-674 (2002)

[9] K.A. De Jong, An Analysis of the Behavior of a Class of Genetic Adaptive Systems, Doctoral Dissertation, University of Michigan, 1975.

[10] Goldberg, D. E. and Richardson, J. (1987). Genetic algorithms with sharing for multimodal function optimization. In Grefenstette, J. J., editor, Genetic Algorithms and their Applications: Proceedings of the Second International Conference on Genetic Algorithms, pp. 41-49, New Jersey.

[11] D.E. Goldberg, L. Wang, Adaptive niching via coevolutionary sharing, Genetic Algorithms and Evolution Strategy in Engineering and Computer Science (1997) 21–38.

[12] Hart, E., Ross, P.: Exploiting the analogy between immunology and sparse distributed memories: A system for clustering non-stationary data. 1st International Conference on Artificial Immune Systems (ICARIS), pp. 49-58, (2002).

[13] J.-K. Kim, D.-H. Cho, H.-K. Jung, C.-G. Lee, Niching genetic algorithm adopting restricted competition selection combined with pattern search method, IEEE Transactions on Magnetics 38 (2) (2002) 1001-1004.

[14] Li, J-P., Balazs, M. E., Parks, G. T., and Clarkson, P. J. (2002). A Species Conserving Genetic Algorithm for Multimodal Function Optimization. Evolutionary Computation, 10(3):207-234.

[15] R. I. Lung, C. Chira, and D. Dumitrescu. An agent-based collaborative evolutionary model for multimodal optimization. GECCO '08, pp. 1969-1976.

[16] S.W. Mahfoud, Crowding and preselection revisited, Parallel Problem Solving from Nature 2 (1992) 27-37.

[17] J. M. Pena, J. A. Lozano, and P. Larranaga. Globally multimodal problem optimization via an estimation of distribution algorithm based on unsupervised learning of bayesian networks. Evolutionary Computation,13(1):43-66, 2005.

[18] A. Pétrowski, A clearing procedure as a niching method for genetic algorithms, IEEE International Conference on Evolutionary Computation, New York, 1996, pp. 798-803.

[19] G. Singh and D. Kalyanmoy Deb. Comparison of multi-modal optimization algorithms based on evolutionary algorithms, GECCO '06, pp. 1305-1312.

[20] Thangaraj, R., Pant, M., Abraham, A., and Badr, Y. (2009). Hybrid evolutionary algorithm for solving global optimization problems. Lecture Notes in Computer Science, 5572:310–318.

[21] Timmis, J., Knight, T., de Castro L.N., Hart, E.: An Overview of artificial immune systems. In: Computation in Cells and Tissues: Perspectives and Tools Thought. Natural Computation Series, Springer-Verlag, 51-86 (2004)

[22] E. L. Yu, P. N. Suganthan, Ensemble of niching algorithms, Information Sciences: an International Journal, 180(15):2815-2833, (2010)

[23] H. Wang ; N. Wang; D. Wang, A memetic particle swarm optimization algorithm for multimodal optimization problems, CCDC, 2011, pp. 3839-3845

[24] J. Zhao; C. Yan; A Bottleneck Assigned Binary Ant System for multimodal optimization, Conference on Decision and Control, 2009, China, pp.6195-6200

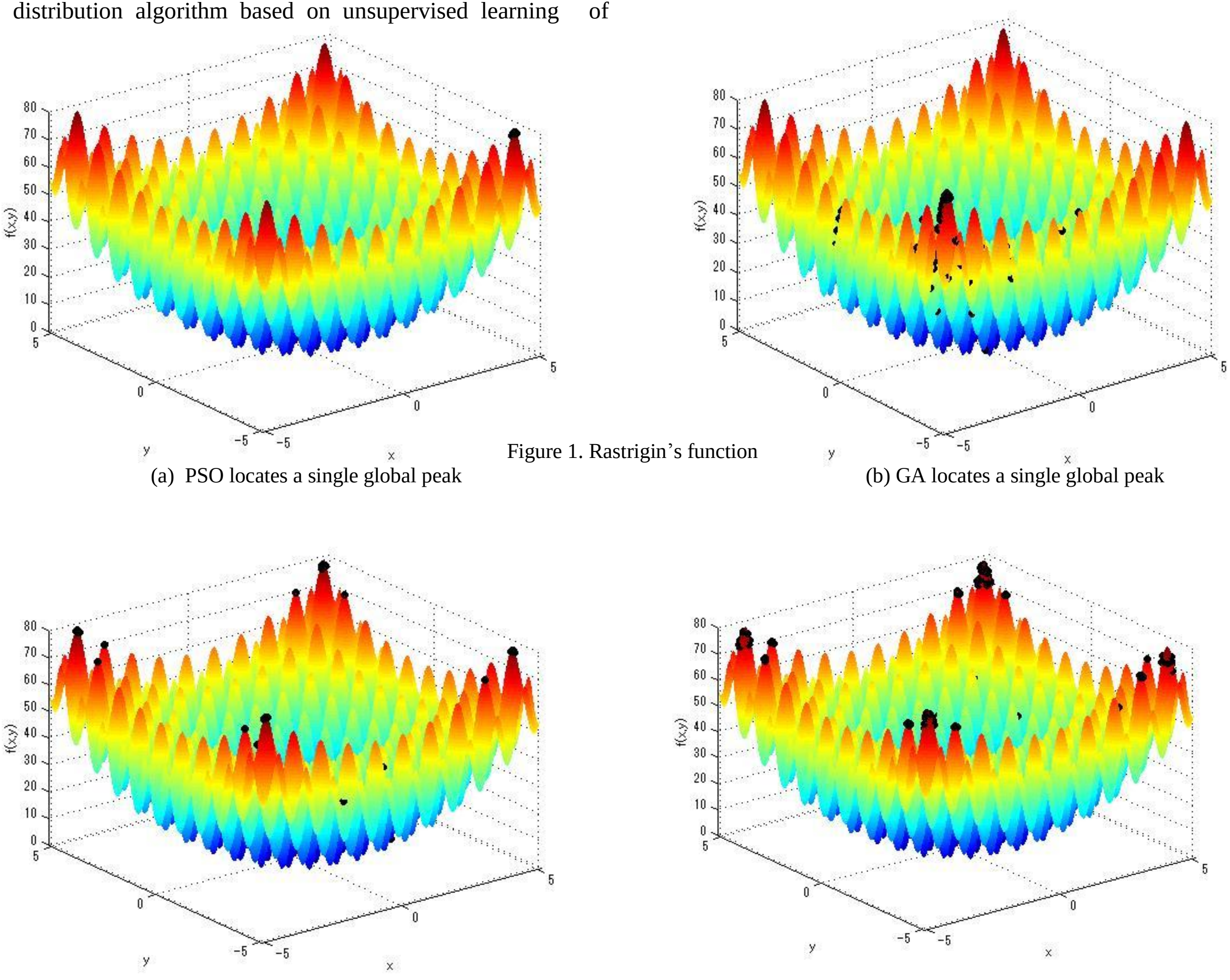


Figure 1. Rastrigin's function

(a) PSO locates a single global peak

(b) GA locates a single global peak

(c) CLONALG (binary) locates all global and a few local optima

(d) CLONALG (real) locates all lobal and a few local optima

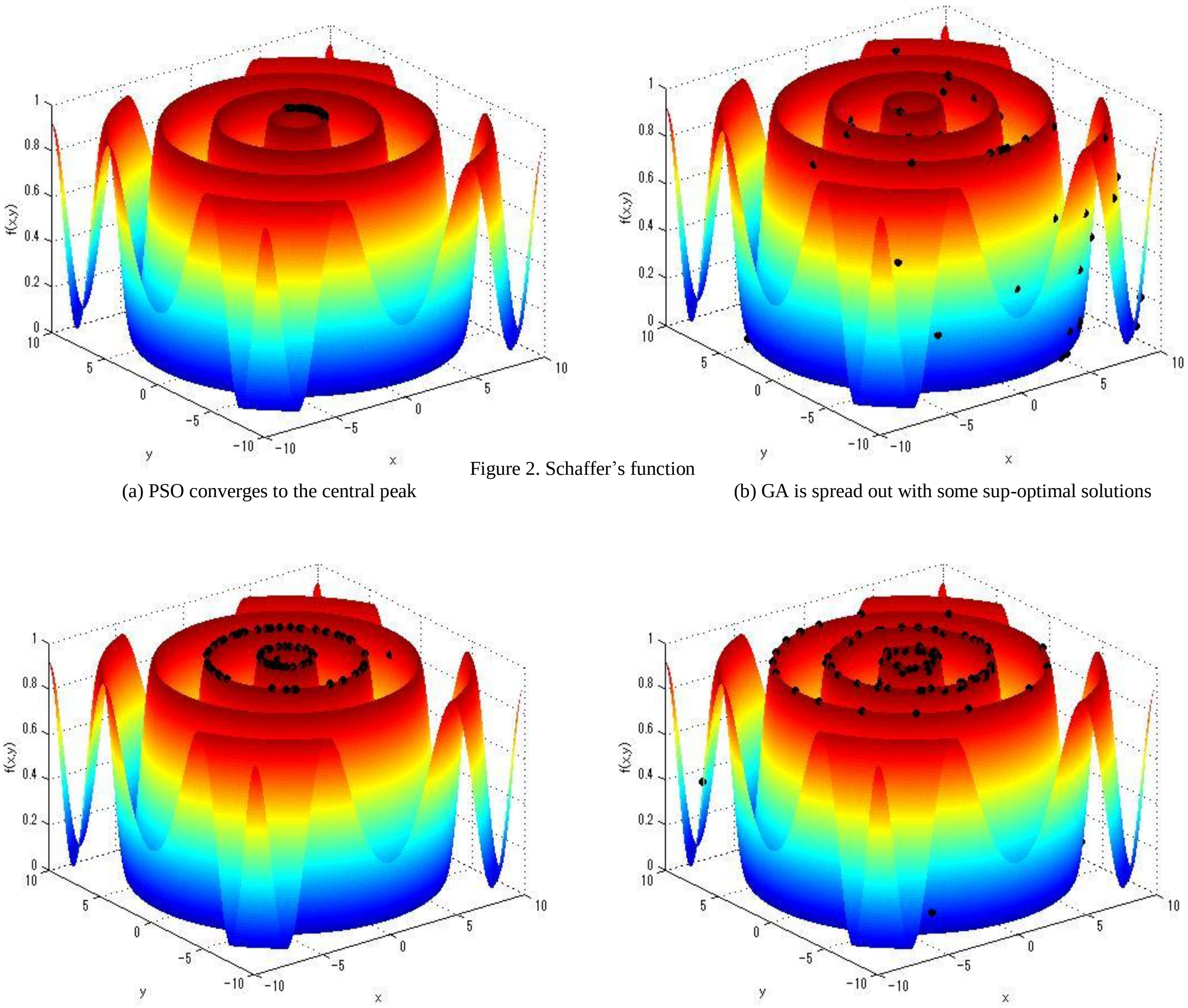


Figure 2. Schaffer’s function

(a) PSO converges to the central peak

(b) GA is spread out with some sup-optimal solutions

(c) CLONALG (binary) locates 2 global maxima

(d) CLONALG (real) locates 3 global maxima

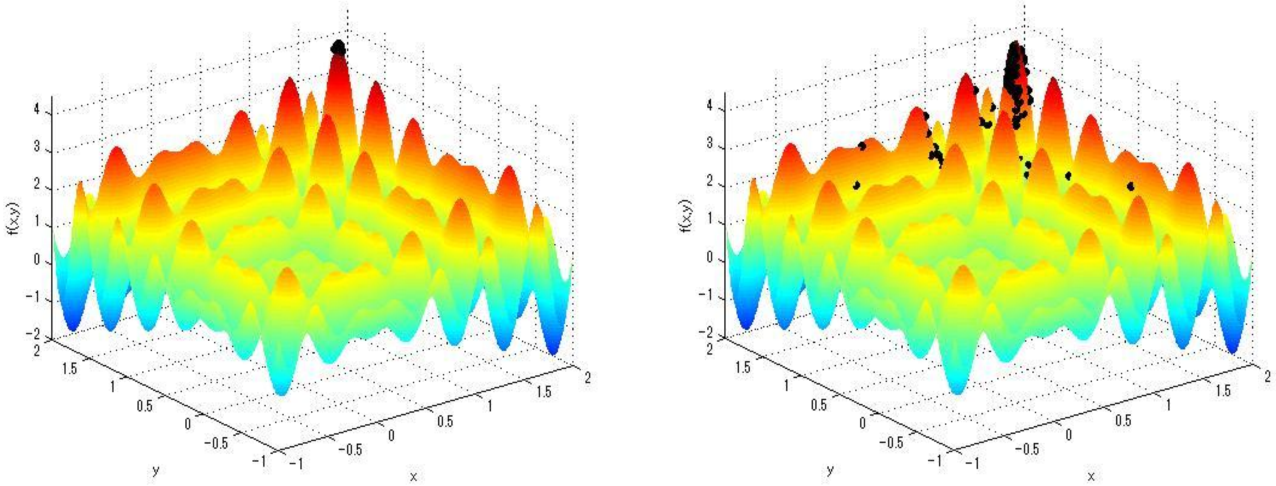
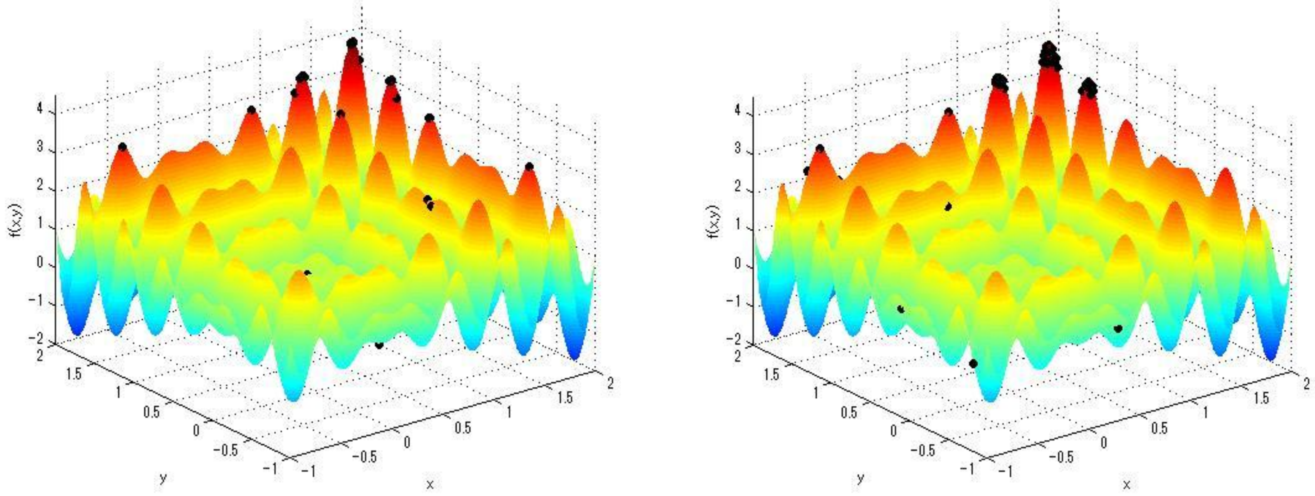


Figure 3. Multi-function

(a) PSO converges to the global optima

(b) GA surrounds the global optima and is spread out



(c) CLONALG (binary) locates the global and some local optima

(d) CLONALG (real) locates the global and some local optima

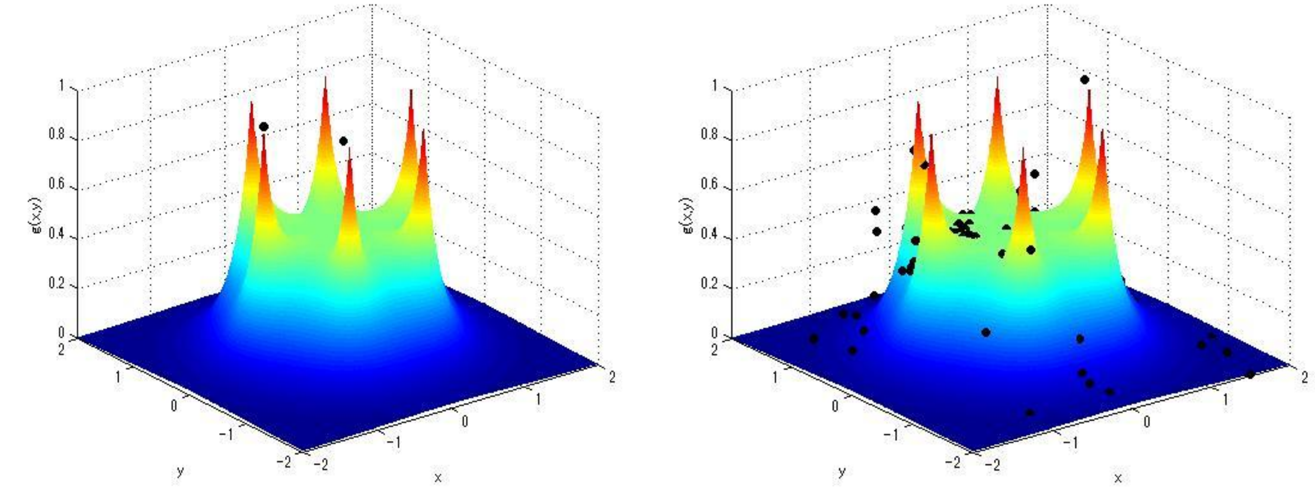
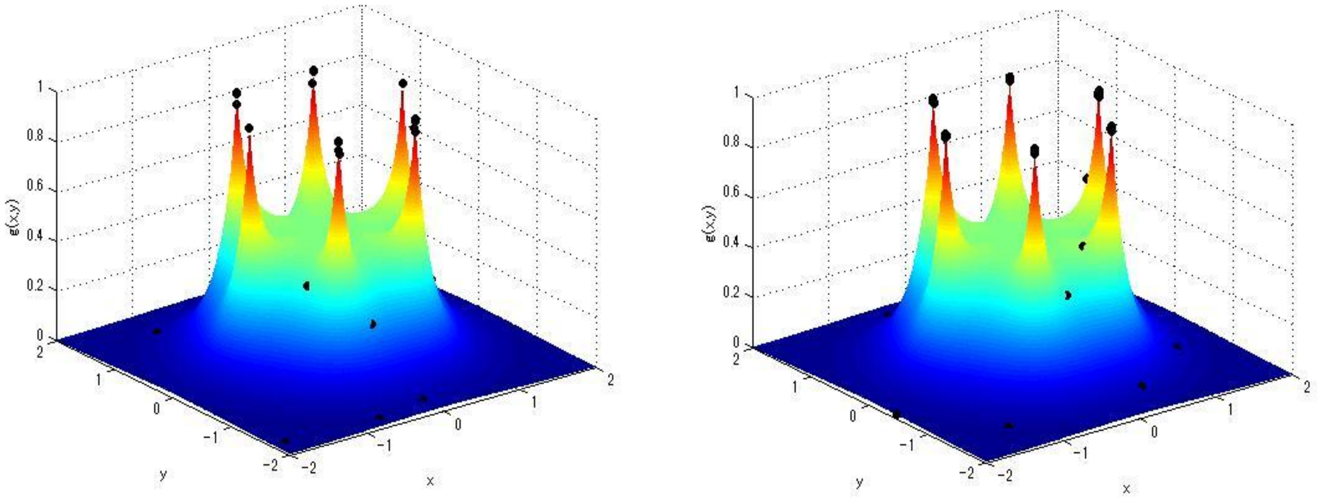


Figure 4. Roots function

(a) PSO locates 2 global optima

(b) GA locates 1 global optimum and is spread out



(c) CLONALG (binary) locates all the global optima

(d) CLONALG (real) locates all the global optima

Data Clustering Using Bee Colony Optimization

Khadijeh Keshtkar mizooji
Department of Electrical & Computer Engineering,
Islamic Azad University, Qazvin Branch,
Qazvin, Iran
kh.keshtkar@gmail.com

Abolfazl Toroghi Haghighat
Department of Electrical & Computer Engineering,
Islamic Azad University, Qazvin Branch,
Qazvin, Iran
haghighat@qiau.ac.ir

Rana Forsati
Department of Electrical & Computer Engineering,
Islamic Azad University, Karaj Branch
Karaj, Iran
r_forsati@sbu.ac.ir

***Abstract*— The paper presents a comparative analysis of data clustering by Bee Colony Optimization (BCO) technique. Experiments over a standard benchmark demonstrate that applying Bee Colony Optimization in the context of clustering is a feasible approach and improves the clustering results. Superiority of the proposed algorithm is demonstrated by comparing it with some recently developed partitional clustering techniques.**

Keywords-Clustering; Swarm Intelligence; Function Optimization.

I. INTRODUCTION

Cluster analysis seeks to divide a set of objects into a small number of relatively homogeneous groups on the basis of their similarity over N variables [1]. Cluster analysis can be viewed either as a means of summarizing a data set or as a means of constructing a topology [12]. Patterns within a valid cluster are more similar to each other than to a pattern belonging to a different cluster. Clustering is useful in several exploratory pattern-analysis, grouping, decision-making, data mining, document retrieval, image segmentation and pattern classification [33].

Our concern in this paper is based on partitioning clustering [7] methods which relocate instances by moving them from one cluster to another, starting from the initial partitioning.

Partitioning methods try to partition a collection of objects into a set of groups, so as to maximize a pre-defined objective value. The most popular partitional clustering algorithms are the prototype-based clustering algorithms where each cluster is represented by the center of the cluster and the used objective function is the sum of the distances from the patterns to the center [8].

The most popular class of partitional clustering methods is is K-means algorithm [11], where K denotes the number of clusters. The reasons for the algorithmic popularity is its ease of interpretation, simplicity of implementation, speed of convergence, adaptability to sparse data and works fast in most situations [1].

The disadvantages of this algorithm lie in the fact that the number of clusters, K, must be specified prior to application. Also, since the summary statistic is mean of the values for each cluster, so, the individual members of the cluster can have a high variance and mean may not be a good summary of the cluster members. In addition, as the number of clusters grow, for example to thousands of clusters, K-means clustering becomes untenable, approaching the O (n^2) comparisons where n is the number of documents. However, for relatively few clusters and a reduced set of pre-selected words, K-means can do well [12]. The other major drawback of K-means algorithm is sensitivity to initial states. Finally, K-means algorithm converges to the nearest local optimum from the starting position of the search and the final clusters may not be the optimal solution.

In order to overcome these problems that exist in traditional partition clustering methods new techniques have been proposed in this area by researchers from different fields. One of these techniques is optimization methods that tries to optimize a pre-defined function that can be very useful in data clustering.

Optimization techniques define a global function and try to optimize its value by traversing the search space.

Bee Colony Optimization (BCO) [26] is a nature-inspired metaheuristic optimization method, which is similar to the way bees in nature look for food, and the way optimization algorithms search for an optimum in combinatorial optimization problems. The performance of the BCO algorithm has been compared with those of other well-known modern heuristic algorithms such as genetic algorithm, differential evolutional algorithm, and particle swarm optimization algorithm for unconstrained optimization problems. The BCO belongs to the class of population-based and Swarm Intelligence techniques [26], which is considered to be applied to find solutions for difficult combinatorial optimization problems. The major idea behind the BCO is to create the multi agent system (colony of artificial bees) capable to efficiently solve hard combinatorial optimization problems. These features increase the flexibility of the BCO algorithm and produce better solutions. The artificial bee colony behaves to some extent similar and to some extent in a different way from bee colonies in nature. They explore through the search space looking for the feasible solutions. In order to discover superior and superior solutions, artificial bees cooperate with each other and exchange information. Also, they focus on more promising areas and gradually discard solutions from the less promising areas via collective knowledge and giving out information among themselves.

In this paper, by modeling the partitioning problem as an optimization problem, a BCO-based clustering algorithm is proposed. The performance of the proposed algorithm by applying it to standard benchmark functions and also for clustering real-world data sets is evaluated. The reminder of this paper is organized as follows. In Section 2, some previous related works are summarized. In Section 3, the BCO-based clustering algorithm is described. Section 4 presents data sets used in our experiments, the performance evaluation of the proposed algorithm compared to K-means and GA-based and PSO-based clustering algorithms. Conclusion is discussed in Section 5.

II. DATA CLUSTERING METHODS: A BRIEF OVERVIEW

Data clustering can be hierarchical or partitional [2][3]. A hierarchical algorithm [4][5] creates a hierarchical decomposition of the given dataset forming a dendrogram—a tree which splits the dataset recursively into smaller subsets and represent the objects in a multi-level structure.

Hierarchical clustering algorithms can be agglomerative (bottom-up) or divisive (top-down) [6]. Agglomerative algorithms begin with each element as a separate cluster and merge them into larger clusters. Divisive algorithms begin with the whole set of data objects and proceed to divide it into successively smaller clusters [6].

Partitional clustering algorithms relocate instances by moving them from one cluster to another, starting from the initial partitioning. Such method requires the number of clusters to be preset by the user [1].

Although hierarchical methods are often said to have better quality in clustering, they usually do not provide the reallocation of objects, which may have been poorly classified in the early stages of the analysis [3] and the time complexity of them declared to be quadratic [9]. On the other hand, in recent years the partitioning clustering methods showed a lot of advantages in applications involving large datasets due to their relatively low computational requirements [٩][١٠]. The time complexity of the partitioning technique is almost linear, which makes it widely used.

In addition to the algorithms mentioned above, several heuristics algorithms, such as statistics [13], graph theory [14], expectation-maximization algorithms [15], evolutionary algorithms 18][30-32] and swarm intelligence algorithms [19-25][27] have been proposed for data clustering.

As the behavior of the K-means algorithm mostly is influenced by the number of clusters specified and the random choice of initial cluster centers, in this study, we present a novel algorithm based on the Bee Colony Optimization. BCO is applied in the clustering problem because of its robust, adaptive search method for performing global search. The proposed algorithm, called Bee Colony Clustering, which is good at finding promising areas of the search space but not as good as K-means at fine-tuning within those areas. To demonstrate the effectiveness and speed of proposed algorithm, we have applied these algorithms on various standard datasets and got very good results compared to the K-means and PSO-based clustering algorithm [22]. BCO and PSO algorithms fall into the same class of artificial intelligence optimization algorithms, population-based algorithms, and they are proposed by inspiration of swarm intelligence. Beside, comparing the BCO algorithm with PSO algorithm, the performance of BCO algorithm is also compared with a wide set of classification techniques. The evaluation of the experimental results shows considerable improvements and robustness of the proposed algorithm.

III. THE BASIC BEE COLONY BASED ALGORITHM TO DATA CLUSTERING

In order to cluster data using bee colony algorithm, we must first model the clustering problem as an optimization problem that locates the optimal centroids of the clusters rather than to find an optimal partition. This model offers us a chance to apply bee colony optimization algorithm on the optimal clustering of a collection of data. The following subsections describe the proposed algorithm.

A. Representation of Solutions

In order to apply BCO to solve clustering problem, we have used floating point arrays to encode cluster centers. The assignment matrix has the properties that each data must assigned exactly to one cluster. An assignment that represents K nonempty clusters is a legal assignment. In this model, each food source discovered by each bee is a candidate solution and corresponds to a set of K centroids. Let us denote by a finite set of pre-selected stages, where K is the number of stages. By B, we denote the number of bees to participate in the search process and by I the total number of iterations.

At each forward pass, bees are supposed to visit a single stage. All bees are located in the hive at the beginning of the search process. Each artificial bee allocates some of the data to the corresponding cluster with special probabilities in each stage, and in this way constructs a solution of the problem incrementally. Bees are adding solution components to the current partial solution until they visit all of the K stages. The search process is composed of iterations. The first iteration is finished when bees create feasible solutions. The best discovered solution during the first iteration is saved, and then the second iteration begins. In each iteration of proposed algorithm, for each cluster (stage), all the bees leave the hive to allocate some of the data to that cluster with special probabilities and come back to the hive to see the work of other bees until that time and decide whether to continue its way or select one of the other bees' solution and continue on that way.

B. Evaluation of solutions

A key characteristic of most clustering algorithms is that they use a global criterion function whose optimization drives the entire clustering process. For those clustering algorithms, the clustering problem can be stated as computing a clustering solution such that the value of a particular objective function is optimized. Our objective function is to minimize intra-cluster similarity while maximizing the inter cluster similarity.

Fitness value of each solution is measured by equation:

$$f = \sum_{j=1}^{K} \left(\sum_{i=1}^{n_j} D(d_{ij}, C_j) \right) \tag{1}$$

A food source represents a possible solution to the problem. The quantity of existing sources of pollen, nectar in the areas is explored by the bees corresponds to the quality of the solution represented by that food source. Bees search for food sources in a way that minimize the ration *f* where *f* is the proportional to the nectar amount of food sources discovered by bees. In this problem, the goal is to find the minimum of the objective function.

The each iteration of the proposed algorithm is detailed in the following steps:

Step 1. **Initialization**: If this is not the first iteration of the algorithm and the best discovered cluster centers during the previous iterations are available, the initial cluster centers for all the stages are set to the best answer of the previous iteration. Else if this is the first iteration, a set of initial cluster centers generated randomly from the dataset points will be set for each stage. There is a loop from 1 to *K* where in each loop the following two steps are performed:

Step 2. **Constructive moves in the forward pass**: In each forward pass, every artificial bee visits one stage, allocates the data to the corresponding cluster, and after that returns to the hive as detailed in step 3. For each cluster, the probability of a bee choosing the data *i* as a member of jth cluster (c_j), p_{ij}, is expressed as follows:

$$p_{ij} = \frac{e^{-D(d_i - C_j)}}{\sum_{m=1}^{n} e^{-D(d_m - C_j)}}, \; j = 1,2,...,K \tag{2}$$

where $D(d_i - C_j)$ denotes the distance of ith data to jth cluster and *n* denotes the number of not previously chosen data. Within each forward pass a bee visited a certain number of nodes and created a partial solution. After solutions are evaluated (and normalized) the loyalty decision and recruiting process are performed as described in the following subsection.

Step 3. Backward pass (Bees' partial solutions comparison mechanism): After all of the bees completed step 2, they will be back to hive to compare their partial solutions with themselves. We assume that every bee can obtain the information about solutions' quality generated by all other bees. In this way, bees compare all generated partial solutions. Based on the quality of the partial solutions generated, every bee decides whether to abandon the created partial solution or dance and thus recruit the nest mates before returning to the created partial solution. Depending on the quality of the partial solutions generated, every bee possesses certain level of loyalty to the previously discovered partial solution. Our criterion to decide about the goodness of discovered solution in general is sum of the distance of each vector from its cluster center for all the vectors. We want this criterion to be as minimal as possible. So, as the bees are back at the hive, the probability that b-th bee (during stage *u* and iteration *z*) will be faithful to its previously generated partial solution (loyalty decision) is expressed as follows:

$$p_b(u+1, z) = e^{-\frac{O_b(u,z)}{u \times z}} \quad , \qquad b = 1,2,..., B \tag{3}$$

where

$$O_b(u,z) = \frac{SumDis\tan ce_b(u,z) - SumDis\tan ce_{\min}(u,z)}{SumDis\tan ce_{\max}(u,z) - SumDis\tan ce_{\min}(u,z)} \tag{4}$$

where

$$SumDis\tan ce_b(u,z) = \sum_{i=1}^{u} \sum_{j=1}^{N} D^b{}_{ji}$$

$$D^b{}_{ji} = \begin{cases} \left(\sum_{k=1}^{m} (d^b{}_{jk} - c_{ik})^2\right)^{\frac{1}{2}} & , \; if\ d_j\ has\ been\ selected\ by\ b-th\ bee \\ 0 & , otherwise \end{cases}$$

$$SumDis\tan ce_{\min}(u,z) = \min_i \{SumDis\tan ce_i(u,z)\} \quad i = 1,2,..., B$$

$$SumDis\tan ce_{\max}(u,z) = \max_i \{SumDis\tan ce_i(u,z)\} \quad i = 1,2,..., B$$

We denote by O_b the normalized value of sum distance, with sumDistance$_b$ is sum of the distance of each vector from its cluster center for all the vectors that has been classified by bee number b and minSumDistance is minimum of this sum that exists among all bees.

$SumDis\tan ce_{\max}(u,z)$: the objective function value of the worst discovered partial solution from the beginning of the search process

$SumDis\tan ce_{\min}(u,z)$: the objective function value of the best discovered partial solution from the beginning of the search process

u : the ordinary number of the forward pass (e.g., u = 1 for first forward pass, u = 2 for second forward pass, etc.) that in each forward pass one of the clusters' members are decided and z denotes the iteration number.

Step 4. Recruiting process: In the case when at the beginning of a new stage a bee does not want to expand the previously generated partial solution, the bee will go to the dancing area and will follow another bee. Within the dance area the bee dancers (recruiters) 'advertise' different partial solutions. We have assumed in this paper that the probability that b's partial solution would be chosen by any uncommitted bee is equal to:

$$p_b = e^{-\theta O_b(u,z)}, \; b = 1,2,...,RC \qquad (5)$$

where θ is a coefficient which is a double between 0 and 1 and RC denotes the number of recruiters and O_b denotes the normalized value for the objective function of partial solution created by the bth bee advertised partial solution.

$$O_b = \frac{SumDis\tan ce_b(u+1,z) - SumDis\tan ce_{\min}(u+1,z)}{SumDis\tan ce_{\max}(u+1,z) - SumDis\tan ce_{\min}(u+1,z)} \qquad (6)$$

where $maxSumDistance$ is the maximum sum of the distance of each vector from its cluster centers for all the data that has been classified until now that exists among all the bees.

This probability p_b is used in a roulette wheel selection or tournament selection algorithm and one of the bees is selected.

Using Eq. (5) and a random number generator, every uncommitted follower joins one bee dancer (recruiter). Recruiters fly together with their recruited nestmates in the next forward pass along the path discovered by the recruiter. So the bee that wants to continue another partial solution will set its partial solution exactly as the selected bee but will continue the algorithm independently. At the end of this path, all bees are free to independently search the solution space and generate the next iteration of constructive moves.

Step 5. **Set the cluster centers** (compute the Centroid of Clusters): At last, the cluster centers as the centroid of the vectors belong to each cluster for each bee are computed as follows: Each solution extracted by each bee corresponds to a clustering with assignment matrix A. Let $C = (c_1, c_2, ...c_i, ..., c_K)$ is set of K centroids for assignment matrix A. The centroid of the kth cluster is $c_k = (c_{k1}, c_{k2}, ..., c_{Km})$ and is computed as follows:

$$c_{kj} = \frac{\sum_{i=1}^{n} (a_{ki}) d_{ij}}{\sum_{i=1}^{n} a_{ki}} \qquad (7)$$

where m is the number of dimensions in all data.

Step 6. **Selecting the best answer**: In this phase, among all generated solutions, the best one is determined and is used to update the global best. The global best will be used for setting the cluster centers for all the stages in next iteration. At this point, all B solutions are deleted, and the new iteration starts. The BCO runs iteration by iteration until a stopping condition is met.

IV. EXPERIMENTAL RESULTS

In this section, we present the experimental evidences and results that were made on several standard datasets, and the comparisons that were made with other relevant works.

A. *Dataset Description*

In this work, five clustering problems from the UCI database [28], which is a well-known database repository are used to evaluate the performance of the proposed algorithm.

Data Set 1: Fisher's Iris plants database (n = 150, d = 4, K = 3): It is perhaps the best-known database to be found in the pattern recognition literature.

The data set contains four inputs, three classes, and150 data vectors.

Data Set 2: Glass (n = 214, d = 9, K = 6): The data were sampled from six different types of glass: 1) building windows float processed (70 objects); 2) building windows non float processed (76 objects); 3) vehicle windows float processed (17 objects); 4) containers (13 objects);5) tableware (9 objects); and 6) headlamps (29 objects).Each type has nine features: 1) refractive index; 2) sodium; 3) magnesium; 4) aluminum; 5) silicon;6) potassium; 7) calcium; 8) barium; and 9) iron.

Data Set 3: Wisconsin breast cancer data set (n = 683, d=9, K=2):The Wisconsin breast cancer database contains nine relevant features: 1) clump thickness; 2) cell size uniformity;3) cell shape uniformity; 4) marginal adhesion; 5) single epithelial cell size; 6) bare nuclei; 7) bland chromatin;8) normal nucleoli; and 9) mitoses. The data set has two classes. The objective is to classify each data vector into benign (239 objects) or malignant tumors (444 objects).

Data Set 4: (n = 178, d = 13, K = 3): This is a classification problem with "well-behaved" class structures. There are13 features, three classes, and 178 data vectors.

Data Set 5: Vowel data set (n = 871, d = 3, K = 6): This data set consists of 871 Indian Telugu vowel sounds. The data set has three features, namely F1, F2, and F3, corresponding to the first, second and, third vowel frequencies, and six overlapping classes {d (72 objects), a (89 objects), i (172 Objects), u (151 objects), e (207 objects), o (180 objects)}.

B. *Experimental setup*

In the next step, the K-means and the proposed algorithm are applied to the above mentioned data sets. The cosine correlation measure is used as the similarity metrics in each algorithm. The results shown in the rest of paper, for every dataset, are the average of over 20 independent runs of the algorithms (to make a fair comparison), each run with randomly generated initial solutions and different seeds of the random number generator. Also, for an easy comparison, the algorithms run 1,000 iterations in each run since the 1,000 generations are enough for convergence of algorithms.

C. *Comparisons and discussions*

In the previous subsection, the structure of datasets were explained. Now, in this section, we evaluate and compare the performances of the proposed algorithm according to its quality of generated clusters with K-mean [11], PSO [22] and a GA-based [29] clustering algorithm. For evaluation of the clustering results' quality, we use SICD metric which has been selected from internal measures. Whereas SICD examines how much the clustering satisfies the optimization constraints. The smaller the SICD value, the more compact the clustering solution is. Table 1 demonstrates the normalized SICD value of algorithms.

Looking at the results in Table 1, we can see that the results obtained by proposed algorithm are significantly comparable by results obtained by the other evolutionary based algorithms.

TABLE 1- SICD COMPARISONS AMONG PROPOSED ALGORITHM AND THE OTHER ALGORITHMS

		GA	TS	SA	ACO	K-means	PSO	Proposed Algorithm
Iris	Average	139.98	97.86	97.13	97.17	106.05	103.51	97.05
	Worst	193.78	98.57	97.26	97.81			97.33
	best	125.19	97.36	97.1	97.1	97.33	96.66	97.22
Wine	Average	16530.5	16785.46	16530.53	16530.5	18061	16311	16449.81
	Worst	16530.5	16837.54	16530.53	16530.5			16461.8
	best	16530.5	16666.22	16530.53	16530.5	16555.68	16294	16433.37
Glass	Average	-	-	-	-	260.4	291.33	225.19
	Worst	-	-	-	-			250.44
	best	-	-	-	-	215.68	271.29	214.85
Cancer	Average	-	-	-	-	2988.3	3334.6	2976.89
	Worst	-	-	-	-			2977.57
	best	-	-	-	-	2987	2976.3	2976.24
Vowel	Average	_	_	_	_	159242.9	168477	150881.16
	Worst	-	-	-	-			154469.62
	best	_	_	_	_	149422.3	163882	149466.61

V. CONCLUSION AND FUTURE WORKS

In this paper, we proposed a swarm-based data clustering technique. In the proposed algorithm, a group of bees created k centroids, as the cluster centers of each cluster and then data are assigned to the clusters. In other words in the proposed algorithm the solutions represented by the bees were considered as initial centroids for each center of the k-means clusters, which led to significant improvements. Also some relevant comparisons were made, to demonstrate the effectiveness of the algorithms. Our experimental results on different datasets showed that proposed algorithm produces better solutions with high quality in comparison with other algorithms and the difference is tremendous. Different improvements can be done to enhance the evaluation metrics. The bee colony based algorithm can be extended by K-means algorithm through different hybridization methods. For example by running k-means and bee colony colony alternatively 2 different procedure would be produced.

REFERENCES

[1] S. Hanuman, V. Babu, A. Govardhan, and S. C. Satapathy, "Data Clustering Using Almost Parameter Free Differential Evolution Technique", International Journal of Computer Applications, vol. 8, no. 13, pp. 1-7, 2010.

[2] J. Han, M. Kamber, and A. K. H. Tung, "Spatial Clustering Methods In Data Mining: A Survey", In Geographic Data Mining and Knowledge Discovery, New York, 2001.

[3] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data Clustering: A Review", ACM Computing Surveys (CSUR), vol. 31, no. 3, pp. 264-323, 1999.

[4] S. Guha, R. Rastogi, and K. Shim, "CURE: An Efficient Clustering Algorithm for Large Databases", In Proc. of ACM-SIGMOD Int. Conf. Management of Data (SIG-MOD98), pp. 73-84, 1998.

[5] G. Karypis, E. H. Han, and V. Kumar, "CHAMELEON: A Hierarchical Clustering Algorithm Using Dynamic Modeling", Computer, vol. 32, pp. 68-75, 1999.

[6] S. Xu and J. Zhang, "A Parallel Hybrid Web Document Clustering Algorithm and Its Performance Study" , The Journal of Supercomputing, vol. 30, pp. 117-131, 2004.

[7] P. S. Bradley, U. M. Fayyad, and C. A. Reina, "Scaling EM (Expectation Maximization) Clustering To Large Databases", Microsoft Research Technical Report, 1998.

[8] B. Mirkin, "Mathematical Classification and Clustering", Kluwer Academic Publishers, Dordrecht, the Netherlands, 1996.

[9] M. Steinbach, G. Karypis, and V. Kumar, "A Comparison of Document Clustering Techniques", KDD2000, Technical report of University of Minnesota,2000.

[10] J. Kennedy, R. C. Eberhart, and Y. Shi, "Swarm Intelligence", Morgan Kaufmann, New York, 2001.

[11] J. B. MacQueen, "Some Methods For Classification And Analysis Of Multivariate Observations", Proceedings of the Fifth Berkeley Symposium on Mathematical Statistic and Probability, University of California Press, Berkley, pp. 281-297, 1967.

[12] S. Vaithyanathan and B. Dom, "Model Selection in Unsupervised Learning with Applications to Document Clustering", In Proceedings International Conference on Machine Learning, 1999.

[13] E. W. Forgy, "Cluster Analysis of Multivariate Data: Efficiency Versus Interpret Ability of Classification", Biometrics, vol. 21, no. 3, pp. 768–769, 1965.

[14] C. T. Zahn, "Graph-Theoretical Methods For Detecting And Describing Gestalt Clusters", IEEE Trans. Comput., pp. 68–86, 1971.

[15] T. Mitchell, "Machine Learning", McGraw-Hill, New York, 1997.

[16] J. Mao and A. K. Jain, "Artificial Neural Networks For Feature Extraction And Multivariatedata Projection", IEEE Trans. Neural Netw, pp. 296–317, 1995.

[17] S. H. Liao and C. H. Wen, "Artificial Neural Networks Classification and Clustering of methodologies and Applications Literature Analysis From 1995 To 2005", ExpertSys. Appl, vol. 32, pp. 1–11, 2007.

[18] S. Paterlini and T. Minerva, "Evolutionary Approaches for Cluster Analysis", Soft Computing Applications, Springer–Verlag, pp. 167–178, 2003.

[19] C. H. Tsang and S. Kwong, "Ant Colony Clustering And Feature Extraction For Anomaly Intrusion Detection", Stud. Comput. Intell, vol. 34, pp. 101–123, 2006.

[20] R. Younsi and W. Wang, "A New Artificial Immune System Algorithm for Clustering", IDEAL 2004, LNCS 3177, Springer, Berlin, pp. 58–64, 2004.

[21] P. S. Shelokar, V. K. Jayaraman, and B. D. Kulkarni, "An Ant Colony Approach for Clustering", Anal. Chim. Acta 509 , pp. 187–195, 2004.

[22] S. Paterlini and T. Krink, "Differential Evolution and Particle Swarm Optimization In Partitional Clustering", Comput. Stat. Data Anal, pp. 1220–1247, 2006.

[23] Y. Kao and K. Cheng, "An ACO-Based Clustering Algorithm", in ANTS, LNCS 4150, Springer, Berlin, pp. 340–347, 2006.

[24] M. Omran, A. Engelbrecht, and A. Salman, "Particle Swarm Optimization Method for Image Clustering", Int. J. Pattern Recogn. Artif. Intell, vol. 19, no. 3, pp. 297–322, 2005.

[25] D. T. Pham, S. Otri, A. Afify, M. Mahmuddin, and H. Al-Jabbouli, “Data clustering using the bees algorithm,” In: Proc. 40th CIRP International Manufacturing Systems Seminar, 2007, Liverpool.

[26] P. Lucic and D. Teodorovic, " Bee System: Modeling Combinatorial Optimization Transportation Engineering Problems by Swarm Intelligence, In preprints of the TRISTAN IV Triennial symposium on Transportation Analysis. Sao Miguel, Azores Island, pp. 441-445, 2001.

[27] K. Krishna and M. NarasimhaMurty, "Genetic K-Means Algorithm", IEEE Transactions on Systems, Man, and Cybernetics Part B: Cybernetics, vol. 29, no. 3, pp. 433-439, 1999.

[28] P. Murphy and D. Aha, "UCI Repository of Machine Learning Databases", 1995, URL http://www.sgi.com/Technology/mlc/db, [retrieved: 03, 2012].

[29] U. Mualik and S. Bandyopadhyay, "Genetic Algorithm Based Clustering Technique", Pattern Recognition, vol. 33, pp. 1455–1465, 2002.

[30] R. Forsati, A. Moayedikia, and B. Safarkhani, "Heuristic Approach To Solve Feature Selection Problem", DICTAP 2011, 2011, pp. 707-717.

[31] R. Forsati, M. Shamsfard, and P. Mojtahedpour, "An Efficient Meta Heuristic Algorithm For Pos-Tagging", Fifth International Multi-Conference on Computing in the Global Information Technology (ICCGI), pp. 93-98, 2010.

[32] R. Forsati, M. Mahdavi, M. Kangavari, and B. Safarkhani, "Web page clustering using Harmony Search optimization", Canadian Conference on Electrical and Computer Engineering, CCECE 2008, pp. 1601-1604, 2008.

[33] D. B. Kenneth, "Cluster Analysis. Sociological Methodology", vol. 6, pp. 59-128, American Sociological Association, 1975.

Evaluation of Preferred Brightness and Detail Levels in 3D and 2D Images Based on HDR Tone Mapping

Zicong Mai, Colin Doutre, Panos Nasiopoulos, and Rabab Ward
Department of Electrical and Computer Engineering
The University of British Columbia, Canada
Vancouver, Canada
{zicongm, colind, panosn, rababw}@ece.ubc.ca

***Abstract*—High dynamic range (HDR) images provide superior picture quality by allowing a larger range of brightness levels to be captured and reproduced. Even with existing 8-bit displays, picture quality can be significantly improved if content is first captured in HDR format, and then tone-mapping is applied to convert from HDR to 8-bit, low dynamic range format. Tone mapping methods have been extensively studied for 2D images. By varying the tone mapping process, the brightness and level of details (i.e., sharpness) of the output image can be altered. In this paper, we present a study that compares: i) the preferred level of brightness, and ii) the preferred amount of details, of tone mapped images when they are displayed in 2D and 3D formats. Images at a large range of different brightness and detail levels were first generated by HDR capturing followed by tone mapping with different parameters. We performed an extensive subjective experiment that allowed participants to vary the brightness and amount of details in the output tone mapped images and select the level they thought gave the best visual quality. The results showed that there is no statistically significant difference between the preferred brightness level for 2D and 3D content. On the other hand, the subjects consistently chose a higher level of details (i.e. a sharper image) for 3D images compared to the level of details they thought was optimal in 2D mode. This result indicates that when processing 3D content, the image should be sharper than the same content viewed in 2D mode for optimal appearance.**

***Keywords*—3D; high dynamic range; HDR; tone-mapping; quality of experience**

I. INTRODUCTION

High dynamic range (HDR) images/videos have been gaining increasing attention in the past decade because of the superior picture quality they are able to deliver. Existing low dynamic range (LDR) content allows only a limited range of contrast which is far below the capability of human eyes. HDR media supports a very large luminance range that is comparable to what the human vision systems is able to perceive. HDR signals need to be encoded with at least 10 bits per color component [1], as opposed to LDR signals which are represented by only eight bits. Although the majority of today's displays can support only LDR content, they can all provide much better picture quality if the content is first captured in HDR and then converted to LDR format. Such production pipeline, i.e., shooting in HDR and then rendering to LDR, has been gaining increasingly interest in movie/television production and high-end photography.

In order to show HDR content on existing/8-bit displays, a process that converts HDR to LDR signals needs to be conducted. This process is called tone-mapping. A number of tone-mapping operators (TMOs) have been developed under different principles [2]-[8]. Combining HDR capturing and tone mapping is beneficial even when an LDR display is used because they together produce higher quality images with much less over and under saturated areas compared to the traditional LDR capturing process. In addition, tone-mapping allows higher degrees of freedom for artists who during postproduction can decide the final effect/style of the resultant LDR image.

In a similar manner, the 3D display technology also aims at providing viewers with a more realistic visual experience by providing a sense of depth. An increasing number of theatres and households have been equipped with 3D display systems. More 3D content is being produced for satisfying the demand of the rapidly growing 3D consumer market. Ideally, content could be captured in a 3D HDR representation and viewed on a 3D HDR display, to achieve a more lifelike picture quality. However, existing 3D displays can support only 8-bit LDR content. In order for HDR content to be displayed on existing imaging systems, LDR signals need to be generated for each view of the 3D HDR pair. That is, tone-mapping needs to be applied to each view. The fact that tone-mapped images produce less under- and over-exposed areas will help the fuse of the 3D depth, and the superior picture quality of HDR tone-mapped content will also add value to the 3D representation.

By varying the tone mapping method and the related parameters, LDR images with different visual effect can be achieved. In particular, the tone mapping process can adjust the brightness and levels of 'details' in the output images. Many tone mapping methods have been developed trying to preserve local contrast, which gives images with a greater levels of details, i.e., images that look sharper and have stronger texture. Brightness and sharpness/texture have also been noted by artists to affect the visual comfort and quality of 3D content [10]. In summary, the optimal tone mapping parameters may be different for 3D images than those for 2D images.

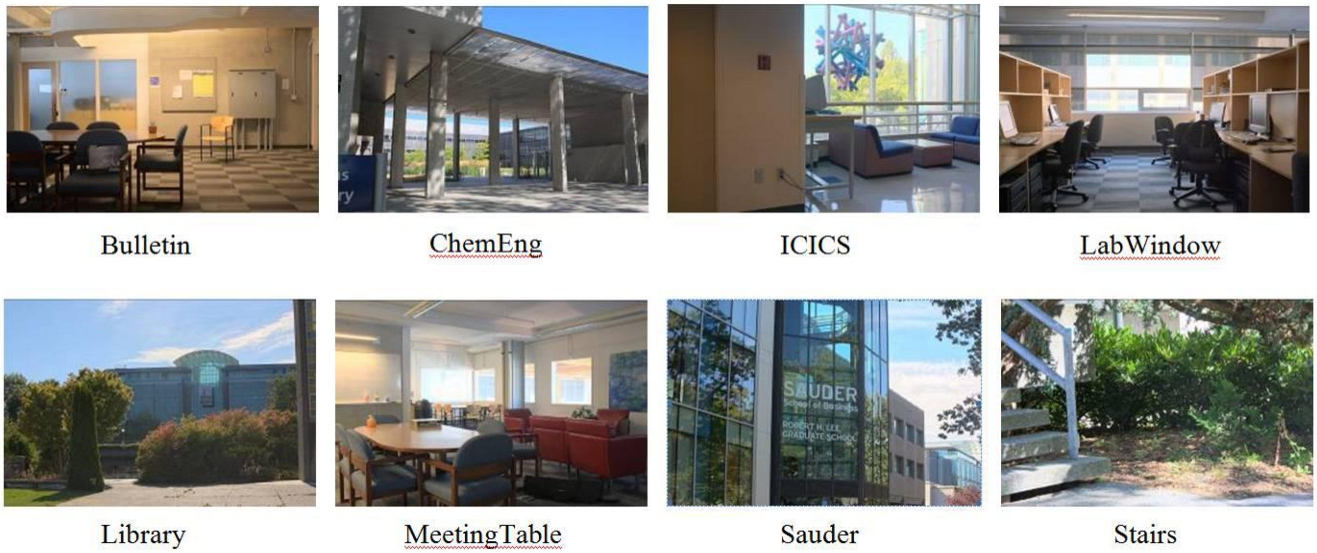


Fig. 1. Scenes used in the subjective test. Since HDR content cannot be shown on the paper or most monitors, all the images displayed above are tone-mapped using the photographic TMO.

In this paper, we present a study on whether the preferred levels of brightness or details are different for tone-mapped 2D and 3D content. 3D content may need to be prepared in a different way from 2D content, in order to provide the best possible picture quality. Here, we address the problem of identifying the preferred brightness and detail levels for 2D and 3D images by performing a set of subjective experiments.

The rest of the paper is organized as follows: Section II describes the experimental setup. Results and discussion can be found in Section III. Section IV concludes the paper.

II. Experimental Setup

Our study focuses on how the preferred level of brightness and the preferred amount of details differ between 3D and 2D images. “Preferred” means the best depth impression for 3D images and the best overall picture quality for 2D images.

A. Image Preparation

Eight scenes, four indoors and four outdoors, were captured in 3D with multiple exposures, and then stereoscopic HDR images were generated by blending these exposures [10]. These scenes were selected to represent different scenarios, such as scenes containing light sources and scenes having only reflectance. Fig. 1 shows an LDR version of the eight scenes. In order to investigate the effect of brightness and the amount of details, 3D images at different levels of brightness and with different amount of details are generated from the eight HDR images respectively.

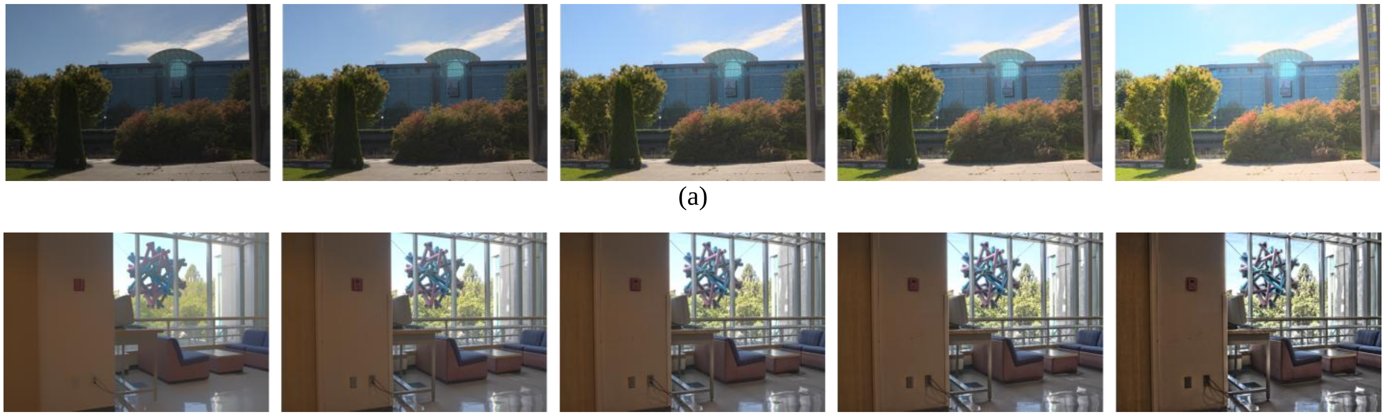
To vary the brightness levels in the LDR version with consistent effect, we chose to apply the popular photographic TMO since it provides a user parameter (key value α) for changing brightness. This tone-mapping operator simulates the dodging and burning techniques in photographic tone reproduction. For each scene (i.e., each HDR image), we altered the value of α such that i) 41 LDR images with different brightness levels were produced, from very dark to very bright, and ii) the difference in the overall image brightness between two consecutive levels is as constant as possible. Fig. 2(a) illustrates tone-mapped images at different brightness levels generated using the above approach.

To produce different amounts of details in the LDR version, we choose a popular TMO based on bilateral filtering [5]. This TMO filters an HDR image into a base (low-pass) layer as well as a detail layer and then combines the modified versions of them to yield its tone-mapped image. The original *bilateral filtering TMO* uses a fixed ratio between the base and the detail layers when binding them. In our modification, this ratio λ may be changed for adjusting the contribution of the detail layer against the base layer as can be seen in (1).

$$I_{LDR} = L_b + \lambda \cdot L_d \qquad (1)$$

where I_{LDR} denotes the resulting tone-mapped image. L_b and L_d represent the base layer and the detail layer, respectively. The effect of changing this ratio λ on the amount of details is demonstrated in Fig. 2(b).

For each of the scenes, four sets of tone mapped versions of the screen were created: i) 3D representation with various brightness levels, ii) 3D representation at various detail levels, iii) 2D representation at various brightness levels, and iv) 2D representation at various detail levels. For each version, we prepared 41 different levels. In total, there are 8 (scenes) x 4 (versions) x 41 (levels) = 1312 images in our system that have to be evaluated.



(a)

(b)

Fig. 2. Demonstration of images at different brightness and detail levels: (a) - the scene “Library” at different brightness levels; (b) the scene “ICICS” at different detail levels.

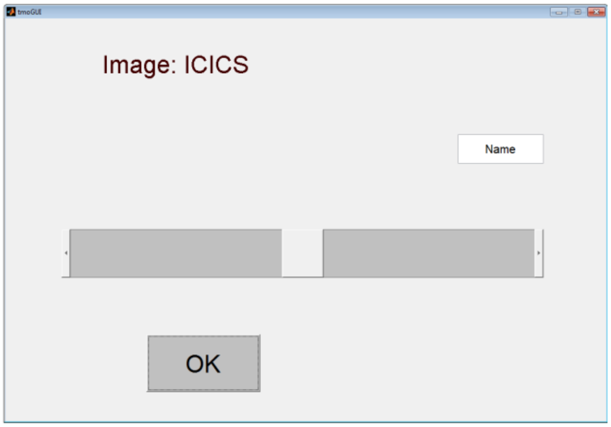


Fig. 3. Illustration of the graphic user interface for subjects to fully navigate the psychophysical experiment on their own pace and to select the images at the preferred brightness or detail levels.

B. *Testing enviorment and procedures*

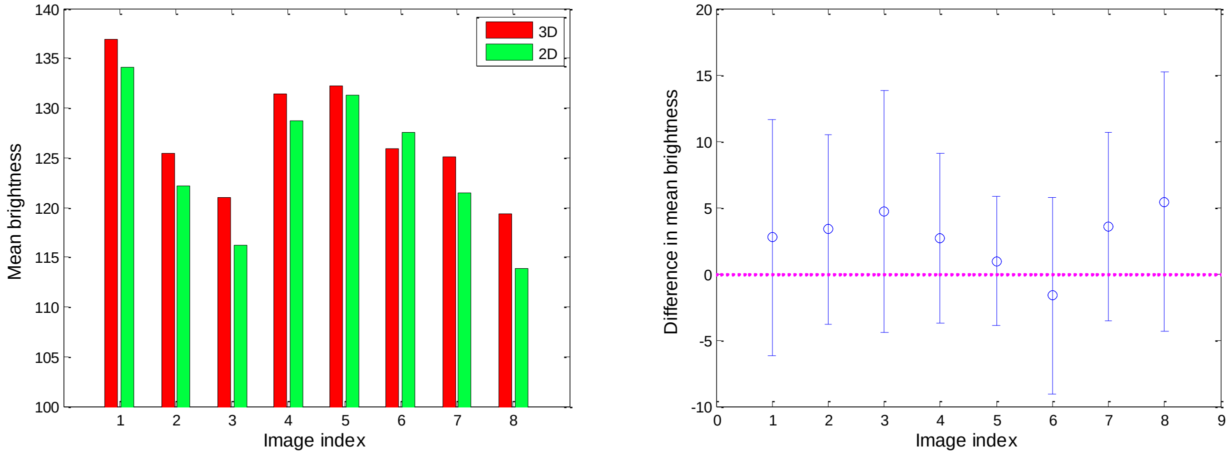
Eighteen subjects (18 – 38 years old) participated in our subjective experiment. All of them had normal or corrected vision with no/marginal experience in 3D technology. The display device used in our test was a 65” Full HD 3D display (©Panasonic, Plasma, TC-P65VT25), which uses active shutter glasses. The viewing conditions of our tests were set based on the ITU-R Recommendation BT.500-11 [13]. Viewers keep having their 3D glasses on when watching 2D videos. This guarantees that the brightness reduction is the same when watching 3D and 2D videos.

A secondary display, a 19” 2D screen, was placed closed to the viewer, where a graphic user-interface was created for subjects to select their preferred image (level) for each of the versions. Fig. 3 demonstrates the user interface. A slider bar can be found near the center of the interface. The slider consists of 41 values and each value corresponds to each brightness/detail level. The position of the slider is reset to the middle of the bar when a new scene is loaded. Once a particular position/value is chosen, an image at the selected level (brightness or amount of detail) is updated and shown immediately. As the slider is moved from left to right, an image will become brighter in the case of evaluating brightness versions and will have more details in the case of evaluating detail versions. An “OK” button sits below the slider and is used by subjects to confirm their selection. The selected value on the slider will then be recorded.

Before the test started, participants were provided with a training section which ensures they were comfortable in using the scoring interface and had their eyes adapted to the viewing conditions. In the test, they were asked to move the slider for selecting the preferred images for each version of each scene. As stated at the beginning of this section, “preferred” means the best depth impression for 3D images and the best overall picture quality for 2D images.

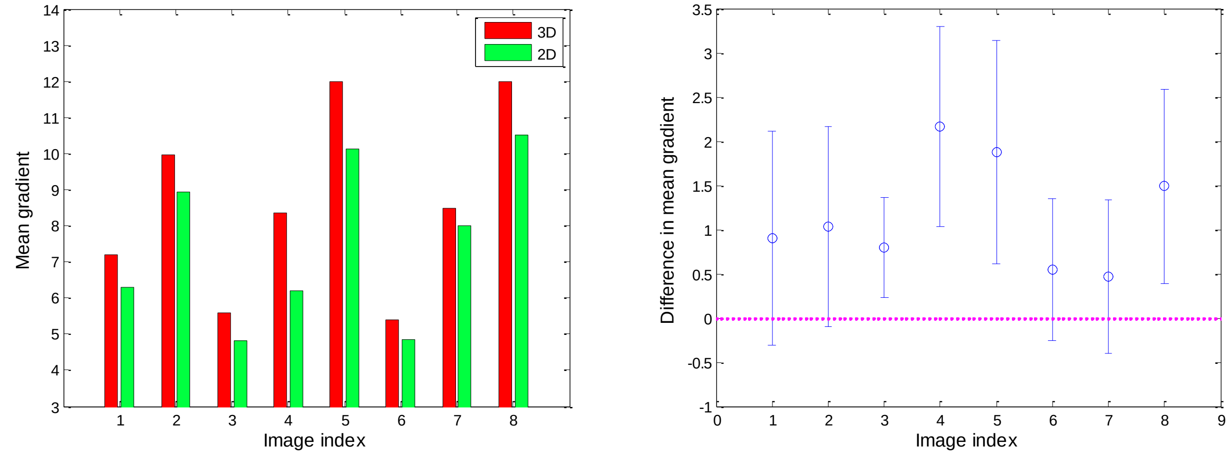
During the test, subjects know in advance the kind of image (3D or 2D) to be viewed. The procedure for each subject would proceed as follows. First, a 3D version of one scene would be shown to the subject, and he/she would adjust the slider on the GUI to select the level of brightness they thought gave the best 3D quality. Then the subject would press the OK bottom on the GUI, after which 5 seconds of grey is shown to allow the subject to rest his/her eyes. Then, the next scene would be displayed and the subject would repeat the process. After the user has done this for all 8 scenes, the slider is changed to control the amount of detail in the images, while still displaying 3D images. Again the user would adjust the slider to select what is thought to be the best amount of details for each scene. After the subject has selected the preferred amount of details for all eight scenes, the entire procedure is repeated in 2D viewing mode. That is, the user first selects what he/she considers the optimal amount of brightness for each scene, and then selects the optimal amount of details for each scene viewed in 2D.

The time of each test was completely controlled by the participants and lasted until they reached their final decision.



(a) (b)

Fig. 4. Comparison between 2D and 3D images in terms of the preferred brightness level: (a) – mean image average, (b) subtraction of the mean image average of 2D from that of 3D images. The horizontal axis denotes the image index, and 1 – 8 correspond to 'ICICS', 'MeetingTable', 'LabWindow', 'Bulletin', 'Stairs', 'SauderBuilding', 'LibraryTree', 'ChemEngEntrance', respectively. The vertical axis denotes pixel value in (a) and the difference of pixel values in (b).



(a) (b)

Fig. 5. Comparison between 2D and 3D images in terms of the preferred detail level: (a) – mean image detail, (b) subtraction of the mean image detail of 2D from that of 3D images. The horizontal axis denotes the image index, and the image order is the same as Fig. 4. The vertical axis denotes gradient value in (a) and the difference of gradient values in (b).

Although the time varies depending on the subject, no subject took more than 25 minutes to complete a test.

III. RESULTS AND DISCUSSION

After recording the results from the psychophysical tests, we performed statistical analysis on the collected data. First, we tested for outliers based on ITU-R Recommendation BT.500-11 [13] and in our case no outlier was identified, so the data for all subjects were used.

In order to quantify the brightness and the detail levels for the different tone mapped images, we use mean image brightness B_{img} and mean image gradient D_{img}. These are defined as:

$$B_{img} = \frac{1}{m \cdot n}\sum_{i=1}^{m}\sum_{j=1}^{n} I(i,j)$$

$$\begin{aligned} D_{img} &= \frac{1}{m \cdot n}\sum_{i=1}^{m}\sum_{j=1}^{n} \left| I(i,j) - I(i+1,j) \right| \\ &\quad + \frac{1}{m \cdot n}\sum_{i=1}^{m}\sum_{j=1}^{n} \left| I(i,j) - I(i,j+1) \right| \end{aligned} \tag{2}$$

where $I(i, j)$ denotes the pixel value at the location of (i, j) in an image I, and m and n are the dimensions of the image I.

The plot in Fig. 4(a) shows the mean image brightness (B_{img}) of the preferred 3D and the preferred 2D representations for each of the eight scenes (horizontal axis). The height of each bar is obtained by first calculating the mean image brightness of the preferred level from each of the 18 subjects and then averaging these 18 values. A general observation is that the mean brightness of the preferred 3D images is slightly higher than that of the preferred 2D counterparts for all the scenes except scene six ("Sauder"). All the mean brightness values fall in the interval between 115 and 135.

In order to gain a better understanding of the differences between 3D and 2D viewing, for each subject we calculated the difference between the average brightness of their preferred images in 3D and 2D modes as:

$$\Delta_B = B_{pref,3D} - B_{pref,2D} \tag{3}$$

where $B_{pref,3D}$ is the average brightness of the image the user selected with the slider in 3D mode (their 'preferred' image), and $B_{pre,f2D}$ is the corresponding average brightness of the preferred image they selected in 2D mode. A positive value for Δ_B indicates the subject prefers a brighter image when viewing in 3D compared to viewing in 2D. Fig. 4(b) shows the difference averaged over the eighteen participants, and also the 95% confidence interval for the differences. It is seen that although the average values of seven out of the eight scenes are above zero, all of the 95% confidence bars cross the zero axis. Therefore, there is no statistically significant difference in the preferred brightness level between a 3D image and its 2D counterpart.

The results of the preferred mean image gradient of the 3D and the 2D representations are shown in Fig. 5(a). Each point is the average of the preferred gradient selected by each of the subjects for each of the scenes. For all scenes people selected a higher level of details in 3D viewing mode compared to 2D viewing. Similar to the brightness case, we compute the difference in mean gradient between the preferred 3D and the preferred 2D images for every single subject and every scene. Fig. 5(b) shows such difference averaged over 18 subjects for each of the scenes, and the 95% confidence intervals are also provided. Positive values on the vertical axis mean that people favor more details in 3D images than their 2D counterparts. It is observed from the plot that all the average difference is above zero. Moreover, for many of the scenes, the 95% confidence interval is either completely (scenes 3, 4, 5 and 6) or majorly (scenes 1, 2, 7 and 8) above zero. A reliable conclusion can thus be drawn that people prefer a higher level of details (i.e., a sharper image with more texture) when viewing in 3D than when viewing in 2D. Since our test images cover a wide range of content (four indoor and four outdoor scenes), this conclusion will hold regardless of the nature of the content. This could be explained by the conclusion in Cormack et al. [12], that the 3D effect can be improved when the contrast is near the visibility threshold. By adjusting the tone-mapping process to produce more detail, the 3D effect in image regions where the contrast is near the threshold may be improved.

IV. Conclusions

In this paper, we studied how different are 3D and 2D HDR tone-mapped images in terms of i) the preferred brightness levels and ii) the preferred detail levels. Images at a large range of different brightness and detail levels were generated from high dynamic range capturing followed by the tone-mapping process. With such a great variety of images, we conducted an intensive subjective experiment that allows participants to select 3D and 2D images with their preference on brightness and details, respectively. Our results show that while people selected slightly brighter images in 3D viewing compared to 2D, the difference is not statistically significant. However, compared to 2D images, the subjects consistently preferred having a greater amount of details when viewing in 3D.

References

[1] R. Mantiuk, G. Krawczyk, K. Myszkowski, and H.-P. Seidel, "Perception-motivated high dynamic range video encoding," *ACM Transactions on Graphics*, vol. 23, no. 3, pp. 730–738, 2004.

[2] E. Reinhard, M. Stark, P. Shirley, and J. Ferwerda, "Photographic tone reproduction for digital images," *ACM Transactions on Graphics (Proc. SIGGRAPH)*, vol. 21, no. 3, pp. 267–276, 2002.

[3] R. Mantiuk, S. Daly, and L. Kerofsky, "Display Adaptive Tone Mapping," *ACM Transactions on Graphics (Proc. SIGGRAPH)*, vol. 27, no. 3, pp. 68–68, 2008.

[4] F. Drago, K. Myszkowski, T. Annen, and N. Chiba, "Adaptive Logarithmic Mapping for Displaying High Contrast Scenes," *Computer Graphics Forum (Proc. of Eurographics)*, vol. 22, no. 3, pp. 419–426, 2003.

[5] F. Durand, and J. Dorsey, "Fast Bilateral Filtering for the Display of High-Dynamic-Range Images," *ACM Transactions on Graphics (Proc. SIGGRAPH)*, vol. 21, no. 3, pp. 257–266, 2002.

[6] Z. Mai, H. Mansour, R. Mantiuk, P. Nasiopoulos, R. Ward, W. Heidrich, " Optimizing a Tone Curve for Backward-Compatible High Dynamic Range Image/Video Compression ", *IEEE Transactions on Image Processing (TIP)*, vol. 20, no. 6, pp. 1558-1571, 2011

[7] R. Fattal, D. Lischinski, and M. Werman, "Gradient Domain High Dynamic Range Compression," *ACM Transactions on Graphics (Proc. SIGGRAPH)*, vol. 21, no. 3, pp. 249–256, 2002.

[8] E. Reinhard, and K. Devlin, "Dynamic Range Reduction Inspired by Photoreceptor Physiology," *IEEE Transactions on Visualization and Computer Graphics*, vol. 11, no. 1, pp. 13-24, 2005.

[9] B. Mendiburu, "3D Movie Making – Stereoscopic Digital Cinema from Script to Screen." Elsevier, 2008.

[10] P. E. Debevec and J. Malik, "Recovering High Dynamic Range Radiance Maps from Photographs," Proceedings of the 24th annual conference on Computer graphics and interactive techniques (*Proc.* SIGGRAPH '97), pp. 369-378, 1997.

[11] ITU-R, "Methodology for the subjective assessment of the quality of television pictures," ITU-R, Tech. Rep. BT.500-11, 2002.

[12] L.K. Cormack, S.B. Stevenson, and C.M. Schor, "Interocular correlation, luminance contrast and cyclopean processing," *Vision Research*, vol.31, no.12, pp. 2195–2207, 1991.

A$'$ARAF: An Asynchronous Router Architecture using Four-Phase Bundled Handshake Protocol

Syed Rameez Naqvi
Department of Computer Engineering
Vienna University of Technology
Vienna, Austria
rnaqvi@ecs.tuwien.ac.at

***Abstract*—The problem of the global clock distribution on multicore systems has rightly abandoned the use of synchronous interconnection networks, but increased the design complexity of the clocked islands. In this work, the design of a completely asynchronous router, with multiple arbitration paths, is presented in detail. The data flow between all the heterogeneous components is based on four-phase handshake protocol. We simulate two 2D mesh networks of different sizes, and propose an evaluation methodology for each of the two important properties: deadlock freedom and reachability. Simulation results show that our networks satisfy both of these properties even against a reasonable flit-injection rate. The inter-router communication is also based on the same single rail, return to zero protocol.**

***Keywords*-Asynchronous; Networks-on-Chip; Router architecture; Four-phase bundled; Single rail; Return to zero (RZ);**

I. Introduction

Until recently, the multicore systems made use of the standard bus architecture to allow communication between the cores. Although such systems did not pose a great threat to performance, it is expected that in near future buses are going to become a bottleneck with a tremendous increase in the number of processing cores integrated on a single chip. The Networks-on-Chip (NoC) approach proves to be an efficient solution to communication problems, reducing wiring complexity, and thus, power consumption. However, global clock distribution in NoC, in real time bounds, may also not be possible with the billion transistor era approaching fast [1]. The Asynchronous Networks-on-Chip (ANoC) design, a special case of Globally Asynchronous Locally Synchronous (GALS) systems, has gained fame in the recent years. It not only eliminates the need of a global clock signal, but promises to provide power efficiency and higher modularity compared to its synchronous counter parts [2].

The primary job of the NoC, whether clocked or not, is to provide a communication infrastructure between numerous cores (may be processors or IP modules). Normally these cores are clocked, as a result of which, there has to be an interface between the core and the ANoC. Either this interface is made a part of the router, resulting in clocked architecture, or kept separately as a Sync-Async converter between the core and the network. While AEthereal [3] is one of the most famous NoCs that adopt the former approach, Sheibanyrad and Greiner [4] have proposed two efficient Sync-Async converters for an ANoC. In this work we assume the latter approach which seems more attractive, as it significantly reduces the design complexity of the router by eliminating the need of synchronizer circuits, and thus minimizes power consumption at the routing level.

Zeferino et al. [5] have proposed SOCIN NoC, which uses a handshake signal based flow control, where a VALID signal is sent whenever a new flit is transmitted. In this work we present the design of an asynchronous router architecture, which uses the same single rail, return to zero (RZ) handshake protocol. A 2D mesh network, with 2x2 and 4x4 structures, has been simulated. Our simulation results show deadlock freedom even with all interconnects being exercised simultaneously. Furthermore, the novelty of the design, besides being completely asynchronous, rests with the deployment of multiple arbiters per tile. While a single arbiter in a switch would allow only one input port to access the output port at a time, our scheme allows multiple pairs of tiles to communicate independently. For instance, data at the east input port being switched to the west output port, does not hinder switching between the north and south ports pair; as well as data at the west input port can proceed to the east output port simultaneously. This tradeoff between the wiring complexity and the throughput is made in order to compensate the slow nature of four-phase handshake protocol [6], which requires two transitions from each of the sender (producer) and the receiver (consumer), alternately for a successful data transfer.

The rest of the paper is organized as follows. The step-by-step design of the router and network is described in Section II. Section III presents the evaluation methodology. The simulation results are given in Section IV. Section V concludes the paper with future work.

II. Router Architecture and Network Design

The ultimate aim of our work is to introduce a novel fault-tolerance/self-repairing mechanism in the ANoC (not the topic of this paper though). Hence, we do not emphasize much on the performance aspects related to the throughput of our network. In fact, we intend to keep the network as simple as possible, so that the issues that the complexity of the design adjoins may be avoided; such as, livelocks which may occur even with the state-of-the-art adaptive routing in place. Having multiple arbiters on our tile also adds to the simplicity of the design by reducing the number of candidates for arbitration, eventually eliminating the possibility of a deadlock due to traffic congestion: the primary focus of this work.

To start-off with, we build our interconnection network as a 2D-Mesh of sizes 2x2 and 4x4, which might be scaled later if required. Wormhole switching [7] alongside XY-Routing [8] has been adopted which allows us to: 1) build completely independent flow-paths for the header-, body- and tail- flits, saving power, 2) reduce the design complexity of the Input Controller, and 3) guarantee deadlock freedom.

Table I presents the packet- size and format for all the packet types. The two little-endian, most significant bits (MSBs) indicate the type of flit. “11” represents a header flit, “10” is reserved for a tail flit, and “0x” for the body flits. Since we keep an explicit specifier for the tail flit, we do not need an additional adder to count the number of flits that have arrived. In addition, this gives a flexibility to transmit and receive packets of variable sizes. The addressing scheme that we have adopted is influenced by MANGO [9], in which each pair of bits, starting from 31 down to 0, in the header flit, indicates the next hop; and thus each pair needs to be removed/rotated on every hop so that the next pair can indicate the next subsequent hop. For instance in Table 1(A), “00” at positions 31:30 tells the switch that the incoming data has to be directed to east. Therefore, “00” after being written into the destination latch, must be rotated, thus bringing “10” at its positions. Subsequently, at the next hop, the switch will direct the incoming data to north corresponding to the pair “10”. In the same manner, “01” corresponds to the west output port, and “11” to the south. If the next hop is identical to the input port, then the packet is assumed to be directed to the core; therefore, backtracking [10] is not supported.

Fig. 1 presents the block diagram of our fully acknowledged asynchronous router. Primarily, the overall design is divided into two parts: Input Handler (IH) and Output Generator (OG). The former one is responsible for capturing the input data, reserving the appropriate output port arbiter, guiding the incoming data to the output unit, and providing the output unit with associated control signals. On the other hand, the latter guides the incoming data to the appropriate output port based on the control signals it received from the IH. In the following we describe each of these modules briefly.

Table I
PACKET FORMAT AND SIZE: (A) HEADER FLIT, (B) TAIL FLIT, (C) BODY FLIT

1	1	1	0	...	0	1	0	0	1	1
Bit-0	1	2	3	...	28	29	30	31	32	33
Dest-16		Dest-15		...	Dest-2		Dest-1		Flit-Type	
(A)										
1	0	0	0	...	0	0	1	0	0	1
Bit-0	1	2	3	...	28	29	30	31	32	33
Payload									Flit-Type	
(B)										
0	0	0	0	...	0	0	1	0	x	0
Bit-0	1	2	3	...	28	29	30	31	32	33
Payload									Flit-Type	
(C)										

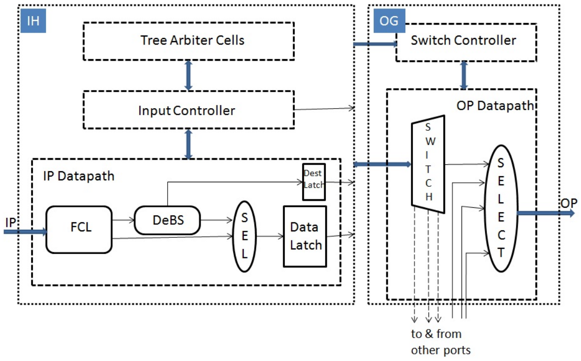


Figure 1. Block Diagram of the Async Router

A. Flit Categorization Logic (FCL)

As explained above, the two MSBs indicate the type of the incoming flit. Depending upon the type, the flit has to be directed to the appropriate unit. The Flit Categorization Logic (FCL) is responsible to: i) identify the flit, ii) report its type to the Input Controller (ICON), and iii) guide it either to DeBS in case of a header flit, or to the Select Module otherwise.

B. Destination Bits Shifter (DeBS)

The DeBS module performs two functions: 1) rotates the bits 31:30 of the header flit to the least significant bit (LSB) places, so that the new pair at places 31:30 indicates the output port of the succeeding node, 2) forwards the rotated bits to the destination latch, fig 2. The latched data then guides the OG in switching the header and the following flits to their appropriate destination ports. In case the incoming flit is not a header flit, DeBS becomes silent (no power consumption, except for the leakage current).

C. Select Module

The Select Module is nothing but a multiplexer without an explicit selection line, whose operation is quite simple. It arbitrates the active input to the output. Therefore, the data-valid control signal associated with every incoming flit

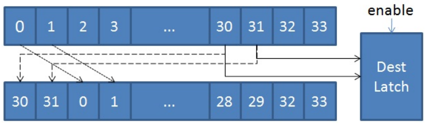


Figure 2. DeBS Module Operation Concept

acts as the selection line of the multiplexer itself. Since the incoming flit can either be a header, body or a tail flit, there is no possibility at the Select module to have a situation where two or more inputs could arrive at the same time. Hence it does not need to comprise any complicated arbitration logic at all.

D. Input CONtroller (ICON)

The Input Controller (ICON) has been modeled as an STG in Workcraft [11] and synthesized using Petrify [12]. Two important functionalities that ICON is made to perform are: 1) on-demand reservation of the mutual-exclusion (MUTEX) element associated with each output port, 2) generating the latch-enable signals both for the destination and the data latches. The destination latch needs to be enabled only with the header flit, while the data latch needs to be enabled with every incoming flit. In fig. 3 we have presented its STG along with explanation of the variables used in Table II. Realizing the level of difficulty in understanding the STG, in the following we briefly describe the operation of the ICON.

At the arrival of the header flit, a request is raised and sent to the ICON, so as indicated by “rh+” in the STG of fig 3. The arrival of the header flit must be followed by the reservation of the MUTEX. As a result, a request is sent to the arbiter associated with the target output port. This is indicated by “rm+”. Once the grant from the arbiter is received “gm+”, the destination bits and the data must be latched. On confirmation of the data being latched, an acknowledgement “ah” and a request “ro” are respectively sent to the previous and the next nodes simultaneously. The destination latch must not be enabled for the body and tail flits, since header flit is the only one to contain the routing information. The body and tail flits proceed similarly except for the release of the MUTEX with the tail flit. This is done by lowering the request to the MUTEX “rm-”, on receiving an acknowledgement from the next node.

Table II
VARIABLES USED IN FIG. 3

Signal	Input/Output	Explanation
rh	Input	request signal from the header flit
rb	Input	request signal from the body flit
rl	Input	request signal from the tail flit
gm	Input	grant/ack signal from the MUTEX
dest_a	Input	ack signal from the destination latch
data_a	Input	ack signal from the data latch
ao	Input	ack signal from the output side demux
rm	Output	req signal to the MUTEX
dest_r	Output	req signal to the destination latch
data_r	Output	req signal to the data latch
ro	Output	req signal to the output side demux
ah	Output	ack signal to the header flit
ab	Output	ack signal to the body flit
al	Output	ack signal to the last/tail flit

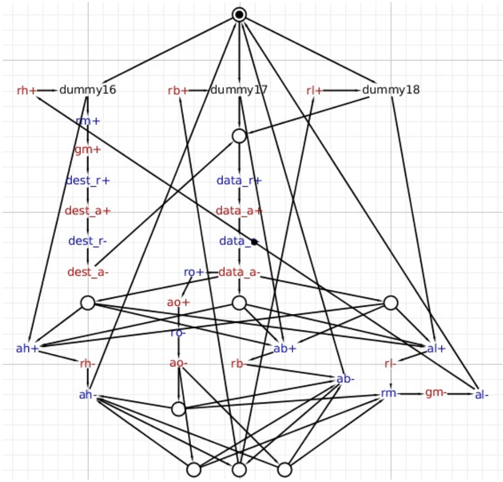


Figure 3. STG of the Input Controller

E. Arbiter Circuit

We have adopted the conventionally used two-input tree-arbiter-cell (TAC) [13], fig. 4, to allow sharing among all the input ports contending for the same output port. Whichever input port requests (C1req or C2req) first, wins the arbitration (C1gr or C2gr) to access the desired output port. A four-input arbitration circuit can be built by making use of two TACs and a MUTEX as shown in fig. 5. Please note that “r1/r2”, “g1/g2” signals in fig. 4 correspond to respective request and grant signals to and from each MUTEX element. An input port keeps hold of the MUTEX until the message is completely transmitted. Although the structure of the tree does not guarantee round robin arbitration, the latter can be achieved by simply replacing a module with priority arbiter if felt necessary. The tree arbiter however, utilizes fewer resources.

The XY-routing algorithm naturally limits the number of permissible switching turns, fig. 6. It is forbidden to switch the data from the north or the south input port to east and/or west output ports. This allows us to keep a different sized arbiter for each output port. For instance, in case of east output port, the number of contending input ports is only two, i.e. west and the core. On the other hand, the number of contending input ports for north and south output ports would be four. However, keeping the generality of the router alive, it is possible to have the four input arbiter module on a reconfigurable partition of the FPGA which allows dynamic

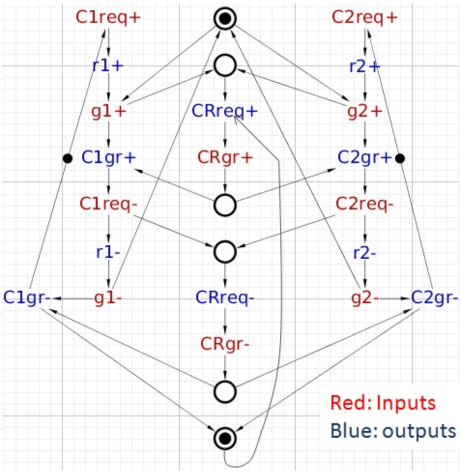


Figure 4. STG of a Two-Input Tree Arbiter Cell

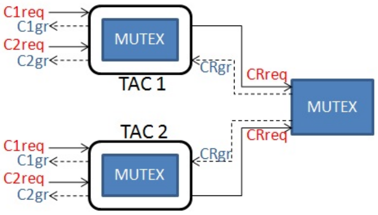


Figure 5. A Four-input Arbiter made from two TACs

exchange of the arbiter circuits as per the routing algorithm (not covered in this paper).

In fig. 7 we have presented the complete circuit diagram of our router with all the important control signals. Some of the wires (req/ack to other input ports) have been deliberately removed from the figure keeping in view the limitation of space. The input demultiplexer together with the three C-gates make up the unit FCL (please refer to Sec. II-A). A C-gate [14] is the most fundamental element for any asynchronous circuit. It works as an AND-gate if the two inputs share the same logic state, and maintains its previous state otherwise. In our router, a number of C-gates are used to ensure the speed independence (SI) [6] property of the circuit; for instance a C-gate placed between DeBS and ICON (please refer to Sec. II-B and Sec. II-D respectively) forces the demultiplexer to keep its output data stable until they are acknowledged by both of the receivers. The output demultiplexer, which is controlled by the destination latch, acts as a switch, and together with the four select modules make up the OG module shown in fig. 1. The four select modules and the arbiters are shared between all the input ports, whereas, the rest of the circuit needs to be replicated for each input port to allow parallel execution.

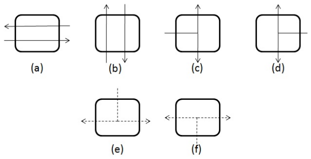


Figure 6. Permissible (a, b, c, d) and Forbidden (e, f) Switching

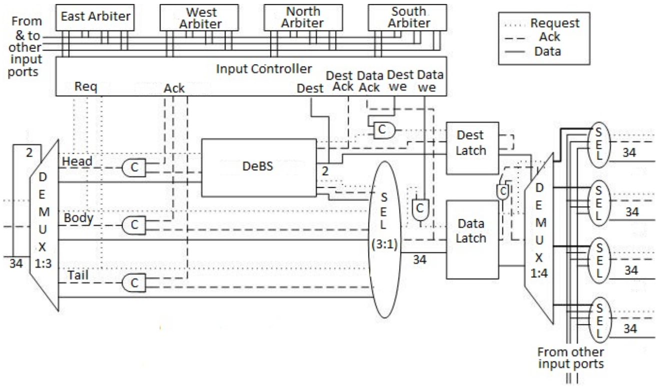


Figure 7. Complete Async Router Architecture

III. EVALUATION METHODOLOGY

Any NoC is expected to guarantee three things: i) deadlock freedom, ii) livelock freedom, iii) reachability to every other node. While any deterministic routing protocol would naturally handle livelocks, we define a methodology to test A′ARAF NoC on both of the other dimensions. The latter is rigorously tested for a 4x4 2D mesh, whereas the former is done for 2x2.

A. *Deadlock Freedom*

Although a network of size 2x2 does not seem to be an impressive test case, the analysis, however, can be very thorough. The point is to ensure that all the 16 nets (34 bit each) are exercised simultaneously, loading the network with maximum possible traffic. Our methodology is adopted from [15] in which Cota et al. have tested the interconnects for possible faults. Fig. 8, reprinted from the same paper nicely describes the methodology, where “core 0” communicates with “core 3” and vice versa, and “core 1” does with “core 2” and vice versa. We maintain two considerations: 1) The communication pattern remains XY-routing, for example “core 0” sends its packets to “core 3” via “router00”, “router01”, and “router11”, and “core 3” sends its packets to “core 0” via “router11”, “router10”, and “router00”. Please note that each type of an arrow represents communication only in one direction, and maintains XY-routing pattern. For instance, bold arrows represent the packet transfer from “core 0” to “core 3”, and so on for the rest. 2) The order of the flits must be header flit to tail flit.

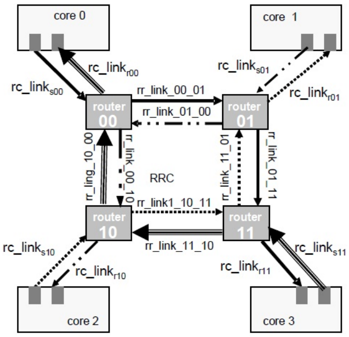


Figure 8. Deadlock Freedom Evaluation Methodology, reprinted from [15]

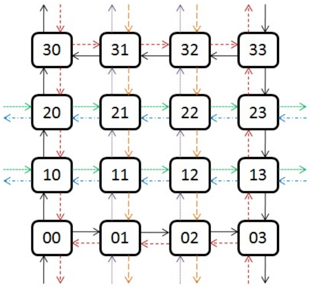


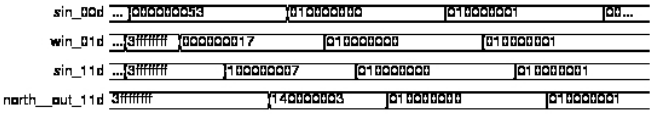
Figure 9. Reachability Evaluation Methodology

B. Reachability

In order to verify that all nodes can access each other, we propose to forward two test packets header, body, tail from each corner node to the nodes at the far ends. For example, “router00” is supposed to forward one packet each to routers “03” and “30” along x+ and y+ axes respectively. In the same manner, “router33” forwards packets to routers “30” and “03” along x- and y- axes respectively. For the two sandwiched rows and columns, each node forwards one packet to the far end. For example, nodes “01” and “31” forward a packet to each other simultaneously, and so on for the remaining pairs. The overall scheme is presented in fig. 9.

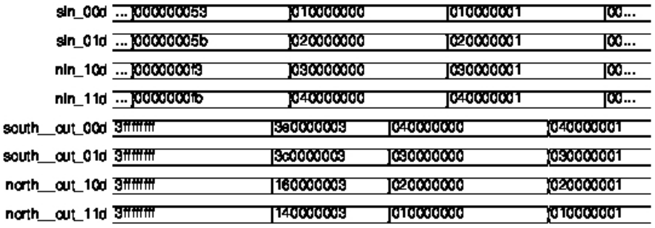
IV. SIMULATION RESULTS

According to the methodology described above, we perform four different simulations to test our NoC. All of the simulations are done in Modelsim using a test bench and macro files. We inject 150 packets, one after the other without halting, on each input of a 2x2 network to verify deadlock freedom. All of the packets reach their respective destinations, and are received correctly. Fig. 10 shows the complete propagation of a packet from the south input port of “router00” to the north output port of “router33”. Please note that the header flit changes on every hop, since the two destination bits keep rotating on every node. For example, in the header flit “000000053” (hexadecimal representation), equivalent to “000...01010011” (34 bits), the last two bits “11” indicate that it is a header flit. The next two bits “00” have to be rotated after the first hop, bringing “10” at their places. So the flit changes to “000...00010111” which is equivalent to “000000017”. Similarly, the flit keeps changing on the rest of the hops. However, the body and tail flits remain the same until they reach the destination. On the other hand, fig. 11 shows a snapshot of the simulation for a few of the initial packets during the deadlock freedom test simulation. It can be seen how packets are correctly transferred between the routers connected on diagonals.



```
sin_00d        ... 000000053     010000000     010000001     00...
win_01d        ... 3fffffff 000000017     010000000     010000001
sin_11d        ... 3fffffff      100000007     010000000     010000001
north__out_11d 3fffffff         140000003     010000000     010000001
```

Figure 10. Propagation Path of a Packet



```
sin_00d        ... 000000053     010000000     010000001     00...
sin_01d        ... 00000005b     020000000     020000001     00...
nin_10d        ... 0000000f3     030000000     030000001     00...
nin_11d        ... 0000000fb     040000000     040000001     00...
south__out_00d 3fffffff     3e0000003   040000000     040000001
south__out_01d 3fffffff     3c0000003   030000000     030000001
north__out_10d 3fffffff     160000003   020000000     020000001
north__out_11d 3fffffff     140000003   010000000     010000001
```

Figure 11. Deadlock Freedom Verification

For the reachability test, once again we inject 150 packets (3 flits each) on every input. This time however, the network is 4x4 2D mesh. Two different simulations are performed. Fig. 12 shows the case where we follow the methodology described in the last section. All packets correctly reach their respective outputs. The final simulation, fig. 13, shows the correct working of the arbitration circuit. We deliberately force all the nodes to forward packets to one destination, north output of “router33” (“north__out_33d”). Once again, each node transmits 150 packets to the same target node. The arbiters give access to all the paths one at a time, and lead to correct and complete reception of all of them.

V. CONCLUSION AND FUTURE WORK

In this paper we have presented the design and implementation of A′ARAF, a router for asynchronous NoCs. Our async router supports wormhole switching, and it has been made generic to support any deterministic and adaptive routing algorithm. We have verified two important properties: deadlock freedom and reachability for XY-routing, by heavily loading the network of two different sizes. Simulation results have been presented and discussed in detail.

Although all the deterministic routing algorithms guarantee deadlock freedom, their major drawback is their inability

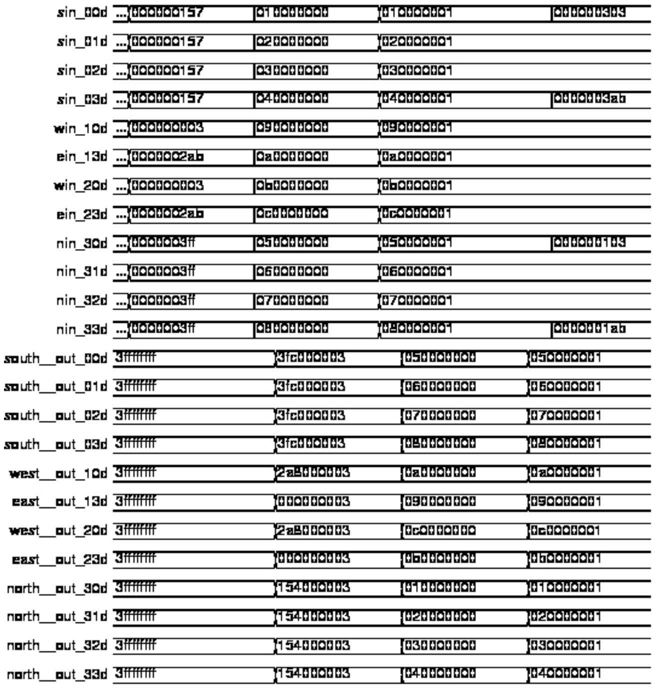


Figure 12. Reachability Verification

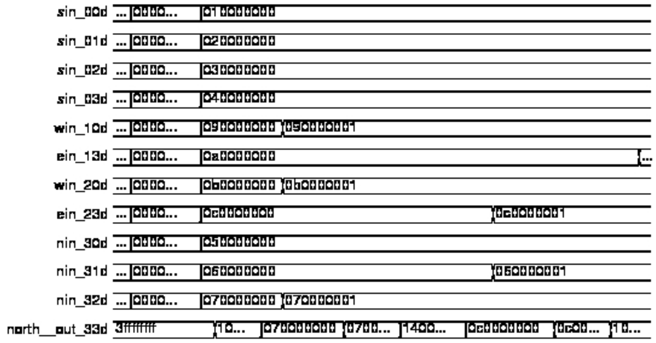


Figure 13. All-to-one Arbitration

to tolerate faults. XY-routing for instance, would immediately result in a complete deadlock once a node fails to forward a packet in the desired direction, or the channel itself becomes permanently fault due to electromigration effect. In future, we aim to address both transient and permanent faults within our routers architecture and the interconnects as well, keeping in view the outstanding problems with the state-of-the-art fault-tolerant NoC designs.

References

[1] A. Agarwal, C. Iskander, and R. Shankar, "Survey of NoC Architectures and Contributions," *Engineering, Computing and Architecture*, vol. 3, no. 1, 2009.

[2] Y. Shi, S. B. Furber, J. Garside, and L. A. Plana, "Fault tolerant delay insensitive inter-chip communication," in *Proc. 15th IEEE Symp. on Asynchronous Circuits and Systems (async 2009)*, ser. ASYNC '09. Washington, DC, USA: IEEE Computer Society, 2009, pp. 77–84.

[3] K. Goossens, J. Dielissen, and A. Radulescu, "Aethereal network on chip: concepts, architectures, and implementations," *Design Test of Computers, IEEE*, vol. 22, no. 5, pp. 414 – 421, sept.-oct. 2005.

[4] A. Sheibanyrad and A. Greiner, "Two efficient synchronous <–> asynchronous converters well-suited for networks-on-chip in gals architectures," *Integr. VLSI J.*, vol. 41, no. 1, pp. 17–26, Jan. 2008.

[5] C. A. Zeferino and A. A. Susin, "SoCIN: A Parametric and Scalable Network-on-Chip," in *Proc. 16th Symp. on Integrated Circuits and Systems Design*, 2003, pp. 169–175.

[6] J. Sparso and S. B. Furber, *Principles of Asynchronous Circuit Design: A Systems Perspective*. Springer, 2001.

[7] K. M. Al-Tawil, M. Abd-El-Barr, and F. Ashraf, "A Survey and Comparison of Wormhole Routing Techniques in Mesh Networks," *IEEE Network*, vol. 11, pp. 38–45, 1997.

[8] C. Neeb, M. Thul, and N. Andwehn, "Network On-Chip-Centric Approach to Interleaving in High Throughput Channel Decoders," in *Proc. IEEE Int. Symp. on Circuits and Systems*, 2005, pp. 1766–1769.

[9] T. Bjerregaard and J. Sparso, "A router architecture for connection-oriented service guarantees in the mango clockless network-on-chip," in *Proc. the conf. on Design, Automation and Test in Europe - Vol. 2*, ser. DATE '05. Washington, DC, USA: IEEE Computer Society, 2005, pp. 1226–1231.

[10] M. Koibuchi, H. Matsutani, H. Amano, and T. Mark Pinkston, "A lightweight fault-tolerant mechanism for network-on-chip," in *Networks-on-Chip, 2008. NoCS 2008. Second ACM/IEEE Int. Symp. on*, april 2008, pp. 13 –22.

[11] I. Poliakov, V. Khomenko, and A. Yakovlev, "Workcraft — A Framework for Interpreted Graph Models," in *Proc. 30th Int. Conf. on Applications and Theory of Petri Nets*, ser. PETRI NETS '09. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 333–342.

[12] J. Cortadella, M. Kishinevsky, A. Kondratyev, L. Lavagno, and A. Yakovlev, "Petrify: A Tool for Manipulating Concurrent Specifications and Synthesis of Asynchronous Controllers," 1996. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.17.8484

[13] D. L. Dill, "Trace theory for automatic hierarchical verification of speed-independent circuits," in *Proc. 5th MIT conf. on Advanced research in VLSI*. Cambridge, MA, USA: MIT Press, 1988, pp. 51–65.

[14] I. E. Sutherland, "Micropipelines," *Commun. ACM*, vol. 32, no. 6, pp. 720–738, Jun. 1989.

[15] E. Cota, F. Kastensmidt, M. Cassel, P. Meirelles, A. Amory, and M. Lubaszewski, "Redefining and testing interconnect faults in mesh nocs," in *Test Conf., 2007. ITC 2007. IEEE Int.*, oct. 2007, pp. 1 –10.

Unsupervised Information-Based Feature Selection for Speech Therapy Optimization by Data Mining Techniques

Mirela Danubianu, Valentin Popa
Faculty of Electrical Engineering and Computer Science
"Stefan cel Mare" University of Suceava
Suceava, Romania
e-mail: mdanub@eed.usv.ro, valentin@eed.usv.ro

Iolanda Tobolcea
Faculty of Psychology and Education Science
"Alexandru Ioan Cuza" University of Iasi
Iasi, Romania
e-mail: itobolcea@yahoo.com

***Abstract*— Data mining was proven to be an efficient way to find new and useful knowledge in data. Since data dimensionality has major implications on the performance of the algorithms used, one of the data pre-processing operations refers to reducing the number of features. One way to do that is feature selection based on their relevance and redundancy analysis. This paper presents a feature selection method which is applied on data provided by TERAPERS – a computer-based speech therapy system for Romanian children suffering of dyslalia.**

Keywords-data mining, feature selection, feature relevance, feature redundancy, speech disorder therapy

I. INTRODUCTION

The development of the informational society, which led to the increased use of the information technology in the most diverse areas of life, has allowed collecting and storing a huge amount of data. For this reason, over the last years we have witnessed the development of a research area designed to analyze large volumes of data in order to discover valuable and unexpected information, called Knowledge Discovery in Databases (KDD).

Defined as the process of identifying "valid, novel useful and understandable patterns from large data sets" [1], KDD can be viewed as a sequence of several steps. A symbolic representation of KDD process is presented in Figure1 [2].

It starts with a business analysis for determining the KDD goals. Then, there is a data understanding stage which aims to collect and describe data and to verify data quality, followed by the data preparation stage. The core of KDD process is the data mining stage. Data mining involves the analysis of large volumes of data using algorithms which, at acceptable efficiency of calculation, produce a particular enumeration of patterns from such data. As an exploration and analysis technique applied on large amounts of data in order to detect patterns or rules with a specific meaning, data mining may facilitate the discovery, from apparently unrelated data, of relationships that are likely to anticipate future problems or might solve the problems under study. It involves the choice of the appropriate data mining task, and, taking into account specific conditions, the choice and the implementation of the proper data mining algorithm. For the next stage, the mined models are evaluated against the goals defined in the first stage. The last stage of the process uses the knowledge discovered in order to simply generate a report or to deploy a repeatable data mining process.

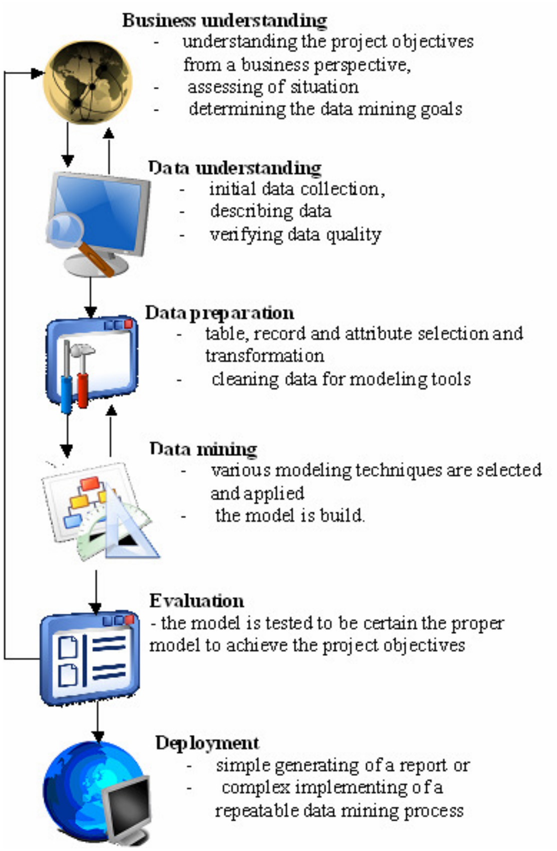

Business understanding
- understanding the project objectives from a business perspective,
- assessing of situation
- determining the data mining goals
Data understanding
- initial data collection,
- describing data
- verifying data quality
Data preparation
- table, record and attribute selection and transformation
- cleaning data for modeling tools
Data mining
- various modeling techniques are selected and applied
- the model is build.
Evaluation
- the model is tested to be certain the proper model to achieve the project objectives
Deployment
- simple generating of a report or
- complex implementing of a repeatable data mining process

Figure 1. Overview of KDD process

Although the stage of applying data mining algorithms is considered the key element of the KDD process, it must be noted that the results provided in this phase are strongly conditioned by several factors such as: data quality and their organization. It is known that in data collected from various primary sources one can find missing values, distortions misrecording or inadequate sampling. Therefore, it is very important to carefully examine the data before carrying out further analyses. Moreover, as one of the most critical operations in the KDD process, the proper preparation and transformation of the initial data set are essential in order to produce useful features for the selected data mining methods.

Data preparation is focused mainly on two issues: firstly, the data must be organized into a standard processing form by data mining algorithms, and, secondly, the data sets used must lead to the best performance and quality for the data mining stage.

II. DIMENSIONAL DATA REDUCTION FOR DATA MINING

Nowadays, huge amount of data are easily collected and stored. The dimensions of a data set are determined both by the number of cases and by the number of features considered for each case. Most data mining techniques may not be effective for high-dimensionality data, so the solution consists in data dimensionality reduction. To analyze the opportunity of data reduction we need to know what are the gains and losses, and therefore, we must compare computing times and predictive or descriptive accuracy for the model built for the whole dataset with those built for reduced data sets.

In order to reduce the number of cases, sampling or filtering can be used. By filtering, the cases that do not satisfy an imposed condition can be removed from the analyzed data set; by sampling, a subset of cases with a similar behavior to the whole population can be built. In the last case, a sampling error always occurs: it decreases with the increase in the size of subset, and it becomes zero when the complete data set is considered. The size of a suitable subset is calculated by taking into account the computation cost, the accuracy of the estimator and some data characteristics.

On the other side, feature reduction may be achieved either by feature selection or by feature composition. These methods should produce fewer features, so the algorithms can learn faster and even the accuracy of the built models could be improved [3].

Feature composition involves data transformations that can improve the results and performances of data mining operations. Feature selection aims to detect a subset of features having data mining performances comparable to the full set of features, but with significantly reduced computational costs. This is possible using either feature ranking or minimum set algorithms.

Feature ranking algorithms provide ranked lists of features, ordered according to specific evaluation criteria such as: data accuracy and consistency, information content or statistical dependencies between features. They provide information on the relevance of a feature compared to the relevance of other features, without showing the desirable minimum set of the features. On the other hand, minimum subset algorithms consider that all features have the same relevance and return a minimal set to be used for further analyses.

Feature selection depends on the overall processing goal and its performance evaluation criteria, on the existing data set and the type of model targeted, on the original set of pattern features and on the defined feature selection criterion.

Data dimensionality reduction affects all phases of a data mining process. It must be started in the data preparation stage. In many cases, feature reduction is part of the data mining algorithm and it can also be applied in the evaluation stage for a better evaluation and consolidation of the results obtained.

We can therefore conclude that, by means of the data dimensionality reduction, we aim to improve the performance of the data mining operation, as well as that of the resulted models, to reduce the model dimensionality without affecting its quality, and last but not least, to allow the user to visualize results in fewer dimensions in order to improve the decision making process.

III. FEATURE SELECTION BASED ON RELEVANCE AND REDUNDANCY ANALYSIS

Practice has demonstrated that irrelevant input features induce great computational costs for the data mining process and may lead to overfitting. To avoid these drawbacks, feature selection research has focused on the choice of relevant features from the whole data set [4]. Some results have also revealed the existence and the negative effect of feature redundancy [3] [5] [6]. The conclusion was that it is necessary to reduce the number of redundant features to a minimum level in order not to affect the accuracy of the model built. In [7] it is stated that “features are relevant if their values vary systematically with category membership”. This means that a feature is relevant if it is correlated with the class. This was formally defined in [3] as follows: a feature F_i is relevant iff there are f_i and c for which $P(F_i=f_i) > 0$, so that

$$P(C=c|F_i=f_i) \neq P(C=c) \quad (1)$$

A complete definition of feature relevance takes into account the existence of three disjoint categories of features named: *strongly relevant, weakly relevant* and *irrelevant features* [8].

Let F be the original set of features, F_i a feature, $S_i = F-\{F_i\}$ and C the class associated.

It can be said that:

- F_i is strongly relevant if
$$P(C|F_i,Si) \neq P(C|S_i) \quad (2)$$
- F_i is weakly relevant if
$$P(C|F_i,S_i) = P(C|S_i) \text{ and}$$
$$\exists\, S'_i \subset S_i, \text{ so that } P(C|F_i,S'_i) \neq P(C|S'_i) \quad (3)$$
and, finally,
- F_i is irrelevant if
$$\forall\, S'_i \subset S_i,\ P(C|F_i,S'_i) = P(C|S'_i) \quad (4)$$

A feature with strong relevance is always necessary for an optimal subset and it cannot be removed without affecting the original conditional class distribution. A weakly relevant feature is not always necessary but in certain condition it may become necessary, whereas an irrelevant feature is not necessary at all.

Feature redundancy can be expressed using the feature correlation property, since it is accepted that two features are redundant to each other if they are completely correlated. In

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-202-8

order to define the redundancy of features, it is useful to define the feature's Markov blanket [5].

Let us consider the notation mentioned above, and let be $M_i \subset F$ ($F_i \notin M_i$). M_i is said to be a Markov blanket for F_i if

$$P(F-M_i-\{F_i\}, C|F_i, M_i) = P(F-M_i-\{F_i\},C|M_i) \quad (5)$$

The condition above requires that M_i contains both the information that F_i has about C and about all the other features.

Finally, we could say that a feature F_i is redundant and it should be removed from F if and only if it is weakly relevant and it has a Markow blanket M_i within F.

A short look over a whole set of features reveals that it may contain four disjoint parts. These are: irrelevant features, redundant features as part of weakly relevant features, weakly non-redundant relevant features and strongly relevant features. An optimal subset must contain all relevant features and the weakly relevant but non-redundant ones.

Relevance is usually defined in terms of correlation or mutual information, so the mutual information on the data can be used as a feature selection criterion. In order to define mutual information for two variables (or features) we start from the concept of entropy, as a measure of random variable uncertainty. For a variable X, the entropy is defined as:

$$E(X) = -\sum_i P(x_i)\log_2(P(x_i)) \quad (6)$$

The entropy of a variable X, after observing the values of another variable Y, is defined as:

$$E(X \mid Y) = -\sum_j P(y_i)\sum_i P(x_i \mid y_i)\log_2(P(x_i \mid y_i)) \quad (7)$$

where $P(x_i)$ is the prior probability for all values of X, and $P(x_i|y_i)$ is the posterior probabilities of X given the value of Y. The value by which the entropy of X decreases, estimates additional information about X provided by Y. It is called information gain [9] and it is calculated using the following expression:

$$I(X,Y) = E(X) - E(X|Y) \quad (8)$$

We take into account that for the discrete random variable, the joint probability mass function is

$$P(x_i|y_j) = P(x_i,y_j) / P(y_j) \quad (9)$$

and the marginal probability function p(x) is:

$$P(x_i) = \sum_j P(x_i, y_j) = \sum_j P(x_i \mid y_j)p(y_j) \quad (10)$$

where $p(x,y)$ is joint probability distribution function of X and Y, and $p(x_i)$ and $p(y_j)$ are the marginal probability distribution functions of X, respectively Y**.** Since these are probabilities, we have

$$\sum_i \sum_j p(x_i, y_j) = 1 \quad (11)$$

Finally, for two discrete random variables X and Y, information gain is formally defined as:

$$I(X,Y) = \sum_j \sum_i p(x_i, y_j)\log(\frac{p(x_i, y_j)}{p(x_i)p(y_j)}) \quad (12)$$

According to this expression, we could state that a feature Y is more correlated to feature X than feature Z if:

$$I(X,Y) > I(Z,Y) \quad (13)$$

It can be observed that information gain favors features with more values, so it should be normalized. In order to compensate its bias and to restrict its values to range [0,1], it is preferable to use the symmetrical uncertainty [10], defined as:

$$SU(X,Y) = 2\left[\frac{I(X,Y)}{E(X)+E(Y)}\right] \quad (14\ )$$

A value of "1" for symmetrical uncertainty means that knowing the values of either feature completely predicts the value of the other, whereas a value of "0" implies that X and Y are independent.

There are many feature selection methods that consider the subset evaluation approach. In these cases, feature relevance and features redundancy are handled.

In the traditional framework for feature selection using subset evaluation [11], candidate feature subsets based on a certain search strategy are produced. Each of the candidate subsets is evaluated by a certain measure and it is compared with the previous best one with respect to this measure. If the new subset is found to be better, it replaces the previous best subset. These two stages are repeated until a stopping criterion is satisfied. This method poses difficulties due to the searching through the feature subsets.

A new framework proposed in [12] avoids implicitly handling features redundancy and allows an efficient elimination of redundant features by explicitly handling the features redundancy. This framework, presented in Figure 2, consists of two steps: firstly, the relevance analysis is carried out and the irrelevant features are removed; secondly, a redundancy analysis provides the final subset by eliminating the redundant features from the relevant ones. The advantage of this method consists in the decoupling relevance and the redundancy analyses that lead to an efficient way to find a subset that approximates an optimal subset.

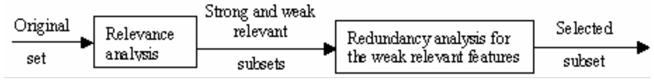


Figure 2. Feature selection through relevance and redundancy analysis

Let us use SU(X,Y) as a correlation measure for both the relevance and redundancy analysis. Such a correlation between any feature F_i and the class C is called C-correlation ($SU(F_i,C)$) and the correlation between any pair of features F_i and F_j ($i \neq j$) is called F-correlation [13].

As we have noted above, the optimal features subset contains those feature which are strongly correlated with the class but are not correlated with each other, and are not redundant. In order to achieve that, C-correlation for each feature must be calculated. Once a relevance threshold γ is established experimentally by the user, one can assume that a feature F_i is relevant if $SU(F_i,C) > \gamma$. After relevant features are selected, they are subject of redundancy analysis. In a natural approach, one could evaluate the correlation between individual features for redundancy analysis, but there are two drawbacks. Firstly, if two features are not completely correlated, it is difficult to determine feature redundancy and which one should be removed. Secondly, this involves calculating the F-correlation for a great number of pairs which it is inefficient for high-dimensional data sets. To avoid these problems it is indicated to approximately determine feature redundancy by approximating Markov blankets for the relevant features found in the previous stage. The basic idea is that a feature with a greater C-correlation value offers more information about the class than a feature with a smaller one. Consequently, when $SU(F_j,C) \geq SU(F_i,C)$, it is necessary to evaluate if F_j can form an approximate Markov blanket for F_i in order to keep more information about the class. For two relevant features F_i and F_j ($i \neq j$), we could say that F_j forms an approximate Markov blanket for F_i if [13] :

$$SU(F_j,C) \geq SU(F_i,C) \qquad (15)$$

and

$$SU(F_i,F_j) \geq SU(F_i,C) \qquad (16)$$

In (15), $SU(F_i,C)$ is heuristically used as a threshold to establish if the F-correlation $SU(F_i,F_j)$ is a strong one.

So, in order to find the appropriate feature subset, those for which there are Markov blankets, which are redundant, should be eliminated from the relevant feature set.

The whole process is presented in Figure 3.

```
Input:  {F,C} ; F={$F_1$, $F_2$, … $F_n$}
        $\gamma$
Output: $S_{opt}$

 1.  S= $\phi$
 2.    for i=1 to n do begin
 3.      calculate $SU(F_i,C)$
 4.        if $SU(F_i,C) \geq \gamma$
 5.          $S=S \cup \{F_i\}$
 6.    end              // S contain all relevant features
 7.  order  S descending on $SU(F_i,C)$        // this aims to
     make easier the comparison  between $SU(F_i,C)$ and
     $SU(F_j,C)$ for $i \neq j$
 8.  $F_j$= FirstElement(S)
 9.   do  begin
10.     $F_i$ =NextElem(S,$F_j$)
11.       if  $F_i$ is not null
12.         do begin
13.           if  $SU(F_i, F_j) \geq SU(F_i,C)$
14.             S = S-{$F_i$}
15.           $F_i$=NextElement(S,$F_i$)
16.         until $F_i$ is not null
17.     $F_j$ = NextElement(S, $F_j$)
18.   until $F_j$ is not null
19.  $S_{opt}$ = S
```

Figure 3. Feature selection method

As it can be observed in the first phase (lines 1-6), one obtains the relevant feature set S. These features are decreasingly ordered (line 7) according to their SU values. Then, the ordered list S is processed (lines 8-19) in order to select the optimum feature subset. This means that the features are filtered based on the presence or the absence of approximate Markov blankets.

IV. Experimental results

A. Data Set Description

We have tested this method on data collected by the TERAPERS system. This is a system which aims to assist the personalized therapy of dyslalia (an articulation speech disorder) and to track how the patients respond to various personalized therapy programs. Implemented in March 2008, the system is currently used by the therapists from Regional Speech Therapy Center of Suceava.

An important aspect of assisted therapy refers to its ability to adapt to the patients' characteristics and evolution. In order to adapt the therapy programs, the therapist must carry out a complex examination of children, through recording relevant data related to personal and family anamnesis. Anamnesis data can provide information on the various causes that may negatively influence the normal development of language. It contains historical data and data provided by the cognitive and personality examination.

The data provided for the personalized therapy programs includes the number of sessions/week, exercises for each phase of therapy and the changes of the original program according to the patient evolution. In addition, the report downloaded from a mobile device collects data on the efforts of child self-employment. The data refers to the exercises done, the number of repetitions for each of these exercises and the results obtained. The tracking of child's progress materializes data indicating the assessing time, and the child's status at that moment. All this data is stored in a relational database, composed of 60 tables.

The data stored in the TERAPERS's database together with the data from other sources (e.g. demographic data, medical or psychological research) compose the set of raw data that can constitute the subject of data mining process. It might be useful, because as it was shown in [14], one can use classifications in order to distribute the people with different speech impairments in predefined classes (if attribute diagnosis contains class label, we can predict a diagnosis based on the information contained in various predictor variables), clustering can be used to group people with speech disorders on the basis of features similarity and to help therapists to understand who are their patients; also, one can use association rules to determine why a specific therapy program has been successful on a segment of patients with speech disorders, while it was ineffective on another segment of patients.

For our experiments, a data set consisting of 102 features with numeric and descriptive values and 400 cases was considered. This is anamnesis data or data derived from complex examinations, based on which classification models will be built, in order to suggest the diagnosis for future cases. Firstly, we have eliminated the features that obviously are not relevant for the objective set (e.g. parents' name and work place) and we obtained 71 features. The feature selection method described above was applied on this data set, and we have compared the performances of the model built on the reduced set of features with those obtained for the model built on the whole data set.

Shown in Figure 4, this experiment is designed and implemented in WEKA [15].

The attribute "*diagnosis*" was considered as class label, and three patients' classification processes were built. There are identical in terms of models, but they differ because the same operators are applied on different datasets.

The first process is carried out on the whole set of features, the second one uses a reduced set which contains all the relevant features, while the third one is applied on the data set formed only by the relevant and non-redundant features.

An experiment containing three processes, each of them using another classification model was carried out (Figure 4). Two rules classification models and a decision tree model (J48) have been considered.

Relevant features are obtained by the C-correlation estimation. As it can be noticed in Figure 5, an ordered list of features is obtained and those for which SU(F_i,C) = 0 are removed. The result consists of 52 relevant features. The final feature subset, obtained by removing those for which the expression (16) is respected, (lines 13-14 in Figure 3), contains 10 features.

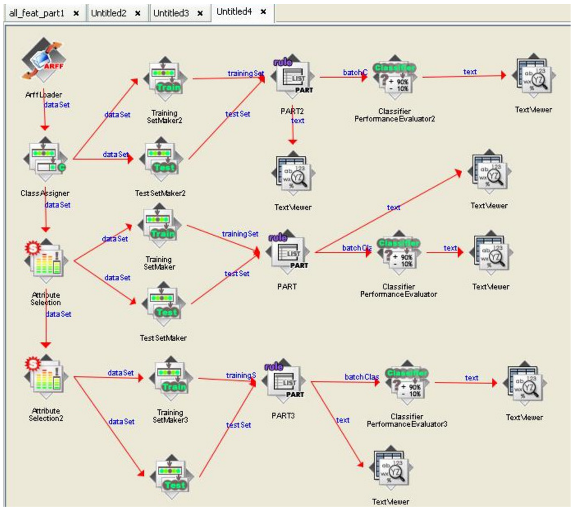
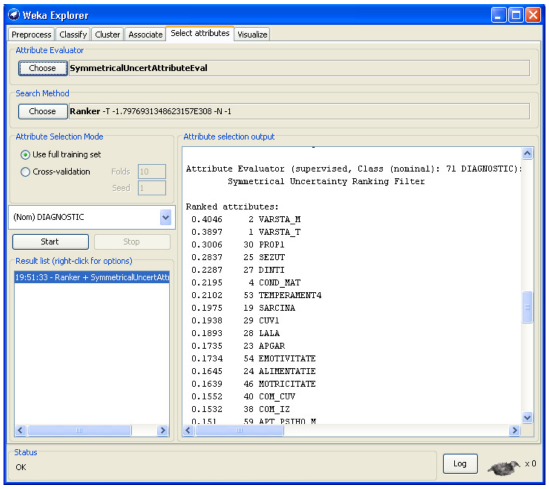


Figure 4. WEKA Knowledge flow for the classification experiment

An analysis of the performances of the three processes, in terms of percent-correct classified cases is shown in Figure 6, and a visual comparison between these performances is presented in Figure 7.

As it can be observed, there are little differences between the percent-correct classified cases for the same classifier applied on the three data sets. For the methods studied, the best results are obtained for the subset of relevant and non-redundant features subset.



Attribute Evaluator: SymmetricalUncertAttributeEval

Search Method: Ranker -T -1.7976931348623157E308 -N -1

Attribute Selection Mode: ● Use full training set ○ Cross-validation Folds 10 Seed 1

(Nom) DIAGNOSTIC

Result list (right-click for options): 19:51:33 - Ranker + SymmetricalUncertAttri

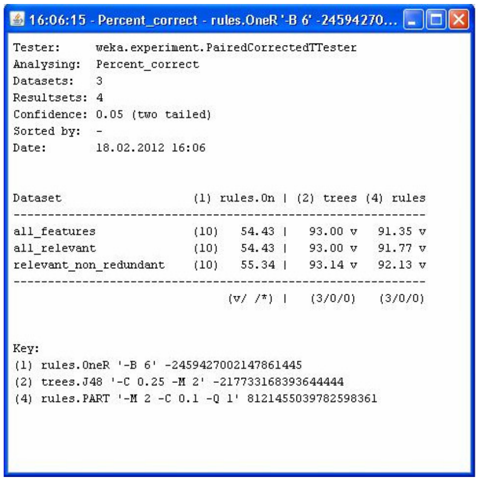
Attribute selection output

```
Attribute Evaluator (supervised, Class (nominal): 71 DIAGNOSTIC):
        Symmetrical Uncertainty Ranking Filter

Ranked attributes:
 0.4046     2 VARSTA_M
 0.3897     1 VARSTA_T
 0.3006    30 PROP1
 0.2837    25 SEZUT
 0.2287    27 DINTI
 0.2195     4 COND_MAT
 0.2102    53 TEMPERAMENT4
 0.1975    19 SARCINA
 0.1938    29 CUV1
 0.1893    28 LALA
 0.1735    23 APGAR
 0.1734    54 EMOTIVITATE
 0.1645    24 ALIMENTATIE
 0.1639    46 MOTRICITATE
 0.1552    40 COM_CUV
 0.1532    38 COM_IZ
 0.151     59 APT_PSIHO_M
```

Status: OK

Figure 5. Partial list of relevant features



16:06:15 - Percent_correct - rules.OneR '-B 6' -24594270...

```
Tester:     weka.experiment.PairedCorrectedTTester
Analysing:  Percent_correct
Datasets:   3
Resultsets: 4
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       18.02.2012 16:06


Dataset                   (1) rules.On | (2) trees (4) rules
------------------------------------------------------------
all_features              (10)   54.43 |   93.00 v   91.35 v
all_relevant              (10)   54.43 |   93.00 v   91.77 v
relevant_non_redundant    (10)   55.34 |   93.14 v   92.13 v
------------------------------------------------------------
                               (v/ /*) |   (3/0/0)   (3/0/0)


Key:
(1) rules.OneR '-B 6' -2459427002147861445
(2) trees.J48 '-C 0.25 -M 2' -217733168393644444
(4) rules.PART '-M 2 -C 0.1 -Q 1' 8121455039782598361
```

Figure 6. Percents of corrected classified cases for the three data sets

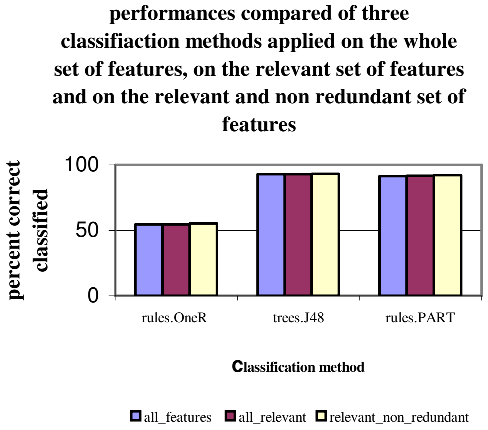


Figure 7. Percents comparison of the corrected classified cases

Significant differences between the three processes have been obtained for the elapsed training time. These results are presented in Figure 8. As it is shown for all the three methods, the best times are achieved for the subset consisting of relevant and non-redundant features.

Practically, for the dataset consisting in 400 cases described above, for the least efficient method (rules.oneR), the training process for the whole set of features lasted 0.37 sec, while for the feature subset containing only relevant and non-redundant features this process it lasted 0.05 sec; for the most efficient method (tree.J48), the elapsed training time for the whole set of features was 0.06 sec and for the relevant and non-redundant features this time was 0.01 sec.

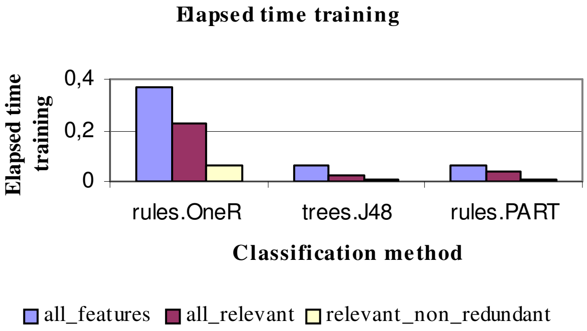


Figure 8. Comparison of elapsed time training

V. CONCLUSIONS AND FUTURE WORK

This work is part of the research that aims to implement a data mining system that will allow the optimization of personalized therapy of speech disorders for children with dyslalia. Combining the feature selection methods with the data mining algorithms is a good practice; therefore, in this paper, we have studied such a method based on the features relevance and redundancy analysis.

This method, applied on the anamnesis data provided by TERAPERS, has shown that both the percent of correctly classified cases and that of the elapsed time for training are better if the considered data mining algorithms are applied on data containing a reduced subset of features.

It must be noted that these results cannot be generalized for all data mining methods and algorithms. This is why we intend to study the impact of feature reduction on clustering and association rules mining.

ACKNOWLEDGMENTS

This paper was supported by the project "Progress and development through post-doctoral research and innovation in engineering and applied sciences– PRiDE - Contract no. POSDRU/89/1.5/S/57083", project co-funded from European Social Fund through Sectorial Operational Program Human Resources 2007-2013.

REFERENCES

[1] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery: An overview. In *Advances in Knowledge Discovery and Data Mining*, U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Eds. AAAI/MIT Press, Cambridge

[2] Danubianu M., Pentiuc S.G., Tobolcea I., Schipor O.A. (2010). Advanced Information Technology - Support of Improved Personalized Therapy of Speech Disorders, *International Journal of Computers Communications & Control*, ISSN 1841-9836, 5(5), pp: 684-692

[3] Kohavi R., John G.(1997). Wrappers for feature subset selection. *Artificial Intelligence, special issue on relevance*, 97(1-2), pp. 273-324

[4] Peng H. Long F., Ding C. (2005). Feature Selection based on mutual Information: Criteria of Max-Dependency, Max-Relevance and Min-Redundancy, *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 27, No. 8,

[5] Koler D., Sahami M. (1996). Towards optimal feature selection. *In Machine Learning: Proceedings of the Thirteenth International Conference on Machine Learning.* Morgan Kaufman

[6] Hall M.A.(2000). Correlation-based feature selection for discrete and numeric class machine learning, *In proceedings of the Seventeenth International Conference on Machine Learning*, p. 359-366, 2000

[7] Gennari J.H., Langley P., Fisher D.(1989). Models of incremental concept formation. *Artificial Intelligence*, (40), p. 11-16

[8] John G.H., Kohavi R., Pfleger P.(1994). Irrelevant features and the subset selection problem. *In Machine Learning: Proceedings of the Eleventh International Conference*, Morgan Kaufman

[9] Quinlan J.R.(1993). C4.5: Programs for Machine Learning, Morgan Kaufmann,

[10] Press, W.H., Teukolsky, S.A., Vetterling. W.T., Flannery. B.P.(1988) Numerical Recipes in C, Cambridge Univerity Press, Cambridge

[11] Liu, H., Motoda, H.(1998). Feature Selection for Knowledge Discovery and Data Mining, Boston Kluwer Academic Publishers, ISBN 0-7923-8198-X

[12] Yu, L., Liu, H.(2004). Redundancy based feature selection for microarray data, *Proc of the Tenth ACM SIGMOD Conference on Knowledge Discovery and Data Mining*, pp. 737-742

[13] Yu, L., Liu, H.(2004). Efficient Feature Selection via Analysis of Relevance and Redundancy, *Journal of Machine Learning Research*, 5, pp. 1205-1224

[14] Danubianu M., Pentiuc St. Gh., Socaciu T. (2009). Towards the Optimized Personalized Therapy of Speech Disorders by Data Mining Techniques, *The Fourth International Multi Conference on Computing in the Global Information Technology ICCGI 2009*, Vol: CD, 23-29 August, Cannes - La Bocca, France

[15] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I., The WEKA Data Mining Software: An Update, *SIGKDD Explorations*, Vol. 11, issue 1.

Software Learnability Evaluation

An Overview of Definitions and Evaluation Methodologies for GIS Applications

Irfan Rafique, Jingnong Weng, Yunhong Wang, Maissom Qanber Abbasi, Philip Lew
School of Computer Science and Engineering
Beijing University of Aeronautics and Astronautics
Beijing, China
e-mail: irfan@cse.buaa.edu.cn, wengjn@buaa.edu.cn, yhwang@buaa.edu.cn, maissom@cse.buaa.edu.cn,
philiplew@buaa.edu.cn

Abstract— **Learnability has been regarded as an important aspect of usability and considered a fundamental usability attribute. Yet, learnability is often overlooked as one of the most influential factors for the success of software applications especially Geographic Information Systems (GIS) applications. GIS Applications have seen a tremendous development during the last decades. Becoming more advanced, the amount, diversity and high turnover demand fast learning from users. Good learnability not only leads to a better productivity quickly but also plays a vital role in initial adoption or rejection of a technology. There are numerous approaches used to define, measure, and evaluate learnability. This paper presents some previously researched definitions along with methodologies for learnability evaluation with a special focus on desktop GIS applications. Our survey of definitions and evaluation methodologies leads us to a conclusion that there is a need of further research for a sound and widely accepted methodology for learnability evaluation of GIS applications.**

Keywords - learnability; GIS applications; learnability evaluation; learnability in use; usability.

I. INTRODUCTION

With the pervasiveness of the software in our everyday lives, the need for quality software systems becomes indispensable. Evaluating and improving quality needs a process of continuous assessment. This evaluation should be based on various functional as well as nonfunctional properties. Non-functional properties as depicted by ISO 25010 [5] such as efficiency, learnability, security, reliability and attractiveness, amongst others, all contribute appreciably to the quality of software systems.

In the last decade, there has been a rapid increase in the use of Geographic Information System (GIS) applications, especially web-based GIS applications (GISApps), in fields like education, transport, criminology, marketing, sociology, business and disaster recovery. Today almost all businesses and government agencies use GISApps as a tool for decision making and problem solving.

Learnability, by some definitions, characterizes how easy is it for users to accomplish basic tasks the first time they encounter the software application. In an increasingly technological world, software, especially GISApps, are becoming more varied and complex. New features are being added quite rapidly to new GISApps, which users are expected to use immediately. The learnability of modern GISApps, especially web-based GISApps, has a distinctive importance. With conventional software systems, users must make an investment (often substantial) in time and effort to install and learn to use an application. However, with web-based applications, users can very quickly switch from one Web application to another with minimal effort. In 2006, Lazar et al. [10] discovered that users reported wasting on average, 42-43% of their time on the computer due to frustrating experiences. When looking at the specific causes of the frustrating experiences that occurred, the study found error messages and missing/hard to find/unusable features were among top five causes closely related with poor usability and more specifically poor learnability. Good learnability will lead to reasonable learning times, adequate productivity during the learning phase, and thus better satisfaction in new users. Improving learnability, thus, has a significant impact on the success of software applications and especially for GISApps, as GISApps involve a different interaction style, three dimensional interface designs and the need of grasping spatial concepts, making them more difficult to learn. But, improvement first requires identifying and understanding learnability issues. Also, learnability issues can only be exposed by clearly defining, and then evaluating it in systematic and consistent way.

Although researchers recognize the importance of learnability, the consensus among researchers regarding defining and evaluating learnability seems lacking, leading to the conclusion that software systems still pose learnability problems. The main objectives of this research are to: 1. Understand learnability in detail with respect to GISApps and the special characteristics of GISApps that need consideration. 2. Analyze the related research in learnability evaluation with reference to GISApps. Although GISApps are being widely utilized in different devices like desktops, mobile devices and cellular phones, etc., this research is focused on learnability related to desktop applications only.

Following this introduction, Section II provides background on state of the art research in learnability. Section III highlights the importance of learnability for GISApps. Section IV provides evaluation schema for learnability with respect to GISApps. The subsequent section emphasizes quality in use (QinU) aspect of learnability and Section VI concludes the paper and outlines future work.

II. Characterizing Learnability

This section of the paper provides background on learnability while examining the existing research and delineating areas for improvement regarding clarity in its definition.

A. Definiing Learnability

In order to evaluate learnability, first we have to define and understand it clearly. There have been a number of different definitions proposed. Table I summarizes some of these definitions. The tabulated definitions are among the many different definitions used by different researches over the last two decades. For the purpose of brevity we have only included representative definitions involving some unique types of measures in defining learnability.

Nielsen [1], Holzinger [19], Shneiderman 1995 [20], and Chapanis [21] define learnability in terms to time that is how quickly users learn to operate the software. Dix et al. [3] and Stickel et al. [22] define learnability in terms of ease with which new users can begin effective interaction with the system. Santos and Badre [6] define learnability in terms of measure of effort required to achieve a defined level of proficiency. Hart and Steveland [8] and Linja Aho [23] define learnability in terms of time and ease with which user starts efficient interaction with the product. Rieman [24], Butler [13] and MUMMS Questionnaire [25] define learnability in terms of user performance without formal training. Ziefle [26] and ISO 9126-1 [16] define it in terms of software product properties that enable user to learn its application. Hart and Steveland [8], Kirakowski and Claridge [4] and MUMMS Questionnaire [25] highlight subjective aspects of learnability by judging it from user's feelings about learning process. Most of the definitions refer to the performance of user relevant to their first interaction with the software (initial learnability), but some researchers have also taken note of extended learnability, that concerns improvement in performance over time ([3][4][11][13]). Extended learnability or advanced learnability has been characterized by learning of new or advanced features, ability to adoption alternate model that involved fewer screens or keystrokes, ability to master the software and ability to achieve maximal performance. This description clearly shows the diversity in the use of measures among researchers regarding defining learnability.

TABLE I. Summary of Learnability Definitions

No.	Source	Definition
1.	Jakob Nielsen (1993) [1]	Novice user's experience on the initial part of the learning curve.
2.	Dix (1998) [3]	Ease at which new users can begin effective interaction and achieve maximal performance
3.	Santos and Badre (1995) [6]	Measure of the effort required for a typical user to be able to perform a set of tasks using an interactive system with a predefined level of proficiency.
4.	Hart and Steveland (1988) [8]	The speed and ease with which users feel that they have been able to use the product or as the ability to learn how to use new features when necessary.
5.	Bevan and Macleod's (1994) [11]	A measure of comparison the quality of use for users over time.
6.	Butler (1985) [13]	Initial user performance based on self instruction" and "[allowing] experienced users to select an alternate model that involved fewer screens or keystrokes.
7.	Kirakowski and Claridge (1998) [4]	Within the web context is the degree to which users feel able to manage the product's basic functions during its first use.
8.	ISO 9126-1 (2001) [16]	The capability of the software product to enable the user to learn its application
9.	ISO 25010 (2011) [5]	Degree to which a product or system can be used by specified users to achieve specified goals of learning to use the product or system with effectiveness, efficiency, freedom from risk and satisfaction in a specified context of use

Grossman et al. [27] carried out a survey of 88 research papers related to learning in HCI (Human Computer Interaction), 45 discussed learnability without a definition, and the remainder had conflicting definitions. They classify learnability definitions in eight different categories. Instead of deciding upon a common definition, they developed taxonomy of learnability definitions after highlighting the short comings in current definitions. Key features of the developed taxonomy include the existence of an optimal performance level, the dimension of experience, and the timeline of when the learning takes place.

In general terms "*learnability is a characteristic where performance improves with experience. As tasks are repeated, elements of the task are better remembered, prompts are more clearly distinguished, skills are sharpened, transitions between successive tasks are smoothed, eye-hand coordination is more tightly coupled, and relationships between task elements are discovered. The aggregation of these effects results in faster performance times, fewer errors, less effort, and more satisfied users*" [28].

B. Software Quality perspectives of Learnability

ISO quality models can be used to support specification and evaluation of software from different perspectives by those associated with acquisition, requirements, development, use, evaluation, support, maintenance, quality assurance and audit of software. The ISO 25010 [5] defines

1) A quality in use (QinU) model composed of five main characteristics (Effectiveness, Efficiency, Satisfaction, Freedom from Risk and Context Coverage) that relate to the outcome of interaction when a product is used in a particular context of use.

2) A product quality model composed of eight main characteristics (including usability) that relate to static properties of software and dynamic properties of the computer system.

Many researchers ([1][2][3]) and standards (IEEE standard 610.12 [29], ISO 9126-1 [16] and ISO 25010 [5]) have mentioned learnability as an important attribute of usability. In ISO 9126 the product centered view of usability and learnability was presented but in recent standards both products centered and QinU centered views have been presented. According to guidelines of ISO [5], learnability can be specified or measured in two different ways. The first one is the "*extent to which a product or system can be used by specified users to achieve specified goals of learning to use the product or system with effectiveness, efficiency, freedom from risk and satisfaction in a specified context of use*". This corresponds to QinU aspect of learnability. We further discuss this aspect in more detail in Section V. The second method of specification and measurement of learnability is by product properties corresponding to suitability for learning as defined in ISO 9241-110 [30]: *"software product quality is the cause and QinU is the effect"*. Thus learnability can be seen as *the collective effect of key product attributes that lead to efficient and effective learning of a software product with high end user satisfaction levels in a specified context of use*.

III. GIS Applications and Learnability

This section depicts specific characteristics of GISApps, which put greater emphasis on learnability of GISApps.

Although GISApps work in a graphic user interface, they are quite different from general computer applications. The special functions required to manipulate spatial aspects make the interface complicated and difficult to learn. To make matters worse, GISApps employ unique interfaces, and therefore users must learn a different interface style with each application. GISApps generally require three dimensional (3D) interaction styles. Although we live and act in a 3D world, the physical world contains many more cues for understanding and constraints and affordances for action that cannot currently be represented accurately in a computer simulation. It is quite difficult for new users to transform traditional WIMP (Windows, Icons, Menus, Pointers) interaction styles to three dimensional interaction, leading to learning difficulties.

A GISApp combines query functions and analysis with visualization and geographic features to examine spatial problems. Using, managing, and analyzing spatial data, and enabling a user to analyze spatial questions is distinctive to GISApps, but this leads to usability issues especially in understanding and learning the application. For GISApps, users have a relatively long learning curve due to the need to grasp geographical concepts and different data types. Also, the level of user knowledge of geographical concepts, and task dependency on geographic concepts are special considerations [31]. A report on the Leonardo Pilot Project, E-GIS (about learning of GIS applications) notes difficulty in learning of GISApps as one of the main causes of student drop out during learning course [32].

In addition to different interaction style, GIS visualization poses several challenges. GIS employ a virtual environment (VE) to display and interact with high dimensional geospatial structures and phenomena. Way-finding in such an environment has certain challenges. In the real environment, kinesthetic feedback is available to the user; movement is restricted by physical boundaries. In VE such feedback is not normally given. Navigation in VE is generally controlled indirectly with interaction tools such as keyboard, mouse, joysticks, etc. Since desktop VEs are seldom immersive, navigation in such a VE is even less similar to real world navigation because navigation in addition to being indirect is typically controlled from the "outside" of the environment (like controlling a toy car by remote control) [33]. It is therefore common for a novice user losing orientation (awareness of the space around, including the location of objects and places) during way-finding process.

GIS displays wide regions on a small screen and allows navigation in large spaces. Unlike the bird's eye, overall view map of an area, the user often deals with only a part of a large scale space (not visualized entirely from one viewpoint). It is common for the novice users to get "lost" when zoomed into a small area without reference text (e.g., place names).

In current era of GISApps, learnability has new challenges as software is mostly released online and online help and support are the main customer support mechanisms. Therefore, the existing research in software learnability needs appropriate considerations specifically for GISApps.

IV. Learnability Evaluation

The previous section discussed some of the particular characteristics of GISApps and the importance of requiring a new model for learnability evaluation. This section examines some of the existing methods for evaluating learnability for GISApps.

Usability engineering research literature mentions several usability evaluation methodologies; however, their suitability for evaluating learnability is not well elaborated. Similarly the suitability of methodologies used for non GISApps for GISApps is also not very obvious. We discuss only those methodologies in this section which have been used for evaluation of GISApps.

One of the most common forms of usability testing is the "Think-Aloud Protocol". In this technique respondents are asked to give a verbal account of their thinking as they answer (concurrent) or immediately after answering (retrospective) a draft survey question [25]. Komarkova et al. [34] employed Think Aloud Protocol to find usability problems in 14 Web based GISApps run by the Czech Regional Authorities. They identified learnability related issues like complexity of search tools and lack of interface understandability. Nivala et al. [35] used this methodology to identify the potential usability problems of web mapping sites, including learnability issues like interface crowdedness, lack of conformity to user expectations and absence of map legends, etc. Think Aloud technique provides rich qualitative data and allows first hand insight into the thought processes associated with different tasks. Think Aloud methodology can be useful for identifying learnability issues, but takes place in the unnatural environment of a usability lab [3]. Moreover, people can only report what they are aware of and can report about the components of high level mental processes, like the sequence of steps that leads to the solution of a problem. Furthermore, it is difficult to identify changes in behavior due to learning by this method.

Focus Group [1] is another methodology in which a number of users are brought together to discuss new concepts and identify issues over a certain period of time. Each group is run by a moderator who is responsible for maintaining the focus of the group on whatever issues are of interest. Fuhrmann and MacEachren [33] employed Focus Group for the evaluation of a geovirtual environment and discovered many learnability issues like lack of predictability, difficult to find features, lack of informative feedback from software etc. Harrower et al. [36] employed Focus Group to assess an animated and interactive geovisualization environment Earth System Visualizer and implications of this environment for learning about spatiotemporal processes. They deduced that novel interfaces may not result in improved performance unless sufficient training is provided on how to use them. Focus Groups usually provide immediate ideas for the improvement of particular products or concepts. This method can be useful for identifying learnability issues and proposal of design guidelines, but it is rather subjective and expensive methodology. Moreover Focus Groups are not efficient in covering maximum depth on a particular issue. Additionally moderator bias can greatly impact the outcome of a Focus Group discussion.

For the evaluation of MapTime, a software package for exploring spatiotemporal data associated with point locations, Slocum et al. [37] employed a methodology consisting of a combination of individual interviews and Focus Groups conducted for three distinct groups of participants: novices, geography students, and domain experts, and discovered that individual interviews are particularly useful in obtaining users' reactions to software (as opposed to having them learn the software on their own) because the interviewer can steer the interview based on the user's responses.

Observation is a quite frequently employed method for learnability evaluation. It involves visiting one or more users in their workplaces. Notes must be taken as unobtrusively as possible to avoid interfering with their work [19]. Video recording has also been a very frequently used method for observational data collection. Jones et al. [38] used video analysis for an exploratory task-orientated project workshop with the four project team members, for usability and learnability evaluation of a geographic profiling tool. They measured learnability by video analysis of users' browsing interaction. Hossain and Masud [39] used video evaluation to evaluate "ArcView" GIS software with four participants during two hours of interaction. They found 12 learnability problems using this method including interface understandability, presence of unfamiliar terms, help and error messages inadequacy etc. Video recording is quite comprehensive way of data collection, but analysis required is quite time taking.

Another means of electronic observation is Data Logging, which involves statistics about the detailed use of a system. Meng and Malczewski [40] used a data logging approach to evaluate usability and learnability of a public participatory GISApp named ArgooMap. The users' every move on the website was recorded with a logging software which made it possible to obtain detailed and useful information about the actual usage of the website holding ArgooMap. Although this methodology captures data automatically, it has not been widely applied for learnability evaluation in GISApps.

Lew et al. [41] used C-INCAMI (Contextual-Information Need, Concept model, Attribute, Metric and Indicator) framework to evaluate learnability as a product characteristic of a GISApp named Chinastar. C-INCAMI is a framework which relies on an ontological conceptual base; on a well-established measurement and evaluation process. Using this methodology some learnability issues like lack of predictability were identified. This methodology is model based and provides quantitative results, but there seems to be involvement of subjective judgments while computing the metrics for learnability attributes.

Some researchers employed a combination of several methodologies for investigating different aspects of usability of GISApps. Nivala et al. [35] for example, conducted a series of expert evaluations and user tests. During the expert evaluations, eight usability engineers and eight cartographers examined the web based GISApps including Google Maps, MSN Maps and Directions, MapQuest, and Multimap, by paying attention to their features and functionality. Additionally, eight user tests were carried out by ordinary users in a usability laboratory. User tests used a combination of Think Aloud and video recording method. Kristoffersen [42] used a "traingularization" of observation, interviews and document study to evaluate usability of ArcView used for viticulture purposes and concluded that the user consider learnability and functionality aspects to be top usability issues.

Many aspects of usability can best be studied by querying the users. This is especially true for issues on the subjective satisfaction of the users and their possible anxieties, which are hard to measure objectively. Use of a subjective questionnaire has been a very effective and popular method for usability evaluation. Being a sub characteristic of

TABLE II. STATISTICS OF LEARNABILITY RELATED QUESTIONS IN FAMOUS QUESTIONNAIRES

Questionnaire Name	**Number of Questions**	
	Product Perspective	***QinU Perspective***
Purdue Usability Testing Questionnaire (PUTQ) [2]	34	-
Software Usability Measurement Inventory (SUMI) [4]	4	6
Questionnaire for User Interaction Satisfaction (QUIS) [7]	7	10
The Post-Study System Usability Questionnaire (PSSUQ) [9]	2	2
Practical Heuristics for Usability Evaluation (PHUE) [12]	5	-
SUS (System Usability Scale) [14]	-	4
IUI (Isometrics Usability Inventory) [15]	11	6
WAMMI (Website Analysis and MeasureMent Inventory) [17]	-	2
Usefulness, Satisfaction, and Ease of use. (USE) [18]	-	4

usability, learnability has also been evaluated by using subjective questionnaires. Table II lists some famous questionnaires along with number of questions relevant to product oriented and QinU oriented aspects of learnability.

It must be kept in mind that the software learnability evaluation should provide quantitative or qualitative results that are comprehensible, acceptable and repeatable, in order to prove a key driver for improvement in learnability and consequently in software quality. All referenced methodologies require users to either identify learnability problems or evaluate subjective or objective learnability by measuring time on task, task correctness, error counts or user's subjective responses. Most approaches are not model based and seem to be not easily reproducible. ISO 25010 and ISO 9126 both treat learnability as a product characteristic that is a characteristic of internal or external quality. However, there has been a limited effort regarding evaluating learnability from product quality perspective [31].

Think Aloud Protocol, Focus Group, and interviews are all quite direct methodologies for learnability issues identification, but have some limitations. These methodologies are prone to high level of subjectivity as well as the interplay of a legion of factors, including the characteristics of the users, the environment, the sample size of the user group, etc., leading to problems in isolating individual factors under examination. Similarly data logging method has not been found to be widely applied for GISApps. C-INCAMI, seems to be a model based scheme but has some subjective judgments involved for producing quantitative results. Moreover, its applicability in GIS domain has been very limited.

V. QULAITY IN USE ASPECT OF LEARNABILITY

As mentioned in Section II-B, learnability has been considered as an attribute of usability. There has been an inconsistency even within ISO software quality model regarding concept of usability. In earlier drafts of ISO 9126, usability was defined primarily in terms of product attributes as "*A set of attributes of software which bear on the effort needed for use and on the individual assessment of such use by a stated or implied set of users*". In ISO/IEC CD 25010.3 [43], the product centered view of usability as presented in ISO 9126 was deemed narrow at that time and renamed as operability but in its final release ISO/IEC 25010:2011 it was retained as usability. In ISO/IEC CD 25010.3, usability appeared as a characteristic of quality in use. The recent ISO 25010 standard, regarding usability evaluation states, "*Usability can either be specified or measured as a product quality characteristic in terms of its sub-characteristics, or specified or measured directly by measures that are a subset of quality in use*". We argue that users achieve their intended goals with effectiveness, efficiency, satisfaction and freedom from risk (the sub characteristics of QinU) not only because of usability, but also due to other product quality characteristics like utility (e.g., right functionality or functional suitability), reliability and performance efficiency etc. The usability attributes of a product are thus only one contribution to the quality in use of an overall system. It is, therefore worthwhile to model "in use" part of usability on QinU model side, rather than considering it as a total "in use" aspect of usability. Being a sub characteristic of usability, learnability can further be modeled on QinU side also. Based on our thorough analysis of literature and questionnaires we have noticed several sub characteristics of learnability that can be incorporated in QinU model. Lew et al. [41] has already mentioned learnability in use concept defining it as "the degree to which specified users can learn efficiently and effectively while achieving specified goals in a specified context of use". Most of the learnability evaluations mentioned in Section IV measure QinU aspect of learnability without defining it specifically as such. In true model based evaluation methodologies both aspects of learnability vis-a-vis product centered and QinU oriented should be considered.

VI. CONCLUSION AND FUTURE WORK

Learnability has an increased importance for GISApps because of the need of grasping spatial concepts and different interaction styles. In learnability research, there has been inconsistency in defining learnability and treating it as product characteristic or QinU characteristic. There have been numerous evaluation methods developed and used by researchers during past years, but there seems to be a lack of consistency and cross verification between evaluated results. Although some researchers have developed methods to evaluate learnability in more organized, conceptual and model based ways, their applicability in GIS domain is quite limited. There is a need to further strengthen the area of learnability evaluation in GISApps domain. More research, thus, needs to be done for a sound and widely accepted learnability evaluation methodology for GISApps.

In the future, we will develop a comprehensive evaluation methodology for modeling and evaluation of learnability of GISApps. As a first step we are developing a comprehensive concept model that can be employed for evaluating evaluate GISApps.

ACKNOWLEDGEMENT

Authors gratefully acknowledge support from the lab of Digital earth and GIS, School of Computer science and engineering, Beijing University of Aeronautics and Astronautics (BUAA), P.R. China.

REFERENCES

[1] J. Nielson, Usability Engineering, San Francisco: Morgan Kaufmann, 1993, pp. 16-39.

[2] H.X. Lin, Y.-Y. Choong and G. Salvendy, "A proposed index of usability: a method for comparing the relative usability of different software systems," Beh. & Infor. Tech. vol. 16, pp. 267-278, 1997.

[3] A. Dix, J. Finlay, G. Abowd and R. Beale, Human-computer interaction, 2nd ed., Hertfordshide, UK: Prentice Hall International, 1998.

[4] J. Kirakowski, N. Claridge and R. Whitehand, "Human Centered Measures of Success in Web Site Design," 4th Conf. on Human Factors and the Web, NJ, USA 1998.

[5] "ISO/IEC 25010:2011: Systems and software engineering – Systems and software product Quality Requirements and Evaluation– System and software quality models," 2011.

[6] P.J. Santos and A.N. Badre, "Discount learnability evaluation," GVU Technical Report, Georgia Institute of Technology, 1995, pp. 30-38.

[7] J.P. Chin, V.A. Diehl and K.L. Norman, "Development of an instrument measuring user satisfaction of the human-computer interface," Proc SIGCHI conference on Human factors in computing systems, ACM, 1988, pp. 213-218, doi:10.1145/57167.57203.

[8] S.G. Hart and L.E. Staveland, "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research," in Advances in Psychology, vol 52, A.H. Peter, M. Najmedin (Eds.), North-Holland 1988, pp. 139-183.

[9] J.R. Lewis, "IBM computer usability satisfaction questionnaires: psychometric evaluation and instructions for use," Int. J. Hum.-Comput. Interact. vol. 7, pp. 57-78, 1995.

[10] J. Lazar, A. Jones and B. Shneiderman, "Workplace User Frustration with Computers: An Exploratory Investigation of the Causes and Severity.," Beh. and Infor. Tech. vol. 25, pp. 239-251, 2006.

[11] N. Bevan and M. Macleod, "Usability measurement in context.," Beh. and Infor. Tech. vol. 13, pp. 132–145, 1994.

[12] G. Perlman, "Practical usability evaluation," Proc CHI '97 extended abstracts on Human factors in computing systems: looking to the future, ACM, 1997, pp. 168-169, doi:10.1145/1120212.1120326.

[13] K.A. Butler, "Connecting Theory and Practice: a case study of achieving usability goals.," Proc CHI 85 Proceedings of the SIGCHI conference on Human factors in computing systems 1985, pp. 85-88, doi:10.1145/317456.317472.

[14] J. Brooke, "SUS: A quick and dirty usability scale," in Usability evaluation in industry, P.W. Jordan, B. Weerdmeester, A. Thomas, I.L. McLelland (Eds.), Taylor and Francis, 1996.

[15] G. Gediga, K.-C. Hamborg and I. Düntsch, "The IsoMetrics usability inventory: An operationalization of ISO 9241-10 supporting summative and formative evaluation of software systems," Beh. and Infor. Tech. vol. 18, pp. 151-164, 1999.

[16] "ISO 9126-1:2001 Information technology — Software product quality — Part 1: Quality model," 2001.

[17] "WAMMI questionnaire," Human Factors Research Group in Cork, Ireland 2011.

[18] A.M. Lund, "Measuring Usability with the USE Questionnaire," Usability and User Experience vol. 8, 2001.

[19] A. Holzinger, "Usability engineering methods for software developers.," Comm. of The ACM vol. 48, pp. 71-74, 2005.

[20] L. Slaughter, K.L. Norman and B. Shneiderman, "Assessing Users' Subjective Satisfaction with the Information System for Youth Services (ISYS)," Proc Third Annual Mid-Atlantic Human Factors Conference, March 26-28, 1995, pp. 164-170.

[21] A. Chapanis, "Evaluating usability," in Human factors for informatics usability, Camb. Univ. Press 1991, pp. 359-395.

[22] C. Stickel, J. Fink and A. Holzinger, "Enhancing Universal Access – EEG Based Learnability Assessment," in Universal Access in Human-Computer Interaction. Applications and Services, vol 4556, C. Stephanidis (Ed.), Springer Berlin / Heidelberg 2007, pp. 813-822.

[23] M. Linja-aho, "Creating a framework for improving the learnability of a complex system," Human Technology vol. 2, pp. 202-224, 2006.

[24] J. Rieman, "A field study of exploratory learning strategies," ACM Transactions on Computer-Human Interaction, vol. 3, pp. 189-218, 1996.

[25] MUMMS, "Questionnaire homepage. Berlin States Museum Trial Evaluation Summary," University College Cork , Ireland, 2003.

[26] M. Ziefle and S. Bay, "Mental Models of a Cellular Phone Menu. Comparing Older and Younger Novice Users," in Mobile Human-Computer Interaction – MobileHCI 2004, vol 3160, S. Brewster, M. Dunlop (Eds.): Springer Berlin / Heidelberg 2004, pp. 571-572.

[27] T. Grossman, G. Fitzmaurice and R. Attar, "A Survey of Software Learnability: Metrics, Methodologies and Guidelines," Proc 27th international conference on Human factors in computing systems, 7th April, 2009, pp. 649-658, doi:10.1145/1518701.1518803.

[28] M.R. Lehto and J.R. Buck, Introduction To human factors and ergonomics for engineers, NY: Taylor & Francis, 2008.

[29] IEEE, "Std 610.12-1990 , IEEE Standard Glossary of Software Engineering Terminology ", 1990.

[30] "ISO 9241-110:2006: Ergonomics of human-system interaction — Part 110: Dialogue principles," 2006.

[31] P. Lew, L. Zhang and L. Olsina, "Usability and User Experience as Key Drivers for Evaluating GIS Application Quality," 18th Intl. Conf. on Geoinformatics, Beijing, China, 2010, doi:10.1109/GEOINFORMATICS.2010.5567803.

[32] H. Sponberg, E. Ossiannilsson, P. Pilesjö, U. Mårtensson, E. Onstein and F. Johansen, "Online GIS-Learning," European Assoc. of Distance Teaching Univ., Lisbon, Portugal 2007.

[33] S. Fuhrmann and A.M. MacEachren, "Navigation in Desktop Geovirtual Environments: Usbaility Assessment," Proc 20th ICA/ACI International Cartographic Conference, August 06-10, 2001, pp. 2444-2453.

[34] J. Komarkova, M. Jedlicka and M. Hub, "Usability user testing of selected web-based GIS applications," W. Trans. on Comp. vol. 9, pp. 21-30, 2010.

[35] A.-M. Nivala, S. Brewster and L.T. Sarjakoski, "Usability Evaluation of Web Mapping Sites," The Cartographic Journal vol. 45, pp. 129–138, 2008.

[36] M. Harrower, A. MacEachren and A.L. Griffin, "Developing a Geographic Visualization Tool to Support Earth Science Learning," Cartography and Geographic Information Science, vol. 27, pp. 279-293, 2000.

[37] T. Slocum, R. Sluter, F. Kessler and S. Yoder, "A Qualitative Evaluation of MapTime, A Program For Exploring Spatiotemporal Point Data," Cartographica: The Intl. J. for Geog. Infor. and Geovisualization vol. 39, pp. 43-68, 2004.

[38] C. Jones, M. Haklay, S. Griffiths and L. Vaughan, "A less-is-more approach to geovisualization – enhancing knowledge construction across multidisciplinary teams," Intl. J. of Geographical Infor. Science vol. 23, pp. 1077-1093, 2009.

[39] M.D. Hossain and M.M. Masud, "Evaluating Software Usability of Geographic Information System," Int. J. of Software Engineering vol. 2, pp. 64-86, 2009.

[40] Y. Meng and J. Malczewski, "Usability evaluation for a web-based public participatory GIS: A case study in Canmore, Alberta," J. of Geography, 17th December, 2009.

[41] P. Lew, L. Olsina, P. Becker and L. Zhang, "An integrated strategy to systematically understand and manage quality in use for web applications," Requirements Engineering vol., pp. 1-32, 2011.

[42] I. Kristoffersen, "Usability Evaluation of GIS used for Viticulture Purposes," Department of Informatics, University of Oslo, Oslo, 2008, pp. 1-144.

[43] "ISO/IEC CD 25010.3: Systems and software engineering – Software product Quality Requirements and Evaluation(SQuaRE) – Quality models for software product quality and system quality in use," 2009.

Mix-matrix Method in Problem of Discrete Optimization

Iakov Karandashev and Boris Kryzhanovsky
Center of Optical Neural Technologies
Scientific Research Institute for System Analysis, Russian Academy of Sciences
Moscow, Russia
Yakov.Karandashev@phystech.edu, kryzhanov@mail.ru

***Abstract*—The problem of a quadratic functional minimization in the configuration space of *N* binary states is considered. In order to increase the efficiency of the random-search algorithm, we suggest to vary the attraction area of the deepest minima of the functional by changing the matrix *T* it is based on. The new matrix *M*, called *mix-matrix*, is a mixture of *T* and T^2. We demonstrate that this brings about changes of the energy surface: deep minima displace very slightly in the space (the Hemming distance of the shift is of about 0.01**N*), they become still deeper and their attraction areas grow significantly. The experiment shows that use of the approach results in a considerable displacement of the spectrum of sought-for minima to the area of greater depths, while the probability of finding the global minimum increases abruptly (by a factor of 10^3 in the case of a two-dimensional Ising model)**

Keywords-quadratic optimization; binary optimization; combinatorial optimization; area of attraction; local search; random search; energy landscape transformation; mix-matrix

I. INTRODUCTION

The goal of this paper is to improve the efficiency of a random search procedure used to solve binary minimization problems. In this class of problems, the solution is reduced to the minimization of the quadratic functional $E(S)$ constructed from a given $N\times N$ matrix T in the N dimensional configuration space of states $S=(s_1, s_2, ..., s_N)$ with discrete variables $s_i=\pm 1$, $i=1,2,...,N$. Many discrete programming problems, such as graph partitioning, graph coloring, traveling salesman problem etc., are reduced to this problem [1-2]. In addition, this problem arises in condensed matter physics where the search of the ground state is important for understanding of a disordered system structure [2].

It is well known that there is no polynomial algorithm for solving this problem, i.e., it is impossible to find a global minimum in polynomial time (the problem is *NP*-hard). Attempts are usually made to improve the efficiency of the random search procedure by modifying the dynamics of a descent over the landscape [1–3] described by $E(S)$. In contrast to this approach, we propose not to change the dynamics of landscape descent but rather to transform the energy landscape itself so as to increase the radius of the attraction domain of the global minimum (and of other minima comparable in depth with the global one).

In previous work [9], we consider the simplest transformation, namely, the raising of T to the power $k=2,3,...$. This approach was found to be productive: due to the landscape transformation, the spectrum of found minima is strongly shifted towards the deep side and the probability of finding the global minimum increases by 10^3 times. It was shown that the optimal value of power is $k=3$. But the algorithm is unstable at $k\geq 3$: for the most part (about 70% of instances) the probability of finding global minima increases more than by 3 orders of magnitude in average, but sometimes (the rest 30%) it may decrease up to zero.

In present paper, we suggest to use a mix-matrix M , i.e., a mixture of T and T^2. We claim that this yields a more reliable approach.

The efficiency of the algorithm proposed is rigorously substantiated only for “random” matrices, whose elements generated as independent random variables. The application of the algorithm to matrices of other types is heuristic.

The paper is constructed as follows. Section 2 includes the problem definition. Some preliminaries concerned the energy landscape of quadratic functional are given in Section 3. We describe the suggested idea regarding mix-matrix in Section 4. In Section 5, it is shown how the mix-matrix transforms the energy landscape of quadratic functional. Section 6 contains the obtained results for matrices of two types (uniform matrices and matrices of 2D Ising model).

II. PROBLEM DEFINITION AND MINIMIZATION PROCEDURE

The standard statement of the binary minimization problem is as follows. Given an $N\times N$ matrix T , find an N -dimensional configuration vector $S_m=(s_{m1}, s_{m2}, ..., s_{mN})$, $s_{mi}=\pm 1$, $i=1,2,...,N$, that minimizes the energy functional $E(S)$:

$$E(S)=-\frac{1}{\sigma_T N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}T_{ij}s_i s_j \ , \qquad (1)$$

where σ_T is the standard deviation of the matrix elements T_{ij} . Functional (1) can be symmetrized. For this reason, without loss of generality, we assume that the matrix T_{ij} is symmetric and its diagonal elements are zero ($T_{ii}=0$).

The minimization procedure is based on the Hopfield model [4], which is the core of most binary minimization algorithms. This is a one-dimensional system of N spins, whose interaction is defined by the energy functional $E(S)$. The standard (asynchronous) dynamics of the model can be described as follows (The full description is shown in Listing 1). The local field $h_i = -\partial E(S)/\partial s_i$ acting on the arbitrarily chosen *i*-th spin is calculated as

$$h_i = \frac{1}{\sigma_T N^2} \sum_{j \neq i}^{N} T_{ij} s_j \qquad (2)$$

If $h_i s_i < 0$, the state of the spin is updated according to the decision rule $s_i = sign\, h_i$. This procedure is sequentially applied to all the neurons until the network converges to a stable state S_m. This dynamics is a descent over the energy landscape $E(S)$, which is a complete analogue of the coordinate-wise gradient descent in a real space.

Listing 1. The program code of the dynamics.

```
algorithm Asynchronous Neural Network Dynamics
    Initialize S = (s_1, s_2...s_N) s_i = ±1
begin
    for i = 1 : N
        h_i = $\sum_{j \neq i}$ T_ij s_j          %calculate local fields
    end for
    flip = 1
    while (flip > 0)
        flip = 0
        for i = 1 : N
            if(h_i s_i < 0) then
                s_i = -s_i                  %reverse spin
                for j = 1 : N,  j ≠ i
                    h_j = h_j + 2T_ij s_i %refresh fields
                end for
                flip = flip + 1
            end if
        end for
    end while
end
```

NP-complete problems are known to have a huge number of local minima. In order to find a global one we have to use the random search. The random search procedure is described as follows. Given an arbitrarily initial state of the network, the nearest local minimum is found. This procedure is repeated until a minimum with an acceptable depth is found. The efficiency of the random search procedure is evaluated by the probability of finding the global minimum, by the rate of finding a minimum with a given depth, or by the mean depth of the minima found.

III. PRELIMINARIES

Before transforming the energy landscape, we establish the basic relations associated with the depth of the global (local) minimum, which underlie the subsequent argument.

The first relation is a constraint on the depth of the minimum. Let $S_0 = (s_{01}, s_{02}, ..., s_{0N})$ be the configuration corresponding to the global minimum $E_0 = E(S_0)$. We extract from T the term T_0 that is responsible for the formation of this minimum. To this end, T is represented as

$$T = T_0 + T_1 \quad , \quad T_0 = r_0 \sigma_T S_0^+ S_0 \qquad (3)$$

The statistical weight r_0 is found from the condition that the elements of T_0 and T_1 do not correlate. Calculating the covariance of the matrix elements and setting it equal to zero, we obtain

$$r_0 = -\frac{E_0 + \bar{T}\delta}{1-\delta^2}, \quad \delta = \frac{1}{N^2}\left[\left(\sum_{i=1}^{N} s_{0i}\right)^2 - N\right], \qquad (4)$$

where $\bar{T}$ is the mean of the elements of T and δ is a variable with a zero mean and a small standard deviation $\sigma_\delta = \sqrt{2}/N$. For simplicity, we set $\bar{T} = 0$ and $\delta = 0$ (the generalization to other cases is obvious). Then (4) yields the relation

$$E_0 = -r_0 \qquad (5)$$

The variances of the elements of T_0 and T_1 are $\sigma_0^2 = r_0^2 \sigma_T^2$ and $\sigma_1^2 = \sigma_T^2 - \sigma_0^2$. Therefore, we have managed to present the random matrix T as the sum of two independent random matrices T_0 and T_1. Moreover, (3) and (4) imply that $S_0 T_1 S_0^+ = 0$, which suggests that the contribution of T_1 to E_0 is strictly zero; i.e., the minimum in S_0 is caused only by the contribution of T_0.

Following [5], we continue decomposition (3) and represent the matrix as a weighted sum of exterior products of random vectors:

$$T = \sigma_T \sum_0^\infty r_m S_m^+ S_m \ , \qquad \sum r_m^2 = 1 .$$

For this type of matrices, it was shown in [6] that any of the vectors S_m present in the decomposition of T is a minimizer of functional (1) if and only if its weight r_m is larger than the critical value

$$r_c = \frac{1}{2\sqrt{0.138N}} \qquad (6)$$

This assertion is concerned primarily with the point S_0, which by definition is a minimizer of functional (1) and satisfies the relations

$$1 \geq r_0 \geq r_c\, , \ E_c \geq E_0 \geq -1\, , \quad E_c = -r_c$$

ISBN: 978-1-61208-202-8

The second necessary relation obtained in [7] is that, as the depth of minimum E_0 increases, its width increases as well and, accordingly, the probability of finding this minimum grows as $P(E_0) \sim \exp\left(-NE_c^2 / E_0^2\right)$

As a result, we have established the following two relations:

– For a larger weight r_0, the minimum E_0 is deeper and the probability of finding it is higher.

– S_0 can be a minimum only if $r_0 \geq r_c$; i.e., the depth of the minimum is larger than the critical value $|E_c|$.

These relations suggest the direction of improving the efficiency of the random search algorithm: the energy landscape (1) has to be transformed so as to increase the depth of the global minimum and, accordingly, to increase the probability of finding it.

IV. THE ALGORITHM

In this section we describe the proposed minimization algorithm. The main idea underlying the algorithm is the transformation of energy landscape of the functional. The surface described by the quadratic form $E(S)$ can be transformed only by transforming the underlying matrix.

Let us define the *mix-matrix M as*:

$$M = \frac{1-z}{\sigma_T} T + \frac{z}{\sigma_{2T}} T^2 , \qquad (7)$$

where T^2 is obtained by raising T to the second power and setting the diagonal elements equal to zero, σ_T and σ_{2T} are the standard deviations of matrices T and T^2 respectively. Substitute the new matrix into (1). Changing the parameter z from 0 to 1, we pass from the matrix T to T^2. Accordingly, the landscape described by $E(S)$ is transformed into that described by:

$$E_z(S) = -\frac{1}{\sigma_M N^2} \sum_{i=1}^{N} \sum_{j \neq i}^{N} M_{ij} s_i s_j \qquad (8)$$

where σ_M is the standard deviation of M_{ij}. Obviously, under the landscape transformation, the global minimum is shifted in space and its depth and the width of the attraction domain change as well.

Accordingly, we propose the following minimization algorithm. Firstly, we choose a value z, then construct the mix-matrix (7) and accordingly the functional $E_z(S)$. Then we start the minimization procedure consisting of two steps:

– At the first step, a descent over $E_z(S)$ is performed and a configuration S_{zm} is found that minimizes $E_z(S)$.

– The second step involves correction, namely, from the point S_{zm}, we descend over $E(S)$ to the nearest minimum S_m of $E(S)$.

Listing 2. The program code of the proposed algorithm.

```
algorithm Mix - matrix algorithm
    Initialize S = ($s_1$, $s_2$...$s_N$) $s_i$ = ±1
    Initialize the mix - matrix $M$ with certain $z$
begin
         %1.Descent over transformed landscape
    for i = 1 : N
        $h_i = \sum_{j \neq i} M_{ij} s_j$          %calculate local fields
    end for
    flip = 1
    while (flip > 0)
        flip = 0
        for i = 1 : N
            if($h_i s_i$ < 0) then
                $s_i$ = -$s_i$                 %reverse spin
                for j = 1 : N, j ≠ i
                    $h_j$ = $h_j$ + 2$M_{ij} s_i$ %refresh fields
                end for
                flip = flip + 1
            end if
        end for
    end while
         %2.Descent over initial landscape
    for i = 1 : N
        $h_i = \sum_{j \neq i} T_{ij} s_j$          %calculate local fields
    end for
    flip = 1
    while (flip > 0)
        flip = 0
        for i = 1 : N
            if($h_i s_i$ < 0) then
                $s_i$ = -$s_i$                 %reverse spin
                for j = 1 : N, j ≠ i
                    $h_j$ = $h_j$ + 2$T_{ij} s_i$ %refresh fields
                end for
                flip = flip + 1
            end if
        end for
    end while
end
```

The descent over $E_z(S)$ is performed as described above: we calculate the local field of the ith spin $h_i^{(z)} = -\partial E_z(S) / \partial s_i$ and, if $h_i^{(z)} s_i < 0$, the state of the spin is updated according to the decision rule $s_i = sign\, h_i^{(z)}$. The full description of the algorithm is given in Listing 2.

In previous work [9] we consider the simplest transformation, namely, when $M = T^k$, $k = 2,3,4,5$. It was shown that the optimal value of power is $k = 3$. In this case the probability of finding global minima increases by 3 orders of magnitude for the most part (about 70% of instances). But sometimes (the rest 30% ones) it may decrease up to zero due to vanishing a minimum near S_0.

As a result of this, in present paper, we introduce a mix-matrix (7), i.e., a mixture of T and T^2, and vary the parameter z from 0 to 1. This yields a more reliable approach.

We will show that at $z \approx 0.5$ the proposed transformation leads to significant increase of the global minimum depth, while the shift from the minimum is smaller ($1-2\%$ of N) than in case $M = T^3$ (3% of N).

V. Correctness of the Algorithm

The algorithm is substantiated only for “random” matrices, whose elements are independent random variables. The application of the algorithm to matrices of other types is heuristic.

A. The deepening of the minima.

Let us show that the landscape transformation leads to a deeper minimum. Consider the energy $E_{z0} = E_z(S_0)$ at the point S_0. Following (3), the mix-matrix M is represented as

$$M = (1-z)\frac{T_0 + T_1}{\sigma_T} + z\frac{T_0^2 + T_0T_1 + T_1T_0 + T_1^2}{\sigma_{2T}}$$

In view of $S_0 T_1 S_0^+ = 0$ and $\sigma_{_M}^2 = (1-z)^2 + z^2$, we then derive from (8) that

$$E_{z0} = \overline{E}_{z0} + R \tag{9}$$

where

$$\overline{E}_{z0} = -\frac{(1-z)r_0 + zr_0^2\sqrt{N}}{\sqrt{(1-z)^2 + z^2}}$$

$$R = \frac{1}{N^2\sqrt{(1-z)^2 + z^2}}\sum_{i=1}^{N}\sum_{j\neq i}^{N}\left(\frac{(1-z)T_1}{\sigma_T} + \frac{zT_1^2}{\sigma_{2T}}\right)_{ij} s_{0i}s_{0j}$$

In the limit of $N >> 1$, E_{z0} can be viewed as a normally distributed quantity with the mean value $\overline{E}_{z0}$ and the relatively small noise R of standard deviation $\sigma_R = 1/N$. The ratio:

$$\frac{\overline{E}_{z0}}{E_0} = \frac{(1-z) + z\sqrt{N}r_0}{\sqrt{(1-z)^2 + z^2}} \tag{10}$$

shows how many times the average value of the modified functional at point S_0 more than the initial functional value at the same point. Taking into account $\sqrt{N}r_0 \approx 1.35$, it is obvious that at any value of z expression (10) is larger than unit, hence when $N >> 1$ one can be sure that the minimum becomes deeper. Fig. 1 confirms this. The largest deepening ($E_{z0} \approx 1.6E_0$) is observed at $z \approx 0.6$.

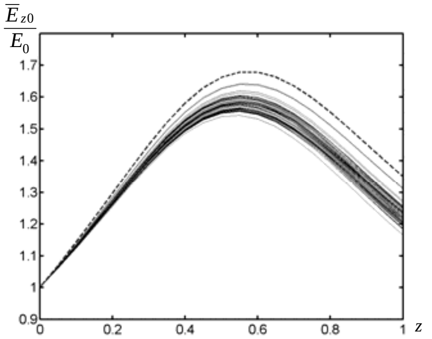


Figure 1. The decrease of energy in the point S_0 (global minimum) due to energy landscape transformation (mix-matrix with T^2). The dashed line is theoretical (10). Other lines are experimental for 50 random instances with uniform matrices.

B. The shift of the minima.

Let us estimate the shift of the minimum under the landscape transformation. The mean shift can be represented as

$$d = N \cdot P\,,$$

where $P = \Pr\{s_{0i}h_i^{(z)} < 0 \mid s_{0i}h_i > 0\}$ is the probability that the directions of the spin s_{0i} and the local field $h_i^{(z)}$ do not coincide. Omitting the unnecessary constants, the value $s_{0i}h_i^{(z)}$ can be represented as

$$h_i^{(z)}s_{i0} = (1-z)Nr_0 + zN^{3/2}r_0^2 + H \tag{11}$$

where

$$H = \sum_{i=1}^{N}\left(\frac{(1-z)T_1}{\sigma_T} + \frac{z(Nr_0\sigma_T T_1 + T_1^2)}{\sigma_{2T}}\right)_{ij} s_{0i}s_{0j}$$

In view of (11), P is expressed in terms of the error function:

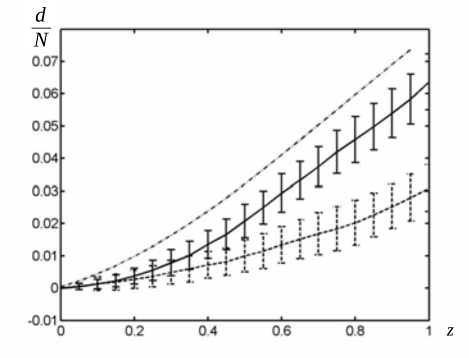
$$P = \frac{1}{2\Phi(\gamma)\sqrt{2\pi}}\int_0^\infty dx\, e^{-\frac{1}{2}(x-\gamma)^2}\left(1-\Phi\left(\alpha\sigma x\right)\right), \tag{12}$$

where $\Phi(\cdot)$ is the probability integral and

$$\gamma = \sqrt{N}r_0 / \sigma \approx 1.9\ ,$$

$$\sigma = \sqrt{\frac{1}{\sigma_T^2 N}\sum_{i=1}^{N}h_i^2} \approx 0.7\ ,$$

$$\alpha = \frac{1-z}{z} + \sqrt{N}r_0$$



.

Figure 2. The shift (in bits) of the global minimum as a function of z (mix with T^2). The curves with error bars were obtained by experiment for two types of matrices: matrix with uniformly distributed elements (solid line) and 2D Ising matrices (dashed lines). The dash-dot line is theoretical (12).

Note that at $z = 0$ the functional $E_{z=0}(S)$ coincides with initial $E(S)$ and therefore the shift is absent, this agrees ($d = N \cdot P = 0$) with (12).

The formula (12) describes a monotone increase of the minimum shift with growing z in view of enlarging functional transformation. This corresponds to a common sense and is proved by experiment (see fig. 2).

Expressions (9)–(12) suggest the following conclusions. With a high probability, the landscape transformation leads to deeper minima and, as a result, to a higher probability of finding them. Moreover, the depth increase (10) is larger for a larger initial depth $|E_0| \approx r_0$. In other words, deep minima become even deeper and the probability of finding them increases, while shallow minima become shallower (or disappear at all) and the probability of finding them is reduced. This means that the spectrum of minima found by the algorithm shifts considerably toward the global minimum, and the probability of finding the latter increases considerably. The spatial minima displacements caused by the transformation are relatively small: it follows from (12) that the smallest shifts are expected for the deepest minima.

VI. RESULTS

The efficiency of the two-step descent algorithm was verified for z ranging from 0 to 1 for matrices of size $N = 100, \ldots 500$ of two types:

- matrices with random elements uniformly distributed within $(-1;1)$;
- matrices of 2-dimensional Ising model with [2].

During numerical experiments we built a mix-matrix for different values of z from 0 to 1 equally spaced with $\Delta z = 0.05$. The results were averaged over 50 random instances of each size and type.

The computational complexity of the algorithm is $O(N^2)$. In experiments, we used the same algorithm realization both for sparse and dense matrices, although it is possible to reduce the complexity up to $O(N)$ in case of 2D Ising matrices.

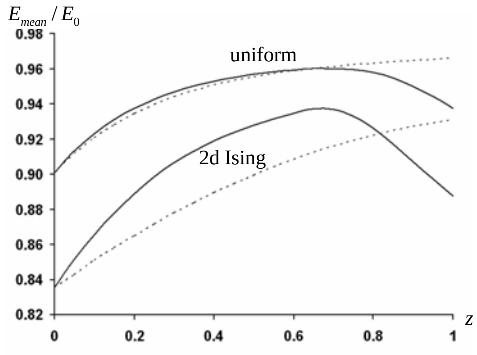


Figure 3. The mean value E_{mean} of energy of local minima found with the proposed two-step algorithm. The solid lines are for mix-matrices with T^2. The dashed lines are for mix-matrices with T^3. The curves are drawn for two types of matrices: uniform matrices (on top) and 2D Ising matrices. The value of E_{mean} is divided by the energy of global minimum E_0 and does not depend on the problem dimension N .

In addition, some part of the running time is spent on matrix multiplication. Nevertheless, each our experiment took no more than one hour for $N \simeq 500$.

Each experiment included $N_{runs} = 10^6$ runs. Each run resulted in a local minimum. We chose two parameters to trace: the mean energy E_{mean} of the minima found and the probability of finding a minimum in the interval of energy close to global one $E \in [-1; -0.99]$, where -1 corresponds to E_0 .

In experiments, we try to use not only the mix-matrix (7) but also mix with T^3 . In this case the mix-matrix was constructed in the same manner (7) but T^2 was replaced with T^3.

The numerical results are shown in Figs. 3 and 4.

Fig. 3 demonstrates how the mean value of energy E_{mean} of minima found for different z changes. It is interesting that the value E_{mean} / E_0 does not depend on the problem dimension but the type of the matrix.

As we can see from Fig. 3, E_{mean} comes near E_0 with increasing z . We observe the maximum of E_{mean} / E_0 at $z \approx 0.7$ for mix-matrix with T^2 and the monotone growth of E_{mean} / E_0 for mix-matrix with T^3 up to $z = 1$.

Fig. 4 shows how many times the probability of finding minima with energy differed from the global one less than 1% increases. For demonstration purpose, we chose the maximal possible problem dimensions, which we can cope with. For 2D Ising matrices the probability of finding minima of energy $E \in [-1; -0.99]$ is not greater than $P_1 = 3 \cdot 10^{-7}$ for $N = 12 \times 12$. For uniform matrices the maximal dimension is $N = 500$ (the probability $P_1 = 3 \cdot 10^{-5}$). The probability obtained with the proposed algorithm was denoted by P_{new} . As we can see from fig. 4, the difference between P_{new} and P_1 turned out to be enormous – approximately 3 orders of magnitude.

An interesting fact is that for uniform matrices the T^2 and T^3 curves almost coincide (see. fig. 3-4), and they start to diverge when $z > 0.7$ only. For Ising matrices, we have another picture: mix with T^2 prevails over mix with T^3 up to $z \approx 0.8$ and after that vice versa.

It can be also seen from Fig. 4 that with increasing z the dispersion rises, and this can lead to the instability of the algorithm, i.e., the transformation may change the search procedure for the worse in some cases.

VII. Conclusion

Finally, we formulate the minimization algorithm proposed.

The preliminary phase consists of the following steps. The original matrix T is symmetrized (if it is initially not symmetric) and its diagonal elements are set to zero. The matrix is raised to the k^th power ($k = 2$ or 3) and the diagonal elements in the resulting matrix T^k are set to zero. Afterwards the matrices T and T^k are normalized on unit dispersion and mixed in accordance with (7), and the mix-matrix M is obtained. It depends on the chosen parameter $z \in (0,1)$. The functional $E_z(S)$ is constructed from M according to (8).

After the preliminary phase, the random search procedure based on the two-step descent algorithm is executed. Specifically, at the first step, a descent over surface $E_z(S)$ is performed from a random initial configuration to the nearest local minimum S_{zm} of $E_z(S)$. The second step involves correction: from the point we descend over surface $E(S)$ to the nearest local minimum S_m of $E(S)$, which is, as a rule, located near S_{zm} .

The simplest Hopfield neural network dynamics [4] was chosen as a descent dynamics (nevertheless, it can be arbitrary).

A comparison shows that the efficiency of the minimization algorithm is improved substantially due to the landscape transformation.

It was shown that we succeeded in decreasing the value $(E_0 - E_{mean})/E_0$ (difference between the mean energy of found minima and global one) by half when $k = 2$ and $z = 0.7$ (see fig. 3).

Due to the proposed method the probability of finding suboptimal solutions with energy differed from the optimum less than 1% increases by 2.5 orders of magnitude for uniform (full) matrices of dimension $N = 500$ and by more than 3 orders for (sparse) matrices of Ising model of dimension $N = 12 \times 12$.

Finally, it seems to be attractive to use the proposed algorithm as a preliminary stage of any sort of genetic algorithms. Indeed, one can run our algorithm with different values of z to obtain a set of minima. Most of the minima must be deep and some of them may lie near the global minimum. Therefore, they are the good candidates for being parents.

Acknowledgment

The work was supported by the program of the Presidium of the Russian Academy of Sciences (project 2.15) and in part by the Russian Basic Research Foundation (grant 12-07-00295).

References

[1] B.W. Kernighan and S. Lin, “An Efficient Heuristic Procedure for Partitioning Graphs”, Bell System Tech. Journal, 49, pp. 291– 307, 1970.

[2] A.K. Hartmann and H. Rieger, New Optimization Algorithms in Physics (Wiley, Weinheim), 2004.

[3] C. Dang, W. Ma, and J. Liang, “A deterministic annealing algorithm for approximating a solution of the min-bisection problem”, Neural Networks 22 (1), pp. 58–66, 2009.

[4] J.J. Hopfield, “Neural Networks and physical systems with emergent collective computational abilities”, Proc.Nat.Acad.Sci.USA. vol. 79, pp. 2554-2558, 1982.

[5] B.V. Kryzhanovsky, “Expansion of a matrix in terms of external products of configuration vectors”, Optical Memory and Neural Networks 6 (4), pp. 187–199, 2007.

[6] D.J. Amit, H. Gutfreund, and H. Sompolinsky, “Spin-glass models of neural networks”, Phys. Rev. A, vol. 32, pp. 1007-1018, 1985; Annals of Physics, vol. 173, pp. 30-67, 1987

[7] B.V. Kryzhanovskii, B.M. Magomedov, and A.L. Mikaelyan, “A Relation Between the Depth of a Local Minimum and the Probability of Its Detection in the Generalized Hopfield Model”, Doklady Mathematics, vol. 72, N3, pp. 986-990, 2005.

[8] B.V. Kryzhanovsky and V.M. Kryzhanovsky, “The shape of a local minimum and the probability of its detection in random search”, Lecture Notes in Electrical Engineering, Vol. 24, pp. 51-61, 2009.

[9] Y.M. Karandashev and B.V. Kryzhanovsky, “Transformation of Energy Landscape in the Problem of Binary Minimization”, Doklady Mathematics, v. 80, No. 3, pp. 927-931, 2009.

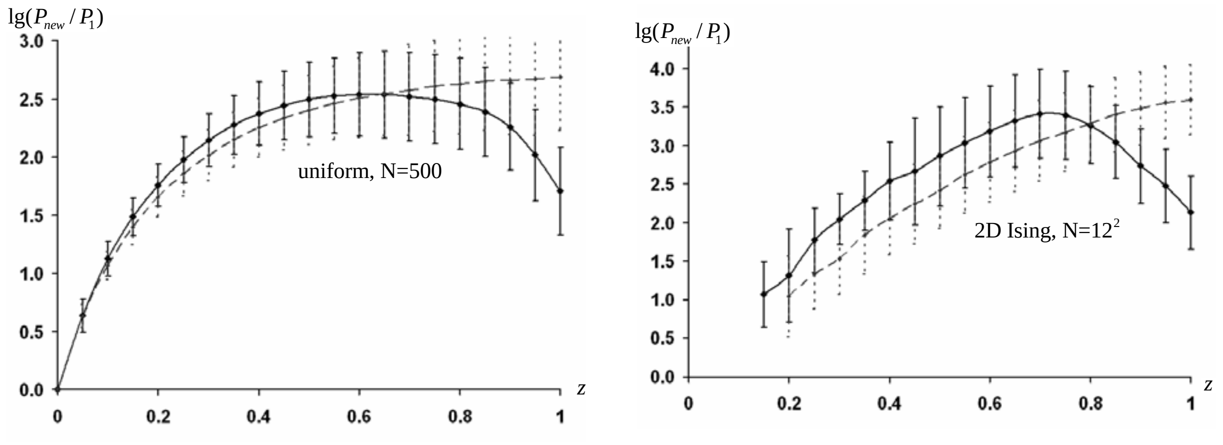


Figure 4. The common logarithm of the ratio of probabilities of hitting the energy interval $E \in [-1, -0.99]$. The solid lines are for mix-matrices with T^2 . The dashed lines are for mix-matrices with T^3 . In the left panel the results for uniform matrices of $N = 500$ (. $P_1 \approx 3 \cdot 10^{-5}$). In the right panel the results for 2D Ising matrices of $N = 144$ (. $P_1 \approx 2.6 \cdot 10^{-7}$). Note, that when z is too small, the algorithm does not find the global minimum in some instances, so the points are missed.

Minimising Expected Misclassification Cost when using Support Vector Machines for Credit Scoring

Terry Harris, Curtis Gittens
Dept. of Computer Science, Mathematics & Physics
University of the West Indies - Cave Hill Campus
Bridgetown, Barbados
terry.harris@mycavehill.uwi.edu, curtis.gittens@cavehill.uwi.edu

***Abstract*— With the gradual relaxation of credit around the world, the cost of losses experienced when extending credit is expected to become increasingly important to financial institutions. In this paper, we offer theoretical and empirical evidence to support the argument that the minimisation of this cost should be the primary objective when developing classification models for credit scoring. This cost can be referred to as the Expected Misclassification Cost. In addition, we present and test a system that builds models to minimise this cost when given varying values for its components. Moreover, we show that using differing values for the components of Expected Misclassification Cost can result in improved performance, in terms of Type I or Type II accuracy, when Expected Misclassification Cost is used as the prime evaluation metric by a support vector machine.**

Keywords- Credit Scoring; Decision Support Systems; Expected Misclassification Cost ; Support Vector Machines

I. INTRODUCTION

The assessment of credit risk is a very important task for financial institutions. This is in part due to the need to avoid losses associated with inappropriate credit approval or rejection decisions [1]. In recent years, credit scoring has emerged as one of the primary ways for financial institutions to assess credit risk [2]. Credit scoring entails the classification of potential customers into applicants with good credit and applicants with bad credit. This is done by analysing the applicant’s data based on a past pattern of customer behaviour [3].

Since Fisher’s [4] seminal paper, numerous models have been proposed, which attempt to differentiate between “good” and “bad” credit applicants. Many of these classification models are based on classical statistical methods such as Discriminant Analysis (DA), Linear or Polynomial Regression (LPR), Logistic Regression (LR), Non-Parametric Models (NPMs), Artificial Neural Networks (ANNs), and Support Vector Machines (SVMs) [5], [6], [7], [8], [9], and [10].

Whatever its form, many existing credit scoring models are built on samples of customer historical data, and their primary objective is to avoid over-fitting while maximising generalisablity from the samples [5]. As a result, improving test accuracy, as in (1), which is the measure of how accurately the model classifies credit applicants from a withheld dataset, known as the test dataset, is of importance [5] and [6]. However, this approach alone can lead to unsatisfactory results if the cost of making one type of error as opposed to another is not considered. We propose that credit scoring models can be improved if they are designed to minimise this type of cost called the Expected Misclassification Cost [11].

$$\text{Test Accuracy} = \frac{\text{True Positive}}{\text{True Positive+False Positive}} + \frac{\text{True Negative}}{\text{False Negative+True Negative}} \quad (1)$$

The remainder of this paper is organised as follows. In Section II, we discuss some of the problems which emerge when using test accuracy as the primary model evaluation metric, before discussing the rationale behind the use of the Expected Misclassification Cost as the model evaluation metric. In Section III, the Support Vector Machine algorithm, which is the classification algorithm implemented in our system, is discussed. The details of the dataset chosen as our case study are presented in Section IV. Described in Section V, is our parameter tuning algorithm and the methodology of the study. Section VI, discusses the results of the study, and Section VII highlights the conclusions and directions for future research.

II. BACKGROUND

A. Skewed Datasets

When a classification model is designed to minimise test accuracy as its main objective, this can prove problematic if the training dataset is skewed in favour of one particular class over another (as is often the case in credit scoring exercises). This is because it becomes difficult to determine if higher test accuracy corresponds to an improved quality classifier. The following example illustrates this point.

Suppose we have a classifier that gives a test accuracy of 99% when determining the creditworthiness of clients. At first glance, this system seems to be a good classification model. However, if the probability of a potential customer being un-creditworthy is 0.5%, it becomes clear that test accuracy tells us nothing about the quality of the classifier because 99.5 % test accuracy can be achieved by classifying all applicants as creditworthy. Without a doubt, this second approach is unacceptable, because by simply approving all applicants, we are not detecting potentially "bad" clients.

To solve this problem, many researchers often use the Precision, as in (2), and Recall, as in (3), evaluation metrics. Precision is the measure of how accurately we have classified our positive predictions (what fraction is correctly categorised), while Recall measures the proportion of the dataset, which was actually positive, that we predicted as positive. Given our previous scenario, the algorithm that simply predicts that the applicant was creditworthy 100% of the time would continue to score 99.5% on test accuracy; however, it would score 0% accuracy on the Recall evaluation metric. As a result, tailoring classification models to improve Precision and/or Recall can help to improve classifier quality when the dataset is skewed.

$$\text{Precision} = \frac{\text{True Positive}}{\text{\# Predicted as Positive}} = \frac{\text{True Positive}}{\text{True positive+False Positive}} \quad (2)$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{\# Actually Positive}} = \frac{\text{True Positive}}{\text{True Positive+False Negative}} \quad (3)$$

B. *Minimising Type I and Type II errors*

Another issue that arises when using total test accuracy as the performance metric to develop credit scoring models, is the problem of minimising Type I error, as in (4), and Type II error, as in (5). If we let the null hypothesis on any credit approval decision be that the credit applicant is un-creditworthy, then a Type I error occurs when we reject the null hypothesis that the potential customer is un-creditworthy and grants them credit when we should have rejected their application. Conversely, a Type II error occurs when we accept the null hypothesis (that the applicant is un-creditworthy) when we should have rejected it, and grant the client credit. Developing a model to maximise Precision and Recall using the F_1 Score, as in (6), which is a type of average for Precision and Recall, can assist with minimising both of these errors. Furthermore, models could be developed to minimise Type I and Type II errors separately and/or jointly. However, focusing solely on effectively minimising Type I and II errors or maximising Precision and Recall does not take into consideration the misclassification cost to the institution of making one type of error over another [11]. We believe that existing credit scoring models could be enhanced if this expected cost is taken into consideration when developing the model.

$$\text{Type I Error} = \frac{\text{False Positive}}{\text{True Negative+False Positive}} \quad (4)$$

$$\text{Type II Error} = \frac{\text{False Negative}}{\text{True Positive+False Negative}} \quad (5)$$

$$F_1\ Score = 2\frac{Recall*Precision}{Recall+Precision} \quad (6)$$

The Expected Misclassification Cost, as in (7), is comprised of two component costs associated with each type of inappropriate credit granting decision or error. Where the variable Z, represents the Expected Misclassification Cost, X the Default Cost, Y the Opportunity Cost, the variable a, the probability of Type I error, and b the probability of Type II error.

$$Z = Xa + Yb \quad (7)$$

The Expected Default Cost is associated with making Type I errors. This type of error can have the most damaging effect on the institution as it often leads to the loss of credit principal and interest. This cost can be quantified as the net present value of the credit principal and interest (base rate plus margin*principal), multiplied by the probability of Type I error. The second error, Type II is associated with the Expected Opportunity Cost of rejecting a potential client who would have been creditworthy. As a result, this cost is simply the net present value of the interest (net interest rate spread*principal) that could have been made, had credit been granted, multiplied by the probability of Type II error.

C. *Motivation*

Intuitively, for credit-granting decisions, Type I errors should be weighted with higher importance than Type II errors [10]. This belief is due to the fact that when a financial institution grants credit to a customer who later defaults, the financial institution potentially loses 100% of the principal and interest on the investment. This is often a higher cost than the opportunity cost of making a Type II error, which is usually limited to the loss of interest on the investment. However, to seek to minimise Type I error while ignoring its impact on Type II error (as they are inversely related) could lead to increased Expected Misclassification Cost to the institution. This can be seen by the following simplified example.

Suppose an institution seeks to minimise Type I error while ignoring its impact on Type II. One way of achieving this would be to simply cease granting credit. However, if this was done, then the institution would face massive opportunity costs because it would not be earning interest. This means that there must be some optimal value for both Type I and Type II errors such that Expected Misclassification Cost to the institution is minimised.

We present a system that produces credit scoring models which classify credit applicants as either creditworthy or un-creditworthy, such that the Expected Misclassification Cost to a financial institution is minimised. In addition, we present a parameter tuning algorithm which selects the parameters *Gamma* and *C* for the SVM (RBF kernel) such that Type I and/or Type II errors are optimised when weighted according to default cost and opportunity cost. We verify our results by testing our system using the LIBSVM (RBF kernel), which is a state of the art SVM by Chang and Lin [12].

III. SVMS AND CREDIT SCORING

The SVM was first developed by Cortes and Vapnik [13] for binary classification. To do this binary classification, SVMs attempt to find the optimal separating hyperplane between classes by maximising the margin (Fig. 1). The points lying on the boundaries are called support vectors, and the middle of the margin is referred to as the optimal separating hyperplane. This margin maximisation characteristic of SVMs is argued to improve the decision boundaries and hence lead to better classifier quality.

A. SVM use in Credit Scoring

Over the past decade, SVMs have been successfully used in many credit scoring systems [14], [15], [16], [17], and [18]. However, the superiority of the SVM when compared to other classifiers remains debatable, as Van Gestel et al. [16] found that even though SVMs showed improved performance, there was no significant difference between SVMs, LR and LDAs. This finding supports a widely held view that modern learning algorithms approximate each other's performance when given large datasets [19]. Consequently, although we use SVMs to implement our credit scoring system we suspect that other classification techniques may approximate or even outperform our system once designed to minimise Expected Misclassification Cost.

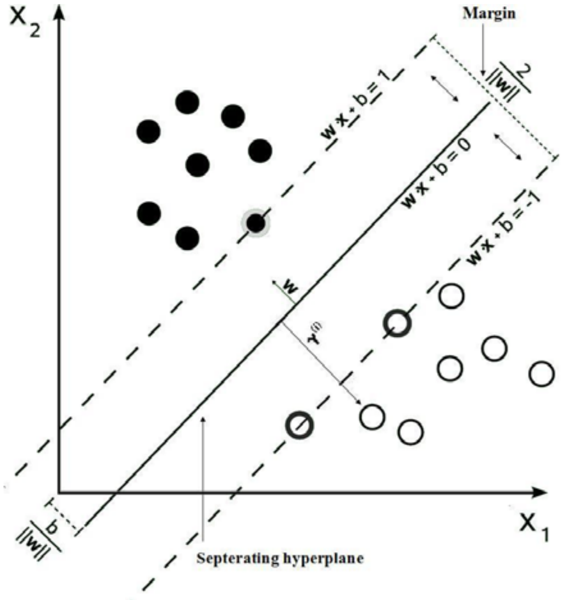


Figure 1: Simplified Depiction of SVM Classification

B. SVMs Development for Credit Scoring

When a financial institution is presented with a new credit applicant, in order to make the credit approval decision the institution seeks to classify the applicant as either "good" or "bad" according to the SVM score. In the case of a linear SVM this score can be represented as the linear combination of the applicant's characteristics (features) multiplied by some weights, as in (8).

$$z = w_1x_1 + w_2x_2 + \cdots + w_nx_n + b \qquad (8)$$

Where *n* represents the number of client features, the *w*'s and *b* are learnt parameters, and the *x*'s are client features. Transforming the *w*'s and *x*'s into column vectors, (8) can be written more concisely as;

$$z = w^T x + b. \qquad (9)$$

The SVM learns the parameters *w* and *b* from training examples of historic client data that the financial institution collected over time. This training dataset will normally consist of a number of example clients; as a result, from a geometric perspective, calculating the value of *w* and *b* means looking for a hyperplane which best separates "good" clients from "bad". To do this, the SVM maximises the margin between the two clouds of data. As a result, when given a training example ($x^{(i)}$, $y^{(i)}$), such that $y \in \{-1,1\}$, the functional margin $\hat{\gamma}$, of (*w*, *b*) can be defined with respect to the training example as;

$$\hat{\gamma} = y^{(i)}(w^T x \ + \ b). \qquad (10)$$

In order to confidently predict the class of the training example the functional margin needs to be large. Thus, if $y^{(i)} = 1$, then for the functional margin to be large $w^T x + b$ must be a large positive number. As a result, if $y^{(i)} = -1$, then $w^T x + b$ needs to be a large negative number. Accordingly, given a training set $S = \{(x^{(i)}, y^{(i)}); i = 1,\ldots, m\}$, the function margin of (*w*, *b*) with respect to *S* is defined as the smallest of the functional margins of the training examples, as in (11).

$$\hat{\gamma} = \ \min_{i=1,\ldots,m} \hat{\gamma}^{(\mathrm{i})} \qquad (11)$$

To find the geometric margin, γ, consider the case of a positive training example where $x^{(i)}$ corresponds to the label $y^{(i)} = 1$. The distance from this point to the decision boundary, $\gamma^{(\mathrm{i})}$, is a straight line (vector) orthogonal to the hyperplane (Fig. 1). To find the value of $\gamma^{(\mathrm{i})}$ the corresponding point on the decision boundary is found. This can be easily determined since *w*/||*w*|| is a unit-length vector pointing in the same direction as *w*. Therefore, the corresponding point on the hyperplane is given by the equation $x^{(i)} - \gamma^{(i)} \cdot w/||w||$, and because this point lies on the decision boundary, it satisfies the equation $w^T x + b = 0$ (Fig. 1), as in (12).

$$w^T \left(x^{(i)} - \ \gamma^{(i)} \frac{w}{\|w\|}\right) + b = 0 \qquad (12)$$

We can simplify (12) as following:

$$w^T x^{(i)} - \gamma^{(i)} \frac{w^T w}{\|w\|} + b = 0 . \qquad (13)$$

Since, $w^T w/\|w\| = \|w\|^2/\|w\| = \|w\|$, we solve for $\gamma^{(i)}$, as is shown in (14);

$$\gamma^{(i)} = (\frac{w}{\|w\|})^T x^{(i)} + \frac{b}{\|w\|}. \qquad (14)$$

Generalising this representation to account for negative training examples, we have;

$$\gamma^{(i)} = y^{(i)} [(\frac{w}{\|w\|})^T x^{(i)} + \frac{b}{\|w\|}]. \qquad (15)$$

Here, if $\|w\| = 1$, then the geometric margin is equal to the functional margin, In addition, the geometric margin is invariant to rescaling of the parameters (w, b). As a result, given a training set $S = \{(x^{(i)}, y^{(i)}), i = 1,\ldots,m\}$, the geometric margin is the smallest of the geometric margins on the individual training examples (16).

$$\gamma = \min_{i=1,\ldots,m} \gamma^{(i)} \qquad (16)$$

Accordingly, when given a training dataset of past clients, it seems natural that the financial institution would want to find a decision boundary that maximises the geometric margin, since this would reflect a very confident set of predictions on the training data. Specifically, this will result in a SVM classifier that separates “good” and “bad” past clients effectively, thus giving the institution reliable information with which to make judgments about future credit applications. As a result, to find the hyperplane that achieves the maximum geometric margin the following optimisation problem is posed:

$$\max_{\gamma,w,b} \gamma,$$

$$s.t.\ y^{(i)}(w^T x^{(i)} + b) \geq \gamma, i = 1, \ldots, m, \qquad (17)$$

$$\| w \| = 1.$$

However, because the $\|w\| = 1$ constraint is non-convex, the problem is transformed into one more suited for optimisation, as in (18). Here, if, $\hat{\gamma} = 1$, then $\hat{\gamma}/\|w\| = 1/\|w\|$, and maximising this is the same thing as minimising $\|w\|^2$.

$$\min_{\gamma,w,b} \frac{1}{2} \| w \|^2,$$

$$s.t.\ y^{(i)}(w^T x^{(i)} + b) \geq \hat{\gamma}, i = 1, \ldots, m. \qquad (18)$$

At this point, a regularisation term ξ, is added to the optimisation problem posed in (18) to modify the algorithm so that it works for non-linearly separable datasets, as is often the case with credit scoring data. The term C is a turning parameter which weights the significance of a classification error to the overall model.

$$\min_{\gamma,w,b} \frac{1}{2} \| w \|^2 + C \sum_{i=1}^m \xi_i,$$

$$s.t.\ y^{(i)}(w^T x^{(i)} + b) \geq 1 - \xi_i, i = 1, \ldots, m, \qquad (19)$$

$$\xi_i \geq 0, i = 1, \ldots, m.$$

Equation (19) represents the primal from of the optimisation problem for finding the optimal margin classifier to separate “good” and “bad” clients. Given that this equation satisfies the Karush-Kuhn-Tucker (KKT) conditions, the condition $g_i(w) \leq 0$ is an active constraint. As a result, the constant to the primal problem can be rewritten as follows:

$$g_i(w) = -y^{(i)}(w^T x^{(i)} + b) + 1 - \xi_i \leq 0. \qquad (20)$$

To develop the dual form of the problem, the Lagrangian for the optimisation problem is constructed, as in (21). Where the α_i ’s and the r_i’s are Lagrangian multipliers.

$$L(w,b,\xi,\alpha,r) \frac{1}{2} \| w \|^2 - c \sum_{i=1}^m \xi_i - \sum_{i=1}^m \alpha_i [y^{(i)}(w^T x^{(i)} + b) - 1 + \xi_i] - \sum_{i=1}^m r_i \xi_i \qquad (21)$$

Equation (21) is minimised with respect to w and b by taking partial derivatives with respect to w and b and setting them to zero. The equations derived are as follows:

$$\frac{\partial}{\partial w} L(w,b,\xi,\alpha,r) = w - \sum_{i=1}^m \alpha_i y^{(i)} x^{(i)} = 0, \qquad (22)$$

$$\frac{\partial}{\partial b} L(w,b,\xi,\alpha,\mathrm{r}) = \sum_{i=1}^m \alpha_i y^{(i)} = 0. \qquad (23)$$

Solving (22) for w produces;

$$w = \sum_{i=1}^m \alpha_i y^{(i)} x^{(i)}. \qquad (24)$$

Therefore, substituting the definitions of w (24) and b (23) in (21) and including the constraints $0 \leq \alpha_i \leq C$ and $\sum_{i=1}^m \alpha_i y^{(i)} = 0$ the dual optimisation problem is derived as;

$$W(\alpha) = \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j=1}^m y^{(i)} y^{(j)} \alpha_i \alpha_j < x^{(i)}, x^{(j)} >,$$

$$s.t.\ 0 \leq \alpha_i \leq C, i = 1, \ldots, m, \qquad (25)$$

$$\sum_{i=1}^m \alpha_i y^{(i)} = 0.$$

This dual form (25) can be solved in lieu of the primal problem, in order to derive the parameters α_i's that maximise $W(\alpha)$ subject to the constraints. These parameters can then be used in (24) to find the optimal *w*'s. Having found *w**, the primal problem can be used to find the optimal value for the intercept term *b*.

Accordingly, after the classification model has been trained, when presented with a new credit applicant the equation $w^T x + b$, would calculate and predict $y = 1$ if and only if this quantity is bigger than zero.

$$(w^T x + b) = (\textstyle\sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)})^T x + b \qquad (26)$$

Equation (26) can be rewritten as;

$$\textstyle\sum_{i=1}^{m} \alpha_i y^{(i)} < x^{(i)}, x > +b. \qquad (27)$$

This representation allows for the inclusions of kernels to deal more effectively with datasets which have multiple dimensions. Kernels map attributes to higher order feature spaces, and this is represented by replacing the *x*'s in the equation with the feature vector ϕ(*x*), as shown in (28).

$$\textstyle\sum_{i=1}^{m} \alpha_i y^{(i)} K(x^{(i)}, x) + b \qquad (28)$$

Where,

$$K\left(x^{(i)}, x^{(j)}\right) =< \phi(x^{(i)}), \phi(x^{(j)}) > . \qquad (29)$$

IV. Data

A German credit scoring dataset was taken from the UCI Machine Learning Repository [20]. This dataset was provided by Prof. Hofmann of Hamburg University and consists of 700 examples of creditworthy applicants and 300 un-creditworthy applicants. This dataset has been widely used in credit scoring research to evaluate the performance of classification models. The dataset measured twenty (20) features for each credit applicant comprising the following categories: the status of the client's existing checking account, the duration of the credit period in months, the client's credit history, the purpose for the credit, the credit amount requested, the client's savings account/bonds balance, the client's present employment status, the client's personal (marital) status and sex, whether the client is a debtor or guarantor of credit granted by another institution, the number of years spent at present residence, the type of property possessed by the client, the client's age in years, whether the client has other installment plans, the client's housing arrangements (whether they own their home, rent, or live for free), the number of existing credits the client has at the bank, the client's job, the number of people for whom the client is liable to provide maintenance for, whether the client has a telephone, and whether the client is a foreign worker.

The data was pre-processed so as to transform all categorical data into numerical data for analysis. In addition, the data was normalised so as to improve the performance of the SVM.

V. Algorithm and Methodology

A. *Parameter Tuning Algorithm*

```
Begin
1.  Randomly sort sample applicant dataset.
2.  Split sample dataset into 3 sub datasets.
    a.  Sub-dataset 1: Training (60%)
    b.  Sub-dataset 2: Cross Validation (20%)
    c.  Sub-dataset 3: Test (20%)
3.  For the # of parameters conduct grid-search
        Select the pair of parameters (C and
        Gamma) based on how well they minimise
        expected misclassification cost on the
        Training dataset using the CV dataset.
    End for
4.  Use the pair of parameters from part 3 to train the
    model using Training dataset.
5.  Test the model for overall Test, Type I, and Type II
    accuracies using the Test dataset (reported results).
6.  Re-train the model using the full dataset and the
    pair of parameters selected in part 3.
End
```

B. *Method*

Our empirical testing began by randomly sorting the dataset before splitting it into 3 sub-datasets; the training dataset, the cross validation dataset, and the test dataset. The initial step of randomly sorting the dataset was done in order to increase the probability of an equal distribution of clients across the 3 sub-dataset. To train for the minimisation of the components of the Expected Misclassification, we further subdivided the cross validation dataset into two data-files, each only containing positive or negative examples. To test for Type I and Type II accuracy the test dataset was also subdivided into two data-files, one with all positive and another with all the negative test examples.

We implemented our system in OCTAVE 3.2.4 and used it to repeatedly train models using the LIBSVM package fitted with a RBF Kernel. These models where built using the training dataset and certain values for the parameters *Gamma* and *C*. We used a grid search technique to find the parameters *Gamma* and *C* which minimised Expected Misclassification Cost using the cross validation dataset. When deciding on the search ranges for *C* and *Gamma* care was taken to ensure that ∃ *C* and ∃ *Gamma*, within the search ranges, which produced models that have zero Type I error, and zero Type II error (on two separate models). This was an important step to ensure that each component of Expected Misclassification Cost could be minimised to zero. The usual approach when selecting the parameter ranges is to use known benchmarks. However, these ranges may not be well-suited to every dataset and do not guarantee perfect Type I or Type II accuracy on any of the possible models.

Having found the pair of parameters which minimised the Expected Misclassification Cost on the cross validation dataset, we used them to build our models. Three models were built using varying assumptions for Default Cost and Opportunity Cost. This was done in order to illustrate the dynamic nature of our system. The results are presented in TABLE I.

VI. Results and Analysis

The first model shown in the TABLE I was built weighting Default Cost and Opportunity Cost equally. As a result, the minimisation of Expected Misclassification Cost equated to the minimisation of overall test accuracy. We use this model as a control to illustrate the variations in performance achievable if different weights are used when setting Default Cost and Opportunity Cost. This first model surpassed most contemporary classifiers in terms of Type I accuracy on this dataset (TABLE II). This performance is interesting because many existing SVM classifiers that have reported results on this dataset were highly optimised for performance while our system is not. The reason for our relatively superior performance could be attributed to the fact that we selected the parameter ranges to ensure that errors on both Type I and Type II error metrics could be minimised as low as possible. However, further investigation into this hypothesis needs to be conducted to confirm our intuition.

TABLE I. MODELS AND ACCURACIES

Model	Parameters		Accuracies (%)				
	Gamma	*C*	*Train*	*CV*	*Type I*	*Type II*	*Test*
1	2^{-50}	2^{57}	74.83	73.13	66.66	73.24	71.36
2	2^{-50}	2^{49}	71.16	69.84	75.45	66.20	68.84
3	2^{-50}	2^{41}	76.66	76.12	40.35	90.14	75.88

TABLE II. PERFORMANCE COMPARISONS

Models	Accuracies (%)		
	Type I Accuracy	*Type II Accuracy*	*Total Accuracy*
Model 1	66.66	73.24	71.36
Model 2	75.45	66.20	68.84
Model 3	40.35	90.14	75.88
Yu et al. [10]	53.57	90.33	78.46
Wang et al. [15]	45.62	89.44	76.30
Ahmad et al .[21]	66.66	88.08	81.42

The second model (TABLE I) was built with the objective of reducing Expected Default Cost (weighted Type I error), while placing less emphasis on Expected Opportunity Cost (weighted Type II error). To achieve this, Default Cost was set to one while Opportunity Cost was set to one-half. As a result, when the system selected parameters to minimise Expected Misclassification Cost, the Expected Default Cost was weighted twice as significant as the Expected Opportunity Cost. This process successfully achieved better performance (75.45%). As shown in TABLE II, this result surpassed the performance in terms of Type I accuracy of many of the known published SVM systems on this dataset, while still remaining relatively generalisable at 68.84% test accuracy. We attribute this performance to the fact that when given the input values for Default Cost and Opportunity Cost our system selected parameters for the model which placed more emphasis on the reduction of Expected Default Cost which is calculated based on Type I error. Focus was placed on Expected Default Cost because it was the primary contributor to Expected Misclassification Cost in this model.

The third model presented in TABLE I was built with the intention of reducing Expected Opportunity Cost (weighted Type II error), while weighting the impact of Expected Default Cost (weighted Type I error) with less importance. To achieve this, Default Cost was set to one-half, while Opportunity Cost was set to one. As a result, this model showed a 16.9% improvement in terms of Type II accuracy when compared to the control (Model 1). In addition, this model showed an improvement of 4.52% over the Model 1 in terms of test accuracy (75.88%). However, this model resulted in a 26.31% fall in terms of Type I accuracy. We attributed this occurrence to the fact that the model is weighted to select those parameters for *C* and *Gamma* which minmise the Expected Opportunity Cost since it had a greater impact on Expected Misclassification Cost in this model.

VII. Conclusion and Future work

In this paper, we presented a system for the minimisation of the expected cost to financial institutions when making credit granting decisions. We showed that the minimisation of this cost, which is referred to as the Expected Misclassification Cost, can be achieved by considering its components when building classifier models. In addition, we showed that this approach can lead to performance gains by increasing Type I and Type II accuracy.

Future work will consider the generalisablity of this approach to other classifiers and classification problems. In addition, other studies will investigate the advantages and disadvantages of using Expected Misclassification Cost as the primary model evaluation metric in combination with ensembles, bagging, boosting and other SVM performance enhancing techniques.

Acknowledgment

This research was partially funded by a Graduate Scholarship from the University of the West Indies, Cave Hill Campus. In addition, we would like to acknowledge fellow members of our research group (Sequitur), Mr. Gamal Crichton and Mr. Xavier Caddle, for their valuable comments and suggestions. Finally, we would like to thank the anonymous reviews, who helped to improve the readability of this paper.

REFERENCES

[1] L. Yu, S. Wang, and K. K. Lai, "Credit risk assessment with a multistage neural network ensemble learning approach," *Expert Systems with Applications,* vol. 34, pp. 1434-1444, 2008.

[2] C. L. Huang, M. C. Chen, and C. J. Wang, "Credit scoring with a data mining approach based on support vector machines," *Expert Systems with Applications,* vol. 33, pp. 847-856, 2007.

[3] L. Thomas, R. Oliver, and D. Hand, "A survey of the issues in consumer credit modelling research," *Journal of the Operational Research Society,* vol. 56, pp. 1006-1015, 2005.

[4] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Human Genetics,* vol. 7, pp. 179-188, 1936.

[5] Z. Huang, H. Chen, C. J. Hsu, W. H. Chen, and S. Wu, "Credit rating analysis with support vector machines and neural networks: a market comparative study," *Decision support systems,* vol. 37, pp. 543-558, 2004.

[6] Y. Wang, S. Wang, and K. Lai, "A new fuzzy support vector machine to evaluate credit risk," *Fuzzy Systems, IEEE Transactions on,* vol. 13, pp. 820-831, 2005.

[7] H. Li and J. Sun, "Predicting business failure using multiple case-based reasoning combined with support vector machine," *Expert Systems with Applications,* vol. 36, pp. 10085-10096, 2009.

[8] L. Yu, S. Wang, and J. Cao, "A modified least squares support vector machine classifier with application to credit risk analysis," *International Journal of Information Technology & Decision Making,* vol. 8, pp. 697-710, 2009.

[9] L. Zhou, K. K. Lai, and L. Yu, "Least squares support vector machines ensemble models for credit scoring," *Expert Systems with Applications,* vol. 37, pp. 127-133, 2010.

[10] L. Yu, X. Yao, S. Wang, and K. Lai, "Credit Risk Evaluation Using a Weighted Least Squares SVM Classifier with Design of Experiment for Parameter Selection," *Expert Systems with Applications,* pp. 15392-15399, 2011.

[11] D. J. Hand and W. E. Henley, "Statistical classification methods in consumer credit scoring: a review," *Journal of the Royal Statistical Society: Series A (Statistics in Society),* vol. 160, pp. 523-541, 1997.

[12] C. C. Chang and C. J. Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST),* vol. 2, p. 27, 2011.

[13] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning,* vol. 20, pp. 273-297, 1995.

[14] B. Baesens, T. Van Gestel, S. Viaene, M. Stepanova, J. Suykens, and J. Vanthienen, "Benchmarking state-of-the-art classification algorithms for credit scoring," *Journal of the Operational Research Society,* vol. 54, pp. 1082-1088, 2003.

[15] G. Wang, J. Hao, J. Ma, and H. Jiang, "A comparative assessment of ensemble learning for credit scoring," *Expert Systems with Applications,* vol. 38, pp. 223-230, 2011.

[16] T. Van Gestel, B. Baesens, J. A. K. Suykens, D. Van den Poel, D. E. Baestaens, and M. Willekens, "Bayesian kernel based classification for financial distress detection," *European journal of operational research,* vol. 172, pp. 979-1003, 2006.

[17] Y. C. Lee, "Application of support vector machines to corporate credit rating prediction," *Expert Systems with Applications,* vol. 33, pp. 67-74, 2007.

[18] T. Bellotti and J. Crook, "Support vector machines for credit scoring and discovery of significant features," *Expert Systems with Applications,* vol. 36, pp. 3302-3308, 2009.

[19] E. Brill, J. Lin, M. Banko, S. Dumais, and A. Ng, "Data-intensive question answering," 2001.

[20] A. Frank and A. Asuncion. UCI Machine Learning Repository [Online]. Available: http://archive.ics.uci.edu/ml [retrieved: January 21, 2012]

[21] G. Ahmad, R. Manoj, P. Dhirendra, S. Ugrasen, G. Neelesh, G. Roopam, K. Verma, K. K. Brajesh, S. Raghuvir, and S. Pushpa, "A Hybrid Support Vector Machine Ensemble Model for Credit Scoring," 2011.

A Comparison of Algorithms to Construct Ontologies from Relational Databases

Sameen Fatima
Faculty of Computer Science
Institute of Business Administration
Karachi, Pakistan
e-mail: shigon_sl@hotmail.com

Quratulain Rajput
Faculty of Computer Science
Institute of Business Administration
Karachi, Pakistan
e-mail: qrajput@iba.edu.pk,
quratulain.rajput@gmail.com

Abstract—**Relational databases (RDB) play a vital role in managing the organization's data but they are dependent on autonomous hardware and software and thus create a problem of data integration. On the other hand Ontologies are considered as one of the most popular solutions in knowledge representation as a universal language to share and integrate knowledge. To overcome the integration problem in database and to realize the vision of semantic web, data has to publish over a web as ontologies. The purpose of this research is to explore algorithms to construct ontologies automatically from relational databases. A comparison is made on the basis of degree of automation and accuracy to transform relational database into ontology. Finally, assessing the strength and weakness of each algorithm and explore the future research directions.**

Keywords- Ontology; Relational database; Mapping rules

I. INTRODUCTION

It is said that Information is power. If the information is separated and isolated from other data it cannot bring any value to the organization. Before the emergence of database systems it was difficult to manage this information. As the organizations engaged their resources in managing a lot of duplicated data, handling data dependencies, dealing with incompatible file formats and representation of data from user's view, they cannot utilize this information to its full potential. Database systems were introduced to manage these autonomous files as a single centralized collection of data. These systems reduce data duplication, avoid data inconsistency, allow sharing of data, increased security and maintain data integrity and make it available on demand [1]. This approach remained successful and sufficient to meet user requirements for several years. However, today's user data processing requirements and capabilities have changed and new applications often involve accessing and maintaining data from several pre-existing databases, which are typically located on autonomous software and hardware platforms distributed over the many sites of a large computer network which leads to heterogeneity and legacy problems, initiating a need for timely and efficient solution by sharing existing knowledge [2].

Besides ongoing advances in database technologies there are still the challenges of uniform and scalable access to multiple information sources including databases and other repositories [3]. Now, (World Wide Web) WWW is playing a more vital role in information sharing for the purpose of education, business, research etc. therefore more and more people are publishing the data over web to share it among large audiences. However, Data is being published in different formats such as PDF, Doc, HTML, etc. Among these different formats of data, most of the information is coming from databases. One of the study reported that "it was determined that Internet accessible databases contained up to 500 times more data compared to the static web and roughly 70% of websites are backed by relational databases"[4]. With the continuous increase in the volume of published data, it is desirable to provide some automatic mechanism to search and integrate information over the Web which is not possible on the existing web. In recent years, with the advent of semantic web technologies (RDF [5], RDFS [5], and OWL [5]) that have been standardized under W3C group, has proven to be a powerful support for the techniques used for managing data and for the problems of data heterogeneity and semantic interoperability [5]. Ontologies (RDFS or OWL) have been suggested as a way to solve the problem of information heterogeneity by providing formal, shared and explicit definitions of data called semantics. The addition of such semantics also improves the query processing by providing more meaningful answer. Additionally, ontologies also have reasoning ability to infer new knowledge and to identify inconsistencies. An ontology-based access to relational data reduces the barriers for data exchange and integration. The expressive and formal semantics increases the value of the existing data and enables new applications on that data [3].

Recently different projects have been developed over Web using semantic web technologies such as DBpedia, Semantic wikis. Moreover, due to the popularity of ontologies, now commercial relational databases (such as Oracle) also provide support of ontologies. However, the construction of ontology is still manual [6]. Thus, it is highly desirable to transform databases into ontologies mainly because of two reasons, first to publish relational data as RDF/OWL on the web and secondly to combine a relational data with existing RDF/OWL for data integration.

This paper compares the existing work by comparing three different algorithms to automatically/semi-automatically construct ontologies from relational databases that can provide a conceptual view over the data. Therefore we can take advantage of both technologies. To construct the ontology model these algorithms established rules between ontological constructs (concept, relation, individual, etc.) and relational databases (tables, attributes, attribute values, etc.). Recently, several approaches have been proposed in literature to transform databases into ontology [7][8][9][10].

The rest of the paper is outlined as follows. The next section presents a brief overview of three recently proposed algorithms to construct ontology from Relational databases with example. Section III compares these algorithms to indicate the challenges involved in this research. Finally, Section IV provides conclusion and future research directions.

II. Algorithm for Ontology Construction from Relational Databases

The construction of ontologies from relational databases and the development of such type of tools has been a major field of interest of the researchers with the evolution of semantic web [11]. In this paper, we have selected three most recently proposed algorithms and describe each algorithm separately with example to investigate the challenges in this field. In general, following are the main components of relational database and ontology that are being considered in the selected algorithms.

Components of Relational database:

- Relations (Tables)
  - ○ Entity tables
  - ○ Relationship tables
- Attributes of relation
  - ○ Key attributes (Foreign key, Primary key)
  - ○ Non-key attributes
- Restriction
  - ○ Limit on the attribute value
  - ○ UNIQUE value attribute (Primary key)
  - ○ NOTNULL (attribute value cannot be null)

  …

Components of ontology:

- Concepts/Classes
- Relationships
  - ○ Taxonomy relation (Class hierarchy)
  - ○ Non-Taxonomy relation (Object and Data type properties)
- Restrictions (Axioms)
  - ○ Cardinality restrictions
  - ○ Functional property

  …

The above mentioned components are analyzed to identify the associations between ontological and relational component. These associations would result in the development of rules to construct ontology automatically from relational databases. Furthermore, to understand the working of rules of each algorithm an example of relational database has been selected as shown in Figure 1.

Student

StudId	StudName	TownId
S001	Saad	T311
S004	Kashif	T119
S102	Faisal	T108

Department

DeptId	DeptName
D008	Computer Science
D119	Electronics
D203	Mathematics

Town

TownId	TownName	TownPcode
T311	Gulshan	73500
T108	Sadder	73400
T119	Defence	77400

PhDStudent

StudId	Reseachfield
S001	Database
S004	Semantic Web
S009	Networking

Employee

EmpId	EmpName	EmpSex
E003	Ahmed	male
E112	Sana	female
E203	Sadaf	female

Studies

StudId	DeptId
S001	D008
S004	D008
S102	D203

Affiliates

EmpId	DeptId
E003	D008
E112	D119
E203	D203

Figure 1: Example of relational database

A. *Algorithm 1*

This algorithm [7] was presented by Peng et al. emphasizes the problem of efforts and cost involved in the manual construction of ontologies. They suggest reducing the cost through ontology learning of structured data. The relational databases (RDB) come under structured data which is a domain specific model. To construct ontology they established correspondence rules between components of ontology and RDB where the tuple in the table shows the instance of Relational schema.

This algorithm divides the relations (tables in RDB) into two types of relations one is Correlative Relation and another one is Basic Relation. Correlative relations are those which do not have any non-key attributes. In contrast, the relations which are not correlative are considered as Basic relations. Following are the rules developed to construct ontology.

Rules for Ontology Concept/Class:

- If a relation is a basic relation then it will be converted into class of ontology. For example, the database shown in Figure 1, tables of Student, PhdStudent, Town, Department and Employee are created as classes in ontology.

Rules for Ontology Relationships:

- If a relation is a correlative relation then it will be converted into two Object properties in ontology that show relation between entities. Example, consider the

Studies and affiliate tables of the database as shown in Figure 1 two Object properties will be created for each relation where one is inverse of the other.

- A primary key of correlative relation will be converted into Object property with their referenced tables. For the database example, stuId in table Studies.
- If a relation attribute is not a foreign key attribute then it will become Data type property in ontology. For example, consider the database example of Figure 1, the attributes of the relations such as StudName in relation Student, TownName and TownPcode in relation Town ResearchField in relation PhdStudent, and so on are created as Data type property in ontology.
- All the foreign keys in a basic relation will also create a relation of Object property. For a given example, attribute TownId is a foreign key in Student relation, thus create an Object property corresponding to TownId between classes Student (as domain) and Town (as range).

Rules for Ontology Restrictions:

- If a relation is a basic relation and has only one Primary Key then it will be converted to the data type functional property with minCardinality or maxCardinality equals to 1. For the example in Figure 1 consider primary keys of each basic relation.
- If a relation is a basic relation and has more than one Primary Key then each primary key will have a restriction with minCardinality or maxCardinality equals to 1.
- If the property is set as NOTNULL then it will take the restriction minCardinality as 1.
- If the attribute of relation is set as UNIQUE then restriction will be created as functional property.

Thus, the algorithm applied rules discussed above by getting MetaData of each table and construct ontology for relational database.

B. Algorithm 2

This algorithm [8] was presented by Zdenka Telnarova, like previous paper this paper also discusses the importance of automatic construction of ontologies from relational databases.

This algorithm transforms relational model in to conceptual model (ontology) by considering the reverse of transformation rules used in the transformation of conceptual model (ERD) to relational model (RDB). This algorithm proposed following set of rules that transform relational model into conceptual model (ontology).

Rules for creating Classes:

- If we have multiple relations in database and all of those relations have same Primary key then it is possible to integrate all of these relations under the single class/concept of ontology. This rule expresses that same primary key in different relation corresponds to same entity and for the purpose of normalization it was divided into more than one relation. Thus integration is a reverse of normalization. Consider an example shown in Figure 1 where primary key of student is also a primary key of phdStudent relation thus these two relations can be integrated into a single concept student in ontology.
- If we have relations in the database and no other relation could be integrated with it according to rule 1. Moreover, the attribute is primary key but not a foreign key in a given relation then this relation is created as a concept in ontology. Relations Town, Department, and Employee are example of such relations.

Rules for creating Relationships:

- The two relations, where attribute of one relation is equal to attribute of another relation and at the same time the common attribute is not the primary key in one relation (similar to foreign key), then Object property can be created. This rule reveals that to incorporate relationships between entities, foreign keys are being added in RDB thus it can be created as Object properties in ontology. For example TownId in student is created as Object property where domain concept is Student relation while range is Town concept.
- If we have two relations, then it is possible to create two Object properties if the following two conditions are fulfilled: a) a Relation has more than one Primary Key (as in relationship table) and b) Foreign Key of one relation belongs to Primary Key of another relation (similar to correlative relation according to algorithm 1). This rule is used to convert many to many relationships in RDB. For the database example two Object properties are created corresponding to studies relation where one property is an inverse of other property.
- All the other attributes of relations in database which cannot be converted into Object property according to the above rule become Data type property in ontology. For example StudName, ResearchField, TownName, TownPcode, EmpName, EmpTitle DeptName are created as Data type property.
- If we have two relations and have same primary key in both relation as well as it is also a foreign key in one relation. To express is-a relationships in RDB same identification key is being used in both super-type and sub-type, thus same identification key in different relation indicate hierarchical relationship which can be exploited to create hierarchy of concept in ontology.

For example, student with primary key stuId and phdStudent with primary key stuId which is also a foreign key. Therefore phdStudent become a subclass of student.

Rules for Restrictions:

- If the attribute in a relation is the primary key then it will have restriction with minCardinality and maxCardinality equals to 1.
- If the attribute in the relation is declared as NOTNULL then the minCardinality equals to 1.
- If the attribute in the relation is declared as UNIQUE then the maxCardinality is equals to 1.

Rules for creating Instances:

- If C is the corresponding class to database relations R_1 or integration of more than one relation then each tuple is considered as an instance of Class C.

C. Algorithm 3

This is a recently proposed algorithm [9] that has been proposed by Zhou et al., to construct ontology for relational database and used WordNet to further extend/reorganize the ontology. Similar to algorithms 1 and 2, this algorithm also described the rules to obtain ontology for a given database.

This algorithm also divides the table in to two types of table one represent entity and one represent relationship between entities. They also suggest the finding of inheritance between concepts.

Rules for creating Classes:

- If the table is entity table (similar to basic relation in algorithm 2) then create corresponding concept/class in ontology. As the example shown in Figure 1, studies and affiliates are relationship tables while others are entity tables for which concepts are created.

Rules for creating relationships:

- If table T1 is a relation table, T2 and T3 are entity tables that correspond to C2 and C3 concepts in ontology then two Object properties will be created; in which one is the inverse property of other. For example, Studies is transformed in to two Object property between concepts Student and Department.
- If a column is not a foreign key, it will be transformed to the Data type property of related table.
- If table T1 has T2’s foreign key, then foreign key will be transformed into an Object property, its domain is T1’s corresponding concept, range is T2 corresponding concept. If a column is not a foreign key in a relation table, it will become the common property of related concepts.
- If there is a column within an entity table that could have several values, no matter how many records, some sub concept could be created by the column’s data value. For example, in Employee table, EmpSex has only two possibilities male/female. Thus two sub-concepts can be created such as maleEmployee and femaleEmployee.

Rules for creating Restrictions:

- An entity table with attribute declared as NOT NULL then the corresponding Data type property is restricted to minCardinality equal to 1. However, a foreign key and a primary key corresponded data type properties has minCardinality and maxCardinality are both set as 1.

Rules for creating Instances:

- Each record of entity table will be transformed into related concept’s instance.

Once the ontology has been created by applying above rules, next, this algorithm will use WordNet to extend the ontology. The extension has been done in two ways. First, by adding concepts synonyms, for example, the synonym of department is section, thus, anyone can be used to refer the concept. Second is hierarchy extension, in case of more than one database some of the concept may be found parallel thus by checking the hyponymy to modify the created ontology.

Figure 2 shows the ontology created in general by applying above algorithms on database example shown in Figure 1. In Figure 2, oval shapes represent concepts in ontology, solid line rectangles represent Object properties; dotted rectangles represent Data type properties while head of dotted arrows indicate range and tail indicate domain of properties. However the solid arrow represents is-a relationship between concepts.

III. COMPARISON OF ALGORITHMS

This section analyzes the research challenges in ontology construction from relational databases by comparing algorithms as explained in the previous section. These algorithms are compared to identify issues involved in the construction of new ontology in general and the issues specific to each algorithm. In general, the rules to construct ontology components are based on the design of relational models. These algorithms exploit the semantics of relational database components such as entity table, relational table, attributes, and constraints, etc. The algorithms can be divided in to three categories, first is the construction of ontology from database schema, second is the construction of domain specific ontology and third is the mapping of database with existing ontology [12]. In domain specific ontology, the purpose is to select data that is relevant to the domain rather than to create mirror of database as ontology. The algorithms discussed in this paper come under the first category. Following subsections elaborate the comparison of the selected algorithms.

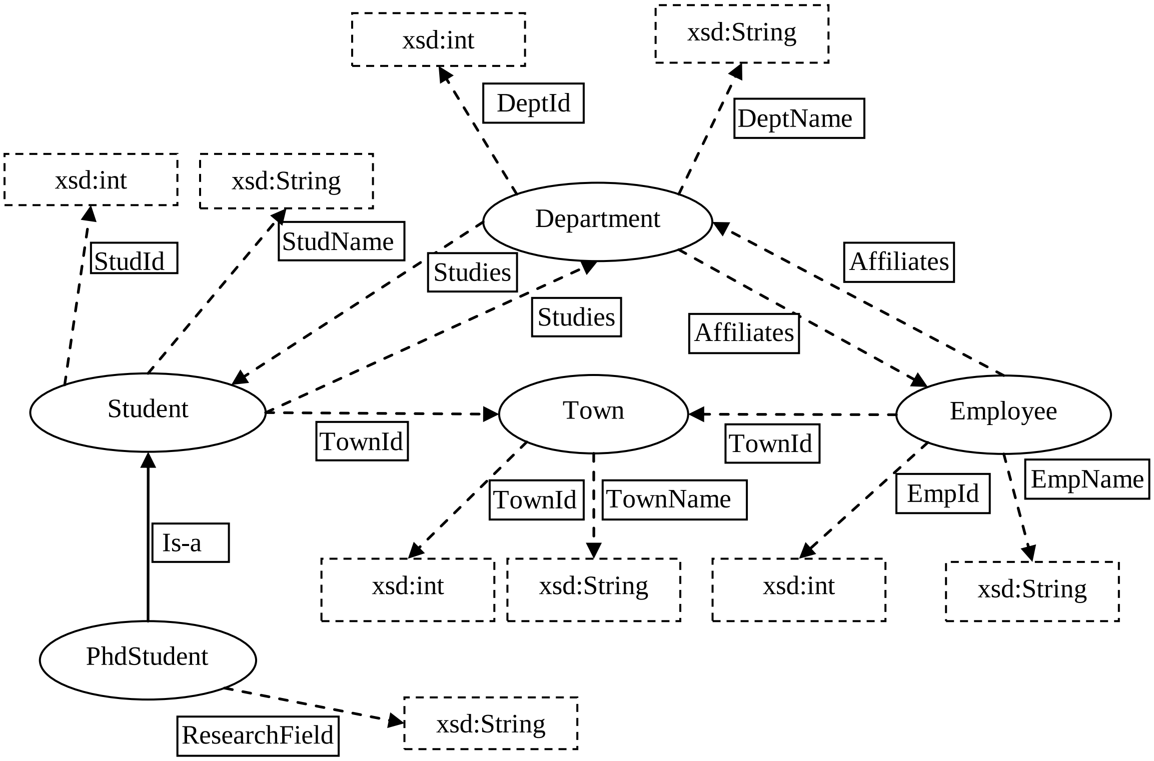


Figure 2: Ontology created from database

A. *Motivation and purpose stated by authors*

The motivations highlighted by authors to construct ontologies from databases are: (i) generate data for semantic web that is processable by machine, and (ii) integration of data from different sources (heterogeneous databases or existing RDF/OWL). To create data for semantic web, databases are rich resource of information and it has been found that large amount of information in dynamic web pages are also generated from databases. Therefore rather than to annotate each dynamic web pages, a better solution would be to create ontology corresponding to a database. Once the ontology has been created it can be used to generate semantic web content that can be processable by machine. Furthermore, it is highly desirable to integrate data from heterogeneous databases or existing RDF/OWL to process the data at large scale for different application development. The authors suggest the use of ontology as the ontologies provide shared and reusable knowledge representation as universal model that support in data integration. Therefore, algorithms are needed to construct ontology from database as discussed in this paper.

B. *Algorithms automation level*

The automation is an important factor to perform transformation from RDB to OWL over a large scale. Therefore authors preferred higher degree of automation in their algorithms using mapping rules. However, mapping rules are defined manually. Once the ontology has been created some post processing is required to formalize the domain vocabularies used in ontology. Furthermore, all three algorithms are limited to basic mapping rules. Addition to this some additional mapping rules need to be defined to extract more semantic information from relational database to be filled in ontology. To extend the use of ontology Peng et al., suggested that the construction of these local ontologies further extended by providing mapping between ontologies and this process could be completely automated.

C. *Algorithm's Ontology language*

This section refers to the selection of appropriate language to represent the ontology. In spite of the fact that there is a demand to transform RDB into OWL or vice versa, there is no standard language has been developed specifically for representing the mapping between RDB and ontology [12]. Therefore, all three algorithms discussed in this paper used ontology languages, such as RDFS/OWL. It has been noted that the use of RDFS is sufficient for components of ontology that are currently being filled by these algorithms. However, the use of OWL provides additional components (not in RDFS) that might require in future or need to fill manually depending on the application.

D. Algorithms Implementation

This section indicates that either these algorithms provide software/tool to practically perform RDB to ontology conversion so that potential user can take benefit from it. Algorithm 1 has been implemented and a prototype is created to perform experiment on Oracle 10g, SQL Server 2000, and MySQL Server 5.0 databases, however, experiment results have not been provided. Algorithm 2 and Algorithm 3 did not provide information about implementation of their algorithms, whereas explained methodology with examples.

E. Issues and Challenges

The algorithms used basic transformation method by creating mapping rules between ontology and relational database components such as table to class, column to predicate etc. These basic mapping rules are generic to apply transformation over a large scale without human involvement. However, database schema is not sufficient information to generate corresponding ontology. Therefore semantically rich ontology can be generated by gathering information from table data, queries, and stored procedures.

The standard database design is based on conceptual model which is then transform into relational model. On the other hand ontology is a conceptual model therefore algorithms develop mapping rule by considering the reverse transformation i-e relational model to conceptual model. However, these reverse transformation most of the time are not possible because databases that have lost original intention in the transformation (conceptual model to relational model) and are very difficult to reverse back. Moreover, most of the transformations in databases development are not documented at all.

F. Recommendations

All three algorithms discussed in this paper highlighted the importance of automatic ontology construction from relational databases and focuses on building domain specific ontologies. This section describes the recommendations and future direction provided by the authors of selected algorithms. One possible recommendation would be to create ontology from conceptual model of database (ERD or UML) and using queries extract data from database to populate into ontology. Secondly, the newly constructed ontology can be further extended by adding more semantics into it and several such ontologies can be integrated to share or exchange knowledge [7]. Moreover, these ontologies would help in building of knowledge warehouse to further extend their utilization [8].

IV. Conclusion and future direction

Database interoperability and semantic reasoning is the ultimate target that the researchers are trying to solve. This considers allowing different databases to be semantically integrated. The paper described importance of ontology construction from database and discussed three recently proposed algorithms. More specifically, these algorithms defined some rules to build a generic approach without human involvement. In spite of all the efforts have been done so far only basic mapping rules are investigated.

Future research directions would focus in extension of the basic mapping rules by adding more mapping rules to create semantically rich ontology.

References

[1] Te-Wei Wang and Kenneth E. Murphy, “Semantic Heterogeneity in Multidatabase Systems: A Review and a Proposed Meta-Data Structure”, Journal of Database Management, Vol. 15, No. 4, pp. 71-87,2004

[2] Natalya F. Noy, “Semantic Integration: A Survey Of Ontology-Based Approaches”, SIGMOD Record, Vol. 33, No. 4, pp. 65-70, 2004

[3] Matthias Hert, University of Zurich, Binzmuehlestrasse 14, CH-8050 Zurich, Switzerland hert@ifi.uzh.ch, “Relational Databases as Semantic Web Endpoints”

[4] B. He, M. Patel, Z. Zhang, and K. C.-C. Chang, “Accessing the deep web,” *Commun. ACM*, vol. 50, no. 5, pp. 94–101, May 2007.

[5] T. Berners-Lee, J. Hendler, and O. Lassila. “The semantic web, a new form of web content that is meaningful to computers will unleash a revolution of new possibilities”. Scientific American, 2001.

[6] Oracle documentation, docs.oracle.com, last access March 25, 2012.

[7] Peng Liu1, Xiaoying Wang, Aihua Bao, Xiaoxuan Wang,”Ontology Automatic Constructing Based on Relational Database”, 2010 Ninth International Conference on Grid and Cloud Computing

[8] Zdenka Telnarova, “Relational Database as a source of ontology creation”, Proceedings of the International Multiconference on Computer Science and Information Technology, pp. 135-139, 2010 IEEE

[9] Xu Zhou, Guoji Xu, Lei Liu, “An Approach for Ontology Construction Based on Relational Database”, International Journal of Research and Reviews in Artificial Intelligence, Vol. 1, No. 1, March 2011,Copyright © Science Academy Publisher, United Kingdom

[10] W3C RDB2RDF Incubator Group,“A Survey of Current Approaches for Mapping of Relational Databases to RDF” ,January 2008-09

[11] Man Li, Xiao-Yong Du, Shan Wang, “Learning Ontology from Relational Database”, Proceedings of the Fourth International Conference on Machine Learning and Cybernetics, Guangzhou, 18-21 August 2005 .

[12] Dimitrios-Emmanuel Spanos, Periklis Stavrou, and Nikolas Mitrou, “Bringing relational databases into the semantic web: A survey,” *Semantic Web*, 2010.

Selection of Adaptive Strategies on Main Agent's Attitude Based on Historical Learning

Guorui Jiang
The Economics & Management School
Beijing University of Technology,
Beijing, China, 100124
E-mail: jianggr@bjut.edu.cn

Hong Guo
The Economics & Management School
Beijing University of Technology,
Bejing, China, 100124
E-mail: guo_hong@emails.bjut.edu.cn

Abstract—**To solve the problems of uncertainty and variability in the current automated commerce negotiation, the intelligence of agent is applied to the process of the commerce negotiation. We propose the adaptive negotiation strategies based on multi-agent negotiation by historical learning algorithm. During negotiation, the main agent, for example buyer agent, obtains historical information of the opponent, as seller, from the third party agent who stores the information of agents participated in and trade information, and then calculates the negotiation attitude values of the opponents by historical learning algorithm. Considering the information of the dynamic market environment, the main agent presents an appropriate strategy by employing the adaptive concession strategy function and the effectiveness evaluation mechanism. The research achievement of this paper is a foundation for developing a real-life Multi-agent-based commerce negotiation system in the future.**

Keywords-adaptive strategy; historical learning; negotiation attitude; multi-agent system

I. INTRODUCTION

E-commerce negotiation based on multi-agent is the general term of the activities performed by agents aiming to reach an agreement on concerned issues [1-2]. At present, negotiation strategy is one of main topics in the field of automatic negotiation, and it is also a method of maximizing the interests of main agent under the constraint of certain negotiation protocol [3]. Due to the asymmetry of information and dynamic change of environment during negotiation, the choice of negotiating strategy becomes the key factor of a successful negotiation. In recent years, the single negotiation strategy has been used more frequently in the agent-based electronic trading system; however, the negotiating strategy combining both the attitude of negotiating agent and the change of dynamic market environment is rarely studied relatively.

Nowadays, the E-commerce negotiation environment is faced with the problems of information incompleteness and finite rationality of E-commerce. For the main agent, there is a problem to be solved urgently, which is how to synchronously improve the existing negotiation strategy, learn and generate adaptive strategy automatically [4]. In related research, the concept of adaptive strategy based on argumentation is proposed by Rahwan et al. [5] [6]. The concession strategies, response strategies, the proposal strategies and relevant utility function are established for realizing adaptive negotiations [7]. Ren et al presented the dynamic adaptability is necessary for agent to negotiate in open and dynamic environment [4]. Adaptive strategy can be used to describe and define decision-making behavior of agent during the process. And agents perform the appropriate adaptive operation for different roles in different environments. This paper focuses on the discussion of the problem that chooses adaptive strategies based on attitude of main agent by historical learning algorithm. The background of negotiation is shown in Section II, the factors affected the adaptive strategy of main agent are shown in Section III, the generation of adaptive strategy of main agent is presented in Section IV, and then the conclusion is given in Section V.

II. BACKGROUD ASSUMPTIONS

Assumption 1. Agents of both sides are selfish, to pursue their own interests for negotiation purposes;

Assumption 2. Agents of both sides have the ability to learn;

Assumption 3. Agents of both sides can react to the impact of environment and the information of opponents;

Assumption 4. The time factor on the negotiation parties is valuable;

Assumptions 5. Agents of both sides have the sincerity to reach agreement by negotiation and there is no fraud in the negotiation process.

III. FACTORS OF ATTITUDE ADAPTIVE STRATEGY OF NEGOTIATING AGENT

A. *Time factor*

Utility theory is a fundamental theory of Western economics， which can provide support for decision makers in a decision-making process by obtaining utility values of an act [8]. During the negotiation, time is one of the most important factors affecting the agents' behaviors. Considering this fact, this paper employs a time-utility function to assess the effectiveness of negotiation time [9]. Some notations are defined as follows:

$U(t)$: denotes time utility;

t : duration of negotiations;

T_b / T_s : limited negotiation time of agent b/s;

$T = \min(T_b, T_s)$: limited negotiation time of both;

σ : utility coefficient, $0 < \sigma < 1$.

The function $U(t)$ is defined as follows:

$$U(t) = 1 - \sigma\left(\frac{t}{T}\right) \qquad (1)$$

With the growth of the time, the utility value is getting smaller and smaller, the chance of negotiation success is getting smaller and smaller.

B. *Factor of Supply-Demand Relationship*

In the real E-commerce trade environment, the number of sellers and buyers has a direct impact on supply-demand relationship. It would be beneficial to the sellers in the negotiation when the number of sellers is less than the number of buyers, and vice versa. Thus, a ration function of the number of buyers and sellers is employed to reflect the supply-demand relationship. The ration function is defined as: B_t / S_t denotes the number of buyers/sellers agent at time t; agent s/b denotes the seller/buyer agent; $C(t)$ is the value of supply-demand relationship at time t, which always reflects the market competition; the larger the value is, the more competitive the market will be. The function $C(t)$ is as follows:

$$C(t) = \begin{cases} S_t / B_t, \textit{ for agent s} \\ B_t / S_t, \textit{ for agent b} \end{cases} \qquad (2)$$

C. *Historical Learning of Negotiation*

In the negotiating process, the negotiation agent knows neither the effectiveness to each other brought by their offers, nor the reasoning principles and constraints of opponent (seller agent), or whether an agreement could be achieved, etc.Thus, learning negotiation history of the opponents continuously is critical for a successful negotiation. The negotiation history can be derived from a third party agent. And the historical prices in the negotiation are taken as the standard in this paper, and then the negotiation attitude of the opponents (seller agents) can be learned from the variable-ration of their two most recent bids [10]. The method of calculating the attitude values of the opponents is discussed respectively, according to three different historical records (historical offers) (assuming that purchase records are stored in the array SP [ ]):

1) Two historical records of negotiation

$SP[K]$: The K-th offer of Agent s, and it is a recent negotiation within the current time. The value of opponent's attitude can be defined as:

$$\text{Attivalue} = SP[K] / SP[K-1] \qquad (3)$$

2) Three negotiating history records

$$R_1 = (SP[K-1] - SP[K-2]) / SP[K-2] \qquad (4)$$

$$R_2 = (SP[K] - SP[K-1]) / SP[K-1] \qquad (5)$$

$$\text{Attivalue} = \begin{cases} R_1 / R_2, R_1 < 0 \textit{ and } R_1 \neq R_2 \\ 1 \qquad R_1 < 0 \textit{ and } R_1 = R_2 \end{cases} \qquad (6)$$

3) Three more negotiating history records

$$\text{Attivalue}_k = \begin{cases} (R_1 / R_2 + \textit{Attivalue}_{k-1}) / 2, R_1 < 0 \textit{ and } R_1 \neq R_2 \\ 1 \qquad R_1 < 0 \textit{ and } R_1 = R_2 \end{cases} \qquad (7)$$

In the algorithm above, (R_1 / R_2) means the rate of changes between two bids of the seller. In this paper, buyer agent is considered as the main agent, and wants to learn the seller agent's attitude of negotiation. It would be the same research if the seller agent is considered as the main agent.

D. *The Attitude Functions of the Main Agent*

The factors which affect the main agent's negotiation attitude include: time factor, supply-demand relationship factor, and the opponent's negotiation attitude. In the attitude function of the main agent, ψ is defined as the attitude value of the main agent; history means the history of the negotiating opponent obtained from a third party agent; ψ denotes the initial attitude value of the main agent. The attitude function is defined as follows:

$$\Psi = f(B_t, S_t, t, T, \text{history}) = \psi \times U(t) \times C(t) / \text{Attivalue} \qquad (8)$$

Different negotiation attitude value of main agent corresponds to different attitude [11]:

1) $2 \le \Psi$ (Patient): The negotiation agent keeps its initial proposal, and it slowly increase or decrease the proposal to its retention until the negotiation time is up;

2) $0.5 < \Psi < 2$ (Calm): The negotiation agent changes its proposed value relatively stable in the whole process of negotiation;

3) $0 < \Psi \le 0.5$ (Eager): The negotiation agent rapidly increases or decreases to its retention in the early negotiation process, and it will keep the value until negotiation ends.

IV. THE GENERATION OF THE ATTITUDE ADAPTIVE STRATEGY BASED ON HISTORICAL LEARNING

In this paper, the whole negotiation process involves three entities: buyer agent, seller agent, third-party agent. Third-party agent stores the historical information of negotiation and agents' credibility, and processes requests from both agents. Before the negotiation, the buyer/seller agent requests a third party agent to provide opponent's trade information, which could be helpful in the output of their negotiation attitude values finally.

A. *The Generation of Attitude Adaptive Strategy of Main Agent*

With aforementioned historical learning and the attitude function of the main agent, an adaptive concession strategy function based on time-bound can be easily established [12]. Let:

$P(t)$: denotes the offer of the main agent at time t;

P_{bmin} / P_{smin} : is the buyer's/seller's minimum offer;

P_{bmax} / P_{smax} : denotes the buyer's/seller's maximum offer;

ψ_b / ψ_s : is defined as the value of negotiating attitude of buyer/seller at time t, and it equals to Ψ ;

$(t / T_b)^{\psi_b}$: is the concession factor;

$(P_{bmax} - P_{bmin})$: means concession interval.

The function $\mathrm{P(t)}$ of intention offer is defined as follows:

$$\mathrm{P(t)}=\begin{cases}\mathrm{P_{bmin}}+\left(\mathrm{P_{bmax}}-\mathrm{P_{bmin}}\right)\times\left(\mathrm{t}/\mathrm{T_b}\right)^{\psi_b}, & \textit{for agent b}\\ \mathrm{P_{smax}}-\mathrm{P_{smax}}-\mathrm{P_{smin}}\times\left(\mathrm{t}/\mathrm{T_b}\right)^{\psi_s}, & \textit{for agent s}\end{cases} \tag{9}$$

Based on the aforementioned function of offer, an effectiveness evaluation mechanism is established. Let p denotes the time at which negotiation issues reached; k denotes the time discount factor of issues. The utility of the issues can be defined as [13]:

$$\rho=\begin{cases}\left(\mathrm{P_{bmax}}-\mathrm{p}\right)\times\left(\mathrm{k}\right)^{\mathrm{t}}, & \textit{for agent b}\\ \mathrm{p}-\mathrm{P_{smin}}\times\left(\mathrm{k}\right)^{\mathrm{t}}, & \textit{for agent s}\end{cases} \tag{10}$$

Given the above, a proposal selection mechanism can be established, of which

$\mathrm{P_{s\to b}(t)}$: means the proposal of Agent s to Agent b at time t;

$\mathrm{P_b(t)}$: the intention offer of Agent b at time t;

$\rho_{\mathrm{b}}(\mathrm{t})$: the utility value of Agent b’s intention bidding;

$\rho_{\mathrm{s\to b}}(\mathrm{t})$: the utility value of Agent s’s bid for Agent b at time t;

$\mathrm{p^t}$: Agent b’s anti-bid at time t.

Evaluation function of the proposal, which Agent b sends to Agent s at time t as follows:

$$\mathrm{Evaluate_b(P_{s\to b}(t),t)}=\begin{cases}\textit{accept}\left(\mathrm{b,s,P_{s\to b}(t)}\right),\ \textit{if}\ \rho_{\mathrm{b}}(t)<\rho_{\mathrm{s\to b}}(t)\\ \textit{reject}\left(\mathrm{b,s}\right),\ \textit{if}\ \rho_{\mathrm{b}}(t)>\rho_{\mathrm{s\to b}}(t)\textit{and}\ t=T\\ \textit{proposal set}\ \mathrm{p^t}=\mathrm{P_b}\left(\mathrm{t}\right),\ \textit{if}\ \rho_{\mathrm{b}}\left(\mathrm{t}\right)>\rho_{\mathrm{s\to b}}\left(\mathrm{t}\right)\textit{and}\ t<T\end{cases} \tag{11}$$

The buyer agent could get the value of attitude from historical learning, and get the offer using concession strategy function given above. Then compare the utility of self-intention offer with the utility of seller agent’s proposal $\mathrm{P_{s\to b}}$ at time t. If the time discount factor of issues k>1, the later the time to reach an agreement the greater the profit will be, the agent b with the value of negotiating attitude $\psi_{\mathrm{b}}\geq 2$ or $\psi_s\geq 2$ could get larger profit. If k<1, the earlier the time to reach an agreement the greater the profit will be, so the agent with $0<\psi_{\mathrm{b}}<0.5$ could get larger profit. At time t, if $\rho_{\mathrm{s\to b}}(\mathrm{t})>\rho_{\mathrm{b}}(\mathrm{t})$, the buyer agent could accept the seller agent’s offer, and then the negotiation succeeds; if $\rho_{\mathrm{s\to b}}(\mathrm{t})<\rho_{\mathrm{b}}(\mathrm{t})$, and the negotiation time is up, then the negotiation terminates. Or else, the buyer agent makes its intention offer $\mathrm{P_b(t)}$ as the re-proposal $\mathrm{p^t}$ to the seller agent, repeat the process above until the negotiation succeeds or terminates.

B. Learning Algorithm for Attitude Adaptive Negotiation Strategy of Main Agent

Based on the theories mentioned above, negotiation learning algorithm process of buyer Agent b is shown in Figure. 1:

1) Initialization. The negotiating participants set could be denoted as A= {Agent b, Agent s}, of which Agent b is the negotiation initiator; Agent s is the negotiation respondent. Negotiation issues set G includes n issues, t=0 means that the negotiating parties are ready;

2) Before the first round of negotiation, Agent b sends request to the third party agent for the information queries of Agent s. After confirmation, Agent b could obtain the history information of Agent s;

3) Agent b gets the value of attitude of Agent s by learning the opponent’s historical negotiating information;

4) Agent b could get its value of negotiating attitude by the attitude function of main agent;

5) Agent b could obtain the intention offer P(t) through adaptive concession strategy function;

6) Agent b sends negotiating request proposal(Agent b,t,G) to Agent s;

7) If Agent s refuses to accept the proposal or the waiting time exceeds T, then the algorithm turns to step 15） negotiation termination. Otherwise, execute the next step;

8) When Agent s receives the negotiating request of Agent b, it gives out its proposal(Agent s, t’, G) for each issue;

9) Agent b evaluates the proposal of Agent s by employing the strategy evaluation function $\mathrm{Evaluate_b(P_{s\to b}(t),t)}$, and makes a strategy choice;

10) If $\rho_{\mathrm{b}}(\mathrm{t})<\rho_{\mathrm{s\to b}}(\mathrm{t})$, Agent b accepts Agent s’ quote, and then turns to step 14） negotiation succeed;

11) If $\rho_{\mathrm{b}}(\mathrm{t})>\rho_{\mathrm{s\to b}}\left(\mathrm{t}\right)$ and t>T, then the algorithm turns to 15）negotiation termination. Otherwise, execute the next step;

12) If $\rho_{\mathrm{b}}(\mathrm{t})>\rho_{\mathrm{s\to b}}(\mathrm{t})$ and t<T, then Act= {re-proposal}, Agent b sends its intention offer which is obtained from the concession strategy function as re-proposal (Agent b, t'' , G) to Agent s. If Agent s modifies its original offer, turn to 3） history study, or else turn to the step 13） confirms quote;

13) If Agent b confirms the offer of Agent s, and then Agent b chooses the Agent s, turn to 14） negotiation succeeds, or else turn to 15） negotiation termination;

14) Act= {accept}, the negotiation succeeds, the current proposal is accepted as the results of the trade;

15) Act= {reject}, the negotiation terminates.

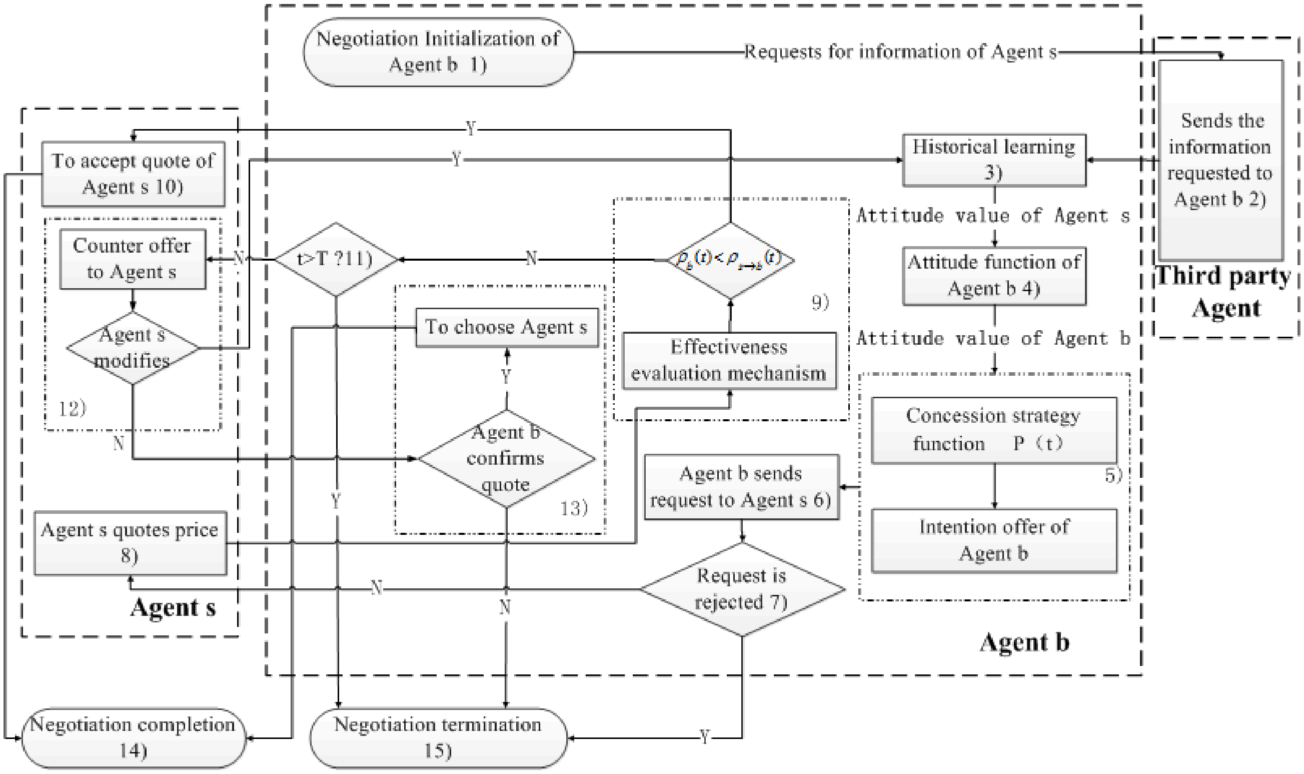

Negotiation Initialization of Agent b 1)
Requests for information of Agent s
Sends the information requested to Agent b 2)
Third party Agent
Historical learning 3)
Attitude value of Agent s
Attitude function of Agent b 4)
Attitude value of Agent b
Concession strategy function P（t）
Intention offer of Agent b
5)
Agent b sends request to Agent s 6)
Request is rejected 7)
Agent b
To accept quote of Agent s 10)
Counter offer to Agent s
Agent s modifies
12)
t>T ?11)
$\rho_b(t) < \rho_{s\to b}(t)$
9)
Effectiveness evaluation mechanism
To choose Agent s
Agent b confirms quote
13)
Agent s quotes price 8)
Agent s
Y
N
Negotiation completion 14)
Negotiation termination 15)

Figure 1. Negotiation learning algorithm process of agent b

V. CONCLUSIONS

The adaptive strategies play a very important role in the process of E-commerce automated negotiation for enhancing efficiency and arriving trade. Based on historical learning algorithm, an adaptive selection strategy based on attitude of main agent was proposed for solving the problems of uncertainty and variability in the current E-commerce negotiations. In this paper, the factors that can influence the adaptive strategy based on attitude of main agent were analyzed, the third-party agent was introduced, and a learning algorithm of the negotiator agent was also proposed.

The main agent could infer the opponent’s attitude value by opponent historical information provided by the third party agent. It could also get its adaptive strategies by negotiation learning algorithm for considering the different attitudes of agents during negotiation process. The main agent could obtain its intention offer by the concession strategies, and then select a proposal by employing the evaluation mechanism of effectiveness to maximize its profit. By the theoretical analysis, it can be seen that the proposed algorithm could be efficiently applied to the “one-to-one “negotiation, as well as “one-to-many” negotiation. The proposed algorithm could help reduce the blindness of proposal and negotiation time, especially when the preferences or behaviors of the opponents are totally unknown during negotiation. Therefore, it has great help for increasing the successful rate of negotiation.

ACKNOWLEDGMENT

This research has been supported by the National Natural Science Foundation of China (#71071005, and #70940005)

REFERENCES

[1] Jennings NR, Faratin P, Lomuscuo AR, et a1. Automated negotiation: prospects, methods andchallenges.International Journal of Group Decision and Negotiation, 2001.10(2): 199-215.

[2] Hai Wang, Yijun Li, Xinpei Hou. E-commerce oriented ANS based on agent.Systems Engineering-theory &Practice, 2005. 11: 14-19.

[3] Wei Shang,Yijun Li.Support systems for Multi-party and Multi-attribute E-business negotiations. Chinese Journal of Management, 2007. 4(3): 279-283.

[4] Fenghui Ren, Minjie Zhang, Kwang Mong Sim. Adaptive conceding strategies for automated trading Agents in dynamic, open markets, Decision Support Systems, 2009. 46: 704-716.

[5] Rovatsos M, Rahwan I, Fischer F. and Weiss G. Adaptive strategies for practical argument-based negotiation. Proc. of the 2nd Int. Workshop on Argumentation in Multi-Agent Systems, 2005. 1-15.

[6] Rahwan I, Sonenberg L, Jennings NR and McBurney P. Stratum: A methodology for designing heuristic Agent negotiation strategies. Applied Artificial Intelligence, 2007.21(6): 489-527.

[7] Lai GM, Sycara K.P.A. generic framework for automated multi-attribute negotiation. Group Decision and Negotiation, 2008. 18(2): 169-187.

[8] Zhiyu Xie. Economic Game Theory (Second Edition). Beijing:Fudan University Press, 2004. 12: 1-55.

[9] Faratin P, Sierra C, Jennings NR. Negotiation decision functions for autonomous agents.Int.Journal of Robotics and Autonomous Systems, 1998. 24 (3-4): 159-182.

[10] Yu Cheng, Ji Gao, Huamao Gu, Zhaoyang Fu. Negotiation decision model based on learning of opponent's attitudes.JournalofZhejiangUniversity(EngineeringScience), 2008. 42(10): 1676-1680.

[11] Hong Zhang, Huacan He. Strategy and algorithm for automated negotiations between multi-agent. Journal of Computer Applications, 2006.26(8): 1935-1937.

[12] Ranran Li, Huamei Sun, Guorui Jiang, Tiyun Huang. A Research on the One-to-many Automated Negotiation Model Adopting Elimination System Based on Multi-agent.China Journal of Information Systems, 2008.2(1): 29-36.

[13] Tianhao Sun, Qingsheng Zhu, Shuangqing Li. Coordinating strategy of one-to-many automated negotiation. Computer Engineering and Applications, 2007. 43(3): 230-233.

Extended Enterprise Integration Model Associated with Automation Index

Rita de Cássia S. Marconcini Bittar *[#], Oswaldo Luiz Agostinho*, Gislaine Fernandes*

*Faculdade de Engenharia Mecânica, Universidade Estadual de Campinas
Campinas, SP, Brasil
[#]Faculdade de Tecnologia, Universidade do Estado do Rio de Janeiro
Resende, RJ, Brasil
Emails {ritabittar@fem.unicamp.br, agostinh@fem.unicamp.br; gislaine@fem.unicamp.br, rmbittar@uerj.br}

***Abstract* — The enterprise of the years 2000 are facing competition in various fronts, like reduction of product life time, increase of diversification, reduction of customer response time, competition internationalization. Due to those factors, it became necessary to reach a higher degree of relationship between the enterprise and suppliers network and customers networks, avoid unnecessary costs and response time between then. Besides that, the utilization of automation is not necessarily done with balanced criteria, creating isles of excellence divorced from the rest of organization. The paper proposes a integration model of the Extended Enterprise that associates automation growth with business process model of the organization in order to create an integration environment that will minimize the application, maintaining the information flows as balanced as possible. The integration status is obtained applying the same level of automation in the various activities of the business processes, measured through automation indexes, resulting in a model which levels of integration of the Extended Enterprise and discussed in a holistic mode.**

Keywords - Extended Enterprise; Automation Index; Business Process; Integration; Information Technology.

I. INTRODUCTION AND OBJECTIVES

In this century, companies are facing changes due to the competition and so a need of fast response previewing the clients' demands is quite mandatory. Besides to focusing on the customer, it is also important to establish partnerships with suppliers and create strategic alliances in order to increase business agility and reliability and to meet market needs. However, fast responsiveness requires changes in both internal and external processes mainly with respect to the relationship with customers and suppliers. In this context, the concept of the Extended Enterprise arises.

Extended Enterprise need new ways of organization and collaborative management, providing them an integration status of their internal and external activities. It was a thought, mainly in the 90s, that this integration status could be achieved through massive application of Information Technology resources. Although the promise that Information Communication Technology (ICT) would connect people, processes and information, current implementations are strongly document-oriented, and the difficulties of integration between companies remain [1]. Nevertheless, the use of Information Systems is a key success factor, since they have potential to integrate business processes inside the company and between related suppliers and customers.

The extended enterprise can be seen as an evolution of the Integrated Supply Chain, because all interrelationships are based on rules and contracts, and nave as principle, that everyone should win whith the transactions, sharing equivalently all the gains and risks of the business, and this success depends on the collective performance rather than isolated parts. To this purpose it is important to manage all the interfaces between organization, consumers and suppliers, in order to ensure a synchronized supply for all the supply chain.

Integration of the Manufacturing System is a prerequisite to obtain the attributes of competitiveness, i.e., leadership and responsiveness. The Business System integration represents an organization state, reflected in the ability to move information synergistically between their process and activities [2].

The business processes integration and the information technology has been one of the key factors for a successful execution of an integrated enterprise system [3].

The objective of this paper is to propose a conceptual integration model for the extended enterprise. The integration model will be based on automation application, through the utilization of automation indexes.

This paper is structured as follows. In Section II, the literature review will be present. In Section III the automation indexes definition will be discussed. In section IV the proposition of the extended enterprise integration model is expanded. Section V presents the possible configuration for the extended enterprise integration model and in Section VI focuses the conclusions.

II. LITERATURE REVIEW

In order to contextualize the Extended Enterprise and the Integration model, it will be presented a review of the main concepts of Collaborative Network, Extended Enterprise, Supply Chain, Business Process, application of automation and definition of Automation Indexes.

Collaborative Network is "a network consisting of a variety of entities that are largely autonomous, geographically distributed, and heterogeneous in terms of their operating environment, culture, social capital and goals" [4].

The concept of Extended Enterprise emerged at the same time as other concepts that emphasize the inter-organizational collaboration such as: virtual enterprise, supply chain and enterprise network [5]. It refers to the collaborative relationship between supply chain members, from which buyers and sellers obtain a competitive advantage and achieve higher customer satisfaction against other supply chains [6]. The extended enterprise is an evolution of the integrated supply chain and can be considered as a complete set of collaborative companies, both upstream and downstream, from raw material to the final consumer, working together to bring value to the market. The advantages of the Extended Enterprise derive from the ability of companies to use their full network of suppliers, vendors, customers and clients [5]. The extended enterprise may also be defined as: "A conceptual business unit or system that consists of a purchasing company and suppliers, who collaborate closely so as to maximize the returns to each partner". The extended enterprise is a philosophy where member organizations (EE actors) strategically combine their core competencies and capabilities to create a unique competency [7].

According to [4], extended enterprise represents a concept typically applied to an organization in which a dominant enterprise ‘‘extends" its boundaries to all or some of its suppliers. An extended enterprise can be seen as a particular case of a virtual enterprise.

During the two last decades, there was a fast growth of the inter-organizational collaboration. Managers and academic researchers representing different areas, as diverse as marketing, economics, marketing, economy strategy and sociology, have carefully studied the inter-organizational network, joint-ventures, coalitions, extended enterprise, partnerships and alliances [5]. In an extended enterprise, which usually involves complex supply chains, the main concern is related to the integration of all members of the supplier and distribution chains, which share a common goal - to obtain a market share through the product realization [8].

In the XXI century, it is expected that the competition will occur between value chains effectively integrated into their competences and resources, in order to compete in a global economy [7]. The extended enterprise also reflects the high level of interdependence between organizations and how to conduct businesses [9]. There are similarities between value chain, supply chain and extended enterprise. At low-levels of integration there are no major differences. The differences emerge when the level of integration and collaboration increase [6]. Managers and researchers recognize that there is a critical interdependence between companies, customers, investors and communities. Such dependence cannot be described only on the basis of a contractual exchange because it involves interactions and network effects. The effective management of the extended enterprise requires both a concept and a approach to practical issues that emerged in this scenario [10]. The Council of Supply Chain Management Professional (CSCMP) considers that: "Supply chain as encompasses the planning and management of all activities. Importantly, it also includes coordination and collaboration with partners, which can be suppliers, intermediaries, third party service providers, and customers" [11]. This is the definition adopted for the present work.

The Association for Operations Management (APICS), has the following definition for supply chain management [12]: "The design, planning, execution, control, and monitoring of supply chain activities with the objective of creating net value, building a competitive infrastructure, leveraging worldwide logistics, synchronizing supply with demand, and measuring performance globally".

Companies need access to accurate information in real time in order to meet the growing challenges of globalization and the reduction in the product life cycle. To a large extent, especially in the last two decades, the use of information technology, mainly based on Enterprise Application Systems (EAS), have allowed companies to respond effectively to the dynamic changes in business [13]. In that sense, it is reasonable to say that information technology is the key to the integration of the extended enterprise. Along with business process reengineering, strategic alliances and management changes, IT can be deployed to enable the planning, control, integration of decisions, information integration and integration of business processes, which may allow companies to operate in the extended enterprise as if it were a single company [6].

Business process management (BPM) is seen as both an IT development, as well as an implementation of quality management, being knowledge and information key factors for its success [14]. According to [14], business processes constitute a "systemic approach to design and continuously improve the organizational processes, by potentialized people and team work, combining emergent technological competences and under a philosophical stance for quality, aimed at delivering value to customers". Also, according to [2], a business process is a set of activities, logically ordered according to precedence rules. The activities and were developed from the definition of manufacturing as a system which can be interpreted as the composition of all business processes and activities.

Business process management is an integral part of today's enterprises, in particular those related to e-business, because the efficiency and effectiveness of the underlying business processes have become a major source of competitive advantage of companies. The process design is the foundation and a critical component of BPM, where new business processes are developed to meet the needs of business problems or existing processes are reviewed to improve company performance [15].

Enterprise integration is the process of ensuring the interaction between enterprise entities necessary to achieve domain objectives. Enterprise integration can be approached in various manners and at various levels as - (i) physical integration, (ii) application integration and (iii) business integration [16].

As a consequence of changes in the competitive environment and marketplaces, in the middle of ‘90s the EE model started to be analyzed in the engineering manufacturing and operations research literature [17].

III. AUTOMATION INDEX APPLIED TO BUSINESS SYSTEM

Automation, when seen as an available technology provides mean for the information flow in each activity of the automation system. In this research the programmable automation is used. Automation is programmable when the human attribute is substituted by the computer program [2].

The automation index developed by [2], applied to quantify automation, is defined as follows:

$$i_A = \frac{n_A}{n_T}$$

where:

i_A = automation index

n_A= number of human activities replaced by devices or instruments with an automation concept

n_T = total number of activities performed by humans

It can be argued that $0 \leq i_A \leq 1$, with the boundaries limits:

$i_{A\,=}\,0$ - all activities are performed by humans

$i_{A\,=}\,1$ - all activities performed by humans are replaced by automation devices,

This definition of automation index is applied to a Business System [2], whose model is shown in Figure 1, considering the main business processes grouped in four groups - Engineering, Shop Floor, Support and Commercial. To facilitate understanding, each group is broken in sub group of business processes, as shown in Figure 1. It is considered that information flows inside each group through the activities of the correspondent business processes, and outside the groups, performing the links between the four groups themselves.

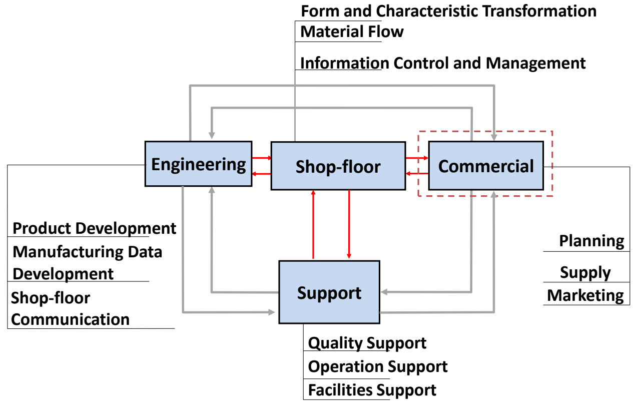


Figure 1 - Business System Model [2]

The automation index for each group and subgroup of business processes will vary as follows:

$$0 \leq i_A \leq 1$$

Each block of process and activities can be represented by a three axles diagram, with the automation index i_A varying from 0 to 1 at each one of the axles. Taking in account the automation indexes, it is defined [2] *that the business systems will have structural integration if the automation index i_A is approximately the same in numerical terms, for each one of the business process groups and sub groups.* Automation indexes with approximately the same numerical values mean the information flow with the same technological degree of automation; consequently with the same interfaces, facilitating the synergies the exchange information flow between the activities. The level of organization maintains coherence with the degree of automation for each architecture.

The four blocks of activities, associating the approximate numerical values of i_A for the twelve axles is shown in Figure 2. This Figure is shows, one condition of structural integration of high the business system.

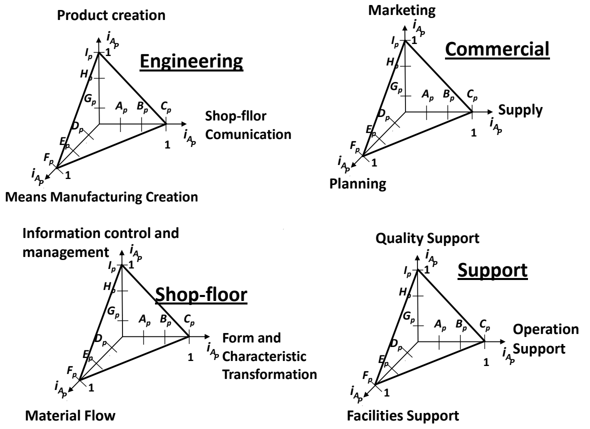


Figure 2 – Manufacturing system with total integration activities [2].

IV. INTEGRATION MODEL ASSOCIATED TO AUTOMATION INDEXES APPLIED TO THE EXTENDED ENTERPRISE

A. Integration Model definition

Integration of the Business System is a prerequisite to obtain the attributes of competitiveness, such as innovation and responsiveness. This article proposes a integration definition, by which the Business System integration represents an organizational state, reflected in the ability to move information synergistically between their activities and sub activities [2]. The information will flow synergistically if the level of automation of the several activities of the Business System is approximately the same in the activities of the Business System. The level of automation will be measured by the automation indexes.

Total Structural Integration Manufacturing System is achieved when it provides harmonic growth of automation, with approximately the same rate in the coordinate axles as shown in Figure 2. One can defined harmonic growth of automation if the indexes i_A increases simultaneously their numeric value at the same rate in the three activities axles [2].

Harmonic growth of automation, measured by their indexes, as well as ensure structural integration in various manufacturing system architectures, provides important strategic tool to obtain the attributes of competitiveness [2].

B. *Business Processes in the Extended Enterprise*

As shown in Figure 1, the block of the business system processes that interfaces with customers and suppliers is the Commercial one, deployed in the sub-set of business process of Planning that interfaces with the Shop Floor business processes.

Each one of the Supplier and Customer set of business processes can be associated to the Business System model shown in Figure 1, as each one represents a Company or Organization, so can be represented according to Figure 3.

Extracting from model in Figure 1 adopts the commercial set of process as link with suppliers, customers business process as Figure 3.

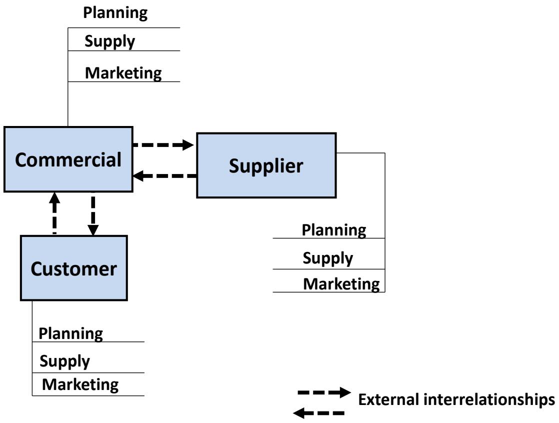


Figure 3 - Business Processes in the Extended Enterprise
Source: Based [2]

a) Commercial – Set of processes that interfaces with both customers and suppliers. It is the place in which the product is assembled and finished so that it can reach the final consumer, through final distribution.

b) Suppliers - Set of business processes that supplies raw material and finished parts to the Organization. The set of all suppliers of the constitutes the supply chain network. Among the suppliers, there are those with high level of automation and integration and those with near-zero automation and low integration.

c) Customers - Set of business processes that allow the Organization market to sell its products, and include the final consumer, chain stores, wholesalers, retailers, utilities, etc. The set of all customers forms the customer network. As in the supply chain network, the automation and integration rates vary among customers. The customer network is responsible for distributing the finished product and deliver it to the consumer.

After the conceptualization of enterprise integration model using automation indexes [2], the same concepts will now be applied to the extended enterprise. In order to perform these concepts, the extended enterprise model will be those from the Figure 3, where the Commercial will interface with the suppliers and customers network. The determination of the extended enterprise model will use the integration concepts developed in [2]. So, the application of automation in the organization business process can be classified in three different discrete points, as shown in Figure 4, represented by the three-orthogonal axles of marketing, supply and planning.

- Automation indexes tend to 0.33, meaning mostly manual activities and use of paper lists;
- Automation indexes tend to 0.66, meaning mostly usage of individual computers without a network connection, running individual software, without integration between them.
- Automation indexes tend to 1, meaning mostly usage of integrated business systems, such as Enterprise Resource Planning (ERP), Customer Relationship Management (CRM), Supply Chain Management (SCM), Engineering software, Manufacturing Execution Systems (MES), etc., running in distributed environment, connecting mid range computers and data storages.

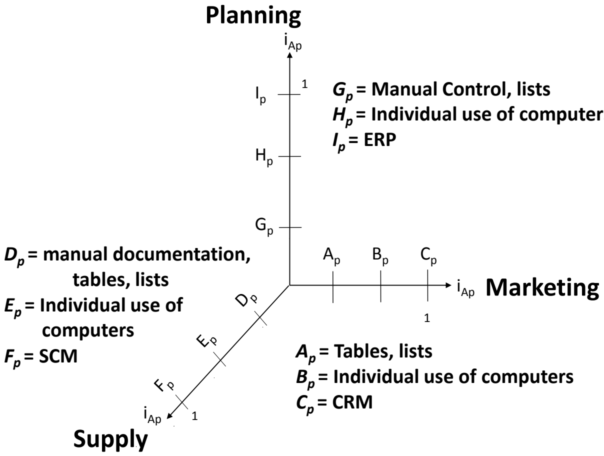

Planning
i_{Ap}
I_p
1
H_p
G_p
G_p = Manual Control, lists
H_p = Individual use of computers
I_p = ERP
D_p = manual documentation, tables, lists
E_p = Individual use of computers
F_p = SCM
A_p
B_p
C_p
i_{Ap}
Marketing
1
D_p
E_p
F_p
i_{Ap}
1
Supply
A_p = Tables, lists
B_p = Individual use of computers
C_p = CRM

Figure 4 - Automation application of commercial business systems [2]

C. *Proposal of Extended Enterprise Integration Model*

The proposal of Extended Enterprise Integration Model associated with automation index achieves its excellence when the three activities – supply, marketing and planning – are in the approximate automation index level and in both internal and external business process of the organization.

To exemplify the integration concept using the automation index harmonic growth see Figure 5. Assuming that the suppliers and customers have the same structural model, the integration of the extended enterprise can be obtained when the automation indexes of the commercial, supplier and customer business processes have automation approximately the same automation index values.

ISBN: 978-1-61208-202-8

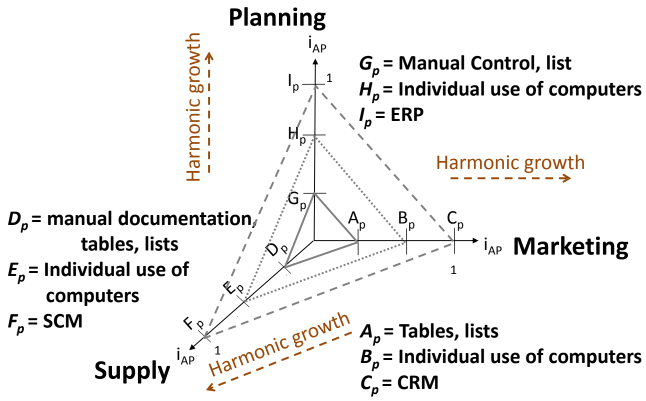


Figure 5 - Commercial business processes level, with approximately the same values of automation indexes in the three axles [2].

Figure 6 shows various configurations of Extended Enterprise Integration, when the commercial has the full integrations, and the Supplier and the Customers present different configuration.

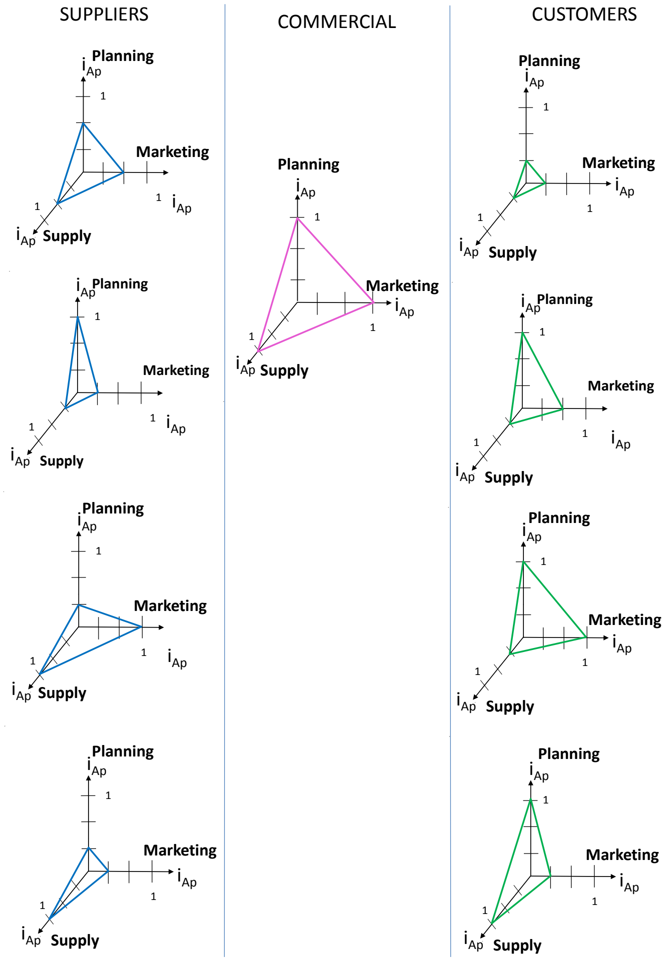
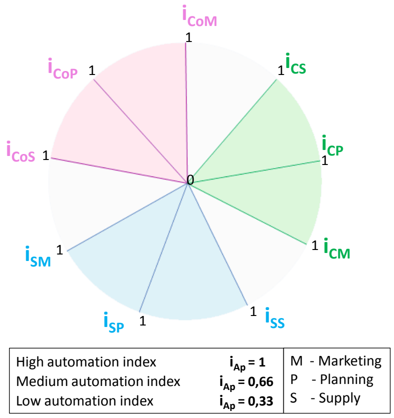


Figure 6 – Possiblities of configurations of Extended Enterprise Integration

To clean up the graphic representation, plan one will be used where the nine axles in “radar form” are shown according to
Figure 7.

Figure 7 describes the plan representation for the extended enterprise integration model associated with the automation index varying from 0 to 1, where:

- i_{CoS} - commercial automation index related to supply activity.
- i_{CoM} - commercial automation index related to marketing activity.
- i_{CoP} - commercial automation index related planning activity.
- i_{SS} - supplier automation index related to supply activity.
- i_{SM} - supplier automation index related to marketing activity.
- i_{SP} - supplier automation index related to planning activity.
- i_{CS} - customer automation index related to supply activity.
- i_{CM} - customer automation index related to marketing activity.
- i_{CP} - customer automation index related to planning activity.



High automation index	$i_{Ap} = 1$	M - Marketing
Medium automation index	$i_{Ap} = 0,66$	P - Planning
Low automation index	$i_{Ap} = 0,33$	S - Supply

Figure 7 - Representation of integration model in “radar form”

Possible configurations of extended enterprise integration model in radar form will be analyzed in the following section.

V. Possible Configurations for Extended Enterprise Integration Model

A. Low integration of extended enterprise and among Commercial, Suppliers and Customers

Where Commercial Business is without integration due to high values for i_{CP} and low values for I_{CS} and i_{CP}, Supplier without integration due to high values of i_{SM} and i_{SP} and low

ISBN: 978-1-61208-202-8

values of i_{SS}, and customers without integration due to high values of i_{CS} and i_{CM} and low values for i_{CP}. Figure 8 shows the plan representation for this condition where dashed lines represent the integration level between activities and the continuous lines represent the integration level between companies.

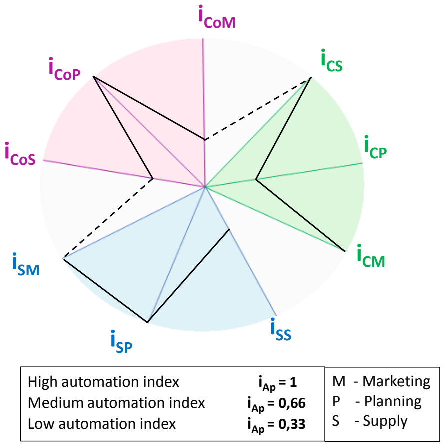


Figure 8 - Automation without structural integration in both: activities and extended enterprise

B. *High integration of Commercial processes, but low integration with Customers and Suppliers.*

The situation in which there is integration among the activities of each company is describe in Figure 9, but there is no transorganizational integration in the extended enterprise taking in account that each organization has the same numerical values for each axle but they are not the same for the three companies.

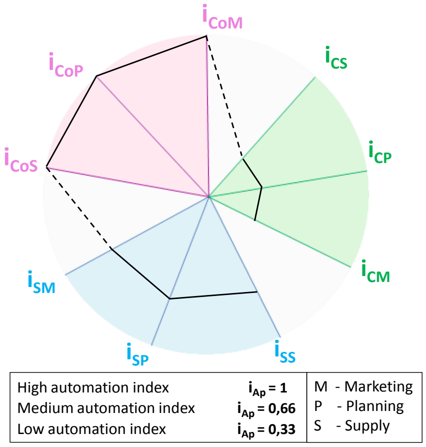


Figure 9 - Integration inside the companies, but without transorganizational integration.

The Commercial processes have integration between the set of activities with automation indexes close to 1, but the customers, even though present internal integration, have low values of automation indexes which does not allow integration with the Commercial processes. It is the same situation with the supplier: even though this supplier is internally integrated, it presents a medium automation index, and does not reach integration with the Commercial processes. In this case, the Commercial provide resources that cannot be shared with customers and suppliers. So, there are situations where there are unequal investments in Information Technology. In this configuration, the Company acquired an ERP system but still uses tables and lists to communicate with customers and suppliers, e.g., via Fax.

Besides the usage of ERP system, the company still uses tables and lists to communicate with customers and suppliers, e.g., via Fax. On the other hand, the customer employs a manual planning but has a CRM to communicate with clients and a SCM to communicate with suppliers. In addition to that, the supplier has an ERP system, uses CRM with its Customers but communicates with its suppliers manually, and in this configuration does not use information systems with their customers.

C. *Extended enterprise with total structural integration*

This configuration is the best situation for the extended enterprise: there are three situations, as shown in Figure 10, where all extended enterprise has the approximately same automation index, both inside the companies and between companies. There are three automaation situations - high (i_A = 1), medium (i_A = 0,67) and low (i_A = 0,33) integration.

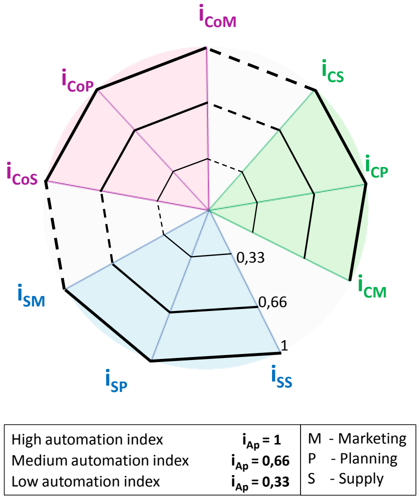


Figure 10 - Extended enterprise with structural total integration

VI. CONCLUSIONS

The proposal of an Extended Enterprise Model with characteristics of integrating the enterprise itself, its customers and suppliers was presented. The integration is obtained through the application of automation indexes with approximately the same numerical value in the nine axles. The integration concept developed in this paper is based upon equivalent numerical values of the automation indexes that will facilitate the information flow, due to the fact that, with equivalent numerical values of the automation, the information exchange interfaces have the same nature and technology.

This concept of integration applied to the frame - Commercial, Supply Chain and Customers networks presents an extended enterprise architeture that will enable the supply chain and customers network to operate in optimal conditions, diminishing the response time and saving resources.

The application of this research will be made in an automotive extended enterprise.

ACKNOWLEDGMENT

Professor Ricardo Suyama for his contributions to this work.

REFERENCES

[1] Mengoni, M., Graziosi, S., Mandolini, M., Peruzzin, M. A knowledge-based workflow to dynamically manage human interaction in extended enterprise. *International Journal on Interactive Design and Manufacturing.* 2011, Vol. 5, 1, pp. 1–15.

[2] Agostinho, O. L**.** *Manufatura como pré-requisito de competitividade.* Tese de Livre Docência - Faculdade de Engenharia Mecânica, Universidade Estadual de Campinas, Campinas: s.n., 1995.

[3] Rocha dos Santos, L., Silva, S. V., de Campos, R., 2007, in IFIP International Federation for Information Processing, Volume 254, Research and Practical Issues of Enterprise Information Systems II Volume 1, eds. L. Xu, Tjoa A., Chaudhry S. (Boston: Springer), pp. 343- 347.

[4] Camarinha-Matos, L., Afsarmanesh, H., Galeano, N., & Molina, A. (2009). Collaborative networked organizations – Concepts and practice in manufacturing enterprises *Computers & Industrial Engineering. doi:10.1016/j.cie.2008.11.024 , 57* (1), pp. 46-60.

[5] Lehtinen, J. and Ahola**,** T**.** Is performance measurement suitable for an extended enterprise? [ed.] Emerald Group Publishing Limited. *International Journal of Operations & Production Management.* 2010, Vol. 30, 2, pp. 181-204.

[6] Davis, E. W. and Spekman, R. E**.** *The Extended Enterprise - Gaining competitive advantage throuhg collaborative supply chains.* Upper Saddle River: Prentice Hall Books, 2004.

[7] Bititci, U. S., Mendibil, K., Martinez, Martinez., Albores, P. Measuring and managing performance in extended enterprises. [ed.] Emerald Group Publishing Limited. *International Journal of Operations & Production Management.* 2005, Vol. 25, 4, pp. 333-353.

[8] Molina, A., Panetto, H., Chen, D., Whitman, L., Chapurlat, V., Vernadat, F. Enterprise integration and networking: challenges and trends. *Studies in Informatics and Control.* December 2007, Vol. 16, 4, pp. 353-36

[9] Jagdev, H., Vasiliu, L., Browne, J., Zaremba, M. A. semantic web service environment for B2B and B2C auction applications within extended and virtual enterprises. [ed.] Elsevier. *Computers in Industry.* 2008, 59, pp. 786–797.

[10] Post, J. E., Preston, Lee. E. and Sachs**, S.** Managin the extende enterprise - The new stakholder View. *California Management Review.* Fall 2002, Vol. 45, 1.

[11] Slone, R. E., Dittmann, J. P. and Mentzer, J. T**.** *The New Supply Chain Agenda: The five steps that drive real value*. Boston: Harvard Business Press, 2010. p 216.

[12] Ayers, J. B**.** *Supply Chain project management: a structured collaborative and measurable approach.* 2a. Boca Raton: Auerbach Publications is an imprint of Taylor and Francis Group, LLC, 2010. 978-1-4200-8392-7.

[13] Madapusi, A. and Miles, G. Routines in enterprise application systems. *Management Research Review.* 2011, Vol. 34, 1, pp. 75-97.

[14] Laurindo, F. J. B. and Rotondaro, R. G. (org.). *Gestão Integrada de Processos e da Tecnologia da Informação.* [ed.] 978-85-224-4507-3. 1 ed. São Paulo: Editora Atlas S.A., 2006. ISSN 978.

[15] Wang, H. J. and Wu, H**.** *Supporting process design for e-business via an integrated process reposit.* s.l.: Springer Science+Business Media, LLC, 2010. Inf Technol Management. DOI 10.1007/s10799-010-0076-z.

[16] Chen, D., Doumeingts, G., Vernadat, F. Architectures for enterprise integration and interoperability: Past, present and future. *Computers in Industry*. 2008, Vol. 59, pp 647–659.

[17] Lauro, V. *The Diffusion of New Collaborative Approaches in Technology Intensive Contexts: Literature Analysis and Empirical Evidences on Virtual and Extended Enterprise Models*. Doctoral Thesis in Economics and Technology Management. Università degli Studi di BergamoVincenzo 2011.

A Method for Finding Similar Time Series and Forecasting with Calendar Constraints – a Commercial Bank Case Study

Krzysztof Kania
Knowledge Engineering Department
University of Economics
Katowice, Poland
krzysztof.kania@ue.katowice.pl

Jerzy Michnik
Operations Research Department
University of Economics
Katowice, Poland
jerzy.michnik@ue.katowice.pl

***Abstract*— In many cases, especially in business activities we can observe constrains regarding an arrangement of the calendar. Sometimes they can be neglected, but sometimes they have a significant impact on the course of events. This article presents a simple method supporting the forecasts of this type of phenomena based on a concept of calendar similarity that may supplement traditional forecasting methods. Presented method has been used in the commercial bank to predict a volume of the documents to process.**

***Keywords* - forecasting; calendar; qualitative methods.**

I. Introduction

Business reality is always time-dependent and forecasting is one of frequently formulated tasks in successful management. Forecasting plays a special role when:

- resources cannot be stored (e.g., energy or work) or storage is very costly,
- shortages of resources may lead to distortions in the functioning of the organization or may lead to major losses,
- storing too large volume of resources creates a risk of wasting resources (a lack of jobs for workers, penalties for unused capacity),
- rapid change in the volume of resources is not possible or is costly.

In such cases, we need an accurate forecast early (sometimes even a few months earlier) and for every subperiod of the forecasted range (i.e., for every day of a whole month in advance). Some business activities and events exhibit variations depending on the specific day of the week or month. Therefore, they can be described as dependent on the calendar. Dependence of the predicted phenomena, events and processes with the calendar may be due to legal regulations or generally accepted norms and customs.

Examples of such phenomena include the different media consumption in the industry on some days (for example, before various holidays, days off, end of the year), the volume of traffic or the number of waiting customers. A similar problem also arises in planning of deliveries to large shops. Another example is volume and type of documents received in offices, banks and post offices during particular days connected with payments, deposits, withdrawals, transfers or the load on the servers for electronic services as well. The specificity of these phenomena is that the relatively stable long cycles (yearly, monthly, weekly) interfere with arrangement of weekdays, public holidays and additional days off.

Forecasting in such conditions requires consideration of factors that are disregarded in the analysis of phenomena that have uniform distribution in time or are insensitive to the arrangement of the calendar. Classical statistical methods do not include a quantitative prediction of the calendar. Methods for finding similarities in time series – very useful in many cases – do not account calendar directly and involved constraints as well. Hence, we need to use additional qualitative methods based on the large amount of data stored in data warehouses. The tasks of this class are in the scope of Business Intelligence systems and to implement them we use a variety of statistical tools and techniques such as neural networks and sequence analysis (see [1][2][3]). The paper presents a method to support prediction with taking calendar constraints under consideration with the example of the task of calculating the volume of processed documents.

The paper is organized as follows. In the next section (II) we define the problem raised in the bank and describe the goal of our research. In Section III, the whole procedure is outlined and presented in details using the example of one forecasted month. The paper ends with the short conclusions.

II. Problem definition

The method proposed has been developed for forecasting a number of people needed to process documents in one of the commercial banks and will be illustrated by the particular example. However, it seems that the method in question can be easily generalized and applied in other fields.

Bank branches receive traditional documents (transfers, fees, taxes, etc.) from their customers. The amount of work to be done by the bank staff depends on the volume of documents and the work structure (processing different types of documents requires a different amount of work). An important limitation is connected with the necessity of working within strict deadline (time of opening sessions of interbank payments). Our problem has to fulfill the following conditions:

- The forecast must be prepared at least one month earlier to plan holidays or prepare other work for employees and to minimize the number of people remaining in readiness to perform work.
- During every month there are two special days (10th and 15th), when the number of documents is the highest (dates for paying taxes and other fees).
- Strong influence of additional factors such as changes in commission fees, opening/closing of branches, changes in types of documents, and their structure, and introducing the new rules for an electronic exchange of information generates some need for cooperation with an experienced user in order to verify our predictions.
- It is necessary to take into account the distinction between weekdays and holidays.

The goal is to predict a number and type of documents to be processed in successive working days of the forecasted month as precisely as possible and, consequently, to determine a number of people needed for processing of the documents. For the procedure we assumed that:

- Total volume for the predicted period is known.
- Distributions of intensity of work from the historical data are known.
- Calendar arrangement causes large changes in the distribution of the phenomenon.

The total monthly volume of documents is predicted with the help of traditional statistical methods such as trend analysis, analysis of the relative and absolute deviations and analysis of the cycles of higher order (in this case - annual). In practice, these methods let us to predict very closely the total volume of documents for specific month.

III. PROCEDURE FOR DIVIDING THE VOLUME OF WORK

The most difficult part of the whole research was to find the distribution of the total volume of documents for every particular day of the month. Fig. 1 shows the percentage distributions of the number of documents in the following days of four different months taken from historical data (gaps in the distributions relate to days off).

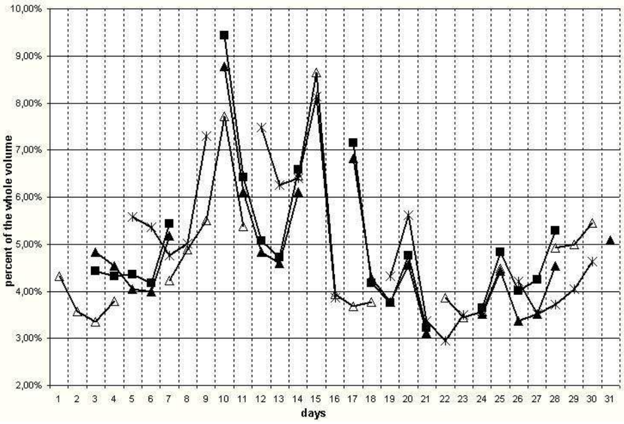


Figure 1. Sample month distributions of the intensity of the work

The monthly distributions of the phenomenon are dissimilar due to the different weekdays arrangement with respect to consecutive days of the month. In some months, the maxima fall to approximately 10th and 15th day, but in the other months distributions are different because the maxima fall on days off. In these cases the months maxima shift to or spread out on the preceding or following working day.

Due to these differences and gaps, a forecast based on the whole set of distributions leads to nowhere. For the same reason other methods of analyzing cyclic time-series like classical statistical analysis, Fourier analysis or wavelet analysis – effective in predicting of the continuous and uninterrupted time series – in this case turn out to be useless (see [4][5][6]). For that reason a new procedure has been proposed. Its outline is shown in Fig. 2.

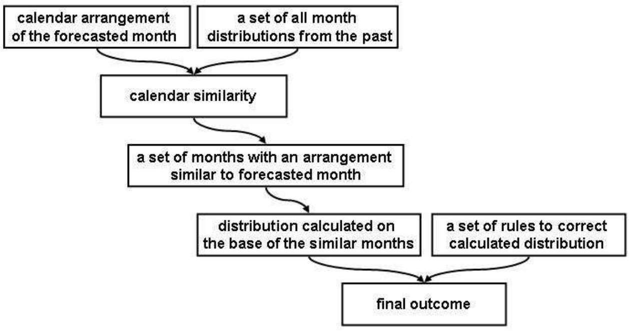

calendar arrangement of the forecasted month
a set of all month distributions from the past
calendar similarity
a set of months with an arrangement similar to forecasted month
distribution calculated on the base of the similar months
a set of rules to correct calculated distribution
final outcome

Figure 2. Outline of the procedure for finding distribution in a particular month

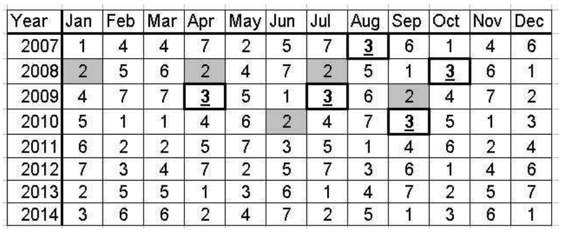
To distribute the volume of work over the working days of the particular month it is necessary:

- To find months with identical or similar calendar arrangement to the forecasted month.
- To incorporate Saturdays, Sundays and additional holidays.
- To improve the method with a set of rules correcting the initial distribution.

The procedure can be supported by Excel solution, so the whole knowledge needed to find similar months and to use additional rules was stored in 4 interrelated matrixes. It was decided to record information in the form of a matrix rather than in procedures or functions, as in a spreadsheet, matrices can be easily operated by users (e.g. introduction of a new day off) without additional tools.

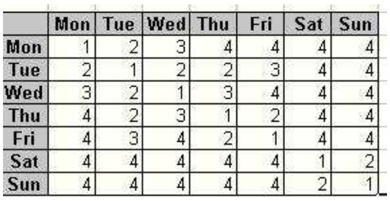
The first step of the procedure – determining the set of months that are similar in the calendar arrangement to the month of the forecast – is based on the content of the two first matrixes (Fig. 3 and Fig. 4). They are associated with a plain observation that the day of the week that starts the month determines the month arrangement until the 28th day. Months have different lengths but this is not significant as the analyzed phenomenon is not volatile at the very end of the month. That means, for example that February, although being 28 days long, may be in terms of a calendar arrangement similar to a longer month (June or July for example).

The second matrix (Fig. 4) contains ranks needed for finding in historical data months with identical or very similar calendar arrangement as in the predicted month. The rank equals to 1, in the matrix means that calendar arrangement of two months is identical (days off and special days 10th and 15th, in the same places). The rank equals to 2 means that calendar arrangement of two months is very similar but not identical, what causes changes in distribution and so on. Ranks are constant and was found through analyzing calendar, interviews with users supported by a graphical analysis of historical data and projections through the simulation.



Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2007	1	4	4	7	2	5	7	**3**	6	1	4	6
2008	2	5	6	2	4	7	2	5	1	**3**	6	1
2009	4	7	7	**3**	5	1	**3**	6	2	4	7	2
2010	5	1	1	4	6	2	4	7	**3**	5	1	3
2011	6	2	2	5	7	3	5	1	4	6	2	4
2012	7	3	4	7	2	5	7	3	6	1	4	6
2013	2	5	5	1	3	6	1	4	7	2	5	7
2014	3	6	6	2	4	7	2	5	1	3	6	1

Figure 3. A part of the matrix containing the number of weekday that starts a month (in Poland Monday is the 1st day of the week)



	Mon	Tue	Wed	Thu	Fri	Sat	Sun
Mon	1	2	3	4	4	4	4
Tue	2	1	2	2	3	4	4
Wed	3	2	1	3	4	4	4
Thu	4	2	3	1	2	4	4
Fri	4	3	4	2	1	4	4
Sat	4	4	4	4	4	1	2
Sun	4	4	4	4	4	2	1

Figure 4. The matrix of months' similarity rank

For example, in respect of the calendar, September 2010 (started at Wednesday) is the same as April and July 2009, October 2008, etc. (bolded and underlined in Fig. 3) as they starts with the same weekday and have rank equal to 1 in the matrix of months' similarity, and is very similar to June 2010, September 2009, January, April and June 2008 (shaded in Fig. 3) as these months started at Tuesday and the rank between Tuesday and Wednesday is equal of 2.

The result of this part of a procedure is a set of the months in which the distribution of intensity of work are actually comparable. In practice only months with rank 1 or 2 were used in forecasting because the distributions from the other months were too different. Since the calendar might be affected by some other factors (such as moving Easter or so called "long weekends"), the final decision on the choice of months is left to the analyst.

In fact, selecting only a few of many months decreases the basis of forecasting but on the other hand leaves only these months that have really similar distributions. Fig. 5 presents graphs for a selected month similar to September 2010. For the rest of the procedure we use an arithmetic mean from values for each day of the month (line with circles in Fig. 5).

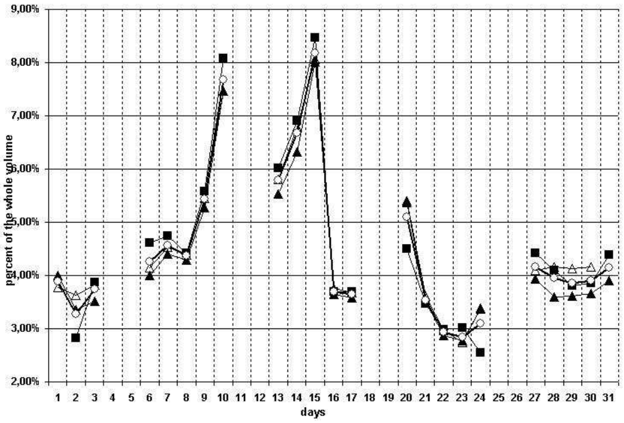
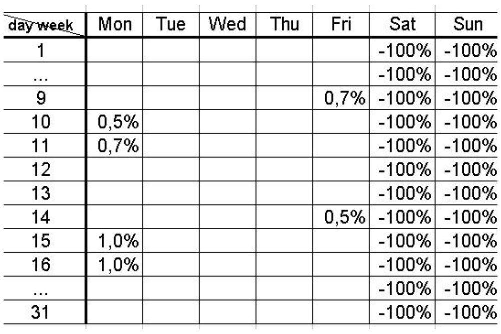


Figure 5. Distributions of selected months and a mean distribution (line with circles) of the intensity of the work for the forecasted month

As the distribution is obtained from different months (shorter and longer), the next step of the procedure is to verify and align the mean distribution obtained for the forecasted month. All verification rules have exactly two arguments in premise and exactly one value in the conclusion. This allowed for writing rules in the form of two-dimensional arrays. The arguments are: a day of the month and a day of the week (Fig. 6) or a month and a day of the month (Fig. 7). The element of the array is a percentage correction that should be made for the combination of arguments it belongs to. A user can use these correction rules to take into account additional factors that are concerned with particular days of week, month or year (Fig. 6 and Fig. 7). All the values in these matrices have been determined empirically and corrected on the basis of the experience of the bank staff, past observations and arrangement of the calendar for a current year.



day week	Mon	Tue	Wed	Thu	Fri	Sat	Sun
1						-100%	-100%
...						-100%	-100%
9					0,7%	-100%	-100%
10	0,5%					-100%	-100%
11	0,7%					-100%	-100%
12						-100%	-100%
13						-100%	-100%
14					0,5%	-100%	-100%
15	1,0%					-100%	-100%
16	1,0%					-100%	-100%
...						-100%	-100%
31						-100%	-100%

Figure 6. A part of the matrix of rules in the week-month relation

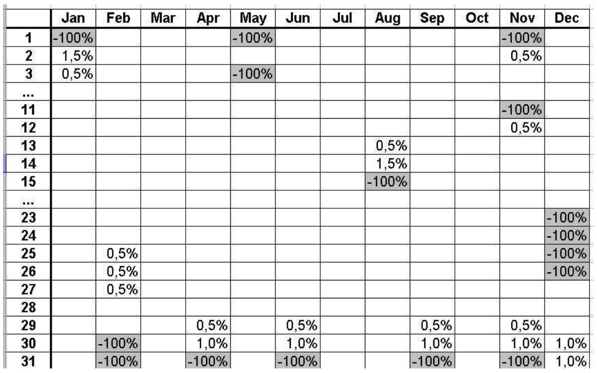
Information, contained in the matrix in Fig. 6, is presented to the user in the form of rules:

```
If    n-th day of the month falls on
                 particular weekday
then  change the value of that day by x%
```

For example, an element (9, Fri), of the matrix has a value 0.7%. This value corresponds to the rule:

```
If    the 9th day of month falls on
      Friday
then  increase the value of forecast on
      that day by 0.7%
```

This example reflects the knowledge that since the 10th (one of the special days) falls on Saturday, the volume of documents will be greater in the day before.



	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1	-100%				-100%						-100%	
2	1,5%										0,5%	
3	0,5%				-100%							
...												
11											-100%	
12											0,5%	
13								0,5%				
14								1,5%				
15								-100%				
...												
23												-100%
24												-100%
25		0,5%										-100%
26		0,5%										-100%
27		0,5%										
28												
29				0,5%		0,5%			0,5%		0,5%	
30		-100%		1,0%		1,0%			1,0%		1,0%	1,0%
31		-100%		-100%		-100%			-100%		-100%	1,0%

Figure 7. A part of the matrix of rules in the year-month relation

The last matrix (Fig. 7) contains values for rules correcting each day of the year especially due to holidays or the different length of the months. At the end of shorter months the volume of documents grows in relation to the mean distribution. Similarly, higher intensity of work is observed right before or after holidays (for example, in Poland August 15th or November 1st). These values are also presented to the user as suggestions and they are as follows:

```
On September 29, the system proposes to
      increase the value by 0.5%.
On September 30, the system proposes to
      increase the value by 1.0%.
On September 31, the system proposes to
      decrease the value by 100.0%.
```

These three particular rules show that the volume of work in September (shorter month) shifts to two previous days.

A user may accept or reject proposals adjusting the distribution to get the final form. Sometimes using the rules may result in the situation that the sum of intensities of work moves away from 100%. In that case an analyst can also manually make changes to the proposed schedule increasing or decreasing all the values throughout the forecast period.

Fig. 8 shows the final result of this procedure – the distribution of the intensity of work during the entire month. Comparing it with the mean distribution we can see that it is slightly different due to correcting rules (circled parts of the distribution).

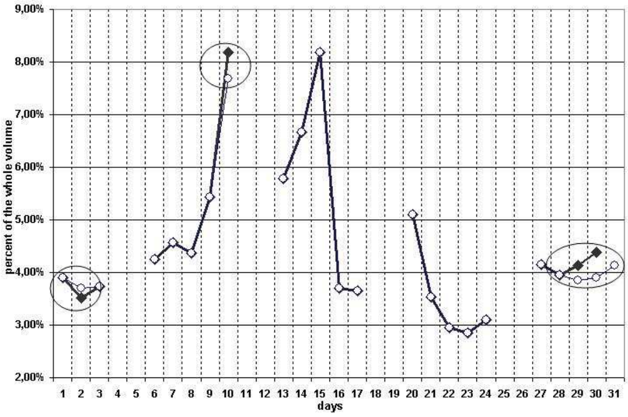


Figure 8. Corrected distribution for September 2010

At the last phase of the procedure, we calculate a number of documents for each day of the forecast by dividing the total monthly volume of documents according to obtained distribution. And finally as an average workers' loading is known we can calculate a number of people needed to process the documents for every day.

IV. Conclusions

The presented method has become a module of a larger system that was implemented in the commercial bank. The procedure described above, has replaced previously used forecasting method based on a simple analogy and improved its results. Moreover, the whole procedure was improved with mechanism for storing data and forecasts in the database. This allowed to connect forecasts with the scenario method and to easily conduct the what-if analysis. Currently, due to introduction of electronic banking, a number of traditional documents processed has declined significantly but it seems that the proposed method could be used for a prediction of other phenomena whose course depends on the specific arrangement of the calendar.

References

[1] Adamo J.M.: Data Mining for Association Rules and Sequential Patterns, Springer-Verlag, New York, 2001.

[2] Han J. and Kamber M.: Data Mining Concepts and Techniques, Academic Press, 2001.

[3] Kovalerchuk B. and Vitayaev E.: Data Mining in Finance, Kluwer Academic Publishers, 2000.

[4] Kania K.: "A New Measure and Symbolic Method that Supports Finding Similarity in Time Series", in: Business Information Systems, W. Abramowicz, G. Klein (eds.), Colorado Springs, USA, 2003, pp. 124-131.

[5] Caraca-Valente J. and Lopez-Chavarrias I.: "Discovering Similar Patterns in Time Series", 6th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data mining, Boston, 2000, pp. 497-505.

[6] Keogh E. and Pazzani M.: "A simple dimensionality reduction technique for fast similarity search in large time series databases", 4th Pacific-Asia Conf. On Knowledge Discovery and Data Mining, Kyoto, 2000, pp. 122-133.

Crowdsourcing Supported Context Detection for Improving Information Search Activities

Michael Beul, Stefan Eicker
paluno - The Ruhr Institute for Software Technology
University of Duisburg-Essen
Essen, Germany
{michael.beul | stefan.eicker}@paluno.uni-due.de

Abstract— **In environments that require the use of software applications, intervals where application functionality, tools, methods and technical systems are changing are often very short. The process of searching for relevant information about a specific issue is frequently executed and time-consuming. Because of the availability of a nearly unlimited amount of data, people spend a lot of time in formulating search queries and evaluating the relevance of the search results. In this paper, we describe a generic approach that improves the information search and retrieval process of different activities with the use of context information. One main goal is the integration of the "crowd" at different stages of this process by combining collective intelligence concepts with context-aware systems. This combination can be used to automatically reduce information overflow by filtering irrelevant data. Furthermore, a real-time information retrieval process without manual search impulses is provided. We also present a prototype as a proof of concept in order to validate feasibility and benefit.**

Keywords - Context Detection; Context Awareness; Information Retrieval; Information Search Process; Collective Intelligence; Crowdsourcing.

I. INTRODUCTION

Nowadays, the importance of information is at a very high level. At the same time, the availability of data (including potentially useful information) is given at any time and location. One major problem is the efficiency of information search and retrieval processes. A study by Delphi Group [1] predicted that employees spent a large amount of time in searching for information (see Figure 1). Only 10% of the respondents disagreed to the statement "Finding the information I need to do my job is difficult and time consuming". The three major impediments finding suitable information are identified as:

- Information changes constantly (41%)
- I don't have good search tools (26%)
- I often don't know exactly what I'm looking for (13%)

Another problem is the fact that a great number of people have suboptimal strategies while using web search engines in order to find relevant information.

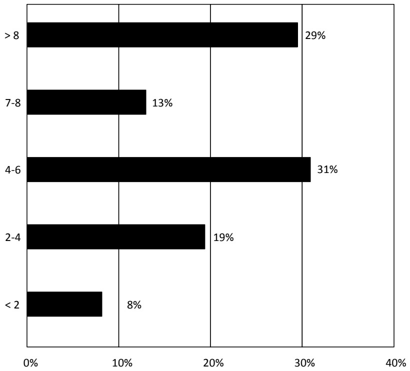


Figure 1. Hours per week spent searching for information [1]

According to a study by Nielsen Norman Group [2], only 1% of the tested persons change their search strategy if the first search results do not fit to the entered issue. One further result of the study is that "users have extraordinarily inadequate research skills when it comes to solving problems on the Web" [2]. Moreover, the study shows that the advanced search features of search engines are not used by most of the test persons "And when they do, they typically use it incorrectly - partly because they use it so rarely that they never really learn how it works" [2].

The main objective of the approach we present in this paper is to handle these impediments by improving the information search and retrieval process using a combination of collective intelligence and context-awareness. Furthermore, the approach addresses passive information retrieval, where people can receive information without even searching for it, respectively receiving solutions for problems users are not aware of. Based on general concepts, the approach can be used in different domains. In Section 4.1, we present three scenarios that show the capability and flexibility of the approach.

The paper is structured as follows. Section 2 introduces current concepts according to context aware systems and our adaption and implementations. In Section 3, we discuss the usage of collective intelligence in order to optimize the information search and retrieval process. Section 4 presents different scenarios and a platform that is used as proof of concept and as a base for the evaluation and validation of the developed approach. Finally, Sections 5 and 6 draw concluding remarks and present related and future work in this area.

ISBN: 978-1-61208-202-8

II. Context Detection and Usage

Several definitions of the word “context” can be found in literature. Some of them primarily refer to location, environment, identity, time or situation [3][4]. An often referenced and more detailed definition is given by Abowd et al.: “Context is any information that can be used to characterize the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves [5].”

Concepts, where concrete context information is used to influence a behavior of systems, devices or environments, have been in the focus of research for several years. Particularly in the field of pervasive or ubiquitous systems context information like location or time is often used to change environment behavior. The implementation of context-aware systems depends on the concrete domain. A suitable approach for architectures of context-aware systems in distributed systems is presented by Chen as the Middleware infrastructure [6]. According to this approach, Baldauf et al. identify a common architecture and present a layered conceptual framework for context-aware systems (see Figure 2) [7].

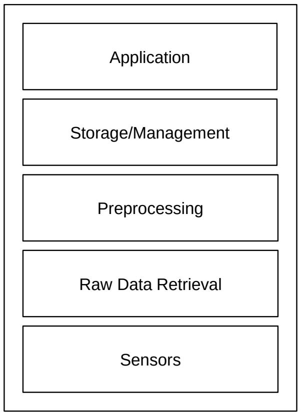


Figure 2. Layered conceptual framework for context-aware systems [7]

At the first layer, sensors collect usable information concerning the concrete context. These sensors can be physical, logical or virtual [8]. The collected data from all sensors (raw data) is transferred to the second layer. The Preprocessing layer includes functionality to convert the raw data to useful information (e.g., reverse geocoding from GPS coordinates). The Storage/Management layer offers the gathered data to the Application layer in an organized form. Finally, the Application layer itself represents the client that uses the detected context information for application functionality. Based on this layered conceptual framework, we developed our novel collective intelligence driven approach.

A. Sensors

As mentioned above context-aware systems use sensors in order to detect useful information inside a concrete environment. In [9], we identified three different environments in the field of software applications:

- *Black-Box Environments* allow no access to the underlying structure (e.g., applications with no access to the application logic or source code)
- *White-Box Environments* allow full access to the underlying structure (e.g., source code)
- *Gray-Box Environments* offer limited access (e.g., support for plug-ins)

The different environments require different sensors for context detection. We differentiate between global sensors and local sensors. Global sensors are high level sensors, e.g., integrated in an operating system to provide global sensor data. Local sensors are low level sensors, e.g., integrated into a concrete application providing application specific sensor data. The data, measured by the sensors, is used in a rule specification process that is described in the following chapter.

B. Rule System

Rules connect sensor data with functionality respectively information, and thus allow mappings between information and concrete user contexts. Considering the crowdsourcing aspect (see Section 3), we created a rule system that is easy to use, but at the same time powerful to support sophisticated context mappings. We use the boolean algebra syntax (with the operations AND, OR, NOT) in combination with fuzzy logic concepts. This provides the creation of simple rules as well as complex rule definitions that are at the same time human readable (see Figure 3).

The structure shows that a rule is a combination of (sub-) rules and sensors. Sensors measure different environment situations. The definition of a sensor includes id, version, value, data type and valid operators. Fuzzy logic information can also be integrated in the sensor definition. Figure 4 shows an XML-representation of the rule presented in Figure 3. One of the sensors (id 78) includes linguistic fuzzy logic expressions in order to compare values to intervals (operator=”equals” value =”high“).

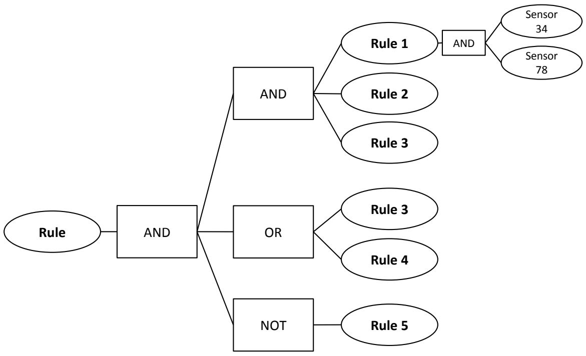


Figure 3. Rule Structure

```
<Rule id=”199”>
  <AND>
    <AND>
      <Rule id=”1”>
        <AND>
          <Sensor id=”34”
             operator=”equals” value=”high” />
          <Sensor id=”78”
             operator=”greaterthan” value=”50” />
        </AND>
      <Rule id=”2” />
      <Rule id=”3” />
    </AND
    <OR>
      <Rule id=”3” />
      <Rule id=”4” />
    </OR>
    <NOT>
      <Rule id=”5” />
    </NOT>
</AND>
</Rule>
```

Figure 4. XML representation of rules

The rule system is decoupled from the sensor system. This allows a separation of sensor creation and rule creation. Sensor creators develop sensors and publish sensor information. Rule creators only need to know information about available sensors and associated sensor data in order to create their rules. Thus the specific know how of the sensor creators (e.g., sensor programming and application integration) as well as the strengths of the rule creators (e.g., expert and process knowledge) can be used in the best way.

C. *Context Mapping*

The relevance of identified information depends to a certain extent on the mapping between sensor data and information. At current status the presented approach provides the mapping types *Explicit Mapping*, *Implicit Mapping* and *Search Query Generation*.

The three mapping types provide both directions of information retrieval, active and passive, respectively information pull and information push. Explicit Mappings support mapping to concrete information, e.g., a document path or a web site. This mapping type provides the best quality of relevance (highest rating). Implicit Mappings map sensor data to application functionality. In a separate process information (e.g., related documents, videos or forum messages) is mapped to application functionality. This concept allows an indirect mapping of context and information. Authors can map their created documents to concrete application functionality. If a rule exists which includes a mapping of sensor data to this functionality the related documents will be displayed.

The third mapping type uses search query generation in order to detect relevant documents. According to the user studies described in [2], we use the collected context information to automatically create queries that use the benefits of a concrete search engine, e.g., advanced search parameters, multilanguage search, use of synonyms, etc. Therefore, in many cases the generated queries are more purposeful than user created queries.

D. *Structure*

Figure 5 shows an extract of structure with relations between context relevant elements. Users act in environments and can own a specific role in this context. They have a need for information that can either be conscious or unconscious. Environments offer a range of functionalities and are observed by sensors. Rules include functionalities, information and sensors. The crowd plays a major role in this structure. It can influence nearly all important elements of the approach. People, that are part of the crowd, develop sensors, define rules, specify functionalities and create information of different types. Advantage of this structure is that the different activities are independent from each other; the particular skills of the crowd members can therefore be used ideally. In the next chapter, we present concepts for the integration of the crowd into the activities.

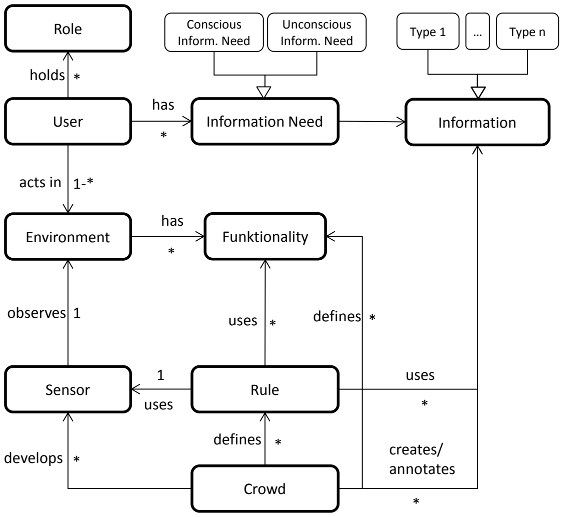


Figure 5. Main Elements and Relations

III. CROWDSOURCING

A major goal of the presented approach is the integration and participation of the mass (crowd). This affects the information retrieval process as well as the context detection and metadata enhancement. Research activities, concerning the “potential of groups”, have been in focus for a couple of years and in different domains, e.g., biology, social science and computer science. The generic term of this research discipline is *Collective Intelligence*.

A. *Collective Intelligence*

Levy defines Collective Intelligence as a „form of universally distributed intelligence, constantly enhanced, coordinated in real time, and resulting in the effective mobilization of skills. [...] The basis and goal of collaborative intelligence is the mutual recognition and enrichment of individuals rather than the cult of fetishized or hypostatized communities [10].” Because of the recent internet infrastructure, collective intelligence has gained in importance in different scenarios. “Collective Intelligence has received a new meaning in recent years, especially through the emergence of new (mostly Web 2.0) applications and user generated content [11].”

Malone et al. identified a small set of building blocks according to most types of collective intelligence systems. In order to classify these building blocks four key questions have to be answered (see Figure 6) [12]: *Who is performing the task? Why are they doing it? What is being accomplished? How is it being done?*

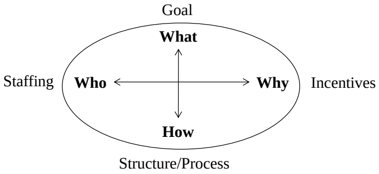


Figure 6. Elements of collective intelligence building blocks or “genes” [12]

According to the approach we present in this paper, the answers to the key questions are as follows:

Who is performing the task? – The crowd, which is represented by an independent mass of people [13]. Participating persons can hold different roles, e.g., author of a document, expert inside a forum/domain, rule creator or information/functionality mapper. They all are part of the crowd and can collectively optimize the entire process.

Why are they doing it? – In most recent collective intelligence systems, the motivation of participation is founded in “Money”, “Glory” or “Love” [12]. In the presented approach another major reason to participate is the own benefit of the results. On the one hand, this applies to the quality and relevance of the received information (reader’s view). On the other hand, the participation provides a targeted distribution of relevant documents (author’s view).

What is being accomplished? – The crowd can affect the quantity, quality and relevance of the detected information in different ways and with the accomplishment of different activities. With regard to the presented approach, main activities concern sensor development, functionality and rule definitions as well as metadata annotations. The next sections describe the main activities in detail.

How is it being done? – We developed a platform that integrates different activities and artifacts of the presented approach. Sensors can be published, rules can be defined and mappings can be provided. The data is stored in a shared database. The local tool for user interaction and information presentation also uses this database (see Section 4) in order to provide real-time changes.

B. *Creation of Sensors*

In order to analyze a concrete situation of users, sensors are required that measure data inside environments. Our approach provides the creation of (virtual) sensors by the crowd. Therefore, we defined a one way interface between sensor and processing component. The communication is realized as a simple REST-request containing key value pairs of sensor data. This concept provides an open and technology independent sensor development. According to Black-Box environments (see Section 2.1), tools can be developed that are able to catch user interaction without intervention to the application itself. Furthermore, the crowd can develop and share plug-ins for using in applications of type Gray-Box. The most efficient way is sending sensor data directly from the application logic. Our approach supports the integration into the application architecture with minimal intervention.

C. *Context Mapping*

As described above, a context describes a concrete situation of objects. Strang and Linnhoff-Popien identified six types of modeling context which cover different requirements: *Key-Value Models, Markup Scheme Models, Graphical Models, Object Oriented Models, Logic Based Models* and *Ontology Based Models*. While the Key-Value Models cover least requirements, the Ontology Based Models support most of them [14]. We use the Key-Value-Model to collect required data of the sensors. In order to map functionality to aggregated sensor data, we implement ontology-based models and adapted the approach from Wang et al. [15] for the creation of concrete context information for a wide range of domains and scenarios. Using a high-level ontology, the crowd can create ontologies for a concrete domain (e.g., application). The ontologies are also stored at the shared database and can be accessed by the crowd in order to share and optimize it.

D. *Enhancement of the Information Source*

An additional way to optimize the search result quality is the enhancement of documents with metadata. Nowadays, relevant documents are often web-based and belong to categories like websites, blogs, expert systems, wikis, forums, tweets or social communities. In order to integrate the crowd into the annotation process of relevant artifacts, an efficient practice has to be available; otherwise the willingness to participate decreases. Metadata annotation concepts like RDFa [16], microformats [17], Microdata [18] and Schema.org [19] cover these requirements. Currently we are developing a schema for forums and expert systems where the different types of entries can be annotated with additional information. These types are, e.g., questions, answers, accepted solutions and non-working solutions.

```
class
      -     HTML4 definition of the 'class' attribute. This
            metadata profile defines some 'class' attribute
            values (class names) and their meanings as suggested
            by a draft of "Hypertext Links in HTML".
      type
                  o     Type of the information. Note: this
                        property may not be repeated (information
                        may not belong in several types). Common
                        values are: QUESTION, ANSWER,
                        ACCEPTED_SOLUTION,
                        NON-WORKING_SOLUTION.
      context
                  o     Context of the information; includes
                        attributes like author, date,
                        application, manufacturer, version; is
                        represented by nested CoDIRContext [9]
                        record.
rel
      -     HTML4 definition of the 'rel' attribute. Here is an
            additional value.

      tag
                  o     Indicates that the referred resource
                        serves as a "tag", or keyword/subject,
                        for the enclosing X element.
```

Figure 7. Microformat profile (draft) for forums and expert systems

The additional metadata allows machines to filter non-relevant information according to a specific context. Figure 7 shows a proposal of a microformat profile for forum-based expert systems. This profile uses the CoDIR-microformat [9] as nested attribute.

IV. Proof of Concept

The following scenario demonstrates the usage and implementation of the presented approach in the context of software engineering activities inside the disciplines SE-Development and SE-Tools.

In order to validate feasibility and benefit of the presented approach, we developed and enhanced a prototype as a proof of concept and added support for tasks in software engineering activities. Figure 9 illustrates the core elements of the prototype.

Virtual Sensors collect context relevant information inside a specific environment. If a sensor detects a change of context, the new context is transferred to the Result & Interaction Interface (RII). The interface transfers the context data and additional parameters to the Information Broker. The Information Broker uses the context information, different search providers and repositories in order to search for relevant information. The results (received from the search providers and enhanced with ranking information) are send back to the RII and displayed to the user. The user can directly access the received resources and (optional) rate the relevance of the information in relation to the current context.

In the presented real-world scenario, the crowd has published a plugin for the IDE Eclipse to track user activities and system behavior. This plugin represents a virtual sensor. This virtual sensor is able to collect different information about the current context based on a domain specific ontology published by the crowd.

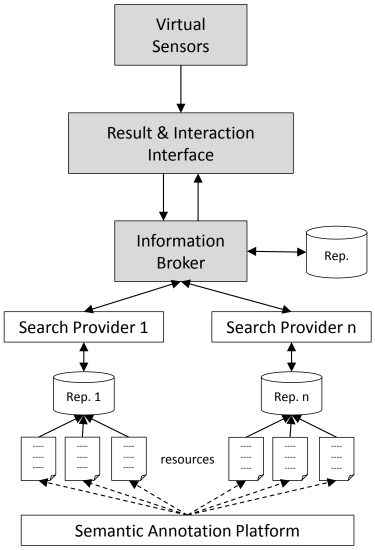
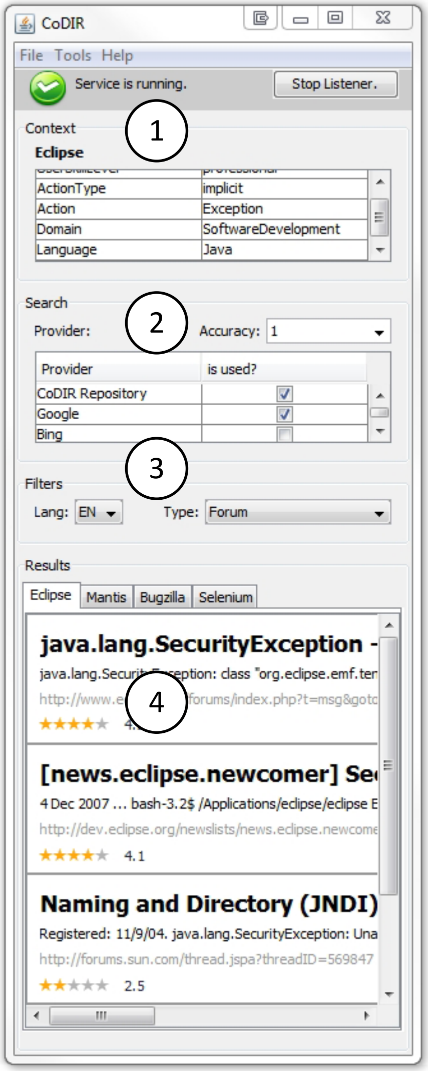


Figure 8. Core elements of the prototype

If the application for example throws a security exception, a context-object is generated and directly transferred to the RII. The RII is implemented as Rich Client Application (see Figure 8). At the top the details of the current context according to the context ontology model are displayed (1). The user can influence the search behavior by customizing different parameters, like the available search provider (2), the language or the document types (3).

The RII transfers all useful information to the Information Broker, which is implemented as a web service. We also implemented a Web Portal for publishing sensors, rules and functionality/information mappings. The Information Broker uses the context information and parameters to search for relevant resources, and transfers the results back to the IRR (4). Inside the result list, the user can rate the relevance of the result items in order to optimize the future search activities for all users.



Service is running.

Eclipse

ActionType	implicit
Action	Exception
Domain	SoftwareDevelopment
Language	Java

Provider: Accuracy: 1

Provider	is used?
CoDIR Repository	☑
Google	☑
Bing	□

Lang: EN Type: Forum

Eclipse | Mantis | Bugzilla | Selenium

java.lang.SecurityException -
java.lang.Securit eption: class "org.eclipse.emf.ter
http://www.e forums/index.php?t=msg&gotc
★★★★★ 4.

[news.eclipse.newcomer] Se
4 Dec 2007 ... bash-3.2$ /Applications/eclipse/eclipse E
http://dev.eclipse.org/newslists/news.eclipse.newcome
★★★★★ 4.1

Naming and Directory (JNDI)
Registered: 11/9/04. java.lang.SecurityException: Una
http://forums.sun.com/thread.jspa?threadID=569847
★★★★★ 2.5

Figure 9. Screenshot of the Result & Interaction Interface (RII)

V. RELATED WORK

Related work can be found in the area of context-aware, ubiquitous and pervasive systems. Several context ontologies have been proposed, mainly with focus on pervasive and mobile computing using physical sensors, e.g., CONON [15], SOUPA [20] and CoBrA Ontology [6]. Our approach addresses domain independent software applications in combination of virtual sensors that are integrated in different types of environments (application systems). According to search activities for relevant documents, several information retrieval approaches have been proposed, e.g., POLAR, a probabilistic object-oriented logical framework for annotation-based information retrieval [21]. Related work in the field of collective intelligence can be found in crowdsourcing approaches [13], Social Web Applications [11] and Web 2.0 technologies. We use the advantages of the different concepts to allow an influence of the crowd in nearly all phases of the process. In [22], Soylu and Causmaecker present an approach of empowering context-aware pervasive computing environments with embedded semantics. Differences to our approach are the missing reference to information retrieval concepts as well as the focused domain.

VI. CONCLUSION AND FUTURE WORK

In this paper, we presented a novel approach that enables the improvement of information search and retrieval processes in software application environments. The focus concentrates on software applications, where the demand on real-time information retrieval is given in different disciplines. The participation of the crowd is a fundamental component of our approach. Hence, an important task for future work is to constantly simplify the process for participation in order to enhance the willingness of the crowd. Using the prototype, we identified the need for a suitable visualization (e.g., graphs or maps) of the results as well as customized recommendation techniques. Currently, we are working on a hybrid and multidimensional recommender system that allows transparency and enhanced filter options. Another goal is the integration of our approach into environments that need security and trust features. Users can then be informed if they are in an unsecure context, or if they share confidential documents.

REFERENCES

[1] Delphi Group (n.d.): Information Intelligence: Content Classification and the Enterprise Taxonomy Practice. Boston, MA (2004)

[2] Nielsen, J.: Incompetent Research Skills Curb Users' Problem Solving. In: Jakob Nielsen's Alertbox (2011)

[3] Ryan, N., Pascoe, J., and Morse, D.: Enhanced reality fieldwork: The context-aware archaeological assistant. In: Computer Applications in Archaeology. Edited by V. Gaffney, M. van Leusen and S. Exxon. (1997).

[4] Hull, R., Neaves, P., and Bedford-Roberts, J.: Towards situated computing. In: Proceedings of the First International Symposium on Wearable Computers, pp. 146-153, Cambridge, MA , USA (1997)

[5] Abowd, G.D., Dey, A.K., Brown, P.J., Davies, N., Smith, M., and Steggles, P.: Towards a Better Understanding of Context and Context-Awareness. In: 1st international symposium on Handheld and Ubiquitous Computing, pp. 304-307. Springer-Verlag. Karlsruhe, Germany (1999)

[6] Chen, H.: An Intelligent Broker Architecture for Pervasive Context-Aware Systems. PhD thesis, University of Maryland, Baltimore County (2004)

[7] Baldauf, M., Dustdar, S., and Rosenberg, F.: A Survey on Context-Aware Systems. In: Inter-national Journal of Ad Hoc and Ubiquitous Computing, vol. 2, nr. 4, pp. 263-277. Inderscience Publishers. Geneva, Switzerland (2007)

[8] Indulska, J. and Sutton, P.: Location management in pervasive systems. In Proceedings of the Australasian Information Security Workshop (CRPITS 03), pp. 143-151, Australian Computer Society. Sydney, Australia (2003)

[9] Beul, M. and Eicker, S.: Don't call us, we call you. A community driven approach for (domain independent) context driven information retrieval (CoDIR). In Proceedings of Fifth Inter-national Conference on Signal Image Technology and Internet Based Systems (SITIS), pp. 458-464, IEEE Press, New York (2009)

[10] Levy, P: Collective Intelligence: Mankind's Emerging World in Cyberspace. Translated by R. Bononno. Perseus Books. Cambridge, MA (1997)

[11] Leimeister, J.M.: Collective Intelligence. In: Business & Information Systems Engineering, vol. 2, nr. 4, pp. 245-248 (2010)

[12] Malone T., Laubacher R., and Dellarocas, C.: Harnessing Crowds: Mapping the Genome of Collective Intelligence, Working Paper No. 2009-001. MIT Center for Collective Intelligence, Cambridge, MA (2009)

[13] Howe, J.: Crowdsourcing - How the power of the crowd is driving the future of business. RH Business Books, London, England (2009)

[14] Strang, T., and Linnhoff-Popien, C.: A Context Modeling Survey. In: Workshop on Advanced Context Modelling, Reasoning and Management. UbiComp, Nottingham, England (2004)

[15] Wang, H.H., Zhang, D.Q., Gu, T., and Gung, H.K.: Ontology Based Context Modeling and Reasoning using OWL. In: Proceedings of the Second IEEE Annual Conference on Pervasive Computing and Communications Workshops. Pp. 18-22, IEEE Press, New York (2004)

[16] Adida, B. and Birbeck, M.: RDFa Primer. W3C, http://www.w3.org/TR/xhtml-rdfa-primer, [retrieved: April, 2012]

[17] Microformats.org (n.d.): About microformats, http://microformats.org/about, [retrieved: April, 2012]

[18] Hickson, I.: HTML Microdata. W3C, http://dev.w3.org/html5/md/Overview.html, [retrieved: April, 2012]

[19] Schema.org (n.d.): Getting started with schema.org, http://www.schema.org/docs/gs.html, [retrieved: April, 2012]

[20] Chen, H., Perich, F., Finin, T.W., and Joshi, A.: SOUPA: standard ontology for ubiquitous and pervasive applications. In: 1st Int. Conf. on Mobile and Ubiquitous Systems: Networking and Services, pp. 258-267, Boston, MA (2004)

[21] Frommholz, I. and Fuhr, N.: Probabilistic, object-oriented logics for annotation-based retrieval in digital libraries. In: Opening Information Horizons - Proc. of the 6th ACM/IEEE Joint Conference on Digtial Libraries (JCDL), pp. 55-64, ACM. New York, NY, USA (2006)

[22] Soylu, A. and Causmaecker, P.D.: Embedded Semantics Empowering Context-Aware Pervasive Computing Environments. In: Symposia and Workshops on Ubiquitous, Autonomic and Trusted Computing, pp. 310-317. Brisbane, QLD (2009)

An Implementation of Discriminative Common Vector Approach Using Matrices

Mehmet Koc, Atalay Barkana
Electrical and Electronics Engineering
Anadolu University
Eskisehir, Turkey
{mkoc6, atalaybarkan}@anadolu.edu.tr

***Abstract*— If one sample per class is available in a face recognition problem, vector-based methods which use within-class scatter will fail. The reason for that is the zero within-class matrix. In this paper a two dimensional extension of the discriminative common vector approach (2D-DCVA) is proposed. The performance of the proposed method is compared with discriminative common vector approach (1D-DCVA) and two dimensional Fisher linear discriminant analysis (2D-FLDA) in ORL, FERET, YALE, and UMIST face databases in one sample problem. Our proposed method outperforms 1D-DCVA and 2D-FLDA in all databases.**

Keywords- one sample problem; common vector; DCVA; two dimensional FLDA

I. INTRODUCTION

Face recognition has many application areas such as security, law enforcement, person identification [1,2]. If only one sample per person is available, then the problem gets difficult. This situation is called one sample problem [3]. Methods which use within-class scatter such as conventional Fisher discriminant analysis (1D-FLDA) will suffer from one sample problem because within-class matrix is a zero matrix. Many algorithms have been proposed to overcome this challenge [3,4,5,6,7]. General tendency at these methods is generating the virtual samples to increase the training set size. But this is not the solution of the singularity problem because in face recognition problems dimension of the feature space is high with respect to the number of feature vectors. This problem is called small sample size problem [8]. One solution to overcome the singularity problem is using the two dimensional variant of one dimensional methods. Two dimensional Fisher discriminant analysis (2D-FLDA) [9] is a solution of the singularity problem in 1D-FLDA. This method was used in [4] and [5] after generating virtual samples. Also discriminative common vector approach (1D-DCVA) which is a variation of FLDA comes up with a solution that overcomes the singularity problem of 1D-FLDA [10].

In this work we proposed a two dimensional extension of the discriminative common vector approach. In order to obtain unique common vector for each class, we use feature vectors instead of feature matrices in the first stage of this method. Then we convert the common vectors into matrices and calculate the discriminative common matrices. In [11], feature matrices are used to obtain common vectors. This method though cannot get unique common vectors. A brief review of the discriminative common vector approach (1D-DCVA) is given in Sec.II. Two dimensional extension of the discriminative common vector approach is given in Sec.III. We used QR decomposition with column pivoting (QRCP) method to generate the virtual samples. QRCP method is given in Sec.IV. We tested the performance of 2D-DCVA in four different databases. Database descriptions and the experiments are given in Sec.V, and finally the results are discussed in Sec.VI.

II. DISCRIMINATIVE COMMON VECTOR APPROACH

Discriminative common vector approach (1D-DCVA) is first introduced in [10]. The method gives a solution to the limitations of methods that use the null space of the within-class scatter matrix.

Let C be the number of classes, N be the number of feature vectors from each class, and let $\boldsymbol{x}_m^i$ be the m^{th} feature vector from i^{th} class. Then the within-class scatter matrix can be written as

$$\boldsymbol{S}_W = \sum_{i=1}^{C}\sum_{m=1}^{N}(\boldsymbol{x}_m^i - \boldsymbol{\mu}_i)(\boldsymbol{x}_m^i - \boldsymbol{\mu}_i)^T \tag{1}$$

where $\boldsymbol{\mu}_i = 1/N\sum_{m=1}^{N}\boldsymbol{x}_m^i$ is the mean of the i^{th} class. The method can be summarized as follows:

- Obtain the projection matrix $\boldsymbol{U} = [\boldsymbol{u}_1\boldsymbol{u}_2 \ldots \boldsymbol{u}_{NC-C}]$ where $\boldsymbol{u}_i, i = 1,2, \ldots NC - C$ are the eigenvectors corresponding to the nonzero eigenvalues of $\boldsymbol{S}_W$.
- Obtain the common vectors by projecting any feature vector from each class onto the null space of $\boldsymbol{S}_W$.

$$\boldsymbol{x}_{com}^i = \boldsymbol{x}_m^i - \boldsymbol{U}\boldsymbol{U}^T\boldsymbol{x}_m^i, m = 1, \ldots, N, i = 1, \ldots, C \tag{2}$$

- Compute the eigenvectors $\boldsymbol{w}_k$ of the scatter matrix of the common vectors $\boldsymbol{S}_{com}$, corresponding to the nonzero eigenvalues and obtain the projection matrix $\boldsymbol{W} = [\boldsymbol{w}_1\boldsymbol{w}_2 \cdots \boldsymbol{w}_{C-1}]$. In here $\boldsymbol{S}_{com}$ is defined as

$$\boldsymbol{S}_{com} = \sum_{i=1}^{C} (\boldsymbol{x}_{com}^{i} - \boldsymbol{x}_{ave})(\boldsymbol{x}_{com}^{i} - \boldsymbol{x}_{ave})^{T} \,. \qquad (3)$$

where $\boldsymbol{x}_{ave}$ is the mean of the common vectors, i.e., $\boldsymbol{x}_{ave} = 1/C \sum_{i=1}^{C} \boldsymbol{x}_{com}^{i}$.

- Obtain the discriminative common vectors by projecting any sample from each class onto the range space of $\boldsymbol{S}_{com}$.

$$\boldsymbol{\Omega}_{com}^{i} = \boldsymbol{W}^{T}\boldsymbol{x}_{m}^{i}, m = 1, \dots, N, i = 1, \dots, C \qquad (4)$$

Let $\boldsymbol{x}_{test}$ be the test vector to be classified. Then classification can be done according to the following decision rule.

$$C^{*} = \underset{j}{\operatorname{argmin}} \{ \| \boldsymbol{\Omega}_{com}^{i} - \boldsymbol{W}^{T}\boldsymbol{x}_{test} \| \},\ \ j = 1, \dots, C \qquad (5)$$

III. Two Dimensional Extension Of DCVA

Let C be the number of image classes, N, be the number of feature vectors in each class and, $\boldsymbol{X}_{m}^{i}$ be the m^{th} two dimensional p by q pixel image of the i^{th} class. We convert the image matrix $\boldsymbol{X}_{m}^{i}$ to a vector $\boldsymbol{x}_{m}^{i}$ in the $n = p \times q$ dimensional space.

It is proved in [10] that the common vectors obtained from total within-class scatter matrix are unique for each class. In the first stage of the proposed method, we use $\boldsymbol{S}_{W}$, to take the advantage of the uniqueness of the common vectors. We apply the eigen decomposition to $\boldsymbol{S}_{W}$ and obtain the projection matrix $\boldsymbol{P}^{\perp}$ of its null space using the eigenvectors corresponding to the zero eigenvalues $\boldsymbol{v}_{i}$, $i = C(N-1)+1, \dots, n$. $\boldsymbol{P}^{\perp}$ can be calculated as follow,

$$\boldsymbol{P}^{\perp} = \sum_{i=C(m-1)+1}^{n} \boldsymbol{v}_{i}\boldsymbol{v}_{i}^{T} \qquad (6)$$

Then the common vector of i^{th}class is calculated as

$$\boldsymbol{x}_{com}^{i} = \boldsymbol{P}^{\perp}\boldsymbol{x}_{m}^{i},\ \ i = 1, \dots, C-1, m = 1, \dots, N \qquad (7)$$

It should be noted that (2) and (7) give exactly the same results. We convert the common vectors $\boldsymbol{x}_{com}^{i}$ into p by q matrices, $\boldsymbol{X}_{com}^{i}$. The covariance matrix of the common matrices can be calculated as

$$\boldsymbol{S}_{com} = \sum_{i=1}^{C} (\boldsymbol{X}_{com}^{i} - \boldsymbol{X}_{ave})^{T}(\boldsymbol{X}_{com}^{i} - \boldsymbol{X}_{ave}) \,. \qquad (8)$$

where $\boldsymbol{X}_{ave} = 1/C \sum_{i=1}^{C} \boldsymbol{X}_{com}^{i}$ is the mean of the common matrices. We are trying to find the optimal projection vectors $\boldsymbol{W} = [\boldsymbol{w}_{1} \vdots \boldsymbol{w}_{2} \vdots \cdots \vdots \boldsymbol{w}_{d}]$ which maximizes the criterion $J(\boldsymbol{W}) = \boldsymbol{W}^{T}\boldsymbol{S}_{com}\boldsymbol{W}$. Here d can be at most $\min(C-1, n)$.

We use the nearest neighbor classifier for classification. The discriminant features of an image $\boldsymbol{X}_{r}$ is calculated as

$$\boldsymbol{Y}_{r} = \boldsymbol{X}_{r}\boldsymbol{W} = [\boldsymbol{y}_{1}^{r} \vdots \boldsymbol{y}_{2}^{r} \vdots \cdots \vdots \boldsymbol{y}_{d}^{r}]. \qquad (9)$$

Let $\boldsymbol{X}_{test}$ be the test image to be classified. The optimal projection vectors of the test image can be given as $\boldsymbol{Y}_{test} = \boldsymbol{X}_{test}\boldsymbol{W} = [\boldsymbol{y}_{1}^{test} \vdots \boldsymbol{y}_{2}^{test} \vdots \cdots \vdots \boldsymbol{y}_{d}^{test}]$. Then the test image is classified according to the following decision rule.

$$C^{*} = \underset{i}{\operatorname{argmin}} \left\{ \sum_{k=1}^{d} \| \boldsymbol{y}_{k}^{test} - \boldsymbol{y}_{k}^{i} \| \right\} \qquad (10)$$

IV. Image Decomposition With QR

QR decomposition is a well-known matrix factorization method [12]. If $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, then it can be decomposed as $\boldsymbol{A} = \boldsymbol{QR}$ where $\boldsymbol{Q} \in \mathbb{R}^{m \times n}$ with orthogonal columns which span the same subspace with the columns of $\boldsymbol{A}$, and $\boldsymbol{R}$ is an upper triangular matrix. QR-decomposition with column pivoting (QRCP) [13,14] is a modified version of QR. In this method the column of the matrix $\boldsymbol{A}$ are sorted such that the absolutes values of the diagonal elements of the matrix $\boldsymbol{R}$ are sorted in descending order. In this way, most of the energy of an image is concentrated into some basis images [5]. The basis images of $\boldsymbol{A}$ can be calculated as $\boldsymbol{q}_{i}\boldsymbol{\tau}_{i}$ where $\boldsymbol{q}_{i}$ is the i^{th} column of $\boldsymbol{Q}$ and $\boldsymbol{\tau}_{i}$ is the i^{th} row of $\boldsymbol{R}$. The orders of columns of $\boldsymbol{A}$ are stored in a permutation matrix $\boldsymbol{P}$ such that the equation $\boldsymbol{Q}^{T}\boldsymbol{AP} = \boldsymbol{R}$ holds. Let the approximation of an image matrix $\boldsymbol{A}$ be $\widehat{\boldsymbol{A}}$. Then it can be calculated as

$$\widehat{\boldsymbol{A}} = \sum_{i=1}^{k} \boldsymbol{q}_{i}\boldsymbol{\tau}_{i}. \qquad (11)$$

Here k is selected according to the ratio E given below.

$$\frac{\sum_{i=1}^{k} d_{i}}{\sum_{i=1}^{m} d_{i}} \geq E \qquad (12)$$

$d_{i}, i = 1,2, \dots, m$ are the absolute values of the diagonal elements of $\boldsymbol{R}$. In experiments we selected $E = 97\%$ as in [5]. In Figure 1 a sample image selected from YALE face database and its two approximations evaluated from image and its transpose are shown. The image and two reconstructed images evaluated from the image and its transpose are labeled as the training images of that subject.

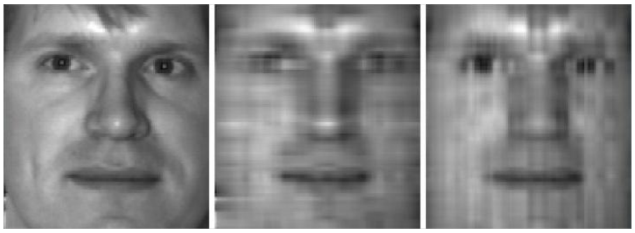


Figure 1. Sample image and its virtual variants evaluated from the image and its transpose.

V. Experiments

In the experimental stage, the performances of DCVA, our proposed method 2D-DCVA, and 2D-FLDA are compared in four face databases namely, ORL [15], FERET [16], YALE [17], and UMIST [18].

ORL face database contains 10 grayscale images from each 40 subjects which are taken in the lab. Images contain different lighting conditions and facial expressions (e.g., closed eyes, glasses, smile). Also images were taken at dark background and subjects are in the frontal position with tolerance to some side movement. The original size of the images is 112 × 92. In the experiments we used the original images of this database. FERET database contains 14,051 grayscale images from 1199 subjects. In the experiments a subset of the database that contains 200 subjects is used. Each subject has two images from f_a and f_b probes. YALE face database contains 11 images from each 15 subjects. Database includes different facial expressions and illumination conditions (i.e., with/without glasses, happy, sad, sleepy, surprised, wink, center-light, right-light, normal). UMIST database contains 20 individuals. The number of pictures per person varies from 19 to 36. Images were taken at various angles from left profile to right profile.

We preprocessed the images by cropping, scaling, resizing. In TABLE I. , the number of subjects, the number of images from each subject, and the size of the images taken from ORL, FERET, YALE, and UMIST databases after the preprocessing operations are summarized.

TABLE I. The Summary of The Databases After The Preprocessing Step

Database	Number of classes	Number of images per class	Dimension
ORL	40	10	112x92
FERET	200	2	100x100
YALE	15	11	120x110
UMIST	20	19	112x92

In the experiments we randomly select an image from each class. Two virtual images are constructed using this image with the QRCP decomposition. The original image and the two virtual images are used to generate the training set images of the subject. The remaining images are used as test images. This procedure is repeated 5 times and the recognition rates are obtained by averaging each run. We implement this process to all databases. The top recognition rates of DCVA, 2D-DCVA, and 2D-FLDA and their standard deviations on the databases are shown in TABLE II.

TABLE II. The Recognition Rates On The Databases

Methods	Databases			
	ORL (%)	FERET (%)	YALE (%)	UMIST(%)
1D-DCVA	69.8± 3.7	88.8± 0.9	58.3± 5.6	55.9± 3.6
2D-DCVA	**76.4**± 2.4	**90.3**± 0.3	**61.6**± 5.2	**64.4**± 4.1
2D-FLDA	76.0± 2.5	90.1± 0.2	59.5± 5.4	61.3± 3.7

VI. Results And Conclusion

One sample problem is an important challenge in face recognition. Methods which use within-class scatter matrix fail. In this work we proposed a two dimensional extension of the discriminative common vector approach. The performance of the proposed method is tested on four different databases namely, ORL, FERET, YALE, and UMIST. 2D-DCVA gave the best recognition results in all databases. 2D-FLDA outperformed 1D-DCVA in all databases. This may be due to fact that the matrix-based methods generally outperform vector based methods [19].

References

[1] W. Zhao, R. Chellappa, P.J. Phillips, and A. Rosenfeld, "Face Recognition: A Literature Survey," ACM Computing Surveys, vol. 35, no. 4, pp. 399-458, 2003.

[2] Z. Daugman, "Face and gesture recognition: Overview," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, no. 7, pp. 675-676, 1997.

[3] X. Tan, S. Chen, Z.-H. Zhou, and F. Zhang, "Face recognition from a single image per person: a survey," Pattern Recognition, vol. 39, no. 9, pp. 1725-1745, 2006.

[4] Q.-x. Gao, L. Zhang, and D. Zhang, "Face recognition using FLDA with single training image per person," Applied Mathematics and Computation, vol. 205, no. 2, pp. 726-734, 2008.

[5] M. Koç and A. Barkana, "A new solution to one sample problem in face recognition using FLDA," Applied Mathematics and Computation, vol. 217, no. 24, pp. 10368-10376, 2011.

[6] H. Yin, P. Fu, and S. Meng, "Sampled FLDA for face recognitionwith single training image per person," Neurocomputing, vol. 69, no. 16-18, pp. 2443-2445, 2006.

[7] M. Apaydın, Ü.Ç. Turhal, and A. Duysak, "An SVD based common matrix method for face recognition: Single image per person," 25th International Symposium on Computer and Information Sciences, 2010, pp. 289-292.

[8] S. Theodoridis, and K. Koutroumbas, Pattern Recognition, Academic Press, USA, 1999.

[9] H. Kong, E.K. Teoh, J.G. Wang, and R. Venkateswarlu, "Two dimensional Fisher discriminant analysis: Forget about small sample problem," Proc. IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing, 2005, pp. 761-764.

[10] H. Çevikalp, M. Neamtu, M. Wilkes, and A. Barkana, "Discriminative Common Vectors for Face Recognition," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 27, no. 1, pp. 4-13, 2005.

[11] V.D.M. Nhat and S. Lee, "Discriminative common images for face recognition," In proceedings of ICANN - Part I, vol. 3696, pp. 563-568, 2005.

[12] T. Kailath, A.H. Sayed, and B. Hassibi, Linear Estimation, Prentice Hall, 1999.

[13] S. Chakroborty and G. Saha, "Feature Selection Using Singular Value Decomposition and QR Factorization with Column Pivoting for Text-Independent Speaker Identification," Speech Communication, vol. 52, no. 9, pp. 693-709, 2010.

[14] S. Ari and G. Saha, "In Search of an SVD and QRcp Based Optimization Technique of ANN for Automatic Classification of Abnormal Heart Sounds," International Journal of Biological and Life Sciences, vol. 2, no. 1, pp. 1-9, 2007.

[15] ORL Face Database, AT&T Laboratories Cambridge1992-1994.

[16] P.J. Phillips, H. Moon, S.A. Rizvi, and P. Rauss, "The FERET Evaluation Methodology for Face Recognition Algorithms," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 10, pp. 1090-1104, 2000.

[17] P. Belhumeur, J. Hespanha, and D. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, no. 7, pp. 711-720, 1997.

[18] D.B. Graham and G.N.M. Allinson, "Characterizing Virtual Eigensignatures for General Purpose Face Recognition," Face Recognition: From Theory to Applications, (Ed: Wechsler, H. ve ark.), NATO ASI Series / Computer and Systems Sciences, 1998, ch. 163, pp. 446-456.

[19] W.-S. Zheng, J.H. Lai, and S.Z. Li, "1D-LDA vs. 2D-LDA: When is vector-based linear discriminant analysis better than matrix-based?," Pattern Recognition, vol. 41, no. 7, pp. 2156-2172, 2008.

Application of Head Tracking for Interactive Data Visualization

Phillip C. S. R. Kilgore
Dept. of Computer Science
LABi @ LSU Shreveport
Shreveport, United States
Email: kilgorep54@lsus.edu

Charles D. McCarthy
Dept. of Computer Science
LABi @ LSU Shreveport
Shreveport, United States
Email: mccarthy20@lsus.edu

Urška Cvek
Dept. of Computer Science
LABi @ LSU Shreveport
Shreveport, United States
Email: ucvek@lsus.edu

Marjan Trutschl
Dept. of Computer Science
LABi @ LSU Shreveport
Shreveport, United States
Email: mtrutsch@lsus.edu

***Abstract*—The utilization of head tracking in gaming and animation applications has increased due to greater availability of relevant libraries and hardware. However, its application to interactive data visualization has not followed the same trajectory. In this paper, we describe an extensible platform permitting the integration of head-tracking interactivity into data visualization software. This platform utilizes Haar classification to provide the recognition of facial features, and uses Kalman filtering to smooth transient input. We also discuss the application of head tracking to data visualization, and address its challenges.**

Keywords- human computer interaction; graphical user interfaces; interaction styles; computer vision; scene analysis; tracking;

I. Introduction

Data visualization focuses on visual presentation of data that is usually highly abstract, high-dimensional and structured, without a "natural" representation on a two-dimensional (2D) plane or three-dimensional (3D) space. One of the simplest data visualization examples is a scatter plot. More complex examples include Radviz, parallel coordinates, multidimensional scaling, or other projections of the data that give insights and uncover previously unknown relationships. Interactivity in high-dimensional data visualizations has been shown to be highly beneficial for the data exploration approach [1]. Interactive data visualization is used for exploration, analysis and presentation of the data [2]. Together with animation, 3D increases the density of information that can be presented on the same screen and thus increases the intrinsic dimensionality of visualizations [3].

3D visualization enables the user to make use of spatial memory. User interface animation in 3D spaces can reveal process and structure (by moving the viewpoint) as previously discussed by Baecker and Small [4]. Investigations of Ware and Franck into motion cues in 3D visualization [5] noted that simple rotation about an axis is effective in interpreting 3D information structures. Three dimensions can bring about problems, including depth perception and occlusion. Occlusion has already been addressed by Elmqvist and Tsigas [6] and others. In this paper, we address the problem of depth perception by using head tracking as a more intuitive interactive approach. Augmented reality, or enriched real environment, has been explored as a tool for interacting with multidimensional information visualizations based on the 3D scatter plots [7] and has been shown to enhance a user's data exploration experience [8].

Head tracking facilitates in determining the location of a user's head relative to a particular focal point. In the context of human computer interaction, this information can be used to change the presentation of application content. Head tracking is common in CAVE and CAVE-like environments that are geared towards groups. [9] Our platform is designed for a single person using a standard desktop computer with an off-the-shelf webcam. It is extensible for the easy integration of head tracking-based interactivity into additional data visualizations.

In this paper, we first describe the detection of features such as face and eyes using the cross-platform Open Computer Vision Library (OpenCV) [10]. Later, we integrate the approach into sample well-known visualizations and address the calibration and optimization. To provide for wide adoption of the system, the system utilizes platform independent tools such as OpenGL [11] for 3D graphics and Qt [12] for the graphical user interface. We conclude the paper with a list of challenges, followed by plans for future work.

II. Approach

In order to adjust the view of the 3D scene, we first have to identify the user's position relative to the monitor. We capture frames in a free-head fashion, with an off-the-shelf webcam (Logitech® Webcam C905) centered at the top of the monitor. We then calculate the user's head position using an object detection algorithm. Many different algorithms exist to perform face detection, although most are computationally expensive [13]. We use the Haar Classifier to detect facial features in frames, and later use Kalman filtering to reduce jitter between frames.

In the reference capture stage, our system extracts a frame from the capture device and analyzes features from the Haar classifier, yielding reference points used to calculate the position of the head. The output of this stage is supplied to a head position calculation, which performs calibration if necessary, and supplies the resulting head position to the visualization for use in the projection matrix calculation. The projection matrix is set using matrix transformations calculated from the head position, and is followed by offsetting the model view matrix by the position of the head. Finally, the visualization is rendered as it would have been otherwise. Each of these stages may be executed concurrently with one another, allowing the process to be run asynchronously on multiple threads.

A. *Head tracking through eye detection*

We chose to track the user's eyes as head reference points due to the relative ease of tracking a user's eyes and their constant presence while exploring a visualization. We assume that the user would be looking directly at the monitor, and thus into the webcam at the top of the monitor. This gives us a frontal face with better view of the eyes.

We considered a number of different eye detection classifiers, from general single eye, to eye pair, to separate classifiers for the left and right eyes [14]. Based on the evaluations, a Haar cascade is run on each region using individualized classifiers for each eye [15].

B. *Open Computer Vision library*

We accomplish eye detection and head tracking through the use of the open source Open Computer Vision Library, (OpenCV) which was developed initially by Intel [10]. We make use of two OpenCV implementations of algorithms, the Haar Classifier [16], and the Kalman filter [17]. We also take advantage of the camera capture functions provided by the library.

C. *Haar Classifier*

Rather than looking at individual pixels, Viola and Jones devised an algorithm called the Haar Classifier to rapidly detect objects, including human faces, using AdaBoost classifier cascades that are based on Haar-like features [16]. Haar-like features are rectangular patterns of black and white areas, which define the change in contrast values between adjacent groups of pixels. The simple rectangular features of an image are calculated using an intermediate representation of an image, called the integral image [16]. The pixels of the entire rectangular subsection of the source image are summed and subtracted from a scaled sum of the pixels masked by the black region in the Haar-like feature.

Many of the Haar-like features will contain common regions of pixels [16]. Calculating the sums iteratively for each feature would result in massive amounts of redundant calculations. The problem can be reduced to a simple four-element summation for each region of the selected feature pattern area by utilizing an integral image [16]. The integral image is calculated by summing all the intensity values of the pixels to the top and left of the (i, j) pixel, and placing that value in the (i, j) pixel of the integral image. It only takes two passes to compute both integral image arrays, one for each array. Calculating a feature is extremely fast and efficient, as it only takes the difference between six to eight array elements forming two or three connected rectangles to compute a feature of any scale. We chose to use one of the frontal face cascades provided by OpenCV due to its high detection rate and low false positive rate [15].

D. *Kalman filter*

Haar classifiers are not perfectly accurate, and occasionally produce false positives. To mitigate these errors we use a Kalman filter [17] implemented by OpenCV. The Kalman filter was developed to predict the state of a system in the presence of noisy measurements.

Creating a Kalman filter for each eye introduces new problems, such as inconsistencies in the eye distance. We chose to filter the center of the eye pairs in order to maintain a more consistent eye separation. This approach still results in noisy eye separation measurements, so we also create a Kalman filter for the eye separation value. The filtered eye position and eye separation are then recombined to form a filtered eye pair.

III. Implementation

We utilize the camera functions provided by OpenCV to configure the webcam and to query frames. The reference capture stage begins by capturing a frame from the webcam, and passing it to the face detection function.

A. *Face detection*

The Haar classifier (Sec. II-C) is passed the region of the camera capture frame that is likely to contain a face and, if a face is found, returns a rectangle describing the face location. Due to minor shifts in the position and size of the face rectangle, we cannot utilize this object as the location of the user's head. Therefore, we proceed to eye detection as a method for more precisely locating the center of the users face as the origin of the head.

B. *Eye detection*

The face region is then subdivided into regions that likely contain the users eyes (Sec. V-B), and each region is scanned with an individualized Haar classifier cascade [15]. If each eye region finds an eye, the eye group is analyzed to determine user distance and position. We reduce the search region, since running eye detection on an entire capture frame would increase the number of false positives and decrease performance.

C. *Update Kalman filters*

The output from eye detection step is more stable than the face detection step, but there is still minor variation in the coordinates for each eye. We attempt to smooth the eye coordinates in both position and relative distance by computing the center of the two points, and then passing

Table I
SYMBOLS USED IN CALCULATIONS

Source	*Symbol*	*Description*	*Unit*
Head Position Calculation			
Input	x_c, y_c	center position	pixel
Input	w, h	capture dimensions	pixel
Config	s	screen height	mm
Config	d_m	actual dot separation	mm
Config	f	field of view	rad
Calc	d_p	measured dot separation	pixel
Calc	r	camera radians per pixel	rad/pixel
Calc	θ	point angle of separation	rad
Calc	β, λ	relative horizontal/vertical angle	rad
Calc	ρ	camera vertical angle	rad
Calc	z'_h	head distance to camera	s
Output	x_h, y_h	head position	s
Output	z_h	head distance	s
Projection Matrix Calculation			
Const	p_n	near plane	s
Const	t_w	virtual room width	s
Const	t_h	virtual room height	s
Output	p_l, p_r	left/right planes	s
Output	p_t, p_b	top/bottom plane	s
Eye Search Region Calculation			
Input	x_f, y_f	face location	pixel
Input	w_f, h_f	face dimensions	pixel
Input	i	eye index: 0 = right, 1 = left	pixel
Output	x_e, y_e	eye location	pixel
Output	w_e, h_e	eye width	pixel

the coordinates from the capture frame to individual Kalman filters. The distance between the user's eyes must also remain constant in order to prevent erratic movement on the z axis. Once filtered, the values are recombined into an eye group with each eye at the same vertical position. A vector could be computed from the raw eyes before filtering in order to restore the eye tilt, but this is unnecessary for our purposes as we only require three degrees of freedom.

D. Head position calculation

The head position calculation stage takes the filtered 2D eye group output from the reference capture stage as parameters to calculate the 3D head location. First, the distance between the eyes (in pixels) and the point at the center of the eye group is calculated. In the beginning of the head position calculation stage, the field of view of the capture device (provided by the configuration file discused in Sec. IV-B1) and the current width of the capture frame are used to compute the radians per pixel of the capture frame (Eq. 1). This is done with each pass to account for changes in capture resolution made by the user.

$$r = f/w \tag{1}$$

Then we calculate half of the angle separation between the reference points. (Eq. 2)

$$\theta = rd_p/2 \tag{2}$$

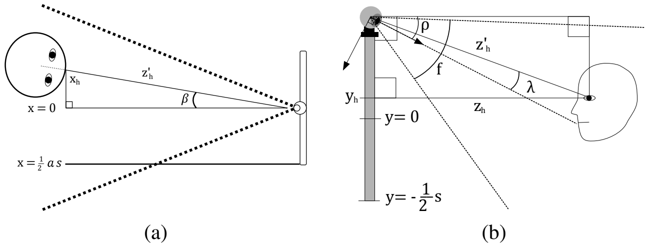


Figure 1. a) Horizontal position of head; b) vertical distance to head

Next, the head distance (screen units) is found using half the actual distance between the user's eyes (mm), the screen height (mm) and the cotangent of the eye angle of separation (Eq. 3).

$$z'_h = \frac{d_m \cot\theta}{2s} \tag{3}$$

The x position of the user's head is then found by taking the sine of the angle between the user's horizontal position (from the reference capture device) and the center of the display surface (Eq. 4, 5). This value is then scaled by the user's distance from the screen. (Fig. 1.a).

$$\beta = r(x_c - \frac{w}{2}) \tag{4}$$

$$x_h = \sin(\beta) z'_h \tag{5}$$

The angle of the camera relative to the user's head is found by calculating the number of pixels that the center point resides above the middle of the frame and multiplying this value by the number of radians in one pixel shift for the current camera (Eq. 6). This value is loaded at startup from the configuration file described in Sec. IV-B1.

$$\lambda = (y_c - \frac{h}{2})r \tag{6}$$

The vertical position of the user can then be calculated for the user's head (Fig. 1.b). In order to account for any vertical tilt that may exist in the cameras view, a calibration routine can be performed at this time if it's required (Sec. IV-B2).

$$y_h = \frac{1}{2} + \sin(\lambda + \rho) z'_h \tag{7}$$

The last step in the head position stage is to calculate the actual distance between the user and the surface of the display (Eq. 8).

$$z_h = \cos(-\lambda - \rho) z'_h \tag{8}$$

ISBN: 978-1-61208-202-8

E. Projection matrix calculation

$$p_l = \frac{p_n(-\frac{1}{2}t_w - x_h)}{z_h} \tag{9}$$

$$p_r = \frac{p_n(\frac{1}{2}t_w - x_h)}{z_h} \tag{10}$$

$$p_t = \frac{p_n(-\frac{1}{2}t_w - y_h)}{z_h} \tag{11}$$

$$p_b = \frac{p_n(\frac{1}{2}t_w - y_h)}{z_h} \tag{12}$$

The camera is positioned at the origin facing in the direction of the negative z axis with an up vector aligned with the y axis. The OpenGL projection matrix is altered to simulate perspective and field of view changes. The camera is positioned at the origin facing in the direction of the negative z axis (the positive z axis points out of the screen [11]), with an up vector aligned with the y axis. This stage takes the 3D head position as input to calculate the values of the left, right, top, and bottom walls of the near plane (Table I, Eq. 9 – 12 [11]). These walls along with values for the near and far plane are set as parameters to the OpenGL view frustum call.

The actual location of the camera in the OpenGL scene is not altered. Instead, we skew the viewport to achieve field of view changes and move the entire scene opposite the head movement to give the appearance of a receding background as the user moves away. In any case, the projection matrix calculation is isolated from the visualization stage to make extensions easier to write; this matrix is calculated ahead of time and can be utilized when updating the visualization.

IV. Calibration

Due to the variance in consumer webcams, reference point separation, and display device aspect ratio, we chose to create a file to store persistent configuration information. This allows us to retrieve previously stored configurations without having to setup the environment every time the system is restarted. In order to calibrate the system, we have to obtain the capture configuration and subsequently calibrate the software.

A. Reference capture device configuration

We generalize reference capture in order to keep our software independent from the type of capture device. Our software can easily adapt to new types of reference capture devices without altering the overall structure. This gives us a framework that is easily adaptable to any capture device, from the webcam to the Wii Remote or other device. We are using the reference point distance, in addition to the height in order to calibrate the system.

Screen height is used as a base unit for all measurements. The head position is given in units of "screen height." The aspect ratio of the capture device is calculated in order to scale the x axis translation of the user and to fit the virtual room to the corners of the monitor. This sets the resolution of the system.

B. Software calibration

The vertical angle of the camera in relation to the display is calculated at run time during a calibration. Initiating the calibration routine causes our software to use the next head vertical position as the user's center view of the screen. Information about the camera specifications and the user's eye distance is loaded from configuration file at runtime.

1) Configuration file: An external configuration file that contains hardware and user data is loaded at startup. This file stores information about the position of the camera, the aspect ratio of the computer monitor, the monitor's physical height, and the distance between the user's eyes from the center of one pupil to the center of the other.

2) Calibration routine: Because of factors such as the height of the user's monitor, the vertical angle of the camera must be determined. Without knowing the vertical angle of the camera, the position of the user's head cannot be reliably calculated. If one could accurately estimate the size of a screen unit, then it would be possible to perform calibration by positioning oneself in any orientation. Instead of asking the user to properly position themselves one screen unit away from the display, we ask the user to position themselves close to the center of the screen, so that the user's y value is known to be half the screen height. Because we have made this assumption and know both the height of the screen and radians per pixel, we can calculate the vertical angle of the user relative to the camera (Eq. 13). We also have enough information to calculate the head distance of the user, and can use this information to derive the vertical angle of the camera.

$$\rho = -\frac{\pi}{2} + \cos^{-1}(\frac{1}{2z'_h}) - \lambda \tag{13}$$

V. Optimization

Performing face detection using a single classification is time-consuming and inefficient due to the redundant nature of the algorithm [18]. Due to the real time nature of head tracking, we performed several optimizations to first set a smaller rough location of the face, which is then used by the classifier. After the classifier is run and the face is detected, it is more efficient to run the eye feature classifier. We use the distances and set formulas that help us identify the eye regions.

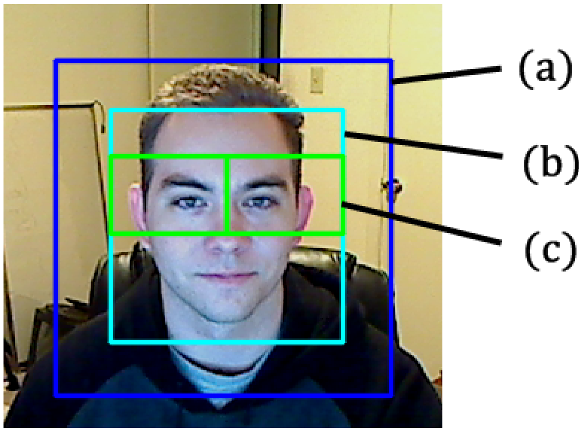


Figure 2. Webcam capture frame: (a) Face search region (b) Detected face (c) Eye search region

A. Face search region reduction

Reducing the face search region can result in improved performance, since the Haar cascade is computationally expensive. We use a simple algorithm developed by [19] to increase face detection performance on the iPhone. The algorithm works by storing the last region that a face was found in and assumes that the next frame will contain a face in the same general region (Fig. 2.a). There is a small amount of padding added to the face region to prevent the face region from becoming smaller with each frame. This happens because as the search region becomes smaller the Haar classifier is more likely to find a face smaller than the previous face, leading to cyclical shrinking. Eventually, the search region becomes too small to contain a face, and the region is reset to the size of the entire frame, therefore degrading performance.

The padding must be as small as possible to maximize the benefits of the algorithm. However, making the padding too small also degrades performance due to the increased probability that a smaller face will be detected with each frame. Reducing the margin also increases the number of times the user's face is lost due to movement outside the bounds of the search region.

B. Eye search region reduction

Once a face is identified, the region is divided into subregions that should contain the user's eyes (Fig. 2.c). The eyes are assumed to lie on approximately the same horizontal line, and the head is assumed to be in an upright orientation. Therefore, the eye regions are defined by the following formulas:

$$x_e = x_f + i(\frac{w_f}{2}) \tag{14}$$

$$y_e = y_f + \frac{h_f}{5} \tag{15}$$

$$w_e = \frac{w_f}{2} \tag{16}$$

$$h_e = \frac{h_f}{3} \tag{17}$$

Reducing the eye search region in this way helps to prevent false positives and increases performance by eliminating regions that are unlikely to contain eyes.

VI. Visualization

Head tracking enabled visualizations allow the user to control the perspective of the view in a natural way. The data is presented in a virtual room to heighten the perception of depth.

A. Objectives and application

Visualization in 3D space can be useful with multidimensional data sets as compared to two-dimensional visualizations by affording an extra spatial dimension for points to lie on. In some cases, this extra spatial dimension can assist in the visual detection of outliers. The Abalone data set [20], a nine-dimensional data set containing 4,177 records, is such an example. Outliers in the Abalone dataset are clearly exposed when ring count is plotted against height, but are obscured when it is plotted against visceral weight.

However, traditional mechanisms of manipulating this space introduce new problems regarding interaction with the visualization. The first problem arises because these visualizations are operated using input devices that are intended for manipulation of objects on a plane, such as a mouse or a trackball. An attempt has been made to address these concerns in specialized 3D input devices, such as 3DConnexion's SpacePilot PRO [21]. However, these devices are not widespread, and may require considerable support.

We have proposed head tracking, via a web cam or similar device, as an alternative to traditional 3D input devices. Devices such as these are inexpensive, and are supported in a variety of modern environments [22]. In a 2007 study published by Dynamics of Institutions and Markets Europe, 31.6% of 2094 participants owned web cams [23]. Because of their widespread usage and availability, optical input devices lend themselves to the navigation of 3D visualizations. Computer vision has previously been used as mechanisms of cursor control for those who have limited motor capability [22], and is currently being implemented by manufacturers of video game consoles [24]. With this research in mind, we considered head tracking as a viable input device to 3D visualizations.

We created an implementation of the traditional scatter plot in three dimensions that allowed for manipulation of the camera via eye tracking. We designed the operation of the visualization viewport to be analogous to looking through a

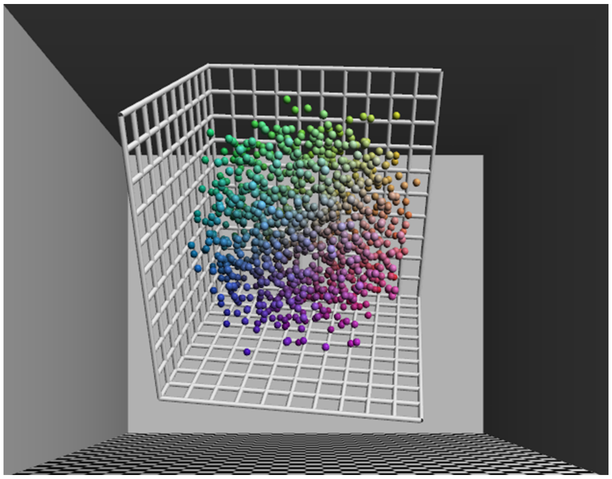


Figure 3. 3D scatter plot interaction utilizing our head tracking framework

window. Eye tracking was not the sole method of input; mouse input was used to manipulate the position of the visualization itself, affording zoom, panning, and rotation. 3D scatter plot (Fig. 3) and PCA projections are trivial examples of 3D visualization, other visualizations may benefit from the addition of head tracking as a navigation method. For instance, volume visualizations could benefit from this input method, allowing users simple access to data obscured by the corners, overlap or pattern in the model. Similarly, projections of classical 2D visualizations into 3D space can benefit from being able to temporarily hone in on data in much the same manner.

B. Challenges

In order to accommodate eye tracking, the visualization viewport is treated as an aperture. As the user moves closer to the aperture, more of the scene is revealed, and the field of view with respect to the scene becomes wider. This variable field of view yields a "dolly-and-zoom" effect, where objects in the background of a scene appear to move into its horizon [25]. Although this kind of perspective distortion is consistent with what is currently known about optics, it appearance seems unnatural, as the eye would never see it the same way, leading to its use in cinematography for dramatic effect. However, this behavior can result in confusion in a visualization system in the absence of a static point of reference, and it was observed in a pre-test of the software on a small group of users. 3D perspective distortion can be described with a 2D distorting visual transfer function, and has been previously detected in information visualizations [26].

Initial testing of the environment revealed that users were mislead into believing that some objects in the visualization were being translated towards the horizon. As a result, we added a room to serve as a static reference point. Like other objects in the background, the back wall of a room will recede into the background as the user moves towards the viewport. The wall is known to be stationary with respect to the user's position, however, and prevents the erroneous conclusion that these objects are in motion.

The notion of a virtual room as static reference point could become ineffective if objects in the visualization were ever to penetrate one of the walls. If one chose to implement zoom by scaling the visualization in question, this problem might be unavoidable, and continuity will be broken. Instead of growing and shrinking the objects in the visualization, we found that an alternative metaphor was bringing the object closer to or further away from the camera. By doing this, one can ensure that the content of the visualization is never large enough to cross the boundaries provided by the virtual room.

Reduction of the search region results in increased performance due to its smaller area that must be analyzed by the Haar classifier, but results in false negatives as the search region approaches face size. We suggest a "user movement vector" to dynamically resize the region based on the user's predicted movement. The user movement vector is a two-dimensional vector that is representative of velocity of the user's head. By utilizing a user movement vector, one can resize the search region to account for quick movements of the head.

VII. Conclusions and Future Work

In this paper, we presented an alternative interactive mechanism that may be used to augment traditional three-dimensional visualizations. This method uses resources that are readily available and may be obtained at relatively little expense. We discussed the implementation of such a system, and steps that are taken in our system to optimize its performance. With these optimizations in place, we were able to implement an interactive 3D scatter plot and PCA projection.

We discussed several challenges that restrict the usability of head tracking in a visualization setting, and how we addressed these issues. One notable challenge we faced was transient input from the input capture system, which we addressed by removing noise generated by Haar classification using a Kalman filter. We also found that problems with maintaining perspective arose, such as perspective distortion introduced by dolly zoom.

We hypothesize that head tracking can be applied, in a general sense, to any 3D visualization. In future work, we would like to apply this research to other types of 3D visualizations, in addition to scatter plots. We would also like to compare the current method of filtering transient input (Kalman filtering) to double exponential smoothing, which claims performance that is two orders of magnitude greater than Kalman filtering.

Implementation of a user movement vector (as described in Sec. VI-B) may increase performance by resizing search regions based on context. We would finally like to perform extensive usability testing to gather evidence to sustain or deny our hypotheses. Though we have performed preliminary testing of our head tracking system, we plan to perform a full study involving users in the future.

VIII. Acknowledgements

This project was supported by grants from the National Center for Research Resources (5P20RR016456-11 and 5P20RR018724-10) and the National Institute of General Medical Sciences (8 P20GM103424-11 and 8 P20 GM103433-10) from the National Institutes of Health. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of NIH.

References

[1] A. Buja, D. Cook, and D. F. Swayne, "Interactive high-dimensional data visualization," *Journal of Computational and Graphical Statistics*, vol. 5, pp. 78–99, 1996.

[2] R. Kosara and H. Hauser, "An interaction view on information visualization," in *State-of-the-Art Proceedings of EUROGRAPHICS (EG 2003)*, Granada, Spain, 2003, pp. 123–137.

[3] G. G. Robertson and S. K. Card, "Information visualization using 3d interactive animation," *Communications of the ACM*, vol. 36, no. 4, pp. 57–71, Apr. 1993.

[4] R. Baecker and I. Small, *The Art of Human-Computer Interface Design*. Addison-Wesley, 1990.

[5] C. Ware and G. Franck, "Evaluation stereo and motion cues for visualizing information nets in three dimensions." *ACM Transactions on Graphics*, vol. 15, Apr. 2006.

[6] N. Elmqvist and P. Tsigas, "View projection animation for occlusion reduction," in *Proceedings of the Working Conference on Advanced Visual Interfaces*, ser. AVI '06, May 2006, pp. 471–475.

[7] B. S. Meiguins, R. M. Casseb do Carmo, A. S. Gonclaves, P. I. Alves Godinho, and M. de Brito Garcia, "Using augmented reality for multidimensional data visualization," in *Proceedings of the the Conference on Information Visualization*, Jul. 2006.

[8] R. M. Casseb do Carmo *et al.*, "Coordinated and multive views in augmented reality environment." in *Proceedings of the 11th International Conference on Information Visualization*, Jul. 2007, pp. 156–162.

[9] J. Jacobson *et al.*, "The CaveUT system: immersive entertainment based on a game engine," in *Proceedings of the 2005 ACM SIGCHI International Conference on Advances in computer entertainment technology*, ser. ACE '05. New York, NY, USA: ACM, 2005, pp. 184–187. [Online]. Available: http://doi.acm.org/10.1145/1178477.1178503

[10] Intel Coproration, *Open Computer Vision Library Reference Manual*, 2001.

[11] Khronos Group, "OpenGL - The Industry Standard for High Performance Graphics," www.opengl.org, Feb. 2012.

[12] Nokia Corporation, "Qt – A cross platform application and UI framework," http://qt.nokia.com, Feb. 2012.

[13] G. Bradski, "Computer vision face tracking for use in a perceptual user interface," *Intel Technology Journal*, 1998, 2nd Quarter.

[14] M. Castrillón Santana *et al.*, "Face amd feature detection evaluation," in *Third International Conference on Computer Vision Theory and Applications*, ser. VISAPP, 2008.

[15] ——, "ENCARA2: Real-time detection of multiple faces at different resolutions in video streams," *Journal of Visual Communication and Image Representation*, pp. 130–140, 2007.

[16] P. Viola and M. Jones, "Rapid object tection using boosted cascade of simple features," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2001.

[17] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Journal of Basic Engineering*, vol. 82D, no. 1, pp. 34–45, 1960.

[18] L. Liuxia, "Research on face detection classifier using an improved adaboost algorithm," in *International Symposium on Computer Science and Computational Technology*, 2008, pp. 78–81.

[19] More Than Technical, "Near realtime face detection on the iPhone w/ OpenCV port,"
http://www.morethantechnical.com, Feb. 2012.

[20] W. J. Nash *et al.*, "The population biology of abalone (haliotis species) in tasmania, i. blacklip abalone (h. rubra) from the north coast and islands of bass strait." Sea Fisheries Division, Tech. Rep. 48, Dec. 1995, UCI Machine Learning Repository. http://archive.ics.uci.edu/ml/datasets/Abalone.

[21] 3DConnexion,
http://www.3dconnexion.com, Feb. 2012.

[22] P. Mauri, T. Ganollers, and M. Garcia, "Computer vision interaction for people with severe movement restrictions," *Human Technology Journal*, vol. 2, no. 1, pp. 38–53, 2006.

[23] F. van Rijnsoever and C. Castaldi, "Perceived technology clusters and ownership of related technologies: the role of path-dependence," in *DIME Workshop on Demand, Product Characteristics and Innovation*, Jena, 2007.

[24] Microsoft Corporation, "Xbox Kinect," http://www.xbox.com/kinect, Feb. 2012.

[25] K. E. Zheng *et al.*, "Parallax compution: Creating 3d cinematic effects from stills," in *Graphics Interface Conference*, Kelowna, British Columbia, Canada, 2009.

[26] J. D. Mackinlay, G. G. Robertson, and S. K. Card, "The perspective wall: Detail and context smoothly integrated," in *SIGCHI '91*, 1991, pp. 173–179.

Chinese Blog Classification Based on Text Classification and Multi-feature Integration

Jianzhuo Yan
College of Electronic Information and Control Engineering
Beijing University of Technology
Beijing, China
yanjianzhuo@bjut.edu.cn

Suhua Yang, Liying Fang
College of Electronic Information and Control Engineering
Beijing University of Technology
Beijing, China
yangsuhua86@126.com, fangliying@bjut.edu.cn

***Abstract*—The Chinese blog has become one of the most important sources of information in China. The content of Chinese blog varies widely, thus its classification is of great significance. The Chinese blog has the features of the title, straight matter, tags and user-defined types, and different features have different lengths. Traditional text classification method of the Chinese blog classification is not ideal. In this paper, the Chinese blog is classified by using a number of Chinese blog features in which traditional text classification technique and short text classification technique will be chosen according to the different length of features. In addition, the feature expansion method is adopted for sparse features of short text, and the features are integrated by linear training. Experimental results show that the proposed method improves the accuracy of classification.**

Keywords- text classification; Chinese blog classification; short text classification; feature expansion; multi-feature integration

I. INTRODUCTION

Chinese blog has become more and more popular in China. In recent years, with the rapid development of Chinese blog, the domains of scientific research and industry have been interested in Chinese blog. If we can make full use of the abundant Chinese blog resources and classify the Chinese blog correctly, it is of great practical and scientific significance to learn the development of internet, improve various internet services and enrich user's internet lives [1].

Chinese blog classification is the core and basis in personalized services. As only blogger's personalized information is well understood, the ideal of personalized services may be achieved. The Chinese blog which is different from the general text has the features of title, straight matter, tags and user-defined types. At present, there has been some related research in the Chinese blog classification; AiXin Sun points out that using tag for Chinese blog classification can improve the classification results [2]. The classification method is only for the whole blog not for the articles of the blog, so the classification granularity of this method is not detailed enough. Singh et al. [3] proposed a method of blog classification by combining the domain ontology, which improves the deficiencies of the traditional "word bag" model in the expression of semantic information. But that method does not combine the blog features, so the blog information can not be expressed well enough. Lin and Nenghai [4] proposed the classification of the multi-feature integration, but the method is lack of the analysis for the blog features. Obviously, the traditional method is not proper for each feature. The text content of title, tags and user-defined types is relatively short, so the description has weak signals. Only digging out more information for the short text, short text classification can be more correctly. Hyponymy relation between words is an important semantic relation, and extending short text feature vector by using the hyponymy relation between words can make the short text information richer.

Chinese blog has some features which have different lengths. For this reason, this paper uses the method of multi-feature integration for the different feature. In the classification process, the long feature and short feature use the traditional text classification technique and the short text classification technique, separately. Thus, the contents of various Chinese blog features can be fully expressed, and integrating the features through linear training. This subject has been widely applied in personalized search, advertisements automatically recommend, the construction of user community, and so on.

II. STATE OF THE ART

For the blog classification, the introduction section shows that many experts classify the Chinese blog by using traditional text classification method, but the result of classification is not very well. The blog has its features [5], and if the features are used in the classification, the result of classification will be more correct. Some papers use the Naive Bayes [6] and k-Nearest Neighbor algorithm (KNN) [7] for the classification. The model of Naive Bayes is a probability classification model based upon two assumptions. It requires that the probabilities of all words are independent and the class of the document has no correlation with its length, but the effect is unstable in practical application. KNN is a method based on lazy and required learning method, and the effect of classification is better. But the time of classification is nonlinear, and when the number of training text increases, the time of classification will sharply increase. Support Vector Machine (SVM) is a new machine learning method which is advanced by Vapnik [8] based on statistical learning theory. It is similar to structure risk minimization principle [9], which has splendidly learning ability and only needs few samples for training a high-performance text classifier.

In this paper, SVM is used for text classification according to the blog features, and experiment results show that the classification is more credibility.

III. CHINESE BLOG CLASSIFICATION

The traditional model of Chinese blog classification includes two modules: pre-processing module and classification module. In this paper, the multi-feature fusion algorithm is added into the classification algorithm, so the module of feature integration is added into the classification module which combines the classification results of each feature. In the classification module, the traditional text classification technique and the short text classification technique are used according to the length of the feature. In addition, the algorithm of feature vector extension is adopted for the classification.

The framework of Chinese blog classification proposed in this paper is shown in Figure 1.

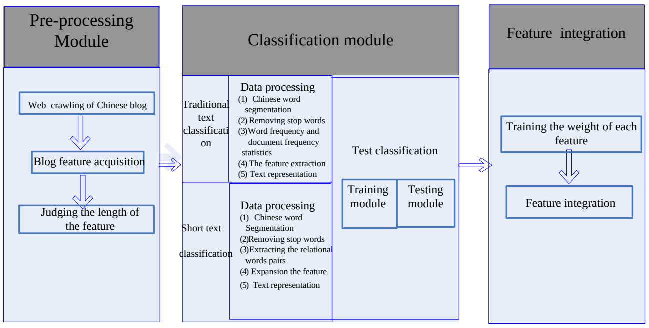

Pre-processing Module
Web crawling of Chinese blog
Blog feature acquisition
Judging the length of the feature
Classification module
Traditional text classification
Data processing
(1) Chinese word segmentation
(2) Removing stop words
(3)Word frequency and document frequency statistics
(4) The feature extraction
(5) Text representation
Short text classification
Data processing
(1) Chinese word Segmentation
(2)Removing stop words
(3)Extracting the relational words pairs
(4) Expansion the feature
(5) Text representation
Test classification
Training module
Testing module
Feature integration
Training the weight of each feature
Feature integration

Figure 1. The framework of Chinese blog classification.

From Figure 1, the aims of the three modules are clear. The first module is to get the feature of the Chinese blog and judge the length of the feature, the second module is to classify the text for each feature of the Chinese blog, and the third module is to integrate the feature.

A. Pre-processing Module

Chinese blog pages are written in Hypertext Markup Language (HTML) which contains a wealth of information, and are semi-structured text files. In addition to plain text, the page also contains some labels and features. Before classification, the features of the straight matter, title, tags and user-defined types should be obtained. Then the content of the features are as regular texts, and the blog can be classified as the texts.

Specific steps include web crawling of Chinese blog, blog feature acquisition and judging the length of the feature, which are as follows:

1) *Web crawling of Chinese blog*: It aims to get the source code of the Chinese blog.

2) *Blog feature acquisition*： By using regular expressions to remove the label of source code of the Chinese blog, the features of the straight matter, title, tags and user-defined types are extracted.

3) *Judging the length of the feature*: The text which has less than 160 characters in length is considered as short text. The title, tags and user-defined types usually belong to the short texts, and the straight matter of the blog which usually has more than 160 characters is taken as the traditional text classified by traditional classification methods.

B. Traditional Text Classification

After text pre-processing, the feature which is judged as the traditional text is to execute data processing. Further data processing is used for further classification. The data processing includes Chinese word segmentation, removing stop words, word frequency and document frequency statistics, the feature extraction and text representation.

1) *Data processing*

The specific steps of data processing are as follows:

a) Chinese word segmentation: The Institute of Computing Technology, Chinese Lexical Analysis System (ICTCLAS) is used for words segmentation.

b) Removing stop words: Create the list of terms which are filtered before the word frequency process started. The list includes mainly conjunctions, prepositions or pronouns.

c) Word frequency and document frequency statistics: Count word frequency for each word which appears in the text. The word frequency *F* is initialized as 1, and added 1 each time to count the document frequency of each category.

d) The feature extraction：By delete the words from the text which has no contribute or very little contribution to the entry category information and taking account of the large amount of information carried by nouns, verbs followed by adjectives and adverbs, this article frame is realized using only the noun.

e) Text representation： In this paper, the term frequency-inverse document frequency (TF-IDF) algorithm [10] is used as vector space model to represent the text. The method of vector space model representation is as follows: each Chinese blog text is represented as a *n*-dimensional vector (w_1, w_2, $w3$, ... , w_n), and the weight of each dimension in the vector of this text should correspond with the weight in this text.

Weight Set: $W = \{w_{i \mid i \in n}\}$

$$w_i = \frac{\sum_{i \in s} (w_i \times tf_i) \times \log(N / n_i)}{\sqrt{\sum_j ((\sum_{i \in n} w_i \times tf_i) \times \log(N / n_i))^2}} \qquad (1)$$

where w_i is the corresponding weight of the *i*-key words, tf_i is the frequency of the *i*-key words in the page, *N* is the total number of text contained in the training set, n_i is the number of the text which contains the characteristics.

2) *Text classification*

The technique of text classification is mainly based on statistical theory and machine learning, such as Naive Bayes, KNN and SVM. The model of Naive Bayes is a probability classification model based upon two assumptions. It requires that the probabilities of all words are independent and the class of the document has no correlation with its length, but the effect is unstable in practical application. KNN is a method based on lazy and required learning method. The effect of classification is better, but the time of classification is nonlinear, and when the number of training text increases, the time of classification will sharp increase. SVM is a new machine learning method advanced by Vapnik according to statistical learning theory. It is similar to structure risk minimization principle, which has splendidly learning ability and only needs few samples for training a high-performance text classifier [11]. The input vector X is mapped to a high-dimensional feature space Z by nonlinear mapping, in which the optimal separating hyperplane is structured. SVM classification function is similar to neural network in form. The output is a linear combination of intermediate nodes, each intermediate node corresponds to a support vector, and the dot product is operated between vectors. The expression of the SVM function for classification of non-linear optimal separating surface is as follow:

$$f(z)=\sum_{supvector} a_i*y_i\varphi(z_i)\varphi(z)+b^*=\sum_{supvector} a_i*y_ik(z_i,z)+b^* \quad (2)$$

Therefore, adopting the kernel function can avoid the high-dimensional feature space for complex operations. The process can be expressed as follows: First, map the input vector X into a high-dimensional Hilbert space H. The kernel function has different forms, and different kernel functions will form different algorithms. In general, the commonly used kernel function has three kinds: Polynomial kernel function, Radial basis function, and Neural network kernel function.

The choice of kernel function has little effect on the accuracy of classification. But polynomial classifier can be applied for the low-dimensional, high-dimensional, large sample, small sample and so on. It is applicable, and has a wider domain of convergence, parameter easy to control, etc. Thus, this paper chooses polynomial classifier as a kernel function.

C. *Short Text Classification*

Traditional text classification method can not be applied in the short text classification very well. The correlation between words and categories is measured when extracting the feature, but the short text has less number of words and weak information, which leads to a serious shortage of short text feature and makes the traditional classifier not accurately classify the text [12]. In this paper, the feature expansion method is used to rich the content of the short text.

The data processing of short text includes Chinese word segmentation, removing stop words, the extraction of relational words pairs, the feature extraction, feature expansion and text representation. The methods of Chinese word segmentation, removing stop words, text representation and the feature extraction are the same as the methods of data processing for the traditional text classification, and the different processes are the extraction of relational words pairs and feature expansion.

1) Extracting the collection of the feature words pairs by HowNet

HowNet is an on-line common-sense knowledge based on unveiling inter-conceptual relations and inter-attribute relations of concepts as connoting in lexicons of the Chinese and their English equivalents. HowNet uses the knowledge representation language to describe the concept, and the words of the knowledge representation language are called as “Sememe” which is the smallest unit of the concept [13].

In this paper, the hyponymy strength of the relational words (A, B) is measured by the semantic distance of sememe.

$$Degree(A,B)=\frac{2\partial}{\partial+d};\ 0<\partial<=1 \quad (3)$$

where ∂ is an adjustable parameter, d is the distance of the Sememe in the Sememe hierarchical tree.

When d is greater than three, the semantic distance of sememe is far, and thus the hyponymy strength is determined as zero.

The collection of the feature words pairs is acquired by the relational words pairs which are extracted from the training corpus and feature items, according to the calculation of hyponymy strength to get the relational words pairs. The threshold of hyponymy strength is set as C, so it needs to meet the following formula:

$$Degree(w_1,w_2)>C \quad (4)$$

After the filtering, we can get the feature words pairs.

2) Expansion the feature of the test corpus

Expansion the feature vector of the test corpus by the relational words pairs which are extracted from the training corpus and feature items, which are described as follows:

Feature vector expansion is show in Figure 2.

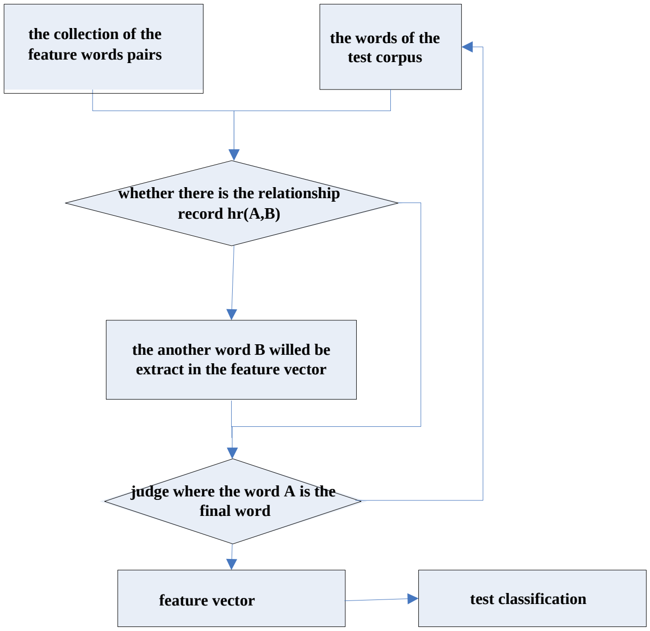

the collection of the feature words pairs
the words of the test corpus
whether there is the relationship record hr(A,B)
the another word B willed be extract in the feature vector
judge where the word A is the final word
feature vector
test classification

Figure 2. Feature vector expansion.

Step1: Judge whether there is the relationship record hr(A,B) by the collection of the feature words pairs and the words of test corpus.

Step2: If there is a relationship record, another word B will be extracted in the feature vector, otherwise go to step3.

Step3: Judge where the word A is the final word. If it is the final word, output the feature vector for text classification, otherwise input the next test word.

After the data processing, the text classification is the same as the traditional text classification.

D. Multi-feature Integration

The importance of each feature for classification is unknown, so we need to get the best weight of each feature. Depending on the different length of each feature, the text is classified by the traditional text classification and short text classification to train the different weight of each feature.

1) Training the weight of each feature

In this paper, the weight of each feature is obtained by linear training which is as follows:

Step1: The classification result of categories which are got by the training of feature t_i is denoted as $\vec{P}_{ti} = (P_{i1}, P_{i2}, ..., P_{in})$.

Step2: If the vector $\vec{a}$ is the vector of the different weight of each feature, the classification weights of the text can be gotten by (5):

$$\vec{f} = \vec{P} \cdot \vec{a} \qquad (5)$$

Step3: By labeling the type for the text of the training corpus, we can get an equation group about $\vec{a}$ using (5) and the value of the vector $\vec{a}$ by linear regression method.

Step4: Solve the average of M training texts, so the final vector of the different weight of each feature can be obtained:

$$\vec{a} = \frac{1}{M}\sum_{i=1}^{M}\vec{a}_i \qquad (6)$$

2)Feature integration

Feature integration is calculated as:

$$\vec{f} = \sum_{i=1}^{m} a_{ti}\overrightarrow{P_{ti}} \qquad (7)$$

where $\vec{P}_{ti} = (P_{i1}, P_{i2}, ..., P_{in})$ is the classification result of categories which is got by the training of feature t_i , $\vec{a}$ is the vector of the different weight of each feature, m is the number of features of the blog article, n is the number of categories.

The classification result is the category which has the highest score in the vector

IV. Experiments and Result Analysis

In the experiment, after the experimental data is collected, three group experiments are made and the results are recorded and, analyzed.

A. Experimental Data

The content of http://blog.sina.com.cn/ is as the reference materials to get the name of the category, and Chinese blog category is the following eight categories: news, sports, finance, entertainment, shopping, reading, travel, and military.

2400 Chinese blog pages as the training data for each category are downloaded from the website of http://blog.sina.com.cn/, and the testing data is 200 Chinese blog pages for each category which are downloaded from the website of http://blog.sina.com.cn/.

B. Experimental Results

There are three experiments. The first experiment only has the traditional text classification method, the second one uses the traditional text classification method and the algorithm of multi-feature integration, and the third one combines the traditional text classification method, the short text classification method and the algorithm of multi-feature integration. Performance evaluation of Chinese blog classification mainly includes the accuracy rate (P), recall rate(R) and $F1$.

$$F1 = \frac{2 \times P \times R}{P + R} \qquad (8)$$

TABLE I. THE RESULT OF BOLG CLASSIFICATION WHICH ONLY HAS THE TRADITIONAL TEXT CLASSIFICATION METHOD

Blog category	**Training corpus/ Testing corpus**	***P* (%)**	***R* (%)**	***F1* (%)**
news	2400/200	84.2	83.0	83.6
sports	2400/200	86.1	85.2	85.6
finance	2400/200	80.3	84.7	82.4
entertainment	2400/200	88.6	87.3	87.9
shopping	2400/200	83.7	84.5	84.1
reading	2400/200	87.1	85.6	86.3
travel	2400/200	88.4	86.3	87.3
military	2400/200	85.6	86.7	86.1

TABLE II. THE RESULT OF CHINESE BLOG CLASSIFICATION WHICH HAS TRADITIONAL TEXT CLASSIFICATION METHOD AND THE ALGORITHM OF MULTI-FEATURE INTEGRATION.

Blog category	**Training corpus/ Testing corpus**	***P* (%)**	***R* (%)**	***F1* (%)**
news	2400/200	85.0	83.4	84.2
sports	2400/200	87.2	87.6	87.4
finance	2400/200	82.1	85.2	83.6
entertainment	2400/200	89.2	88.8	89.0
shopping	2400/200	84.6	85.8	85.2
reading	2400/200	89.2	86.7	87.9
travel	2400/200	89.1	87.2	88.1
military	2400/200	88.7	85.7	87.7

TABLE III. THE RESULT OF CHINESE BLOG CLASSIFICATION WHICH HAS THE TRADITIONAL TEXT CLASSIFICATION METHOD, THE SHORT TEXT CLASSIFICATION METHOD AND THE ALGORITHM OF MULTI-FEATURE INTEGRATION.

Blog category	Training corpus/ Testing corpus	P(%)	R(%)	F1(%)
news	2400/200	87.7	89.3	88.5
sports	2400/200	89.1	88.2	88.6
finance	2400/200	85.5	89.1	87.3
entertainment	2400/200	90.2	90.9	90.5
shopping	2400/200	87.6	88.1	87.8
reading	2400/200	90.2	89.7	89.9
travel	2400/200	93.1	90.6	91.8
military	2400/200	92.2	89.7	90.9

Accuracy rate and recall rate reflect two different aspects of classification quality, while a comprehensive evaluation index of the two aspects is the *F*1 value which is shown in Figure 3.

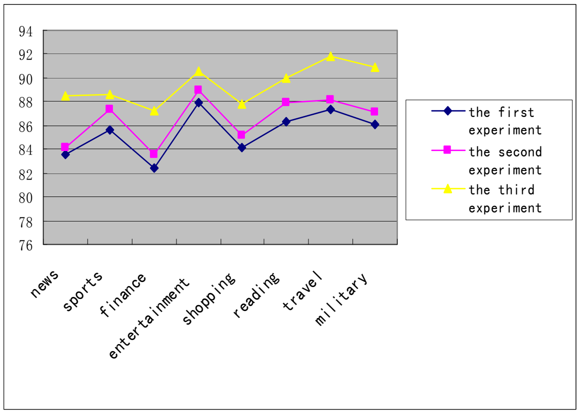


Figure 3. The comparison of comprehensive index *F*1 value

The results are clearly showed in the three tables, the compare results are expressly showed in the figure 2, and the result analysis will be introduced in the Section C.

The running times of the experiments are shown in the Table IV.

TABLE IV. THE RUNNING TIMES OF THE EXPERIMENTS

The times	The first experiment	The second experiment	The third experiment
total time	56min	66min	70min

C. Result Analysis

Comparing the fist experiment and the second experiment, it proves that the algorithm of multi-feature integration makes the Chinese blog classification more effective. The comparison of the second experiment and the third experiment shows that adding the short text classification makes the Chinese blog classification improved.

From the table IV, we can see that the time of the first experiment is shorter than that of the second experiment, and the time of the second experiment is shorter than that of the third experiment.

V. CONCLUSION AND FUTURE WORK

This paper presents a Chinese blog classification method which is based on the algorithm text classification and feature integration. Experimental results show that Chinese blog classification method proposed in this paper makes the classification accuracy of classification improved.

On the other hand, the blogger is the author of the blog. The blogger's interest directly affects the blog category, and the user's interest model will be structured to assist blog classification and more work about the classification will be done to perfect the blog classification.

ACKNOWLEDGMENT

The experiment used Institute of Computer Technology segmentation system interface, the date of HowNet and the data of [14], a special thanks here.

REFERENCES

[1] Hui Y., Bin Y., Xu Z., Chunguang Z., Zhe W., and Zhou C. Community discovery and sentiment mining for Chinese blog [C]. Fuzzy Systems and Knowledge Discovery. 2010, pp. 1740-1745.

[2] Aixin S., Suryanto M. A., and Liu Y. Blog classification using tags :an empirical study [C]. International Conference on Asia-Pacific Digital Libraries. 2007, pp. 307-316.

[3] Singh A. K and Joshi R. C. Semantic tagging and classification of blogs [C]. 2010 International Conference on Computer and ommunication Technology. 2010, pp. 455-459.

[4] Mai L and Nenghai Y. Multi-feature Fusion Method for Blog Post Classification [J]. Journal of Chinese Computer Systems. 2010, pp. 1129-1132.

[5] Luli C. Chinese Weblog Pages Classification Based on Folksonomy [J]. Computer engineering. 2009, pp. 50-52.

[6] Wajeed M. A. Building clusters with distributed features for text classification using KNN [C], in 2012 International Conference on Computer Communication and Informatics. 2012, pp. 583-605.

[7] Lewis D. Naive(Bayes)at forty:The independence assumption in information retrieval [C]. Lecture Notes in Computer Science. Heidelberg:Springer-Verlag, 1998, pp. 4-15

[8] Vapink V. Statistical Learning Theory [M]. New York:Spromger, 1998.

[9] Vapnik V.The Nature of Statistical Learning Theory [M]. New Yorlc:Springer, 1995.

[10] Qian Z., Mingsheng Z., and Min H. Study on Feature Selection in Chinese Text Categorization [J]. Journal of Chinese Information processing. 2004, pp. 17-23.

[11] Qiang N., Zhixiao W., Dai C., and Shixiong X. Web Document Classification Based on SVM [J]. Microelectronics & Computer. 2006, pp. 102-104.

[12] Yahui N., Xinghua F., and Yu W. Short Text Classification Based on Domain Word Ontology [J]. Computer Science. 2009, pp. 142-145.

[13] Huiqing C and Shiping L. A Taxonomic Relation Extraction Method Based on HowNet and Bootstrapping [C]. 2009 Communication theory and new technology development-The Fourteenth National Youth Conference on communication. 2009, pp. 102-108.

[14] http://blog.sina.com.cn/.

Stability Analysis of Cohen-Grossberg Neural Networks With Unbounded Delays

Xuyang Lou
Key Laboratory of Advanced Process Control
for Light Industry (Ministry of Education)
Jiangnan University
Wuxi, China
Email: Xuyang.Lou@gmail.com

Baotong Cui
School of IoT Engineering
Jiangnan University
Wuxi, China
Email: btcui@vip.sohu.com

Qian Ye
School of IoT Engineering
Jiangnan University
Wuxi, China
Email: yeqian85@gmail.com

***Abstract*—The asymptotic stability problem of Cohen-Grossberg neural networks with distributed delays is investigated in this paper. One new uniqueness theorem for the existence of the unique equilibrium of the class of neural networks is presented. Based on the new result, using the Lyapunov stability theory and linear matrix inequality (LMI) technique, and combining Cauchy's inequality, some new conditions for the asymptotic stability of Cohen-Grossberg neural networks with distributed delays are presented. In our results, we do not assume the signal propagation functions to be bounded, differentiable, strictly increasing, and even to satisfy the Lipschitz condition. Moreover, the symmetry of the connection matrix is not also necessary. Thus, we improve some previous works of other researchers. Some examples are also worked out to validate the advantages of our results.**

***Keywords*-Cohen-Grossberg neural networks; asymptotic stability; distributed delay.**

I. INTRODUCTION

In recent years, there has been increasing interest in the potential applications of neural networks in many areas. Many scientists established various types of conditions for the asymptotic stability, absolute stability, complete stability and exponential stability of Hopfield neural networks (HNN), cellular neural networks (CNN), bidirectional associative memory (BAM) neural networks and Cohen-Grossberg neural networks (CGNN) (see [1]–[3] and the references therein).

The Cohen-Grossberg neural network models, initially proposed and studied in Cohen and Grossberg [4], have attracted increasing interest. This class of networks has good application in associative memory, parallel computation and optimization problems, which has been an active area of research and has received much attention. Wang and Zou [5] presented some sufficient conditions for exponential stability of delayed CGNN with asymmetric connection matrix and gave an estimate of the convergence rate. In [6], several sufficient conditions were obtained to ensure a class of delayed CGNN to be asymptotically stable. In [7], based on Lyapunov stability theory and LMI, several sufficient conditions were obtained to ensure delayed CGNN to be robustly stable. Yuan and Cao [9] gave an analysis of global asymptotic stability for a delayed Cohen-Grossberg neural network via nonsmooth analysis. Lu and Chen [8] provided criteria for global stability and global exponential stability with consideration of signs of entries of the connection matrix by using the concept of Lyapunov diagonally stability (LDS) and LMI approach. All of these results above are based on the assumption that the signal propagation functions satisfy either the Lipschitz condition or the boundedness. However, in many evolutionary processes as well as optimal control models and flying object motions, there are many bounded monotone-nondecreasing signal functions which do not satisfy the Lipschitz condition [10]. Therefore, it is important and, in fact, necessary to study the issue of global stability of such a dynamical neural network with non-Lipschitzian activation functions.

Although the use of constant fixed delays in models of delayed feedback provides a good approximation in simple circuits composed of a small number of cells, neural network usually has a spatial nature due to the presence of an amount of parallel pathways of a variety of axon sizes and lengths [2]. In these circumstances, the transmission of signal is no longer instantaneous and cannot be modelled with discrete delays. A more appropriate way is to incorporate distributed delays. Therefore, the studies of the model with distributed delays have more important significance than the ones of model with discrete delays and the distributed delay becomes a discrete delay when the delay kernel is a δ-function, at a certain time (see, Remark 4). However, to the best of our knowledge, few authors [11] have considered Cohen-Grossberg neural network model with distributed delays. Furthermore, the asymptotic stability analysis for CGNN with distributed delays via LMI technique has never been tackled.

Motivated by the above discussions, our objective in this paper is to study further the existence and uniqueness, and global asymptotic stability for the equilibrium point of CGNN with distributed delays, as in [11], but we drop the boundness, differentiability, monotonicity and the Lipschitz condition of the activation functions. Moreover, the symmetry of the connection matrix is not also necessary and the kernel functions need not satisfy the hypothesis $\int_0^\infty sK_j(s)ds < \infty$. Here, a new approach based on LMI technique combining Cauchy's inequality, is developed to obtain sufficient conditions, which guarantee the existence, uniqueness and global asymptotic stability for the equilibrium point of CGNN with distributed delays. The conditions are less conservative than those [11]. Therefore, our proposed results are practical and improve some previous works of other researchers.

II. Model description

In this paper, we consider the following model

$$\begin{aligned}\frac{dx_i(t)}{dt} &= -a_i(x_i(t))\Bigg[b_i(x_i(t)) - \sum_{j=1}^{n} w_{ij} f_j(x_j(t)) - J_i \\ &\quad - \sum_{j=1}^{n} w_{ij}^{\tau} \int_{-\infty}^{t} K_j(t-s) f_j(x_j(s))ds\Bigg], \qquad (1)\end{aligned}$$

where $x_i(t)$ corresponds to the state of the ith unit at time t; J_i, $i = 1, 2, \cdots, n$, denote the constant inputs from outside of the system and w_{ij} represent the connection weights. $a_i(x_i(t))$ and $b_i(x_i(t))$, $i = 1, 2, \cdots, n$, are the amplification functions and the self-signal functions, respectively, while $f_j(x_j)$, $j = 1, 2, \cdots, n$, are the activation functions. $W = (w_{ij})_{n\times n}$ and $W = (w_{ij}^{\tau})_{n\times n}$ are the normal and the delayed connection weight matrix, respectively. The delay kernel K_j is a real value non-negative continuous function defined on $[0, \infty)$ and satisfies, for each j,

$$\int_{0}^{\infty} K_j(s)ds = 1.$$

Throughout the paper, we always assume that

- (H_1) $a_i(x)$ are continuous and positive, i.e., $a_i(x) > 0$, for all $x \in \mathbb{R}, i = 1, 2, \cdots, n$;
- (H_2) each function $b_i(x)$ is locally Lipschitz continuous and there exists $\gamma_i > 0$ such that

$$u[b_i(u + x) - b_i(x)] \geq \gamma_i u^2,$$

for all $x \in \mathbb{R}, i = 1, 2, \cdots, n$;

- (H_3) the functions f_i $(i = 1, 2, \cdots, n)$ satisfy $vf_i(v) > 0$ $(v \neq 0)$, and there exist positive constants μ_i $(i = 1, 2, \cdots, n)$ such that

$$\mu_i = \sup_{v\neq 0} \frac{f_i(v)}{v}, \ \ \forall v \in \mathbb{R}.$$

Remark 1. In [5]–[8], the activation function was required to be bounded, positive and continuous. However, the upper bound of amplification function in this paper is not required. In addition, assumption (H_3) in this paper is as same as that in [5], [9], the condition of differentiability of behaved function in [6]–[8] is not required.

Remark 2. Note that the assumption (H_3) is weaker than the locally and partially Lipschitz condition which is mostly used in literature [5]–[9]. The activation functions such as sigmoid type and piecewise linear type are also the special case of the function satisfying assumption (H_3). Further, if $f_j(\cdot)$ for each $j = 1, 2, \cdots, n$ is a Lipschitz function, then μ_j for each $j = 1, 2, \cdots, n$ can be replaced by the respective Lipschitz constant.

Remark 3. The kernel functions need not satisfy the hypothesis $\int_0^\infty sK_j(s)ds < \infty$ which is required in [11].

Let $\mathcal{C}[X, Y]$ be a continuous mapping set from the topological space X to the topological space Y, and $\mathbb{R}_+ = [0, \infty)$. Especially, $\mathcal{C} \triangleq \mathcal{C}[(-\infty, 0], \mathbb{R}^n]$. Denote A^T and A^{-1} to be the transpose and the inverse of any square matrix A. We use $A > 0$ $(A < 0)$ to denote a positive- (negative-) definite matrix A; and I is used to denote the $n \times n$ identity matrix.

Definition 1. $x(t) = x^* \in \mathbb{R}^n$ is called to be an equilibrium point of system (1), if the constant vector $x^* = (x_1^*, \cdots, x_n^*)^T$ satisfies

$$b_i(x_i^*) = \sum_{j=1}^{n} w_{ij} f_j(x_j^*) + \sum_{j=1}^{n} w_{ij}^{\tau} \int_{-\infty}^{t} K_j(t-s) f_j(x_j^*)ds + J_i$$

for $i = 1, 2, \cdots, n$.

Definition 2. The set $S \subset C$ is called to he a positive invariant set of the system (1) if for any initial value $\phi \in S$, we have the solution $x(t) \in S$, for $t \geq 0$.

III. Existence and uniqueness of the equilibrium point

In order to study the existence and uniqueness of the equilibrium point, we rewrite the system (1) as

$$\dot{X}(t) = F(X(t)), \qquad (2)$$

where
$X(t) = (x_1(t), \cdots, x_n(t))^T$,
$F(X(t)) = (\theta_1(t), \cdots, \theta_n(t))^T$ with

$$\begin{aligned}\theta_i(t) &= -a_i(x_i(t))\Bigg[b_i(x_i(t)) - \sum_{j=1}^{n} w_{ij} f_j(x_j(t)) \\ &\quad - \sum_{j=1}^{n} w_{ij}^{\tau} \int_{-\infty}^{t} K_j(t-s) f_j(x_j(s))ds - J_i\Bigg]\end{aligned}$$

for $i = 1, 2, \cdots, n$.

We consider the initial value problem associated with the autonomous system (2), in which the initial functions are given by

$$x_i(t) = \phi_i(t), \ -\infty < t \leq 0, \ i = 1, 2, \cdots, n, \qquad (3)$$

where $\phi_i(t)$ $(i = 1, 2, \cdots, n)$ are assumed to be bounded and continuous functions on $(-\infty, 0]$. Let Ω be an open subset of $\mathbb{R}^n$. For any $\theta \in \mathbb{R}^n$, we define $\|\theta\| = \sum\limits_{j=1}^{n} |\theta_j|$.

Theorem 1. Let $F : \Omega \to \mathbb{R}^n$ be continuous and satisfy the following condition: corresponding to each point $\theta \in \Omega$

and its neighborhood U, there exists a constant $k > 0$, and functions h_j and Ψ_l $(j, l = 1, 2, \cdots, n)$ such that

$$\|F(\vartheta) - F(\theta)\| \leq k\|\vartheta - \theta\| + k \sum_{l=1}^{n} \Big| \Psi_l(h_j(\vartheta)) - \Psi_l(h_j(\theta)) \Big|$$

on U, where each $h_j : U \to \mathbb{R}$ is a continuously differentiable function in θ satisfying the relation

$$\sum_{j=1}^{n} \frac{\partial h_j(\theta)}{\partial \theta_i} F_i(\theta) \neq 0 \quad \text{on } U$$

and each $\Psi_l : \mathbb{R} \to \mathbb{R}$ is continuous and of bounded variation on bounded subintervals. Then, there exists a unique solution for the initial value problem Eq. (1) or Eq. (2) with (3).

IV. Global asymptotic stability of the equilibrium point

In this section, we consider the global exponential stability for the system (1). Suppose $x^* = (x_1^*, \cdots, x_n^*)^T$ is any equilibrium point of the system (1).

Theorem 2. Suppose Theorem 1 hold for the functions f_j $(j = 1, 2, \cdots, n)$, and assumptions $(H_1) - (H_3)$ are satisfied. The equilibrium point x^* for the system (1) with (3) is globally asymptotically stable, if there exist a matrix $P > 0$, and two diagonal matrices $R > 0, Q > 0$, such that

$$\Omega = \begin{bmatrix} -2P\Gamma + R & PW & PW^\tau \\ W^T P & -RL^{-2} + Q & 0 \\ (W^\tau)^T P & 0 & -Q \end{bmatrix} < 0, \qquad (4)$$

where $\Gamma = \text{diag}[\gamma_1, \gamma_2, \cdots, \gamma_n]$, $L = \text{diag}[\mu_1, \mu_2, \cdots, \mu_n]$.

Theorem 3. Suppose Theorem 1 hold for the functions f_j $(j = 1, 2, \cdots, n)$, and assumptions $(H_1) - (H_3)$ are satisfied. The equilibrium point x^* of the system (1) with (3) is globally asymptotically stable if there exist a matrix $P > 0$, and two diagonal matrices $D > 0, Q > 0$, such that

$$\Theta = \begin{bmatrix} -P\Gamma - \Gamma P & PW & PW^\tau \\ W^T P & \Xi & DW^\tau \\ (W^\tau)^T P & (W^\tau)^T D & -Q \end{bmatrix} < 0 \qquad (5)$$

where
$\Xi = -2D\Gamma L^{-1} + DW + W^T D + Q$,
$\Gamma = \text{diag}[\gamma_1, \gamma_2, \cdots, \gamma_n]$, $L = \text{diag}[\mu_1, \mu_2, \cdots, \mu_n]$.

Remark 4. If delay kernel functions $k_j(t)$ are of the form

$$k_j(t) = \delta(t - \tau_j), \quad j = 1, 2, \cdots, n, \qquad (6)$$

then system (1) reduces to CGNN with discrete delays which has been lucubrated in many literatures. And many crucial results for dynamics of this class of neural networks have been obtained. Therefore the discrete delays can be included in our models by choosing suitable kernel functions.

Remark 5. We can see that the LMI criterion (4) is similar to condition (27) of Corollary 1 in Ref. [7]. However, it should be noted that our result contain that in Ref. [7], because the discrete delays can be included in our models by choosing suitable kernel functions as said in Remark 4. Moreover, the signal propagation functions need not to be bounded and satisfy the Lipschitz condition in this paper, while the assumptions are required in Ref. [7].

Remark 6. For system (1), when $a_i(x_i(t)) = 1$, $b_i(x_i(t)) = b_i(t)x_i(t)$ (in which $b_i(t)$ is not only differentiable but also bounded on interval $(-\infty, +\infty)$, and its maximal lower bound is denoted as $\gamma_i > 0$) and let $W \equiv 0$, the system (1) reduces to a class of pure-delay models with distributed delays which has been studied in [12], but the results derived in this paper are less conservative than those in [12] because of the loose restrictions on the activation functions; when the delay kernel is a δ-function based on the case above, i.e., the distributed delay becomes a discrete delay, system (1) has been briefly indicated in [2].

Remark 7. If the activation functions are bounded and satisfy the Lipschitz condition, Theorem 2 is equivalent to Corollary 1 in Ref. [7]; Theorems 2-3 extend and improve Theorem 3 in Ref. [6].

If the model (1) is simplified to cellular neural networks with time delay, that is, let $a_i(x) = 1$, $b_i(x) = x$, $f_i(x) = 0.5(|x + 1| - |x - 1|)$, then we have $\Gamma = I$, $L = I$. We can have the following corollaries.

Corollary 1. The equilibrium point x^* of the system (1) with (3) is globally asymptotically stable if there exist a matrix $P > 0$, and two diagonal matrices $R > 0, Q > 0$, such that

$$\Omega = \begin{bmatrix} -2P + R & PW & PW^\tau \\ W^T P & -R + Q & 0 \\ (W^\tau)^T P & 0 & -Q \end{bmatrix} < 0. \qquad (7)$$

Corollary 2. The equilibrium point x^* of the system (1) with (3) is globally asymptotically stable if there exist a matrix $P > 0$, and two diagonal matrices $D > 0, Q > 0$, such that

$$\Theta = \begin{bmatrix} -2P & PW & PW^\tau \\ W^T P & -2D + DW + W^T D + Q & DW^\tau \\ (W^\tau)^T P & (W^\tau)^T D & -Q \end{bmatrix} < 0. \qquad (8)$$

Remark 8. The common feature for the asymptotic stability of CGNN with distributed delays is that the conditions are expressed in terms of some nonlinear inequalities, which involve the tuning of some scalar parameters. Although the suitability of these criteria is improved due to these adaptable parameters, it is not easy to check the availability of the scalars since we have no a systematic tuning procedure by now. The criteria in Theorems 2-3 are LMI conditions, which do not require the tuning of parameters.

V. Numerical simulations

In the previous sections, some new sufficient criteria for the global asymptotic stability of the Cohen-Grossberg neural networks with distributed delays have been derived.

In the following examples, for simplicity, some Cohen-Grossberg models with only two neurons are simulated and analyzed.

Example 1. Consider the following CGNN with distributed delays:

$$\begin{cases} \frac{dx_1(t)}{dt} = -(2+\sin(x_1(t))\Big[2x_1(t) \\ \qquad -\frac{1}{4}|x_1(t)| - \frac{1}{4}|x_2(t)| \\ \qquad -\frac{1}{4}\int_{-\infty}^{t} K(t-s)|x_1(s)|ds \\ \qquad -\frac{1}{4}\int_{-\infty}^{t} K(t-s)|x_2(s)|ds + 2\Big], \\ \frac{dx_2(t)}{dt} = -(3+\cos(x_2(t))\Big[2x_2(t) \\ \qquad -\frac{1}{6}|x_1(t)| - \frac{1}{3}|x_2(t)| \\ \qquad -\frac{1}{3}\int_{-\infty}^{t} K(t-s)|x_1(s)|ds \\ \qquad -\frac{2}{3}\int_{-\infty}^{t} K(t-s)|x_2(s)|ds - 2\Big], \end{cases} \tag{9}$$

with initial values

$$\begin{cases} \phi_1(s) = 0.8, s \in (-\infty, 0], \\ \phi_2(s) = 0.5, s \in (-\infty, 0]. \end{cases} \tag{10}$$

One can check $(H_1) - (H_3)$ are satisfied. In this example,

$$W = \begin{bmatrix} \frac{1}{4} & \frac{1}{4} \\ \frac{1}{6} & \frac{1}{3} \end{bmatrix}, \; W^{\tau} = \begin{bmatrix} \frac{1}{4} & \frac{1}{4} \\ \frac{1}{3} & \frac{2}{3} \end{bmatrix},$$

$$J = \begin{bmatrix} -2 \\ 2 \end{bmatrix}, \; \Gamma = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}, \; L = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

For numerical simulation, we choose the delay kernel as $K(r) = e^{-r}$. Applying our Theorem 2, by solving the LMI (4) using the Matlab LMI Toolbox, a feasible solution is

$$P = \begin{bmatrix} 0.9813 & -0.0423 \\ -0.0423 & 0.9311 \end{bmatrix} > 0,$$

$$Q = \begin{bmatrix} 1.2750 & 0 \\ 0 & 1.2750 \end{bmatrix} > 0,$$

$$R = \begin{bmatrix} 2.5499 & 0 \\ 0 & 2.5499 \end{bmatrix} > 0.$$

Therefore, the conditions of Theorem 2 in this paper are satisfied, which implies system (9) has a unique equilibrium point, which is asymptotically stable. Figure 1 shows the time responses of the state variables $x_1(t)$ and $x_2(t)$ with 10 initial states. They have confirmed that by fulfilling the proposed conditions, the existence of a unique equilibrium point $x^* = [-0.5218, 1.9131]^T$, and the global asymptotic stability of system (9) are guaranteed.

Since $f_1(x) = f_2(x) = |x|$ here, we can easily verify that the assumptions of boundedness, monotonicity, and differentiability for the activation functions is not available, so the results in [11] and the references cited therein can not be applicable to system (9).

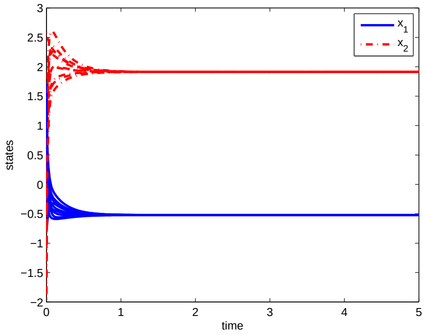


Figure 1. Transient response of state variables $x_1(t)$ and $x_2(t)$ for Example 1.

Example 2. To illustrate Theorem 3, we consider the following Cohen-Grossberg model with distributed delays:

$$\begin{cases} \frac{dx_1(t)}{dt} = -(2+\sin(x_1(t))\Big[2x_1(t) \\ \qquad -\frac{1}{4}f_1(x_1(t)) - \frac{1}{4}f_2(x_2(t)) \\ \qquad -\frac{1}{4}\int_{-\infty}^{t} K(t-s)f_1(x_1(s))ds \\ \qquad -\frac{1}{4}\int_{-\infty}^{t} K(t-s)f_2(x_2(s))ds + 1\Big], \\ \frac{dx_2(t)}{dt} = -(3+\cos(x_2(t))\Big[2x_2(t) \\ \qquad -\frac{1}{6}f_1(x_1(t)) - \frac{1}{3}f_2(x_2(t)) \\ \qquad -\frac{1}{3}\int_{-\infty}^{t} K(t-s)f_1(x_1(s))ds \\ \qquad -\frac{2}{3}\int_{-\infty}^{t} K(t-s)f_2(x_2(s))ds - 1\Big], \end{cases} \tag{11}$$

with initial values

$$\begin{cases} \phi_1(s) = -0.5, s \in (-\infty, 0], \\ \phi_2(s) = 0.5, s \in (-\infty, 0], \end{cases} \tag{12}$$

where f_1 and f_2 are exponentially weighted time averages of the sampled pulse

$$f_j(x_j(s)) = \int_{-\infty}^{s} x_j(\theta)e^{\theta - s}d\theta, \; j = 1, 2, \tag{13}$$

the functions x_1 and x_2 equal one when a pulse arrives at time s and zero when no pulse arrives. Obviously, f_j satisfy (H_3) with μ_j but it does not satisfy the Lipschitz condition.

In this example,

$$W = \begin{bmatrix} \frac{1}{4} & \frac{1}{4} \\ \frac{1}{6} & \frac{1}{3} \end{bmatrix}, \; W^{\tau} = \begin{bmatrix} \frac{1}{4} & \frac{1}{4} \\ \frac{1}{3} & \frac{2}{3} \end{bmatrix},$$

$$J = \begin{bmatrix} -1 \\ 1 \end{bmatrix}, \; \Gamma = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}, \; L = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

By taking $K_j(r) = \frac{2}{\pi(1+r^2)}, j = 1, 2$. Then we have $\int_0^\infty K_j(s)ds = 1$. Clearly, the kernel functions K_j $(j = 1, 2)$ do not satisfy the hypothesis $\int_0^\infty sK_j(s)ds < \infty$. Applying our Theorem 3, by solving (5) using the Matlab LMI Toolbox, a feasible solution is

$$P = \begin{bmatrix} 0.4038 & -0.0174 \\ -0.0174 & 0.3832 \end{bmatrix} > 0,$$

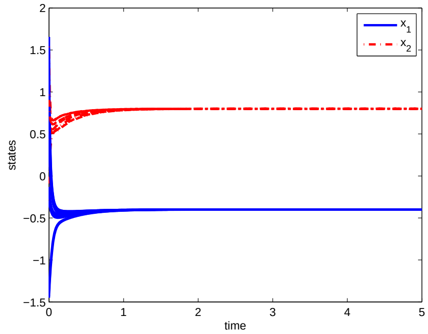


Figure 2. Transient response of state variables $x_1(t)$ and $x_2(t)$ for Example 2.

$$Q = \begin{bmatrix} 1.4488 & 0 \\ 0 & 1.4488 \end{bmatrix} > 0,$$

$$D = \begin{bmatrix} 0.8481 & 0 \\ 0 & 0.8481 \end{bmatrix} > 0.$$

Therefore, the conditions of Theorem 3 in this paper are satisfied, which implies system (11) has a unique equilibrium point. It is easy to verify that $x^* = [-0.4000, 0.8001]^T$ is the unique equilibrium point which is asymptotically stable. Figure 2 shows the time responses of the state variables $x_1(t)$ and $x_2(t)$ with 10 initial states. However, it is very difficult to obtain the result by using the technique in [11] for system (11) with the non-Lipschitzian activation functions.

VI. Conclusions

In this paper, using the Lyapunov stability theory and LMI technique, and combining Cauchy's inequality, we have derived some new sufficient conditions in term of LMI for the existence and uniqueness, and global asymptotic stability for the equilibrium point of CGNN model with distributed delays. The results presented here are more general and easier to check than those given in the related literature because the restrictions of sufficient conditions are less restrictive than those in [5]–[9]. Two examples are provided to illustrate our results.

Acknowledgements

This work is partially supported by National Natural Science Foundation of China (No.61174021, No.61104155), and the 111 Project (B12018).

References

[1] K. Gopalsamy and X. He, "Stability in asymmetric Hopfield nets with transmission delays," *Physica D* 76 (1994) 344-358.

[2] Y. T. Li and C. B. Yang, "Global exponential stability analysis on impulsive BAM neural networks with distributed delays," *J. Math. Anal. App.* 324 (2006) 1125-1139.

[3] O. Faydasicok and S. Arik, "Equilibrium and stability analysis of delayed neural networks under parameter uncertainties," *Applied Mathematics and Computation* 218 (2012) 6716-6726.

[4] M. A. Cohen and S. Grossberg, Absolute stability of global pattern formation and parallel memory storage by competitive neural networks," *IEEE Trans. Syst., Man, Cybern.* 13(5) (1983) 815-826.

[5] L. Wang and X. F. Zou, "Exponential stability of Cohen-Grossberg neural networks," *Neural Networks* 15 (2002) 415-422.

[6] T. P. Chen and L. B. Rong, "Delay-independent stability analysis of Cohen-Grossberg neural networks," *Phys. Lett. A* 317 (2003) 436-449.

[7] L. B. Rong, "LMI-based criteria for robust stability of Cohen-Grossberg neural networks with delay," *Phys. Lett. A* 339 (2005) 63-73.

[8] W. L. Lu and T. P. Chen, "New conditions on global stability of Cohen-Grossberg neural networks," *Neural Comput.* 15 (2003) 1173-1189.

[9] K. Yuan and J. D. Cao, "An analysis of global asymptotic stability of delayed Cohen-Grossberg neural networks via nonsmooth analysis," *IEEE Trans. Circuits Syst. I* 52 (2005) 1854-1861.

[10] B. Kosko, *Neural Networks and Fuzzy System-A Dynamical Systems Approach to Machine Intelligence*, New Delhi, India: Prentice-Hall of India, 1994.

[11] X. F. Liao, C. G. Li and K. W. Wong, "Criteria for exponential stability of Cohen-Grossberg neural networks," *Neural Networks* 17 (2004) 1401-1414.

[12] Q. Zhang, X. P. Wei and J. Xu, "Global exponential stability of Hopfield neural networks with continuously distributed delays," *Phys. Lett. A* 315 (2003) 431-436.

Environmental Codes for Autonomous Position Determination

Kamen Kanev, Hirotaka Kato
Research Institute of Electronics
Shizuoka University
Hamamatsu, Japan
kanev@rie.shizuoka.ac.jp, gs10011@s.inf.shizuoka.ac.jp

Reneta Barneva, Zhuojun Fu
Department of Computer and Information Sciences
State University of New York
Fredonia, USA
barneva@cs.fredonia.edu, fu0317@fredonia.edu

Shigeo Kimura
Institute of Nature and Environmental Technology
Kanazawa University
Kanazawa, Japan
skimura@t.kanazawa-u.ac.jp

Abstract— **This work focuses on pervasive environmental codes that serve as an interface to augmented vision devices and provide support for localization and automated navigation. We begin with a concise overview of methods for automating the localization of humans and autonomous agents including mobile robots. Automated localization is based on mapping where positions, orientations, and other localization parameters are determined either on a plane or in a three-dimensional space. While various devices such as sonar and ultrasound locators, laser scanners, visible light and infrared cameras, etc. are considered for gathering of the necessary mapping information the focus of our work is on the innovative system for environmental semantic encoding that we have developed. In this system, we have successfully implemented semantic surfaces with embedded marking which provide additional information alone, separately and independently from all the visual features and properties of the surrounding physical surfaces.**

Keywords - position determination; mapping; semantic surfaces; navigation; SLAM; surface encoding; CLUSPI.

I. INTRODUCTION

Human activities as well as activities of autonomous agents and mobile robots are essentially connected to the surrounding environment. An interface layer that is responsible for the interactions taking place within such activities could therefore be established and maintained. In this work we attempt to establish such a layer by introducing a special type of environmental semantic encoding that is implemented with *ubiquitous semantic surfaces* [7]. An introduction to semantic surfaces and details about our environmental semantic encoding approach are given in Section II. In this section, we continue with an introduction to a more general localization and mapping. For illustration of autonomous position determination we will refer to robotic mapping, e.g., the building of a map of a local environment surrounding a robot.

The simultaneous robot orientation and map building is an estimation problem known as SLAM (Simultaneous Localization and Mapping) [12]. SLAM is an essential capability for any autonomous agent or a mobile robot traveling in unknown environments where globally accurate position data is not available [9, 10]. High uncertainty often exists in such environments so the capability to map them is essential in order to allow a robot to be deployed with a minimal infrastructure.

A variety of sensors such as sonar and ultrasound, laser, visible light and infrared as well as digital cameras are commonly used to gather information and to “capture” the local environment. For now, we assume that maps are static, that is, no relative movement of environment features exist and no intermittent changes in such features are allowed. However, despite this assumption, the uncertainties of the robot state can become arbitrarily large [12]. This stresses upon the necessity of more reliable tracking of the exact positions of landmarks and other environmental features that can reduce the uncertainty of the robot state. The environmental codes that we consider in this work are specifically designed for environment enhancements facilitating such tracking and potential use in SLAM and FastSLAM [10, 11].

FastSLAM is an efficient SLAM algorithm, which decomposes the SLAM problem into two, namely a robot localization problem and a collection of landmark estimation problems. It uses a modified particle filter for assessing the posterior over robot paths instead of the extended Kalman filter (EKF), which reduces the running time from linear to logarithmical in respect to the number of landmarks [10, 11]. Another advantage of FastSLAM over the EFK is its multi-hypothesis data association. Since each FastSLAM particle represents a specific robot path, data association decision can be performed on a per-particle basis. High weights will be assigned to correct data association in terms of high chances for future resampling. If the data association is incorrect, the weight will be low and then the association will be removed later.

In Section II, we describe the semantic encoding method that we have developed and provide details about the implemented software system that employs it. In Section III, we present experimental measurements and validation of the developed system. In Section IV, we elaborate on the approach of 3D mapping with semantics and SLAM. Finally, we conclude in Section V with some plans for further work.

II. SEMANTIC SURFACES

In linguistics, *semantics* refers to the meaning carried by a language. In the case of mapping, we define *semantic surface* as a surface with embedded marking which provides additional information alone, separately and independently from all the visual features and properties of the surface [7]. Such an embedded surface marking can be implemented, for example, as a (dense) grid of landmarks with links to nodes with specific meaning. We believe that the semantic surfaces as defined here, although applicable outdoors, are best employed in indoor environment and for small scale navigation. In the following sections we discuss in more detail various methods and approaches for surface marking and creation of semantic surfaces that are suitable for position determination of autonomous agents and mobile robots.

A. *Environmental semantic encoding*

In our daily life, we often encounter different codes embedded or attached to various products and equipment. Such codes usually carry digital information, specific to the artifact they are associated with, and are used for its tagging and consequent identification. Typical examples are barcodes employed in shops for merchandise management and control. Although most of the barcodes currently in use are still one-dimensional, more advanced two-dimensional barcodes such as QR and Datamatrix codes, Color codes, and others are being adopted. In addition to their business use, both old-type and newer codes are becoming more accessible for ordinary people through various gadgets, such as camera enabled mobile phones and smart phones, etc. Typical applications involve scanning of a code by a mobile device camera, decoding and extracting information embedded in it, and providing related feedback to the user. This is a very powerful application model that allows on-time delivery of up-to-date context-dependent information, dynamically adjustable to meet the specific need of the current user. It takes advantage of the continuous 3G/WiFi connectivity of the current mobile devices, of their ability to take snapshots of the surrounding environment by a simple press of a button, and of the user profiles containing usage history and preferences information stored on the device.

In a similar way, autonomous agents such as mobile robots can also take advantage of these codes. Context dependent information provided in this way could be further extended and even better targeted if localization information is available. Standard GPS-based tracking and position determination, however, is generally not sufficient for reliable precise localization indoors. To address this issue, we introduce here our research on ubiquitous environmental digital codes for global positioning and navigation [4]. These codes are specifically designed to seamlessly blend in the surrounding environment by ether being practically undetectable by naked human eye due to the size and shape of the employed marks or by becoming part of the environmental patterns naturally covering walls and other surrounding surfaces. Methods for generation of unobtrusive surface codes that blend well with existing printed content have been reported in [1,8] and our work on direct embedding of such codes in surfaces and into the bulk of physical objects have been reported in [3].

In this paper, we focus on larger-scale codes that are integrated with various patterns on surrounding surfaces [4]. The codes do not interfere with the look and feel of the surrounding environment and thus do not disturb the humans. Autonomous agents and mobile robots, however, can extract the codes from their surroundings and employ them for localization (position and orientation) determination.

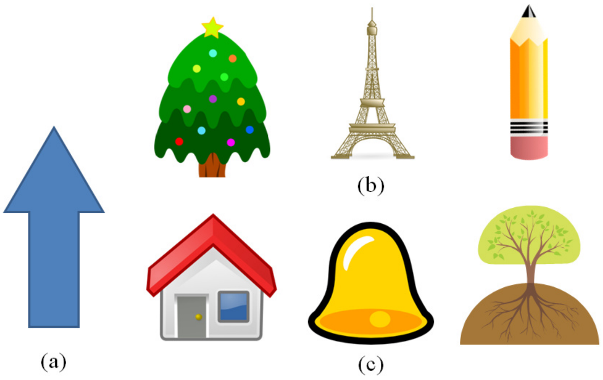


Figure 1. A sample human recognizable arrow (a) and algorithmically recognizable (b) and non-recognizable (c) "A-shaped" objects

There are known approaches addressing interior design patterns [2], where figures with different distinguishable shapes are used for the encoding. As discussed in [2], experimental interior design patterns based on 4-figure and 6-figure digital encoding have been created and consequent figure recognition and decoding test have been conducted. Reported experimental results suggest that better system performance and higher recognition rates would need to be secured before its final adoption.

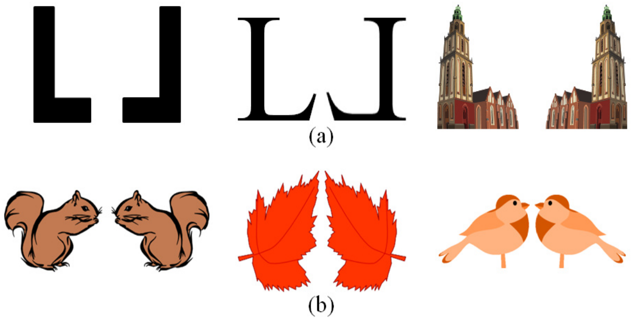


Figure 2. A sample human recognizable (a) and algorithmically recognizable (b) "L-shaped" objects

In contrast to the above discussed interior design patterns, the encoding scheme that we employ in our work uses non-shape based figure discrimination. Its computational complexity is significantly lower than the F-descriptor based method employed in [2]. Our method can work with single figure type patterns where no figure type discrimination is required and digitally encoded data is simply related to the figure rotations. In this way various, differently shaped and sized graphical objects can serve as figures in our patterns where only their orientation matter. For reliable determination of the figure rotational angles, however, graphical objects with easily distinguishable main or up direction should be used.

The arrow shaped figure shown in Fig. 1(a) is a good example of direction recognition by humans, since its pointing direction is unmistakable. But it is also a good example of a graphical object with algorithmically-easy determinable direction. Simple methods to determine the arrow direction would be to find its mass center and to connect it to the most distant arrow point, to calculate and use moments of higher order, etc. Since the figure orientation will be determined algorithmically, it does not necessarily have to match the human judgment. We can, therefore, say that any figure with easily determinable main direction by the employed algorithms could be considered an arrow or an A-shaped object. To humans it may not look like or even remotely resemble an arrow but it can be treated as an A-shaped object as long as its main direction is well determined by the employed software. Some examples of objects that can and cannot be considered as A-shaped are given in Fig. 1(b) and Fig. 1(c), respectively.

A simple extension of this idea would be considering L-shaped objects. Same as for A-shaped objects, the main direction of the L-shaped objects should be well defined and easily determined algorithmically. In addition to this, L-shaped objects and their mirror images should be easily distinguishable. For illustration, examples of human recognizable L-shapes and their mirror images are shown in Fig. 2(a). Further examples of algorithmically recognizable L-shaped objects and their respective mirror images are shown in Fig. 2(b).

B. *Software system*

We have developed an experimental software system (Fig. 3) for basic figure management, for design, generation, and printing of environmental semantic codes, and for code extraction, analysis, and consequent localization in semantically encoded environments.

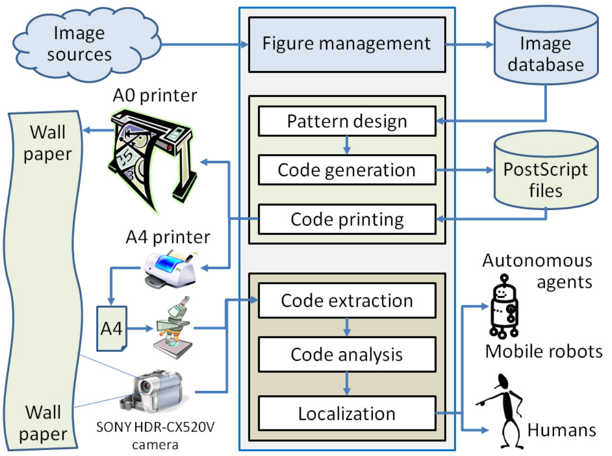


Figure 3. A schematic representation of the developed experiment software system

The system is quite flexible and allows creation of a vast variety of environmental encoding patterns. Depending on the choice of code components or figures, impressive wallpaper patterns close to real artworks could be created.

The elementary figures used by our system are initially organized in a figure database. Employed figures are essentially images in different formats which fall into one of the following two categories:

- raster images: these images are stored in files with extensions bmp, tiff, gif, jpeg, etc. that store arrays of image pixels, and
- vector images: these images are stored in files with extension svg, wmf, etc. and contain drawing information.

Raster images do not scale well, i.e., they get jagged under large magnification, which makes it difficult to use the same designed pattern for differently sized wallpaper codes. Vector images on the other hand scale very well although some standard image processing techniques cannot be directly applied to them.

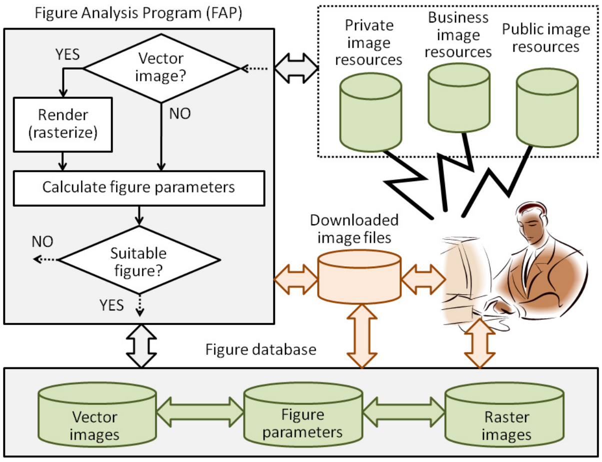


Figure 4. Figure management component of the system

In our system, we support both types of images and we always attempt to employ the most suitable type for the task in question. For example, required figure characteristics such as center of gravity, mass distribution, moments, etc. are calculated based on a raster image file of the figure. If such file is not available, it is automatically generated by scan-conversion (rendering) of the corresponding vector file. In the designed patterns, on the other hand, we use vector representations of the employed figures whenever available.

1) *Figure management*

The process of building the figure database is schematically represented in Fig. 4. First copyright free publicly available, commercial, and private image sources are searched and images that are judged as potentially suitable for environmental semantic encoding are fed into the Figure Analysis Program (FAP). The program accepts all most popular image formats for both raster and vector images. Raster images supplied to the program are immediately evaluated [4] and those found unsuitable are rejected. For positively assessed raster images an attempt is made to locate and download a vector version of the same figure if available. If a vector image is supplied to the program FAP first renders it and then proceeds with its evaluation. For all positively evaluated figures FAP calculates and stores the necessary figure parameters in the database. Once the database is populated with sufficient number of suitable figures the code generation process can be initiated.

2) *Pattern design*

At this stage (Fig. 5), the visual appearance of the environmental semantic code is chosen. As earlier discussed, figures of various shapes, sizes and colors can be employed in the environmental code. Depending on the fundamental code parameters the pattern designer may or may not be limited to certain types of figures, e.g., A-shaped, L-shaped, etc. This, however, is completely transparent to the pattern designer since the software will allow him to use only suitable figures in his design [4]. Figures are arranged on a predetermined grid where the selection of figures for the grid positions is controlled by the pattern designer. Note that other placement parameters of the figures such as figure rotations and small displacements from the grid positions are controlled by the encoding engine. The pattern designer can arrange figures interactively, programmatically, and by combining the previous two methods. This way a fine artistic arrangement can be done interactively and then programmatically replicated to cover large areas.

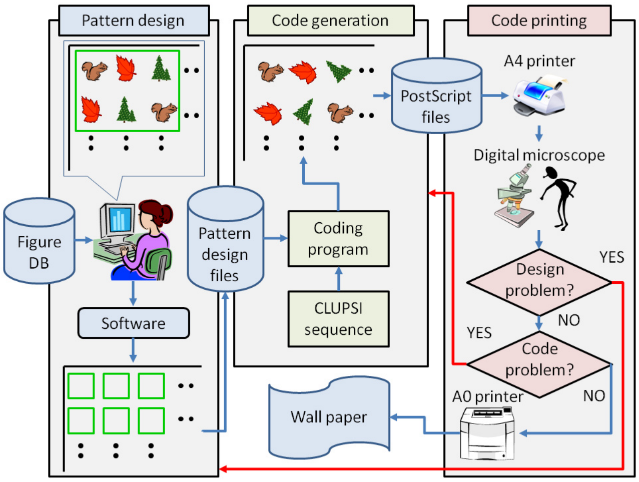


Figure 5. Pattern design, code generation, and code printing components of the system

3) *Code generation*

At this stage (Fig. 5) the Pattern Encoding Engine (PEE) is invoked to calculate the figure rotations that will carry the localization and other semantic information. Some encoding options may also introduce small relative displacement of the figures in respect to their predefined grid positions. Our PEE is based on the Cluster Pattern Interface (CLUSPI) class of codes [5,6]. These codes have no physical margins, blocks, or any other features that may fragment their appearance. CLUSPI encoded semantic surfaces and environmental encoding based on them is a powerful mechanism for creating ubiquitous environments where humans (through specialized devices), autonomous agents, and mobile robots can reliably determine their positions, orientations, and also obtain environment related semantic information.

4) *Code printing*

The environmental semantic codes created by our software are stored in PostScript files that can be directly send to a printing device (Fig. 5). If all images (figures) embedded in the file are of vector type, the code from the file looks well when printed both on small A4 sized and on large A0 sized sheets. This feature of the code is extremely convenient for experimental work. Proofs for visual inspection of the developed codes, for example, can be conveniently printed on A4 sheets. Same proofs can then be used for code consistency checking and structural verification. While the real code is to be used as wall paper (A0 printouts) by using the A4 proof and a digital microscope we are able to verify all essential properties of the printed code.

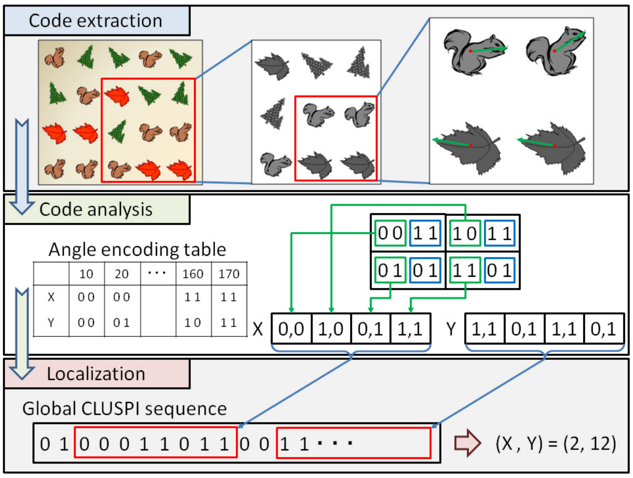


Figure 6. Code extraction, analysis, and localization components of the system

5) *Code extraction*

Typical applications of CLUSPI codes are for digital enhancement of printed materials. In such cases the code reader can be easily tuned to retain a predetermined distance from the encoded surface, e.g., by installing a spacer, which greatly simplifies the code extraction process. However, for large-scale codes embedded in wall paper, for example, positions and orientations of the camera in respect to the encoded surface may vary significantly. With the change of distance, optical parameters of the camera may need adjustment (depending on the focus depth). To tackle this issue we employ an autofocus camera (SONY HDR-CX520V). Another problem is the changing size of the viewing area and thus the total number of figures that can be analyzed which will be discussed in following works. In all cases each frame of the camera video stream is first converted to grayscale with its brightness and contrast normalized (Fig. 6). Then zones containing figures are identified and the figures are separated from the background through standard image processing techniques [15]. Then the barycentre, the orientation (rotational angle), and other parameters (e.g., L-shape mirroring parameters, etc.) are computed for each extracted figure. The obtained data is finally organized in a 2D array corresponding to the figure arrangement on the surface.

6) *Code analysis*

The data contained in the 2D array with extracted figure parameters obtained at the previous step is first converted to short sequences of bits. As shown in Fig. 6, an angle derived from the figure orientation within the pattern carries four bits of information (see the angle encoding table). The first two bits are allocated for encoding the X-coordinate and the remaining two ones are used for encoding the Y-coordinate. The bits for each coordinate from the figures forming a virtual block [4] are arranged in two 8-bit sequences, one for each of the coordinates.

7) *Localization*

Localization information is calculated on the basis of the bit-sequences derived at the previous step (Fig. 6). The CLUSPI scheme encodes coordinates with a global binary sequence having special properties, namely such that any subsequence with a predetermined length (which is a parameter of the code) is unique [1]. Based on that, we can determine the positional coordinates of a set of figures by matching the subsequences that they encode with the global CLUSPI encoding sequence [5,6]. The implementation of this decoding is, of course, more complex and includes redundancy management and error recovery procedures which will be discussed in future works.

III. Experiments and Evaluation

As shown in Fig. 3 and discussed in Section II, generated environmental semantic codes are first printed on A4-sized paper for visual inspection, code consistency checking, and structural verification.

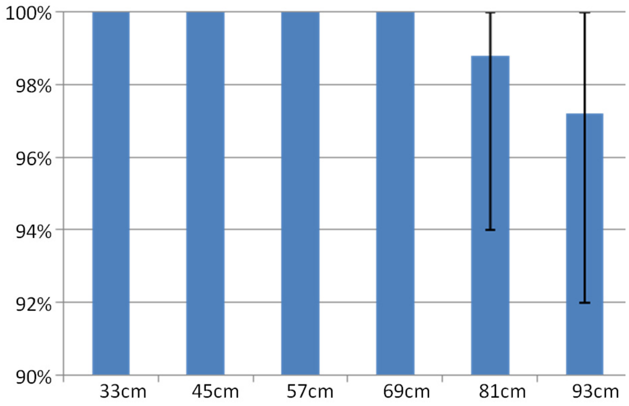


Figure 7. Recognition rate vs. distance (Bars show average; overlaid segments show minimum and maximum values.)

Using such proof pages, we have conducted a range of experiments and measurements of the recognition rates for specific codes at various distances and viewing angles. Obtained results are directly applicable to full-sized environmental encoding patterns by appropriate scaling. The detailed measurement results that we report here are based on a sample coded pattern for position determination with minimal sequences of four figures as shown in Fig. 6. A compact group of four figures (e.g., 2x2) can be viewed with the employed digital microscope from a minimum distance of about 33 cm.

Results for six sets of measurements starting with the minimum distance and incrementing by 12 cm are shown in Fig. 7. At each distance, five spots on the patterned sheet have been randomly chosen and 10 measurements have been done at each spot (50 measurements altogether for each distance). As shown in Fig. 7, no position determination errors have been detected for the distances within the 33-69 cm range and the average success rate remains over 95% for the 69-93 cm range. Above that distance, the recognition rate becomes too low for practical use.

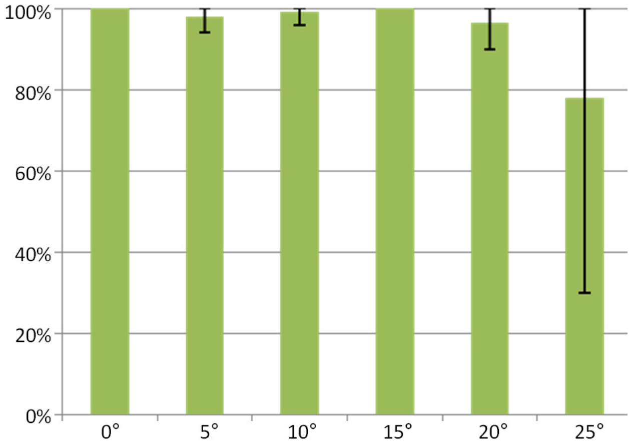


Figure 8. Recognition rate vs. angle (Bars show average; overlaid segments show minimum and maximum values.)

Similarly, results for six sets of measurements at angles from 0° to 25° (in 5° increments) from the vertical are shown in Fig. 8. Again, 10 measurements at 5 random spots have been taken for each of the angles. As shown in Fig. 8 average success rate stays over 95% for all angles up to 20°. Above that angle, recognition rate becomes too low for practical use.

IV. 3D Mapping with Semantics

We consider the pervasive environmental codes in the framework of autonomous position determination as a tool for localization support. The main purpose of a semantic surface in respect to a typical geometric map is to provide "some type of reasoning based on individual entities in the map and/or their classes" [13]. For instance, robots with rescue systems are often designed to assist rescuers to locate victims in various disaster environments such as an earthquake scene. Those systems require reliable maps with specific and detailed object information. In order to identify and properly report the precise localization of injured victims, all available degrees of freedom (e.g., 6DOF) need to be used on top of the available 3D maps. This is an issue that can be addressed by the 6D SLAM method. For real-time applications, however, the time that simple SLAM takes for pose (position and orientation) estimates often becomes a problem. In such situations semantic knowledge could be extracted from the surrounding environment and used to improve the efficiency and to speed up the 6D SLAM method. “A semantic 3D map for mobile robots is a metric map that contains in addition geometric information of 3D data points and assignments of these points to known structures or object classes.” [12,13] Semantic knowledge could, therefore, be provided, through semantic maps implemented as semantic surfaces. The process of 3D mapping with semantics could be divided into four major steps:

- 3D scanning by 6D SLAM,
- 3D scene interpretation,
- objects detection and localization, and
- semantic map presentation.

However, “Prior to the mapping, the object database needs to be initialized and filled with object descriptions both for the 2D and 3D representations” [12,13]. Therefore, due to the 3D nature of such semantic maps, all the data need to be rendered before it can be employed in 6D SLAM computations.

Symbolic level robot planning often relies on such rendered semantic maps and takes advantage of the specific background knowledge embedded in them. Extracted semantic knowledge is then used for reasoning about objects or object classes present in the map. As reported in [14] using semantically labeled points results in a speedup with no loss of quality in computing time for matching of two 3D scans.

V. Conclusion

In this work, we discussed various approaches for autonomous agent localization identifying their advantages and drawbacks. After a concise review of the existing methods, we drew reader’s attention to the benefit of using semantic surfaces consisting of embedded marking with predefined links to specific semantic nodes. Semantic surfaces are especially useful for indoor and small scale navigation where the other methods have some deficiencies.

An experimental software system for environmental coding generation, analysis and processing has been developed providing flexibility and allowing creation of a vast variety of environmental encoding patterns some of which are close to artworks and applicable as decorative wallpaper patterns.

Our intention is using the system for generating semantic surface environments which will be employed in real applications with autonomous agents. In particular, our plans for further work include carrying out thorough experiments to investigate how the recognition will be influenced in an environment where the semantic surfaces will be subjected to wear and tear.

Acknowledgment

This work was funded in part by Cooperative Research Projects at Research Institute of Electronics, Shizuoka University and Grant-in-Aid for Scientific Research (B) 20300269.

References

[1] Barneva, R.P., Brimkov, V.E., and Kanev, K., Theoretical issues of cluster pattern interfaces, Combinatorial Image Analysis, Lecture Notes in Computer Science, No 5852, 2009, Springer-Verlag Berlin Heidelberg, pp. 302-315.

[2] Hiyama, A., Saito, S., Tanikawa, T., and Hirose, M., Design flexibility in seamless coded pattern for localization, In Proc. of the 2007 ACM Symp. on Virtual Reality Software and Technology VRST 2007, Newport Beach, California, USA, November 5-7, 2007, pp. 219-220.

[3] Kanev, K., Gnatyuk, P., and Gnatyuk, V., Laser marking in digital encoding of surfaces, Advanced Materials Research, Vol. 222, 2011, pp.78-81.

[4] Kanev, K., Kato, H., and Koroutchev, K., Encoding of surfaces for global positioning and navigation, The Journal of Three Dimensional Images, Vol. 24, No. 3, 2010, pp. 51-57.

[5] Kanev, K. and Kimura, S., Clustering-scheme-encoded interfaces providing orientation feedback, US Patent No 7991191, 2011.

[6] Kanev, K. and Kimura, S., Digital information carrier, JP Patent No 4368373, 2009.

[7] Kanev, K., Mirenkov, N., Brimkov, B., and Dimitrov, K., Semantic surfaces for business applications, Int. Conf. on Software, Services and Semantic Technologies, Sofia, Bulgaria, 2009, pp. 36-43.

[8] Kanev, K. and Orr, T., Enhancing paper documents with direct access to multimedia for more intelligent support of reading, In Proc. of the IEEE Conf. on the Convergence of Technology and Professional Communication, Saratoga Springs, New York, USA, 23-25 Oct. 2006, pp. 84-91.

[9] Milford, M.J., Robot navigation from nature, STAR 41, pp. 1–7.

[10] Montemerlo, M. and Thrun, S., FastSLAM: a scalable method for the simultaneous localization and mapping problem in robotics, Springer, Berlin, Germany, 2007.

[11] Montermerlo, M., Thrun, S., Koller, D., and Wegbreit, B., FastSLAM: An improved particle filtering algorithm for simultaneous localization and mapping that provably converges, In Proc. of IJCA 2003.

[12] Nüchter, A., 3D robotic map: The simultaneous localization and mapping problem, STAR 52, pp.9-27, 2009, pp. 109-172.

[13] Nüchter, A. and Hertzberg, J., Towards semantic maps for mobile robots, Robotics and Autonomous Systems, 56, 2009, pp. 915-926.

[14] Nüchter, A., Wulf, O., Lingemann, K., Hertzberg, J. Wagner, B., and Surmann., H. RoboCup 2005: Robot Soccer World Cup IX 3D Mapping with Semantic Knowledge. Lecture notes in computer science (0302-9743), 4020, p. 335.

[15] O’Gorman, L., Sammon, M.J., Seul, M., Practical algorithms for image analysis: descriptions, examples, programs, and projects, 2nd ed., Cambridge University Press, New York, USA, 2008.

Policy-Based Autonomic Collaboration for Cloud Management

Omid Mola
Department of Computer Science
University of Western Ontario
London, Ontario, Canada
omola@uwo.ca

Mike Bauer
Department of Computer Science
University of Western Ontario
London, Ontario, Canada
bauer@uwo.ca

***Abstract*—The management of clouds comprised of hundreds of hosts and virtual machines present challenging problems to administrators in ensuring that performance agreements are met and that resources are efficiently utilized. Automated approaches can help in managing such environments. Autonomic managers using policy-based management can provide a useful approach to such automation. We describe different elements of a cloud system and outline how collections of collaborating autonomic managers in cloud can be a step towards better management of clouds. We also give formal definition of different elements in the managed system and show a summary of implementation results.**

***Keywords*-Autonomic Management; Collaboration; Policy-Based Management; Cloud Management.**

I. Introduction

Cloud computing environments often depend on virtualization technology where client applications can run on separate operating virtual machines (VMs), particularly for providers of Infrastructure as a Service (IaaS). Such environments can consist of many different host computers each of which might run multiple VMs. As the number of hosts, virtual machines and client applications grow, management of the environment becomes much more complicated. The cloud provider must worry about ensuring that client service level agreements (SLA) are met, must be concerned about minimizing the hosts involved, and minimizing power consumption. Our focus is on how to better manage the virtual machine and system infrastructure of the cloud provider.

In recent years, there has been a lot of research into Autonomic Computing [1], especially about how to build autonomic elements and managers [2]. Autonomic managers try to monitor and manage resources in real time by building systems that are self-configuring, self-optimizing, self-healing and self-protecting. In the broader vision of autonomic computing, large complex systems will consist of numerous autonomic managers handling systems, applications and collections of services [3]. Some of the systems and applications will come bundled with their own autonomic managers, designed to ensure the self-properties of particular components. Other managers will be part of the general management of the computing environment. The complexity of managing a large system will entail a number of different autonomic managers which must cooperate in order to achieve the overall objectives set for the computing environment and its constituents. However, the relationships between these managers and how they cooperate introduce new challenges that need to be addressed.

We consider the use of policy-based managers in addressing this problem. Our initial focus is on a hierarchy of autonomic managers where policies are used at each level to help managers decide when and how to communicate with each other as well as using polices to provide operational requirements. The ultimate goal is to automatically monitor and manage a larger system by a collective of collaborating local autonomic managers (AMs). In such an environment, we assume that each local AM has its own set of policies and is trying to optimize the behaviour of its local elements by responding to the changes in the behaviour of those elements.

We assume some managers will also be expected to monitor multiple systems and directly or indirectly to monitor other local AMs. We also assume that one of the roles of a higher level manager is to aid other AMs when their own actions are not satisfactory.

The focus of this paper is on collaboration and communication between different managers at different levels of the hierarchy based on the active policies. The core issue addressed is how these local managers should communicate with each other and what information they have to exchange to achieve global performance goals. Finally, we will focus on how to automate the collaboration process itself.

In the rest of this paper, Section II explains the related works, Section III focuses on explaining the cloud architecture and challenges that are being addressed, Section IV gives the formal definitions of our approach towards the problem and finally, we conclude with the future works in Section V.

II. Related Work

Some researchers have already begun to study how the collaboration or cooperation among local autonomic managers can be done in order to achieve a global goal.

A hierarchical communication model for autonomic managers has been used by some researchers. Famaey and

Latre [4] used a policy-based hierarchical model to show how it can be mapped to the physical infrastructure of an organization and how this hierarchy can dynamically change by splitting and/or combining nodes to preserve scalability. They also introduced the notion of context that needs to be accessible in the hierarchy, but do not describe in detail what this context should be and how it should be communicated. In this paper, we focus on what this context should be, how it can be transferred from one manager to the other and when this should happen.

Aldinucci, et al. [5] described a hierarchy of managers dealing with a single concern (QoS). They introduce three types of relationship between components but do not explore the details of how and when such components should interact in actual systems. They used a simulator to evaluate the framework and their main focus was on the concept of a behavioral skeleton where they used autonomic management for skeleton-based parallel programs.

Mukherjee, et al. [6] used a flat coordination of three managers working on three different parts of a system (Power Management, Job Management, Cooling Management) to prevent a data center from going to the critical state. They showed how the three managers can cooperate with each other to keep the data center temperature within a certain limit that is suitable for serving the current workload and at the same time not using more power than required. They showed how these three managers should cooperate based on different business policies.

However, these three managers are fixed and adding new managers to this system will be challenging both in terms of collaboration and scalability. The same approach as in [6] is used in [7], [8] to show the collaboration between a power and a performance manager (only two managers) to minimize the power usage as well as maximizing the performance. This method however does not seem to be generalizable to a larger environment with more autonomic managers involved because of the complexity introduced in terms of interacttions between managers.

Schaeffer-Filho, et al. [9], [10] have introduced the interaction between Self-Managed Cells (SMCs) that was used in building pervasive health care systems. They proposed Role based interactions with a Mission that needs to be accomplished during an interaction based on predefined customized interfaces for each role. This approach is very general and does not address the details of the interactions. In the work presented in this paper, we will address what the policies look like and what specific information needs to be exchanged.

Zhu, et al. [11] has introduced an integrated approach for resource management in virtualized data centres. Their approach is similar to the hierarchical approach we used in our work but the relationship between different controllers are tightly coupled whereas we suggest a loosely coupled communication style to better accommodate failures and heterogeneous autonomic managers. The focus of our work is on policies and how they affect the relationship between managers, but it's not clear how they use policies and if there is any effect on controller's communications.

III. Cloud Management Challenges

In order to describe the challenges, we first explain the cloud architecture.

A. Cloud Architecture

The infrastructure of IaaS providers, is typically composed of data centers with thousands of physical machines organized in multiple groups or clusters. Each physical machine runs several virtual machines and the resources of that server are shared among the hosted virtual machines. Therefore, there are a large number of virtual machines that are executing the applications and services of different customers with different service level requirements (via Service Level Agreement (SLA) parameters).

To have a better understanding of cloud provider environment and architecture, we take a closer look at Eucalyptus [12] (an open-source infrastructure for the implementation of cloud computing on computer clusters). In Eucalyptus, there are three main elements that form the cloud infrastructure in a hierarchical fashion:

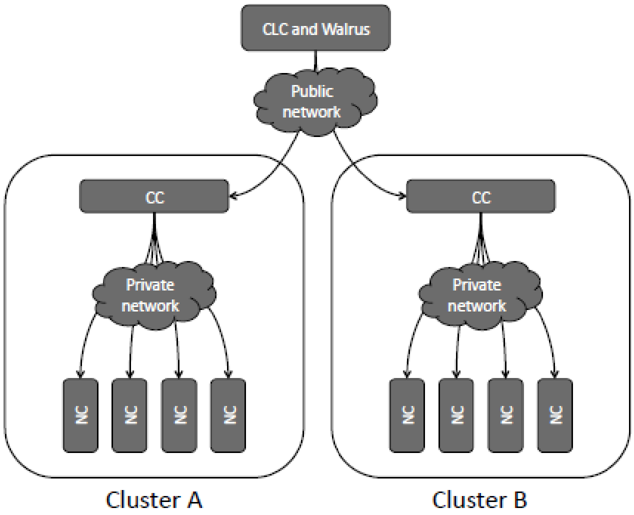


Figure 1. Eucalyptus Hierarchical Design (from [12])

- Cloud Controller (CLC): The CLC is the top level component for interacting with users and getting the requests. The CLC then talks with the Cluster Controllers (CC) and makes the top level choices for allocating new instances of virtual machines.
- Cluster Controller (CC): The CC decides which Node Controller will run the VM instance. This decision is based upon status reports which the Cluster Controller

receives from each of the Node Controllers. CC has three primary functions: schedule incoming instance run requests to specific NCs, control the instance virtual network overlay, and gather/report information about a set of NCs.
- Node Controller (NC): The NC runs on the physical machine responsible for running VMs and the main role of the NC is to interact with the OS and hypervisor running on the node to start, stop, deploy and destroy the VM instances. An NC makes queries to discover the nodes physical resources the number of cores, the size of memory, the available disk space as well as to learn about the state of VM instances on the node. The information thus collected is propagated up to the Cluster Controller in responses to describeResource and describeInstances requests.

B. Challenges

All of the specified elements in the cloud architecture are needed for instantiation of new images or destroying currently deployed VMs and they have some minimal management capabilities. However, the main challenges in managing the cloud environment occur after the VMs start working and receiving loads:

- How should the system respond to the load changes inside one or more virtual machines?
- What should happen to maximize the performance of a specific virtual machine (or an application inside it) according to the agreed SLA?
- How can we scale the system up and down on the fly (change VM parameters)?
- How can one enforce specific operational policies for the entire system?
- How can one make sure that minimum resources are used to perform a task (e.g. minimizing the power usage)?

A deeper look at the cloud architecture and the management needs suggest that providing all these capabilities in real time through a single centralized manager is almost impossible, because of the hierarchical layers in the architecture with different responsibilities at each layer. Also, the dynamics of load change and the need to react to these changes in real time with increasing number of VMs and physical nodes makes it much more difficult to achieve these goals with a traditional centralized manager.

Therefore, a hierarchical approach towards cloud management would be a more efficient way to achieve all of the goals. At the same time, each element in the management hierarchy should act autonomously and manage part of the hierarchy on its own.

IV. Approach and Definitions

Based on the previous discussions, we propose to use a number of different autonomic managers. By using this approach, the problem of managing a large system entails a number of autonomic managers where each one is dealing with smaller or more localized components, and then each manager's job is to focus on managing that component (or small set of components) efficiently based on certain policies.

For example, an AM for an Apache web server should only focus on the behavior of the web server and not the relationship that the webApp might have with a database server or, a Node Controller (NC) AM should only focus on the general performance and the behavior of the VMs inside that specific node.

The hierarchy of autonomic managers might appear as in Figure 2. In the lowest level, the AMs are managing the applications inside the VMs. The AMs at the node controller (NC) level monitor and manage the VMs. Then the AMs at cluster controller (CC) level are responsible for all physical nodes inside that cluster. Similarly the AM at cloud controller (CLC) level monitors and manages all of the clusters.

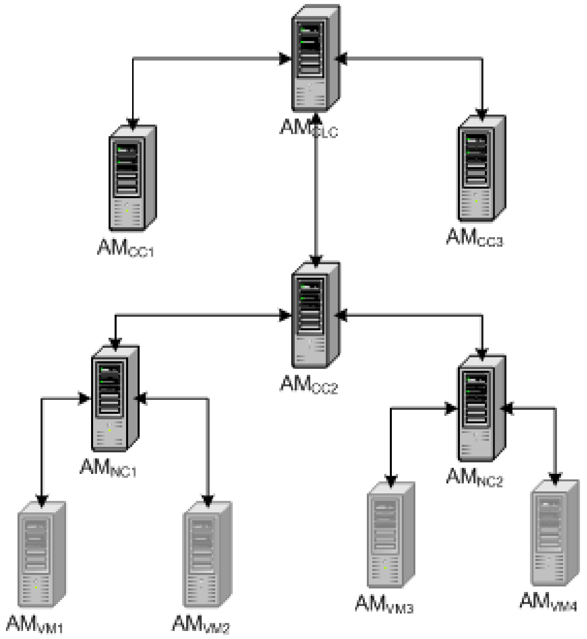


Figure 2. AMs hierarchy based on the cloud architecture

Note that this is a logical organization of autonomic managers and does not necessarily reflect the physical allocation of the AMs, i.e., they do not need to be located on different physical machines. In a large cloud computing provider they could be located on separate machines or some may be located on the same machines. These AMs should then collectively work together to preserve a set of policies for optimizing performance, minimizing resource usage, avoiding SLA violations, etc.

Assuming that the management tasks are specified in terms of policies, this means that we need policies with different granularity deployed at different levels of the infrastructure and we need to ensure that AMs can communicate properly with each other to enforce those policies.

A. Managed System

Our managed system is composed of a set of elements that can be monitored and managed automatically. Each autonomic manager is typically monitoring and managing one or more managed elements(ME). The managed elements will be equivalent to what is found in ordinary cloud infrastructures such as a virtual machine, a physical node, a software resource, or a cluster.

We can define the charactestic and operations of each ME in a definition document called ”ManagedObjectType” which can later be instantiated several times. For example, for modeling a virtual machine ME we can put all properties and metrics of a general VM in the VMManagedObjectType and later we can instantiate two objects of this type called vm1 and vm2.

Other possible managed objects types are: ApacheManagedObjectType, NodeManagedObjectType, ClusterManagedObjectType.

Definition 1. *MOt=*$\langle P, M, A\rangle$ *A ManagedObjectType is a tuple* $\langle P, M, A\rangle$*, where:*

- *P is a finite set of properties, P=*$\{P_1, ..., P_k\}$*,*
- *M is the finite set of metrics, M=*$\{M_1, ..., M_l\}$*, where:* $\forall M_i \in M, M_i = \langle N_i, AC_i\rangle \mid N_i = MetricName, AC_i \in A = RefreshingAction$
- *A is the finite set of actions, A=*$\{A_1, ..., A_m\}$*.*

We denote the set of managed object types by MOT=$\{MOt_1, ..., MOt_n\}$

Actions are operations that can be done on that managed object. For example, actions for a VMManagedObjectType could be startVM(), stopVM(), getVMIP(), refreshCPUUtil(), etc.

Properties of a managed object type are set upon instantiating a new managed object. Examples of properties for VMManagedObject are vmName, vmAllocatedMemory and vmOSType.

The metrics associated with a managed object are those properties that change more often and therefore must include actions specifying how they can be updated/refreshed (e.g. by connecting to another AM and sending a message to get the updated values). Examples of these metrics along with their associated actions are CPUUtil, refreshCPUUtil(), or MemoryUtil, refreshMemoryUtil(), etc.

The actions, metrics and properties defined inside managed objects types can later be used in policies to evaluate a specific condition or to perform an action on that managed object.

Therefore, based on this definition, we can now instantiate several managed objects from a single type. For example, vm1 managed object can be instantiated from VMManagedObjectType, etc.

Definition 2. *Given a set of MOT, a ManagedObject (MO) is a tuple* $\langle p, m, a\rangle$ *where there is a MOt=*$\langle P, M, A\rangle$ *such that*

- *a=A,*
- $p = \{\langle P_1, v_1\rangle, ..., \langle P_k, v_k\rangle\} \mid P = \{P_1, ..., P_k\}$ *and* v_i *is value of the property.*
- $m = \{\langle N_1, V_1, AC_1\rangle, ..., \langle N_l, V_l, AC_l\rangle\} \mid M = \{\langle N_1, AC_1\rangle, ..., \langle N_l, AC_l\rangle\}$ *and* V_i *is the measured value of a metric.*

We denote the set of managed object by MO=$\{MO_1, ..., MO_n\}$

In the rest of this document, whenever we use term managed object, we use this definition.

We assume that inside each AM there is an event handling mechanism for generating events and notifying the interested parties inside the AM. For example, there could be an event bus and different subscribers to certain events (within the AM) and upon raising those events any subscribers will get notified. This event handling mechanism is useful for handling event, condition, action policies and also for communication between managers (both explained later). We assume that for a given system and managed objects, that there are a finite number of event types.

Definition 3. *An event type, Et is a pair* $\langle N, M\rangle$ *where: N is the name of the event type, M=*$\{m_1, ..., m_o\}$*, and* m_i *is the name of a metric from a managed object. We denote the set of event types by ET =*$\{Et_1, ..., Et_o\}$*.*

Definition 4. *Given a set of ET, an event E is a pair* $\langle n, m\rangle$ *where there is an event type Et =* $\langle N, M\rangle$*, n is the name of the event n = N, m =*$\{\langle m_1, v_1\rangle, ..., \langle m_o, v_o\rangle\}$*, where M =*$\{m_1, ..., m_o\}$*, and* v_i *is its value. We denote the set of events by Eve =*$\{E_1, ..., E_o\}$*.*

For a given set of event types, there may be an infinite number of possible events, depending on the value associated with the metrics of that event type. In this respect, an event is an instantiation of an event type with the associated metrics assigned values.

B. Policies

All of the policies expressed as event, condition, action (ECA) policies. In general, all of our policies are of the form:

```
On event: E
if ( Set of Conditions ) then {
    Ordered Set of Actions
}
```

Upon raising an event inside the autonomic manager, then any policy which matches the event will get evaluated. If the conditions in the policy are met, then the policy actions get triggered. We provide examples of policies in the following sections.

Definition 5. *A policy is a tuple $\langle N, E, C, A\rangle$ where N is the policy name, E∈ Eve is one of the events defined for the manager, C is a finite set of conditions, and A is an ordered set of actions defined in MO actions. Each condition, is defined by a tuple $\langle MName, Operator, T\rangle$, where MName is the metric name defined in a MO metrics, Operator is a relational operator and T is a constant indicating a threshold value. Therefore, Pl=$\langle N, E, C, A\rangle$, where:*

- *E∈ Eve,*
- *C=$\{C_1, ...C_p\}$ and C_i=$\langle MName, Operator, T\rangle$ or "true",*
- *A=$\{A_1, ..., A_q\}$, $\forall A_i \in A \exists MO_j \in MO \mid A_i = MO_j.AC_k$*

We denote the set of policies by PLS=$\{Pl_1, ..., Pl_r\}$.

A sample expectation policy for monitoring the Apache response time is:

$Pl_1 = \langle$ "apacheRTPolicy", ManagementIntervalEvent,
apache.responseTime $>$ 500,
apache.increaseMaxClients(+25, 200) $\rangle$

In this policy, ManagedIntervalEvent is an event that gets triggered in a certain time interval (e.g. 1500ms) and it has no metrics associated with it. "apache" is an instance of ApacheManagedObjectType and responseTime is one of the metrics defined in ApacheManagedObjectType. "increaseMaxClients(value, max)" is one of the actions defined in ApacheManagedObjectType and will increase the max client property of the apache web serverby a certain number up to a max (e.g., will not increase it more than 200).

At AM startup there are configuration policies that set up the AM environment, identify the appropriate managed objects and configure them. A sample configuration policy would look like:

$Pl_2 = \langle$ "StartUpConfPolicy", StartUpEvent, true,
{ system.setFatherIP("192.168.31.1"),
system.create(vm1, VMManagedObjectType),
vm1.setIP("192.168.31.3") } $\rangle$

This policy happens on AM startup and configures the parents IP of this AM in the hierarchy and also adds one ManagedObject for managing vm1 (This is happening in AM_{NC1} - see Figure 2). This AM will be responsible for managing physical node 1 which hosts vm1 and will communicate with the manager inside vm1 if necessary. The AM hierarchy can be built this way upon system startup but it can change dynamically throughout their lifetime (e.g. by migration of a VM to another machine). In this example, "system" is an instance of SystemManagedObbjectType which is useful for configuration and management of the manager itself.

C. Structural Relationship Between AMs

In order to explain the relationship between AMs in this system we first need to define the AM itself.

Definition 6. *An Autonomic Manager(AM) is a tuple $\langle MO, Eve, Pol, RI, MI\rangle$ where MO is a finite set of managed objects, Eve is a finite set of events, Pol is a finite set of policies, RI is the refresh interval which determine the time interval for updating the managed objects metrics and MI is the management interval, which determine the time interval for enforcing active policies. These two thresholds can be configured for each AM. We denote the set of AMs by AMS=$\{AM_1, ..., AM_t\}$*

Based on the cloud architecture, we assume AMs are organized in a hierarchical manner to reflect different authority levels in cloud. So, the structural relationship between AMs consists a tree.

Definition 7. *The hierarchy of AMs is a tuple $\langle AMS, Edges\rangle$ where AMS is the set of autonomic managers as the nodes of the tree and Edges=$\{(AM_i, AM_j)|AM_i, AM_j \in AMS\}$ is the set of edges connecting two AMs to each other. The following properties exist in this hierarchy:*

- $\exists AM \in AMS | \nexists AM_i \in AMS, (AM_i, AM)$
- $if(AM_i, AM_j) \in Edges \Rightarrow \nexists AM_k | (AM_k, AM_j) \in Edges$
- $if(AM_i, AM_j) \in Edges \Rightarrow (AM_j, AM_i) \notin Edges$

D. Communication Model

Each manager should be able to receive messages from other managers or send messages to other managers. In previous work [13], [14], we suggested the use a message-based type of communication between AMs. Three different types of messages (NOTIFY, UPDATE_REQ, INFO) were proposed as sufficient for communication between managers.

Since we are dealing with a hierarchy of managers then each manager needs to communicate with either its father or its children. However, it is also possible for an AM to send NOTIFY messages to another AM in some other part of the hierarchy based on a request.

The UPDATE_REQ message is sent from higher level managers to lower level ones. INFO messages are sent in response to the UPDATE_REQ message and NOTIFY messages are sent from one manager to another based on the need. In the previous work [14] we have shown how one can use policies to generate these messages for communication among AMs based on demand.

V. Conclusion and Future Works

Based on the previous discussions, we have introduced an automated collaborative approach towards management of a cloud infrastructure. So far, we have implemented the hierarchy of autonomic managers and did some experiments

Table I
RESULTS OF THREE SCENARIOS

Scenario	SLA Violation(%)
1: No collaboration between AMs	72
2: One-Level collaboration in the hierarchy	42
3: Two-levels collaboration in the hierarchy	24

which confirmed the importance of collaboration between AMs at different layers of the cloud. The complete results can be found in [14], but Table I shows the summary of three scenarios with respect to SLA violation rate.

The main contribution of this paper compared to our previous work is to give formal definition of the managed system and autonomic managers which lead to a better understanding of the problem and developing precise algorithms. The ultimate goal is however to design algorithms that can get the system information (e.g., events, policies and ManagedObjects) and generate the required communication messages automatically. Therefore, the collaboration between AMs will become more automated itself. In this work, we assumed that policies are defined and delivered to managers by system administrators, but as a future work we are planning to make this process more automated.

The next step would then be moving towards developing and evaluating these algorithms, enabling more efficient use of the cloud infrastructure as well as meeting SLA requirements while using fewer resources.

REFERENCES

[1] M. C. Huebscher and J. a. McCann, “A survey of autonomic computingdegrees, models, and applications,” *ACM Computing Surveys*, vol. 40, no. 3, pp. 1–28, Aug. 2008.

[2] J. Kephart, “Research challenges of autonomic computing,” *Proceedings. 27th International Conference on Software Engineering, 2005. ICSE 2005.*, pp. 15–22, 2005.

[3] J. Kephart and D. Chess, “The vision of autonomic computing,” *Computer*, vol. 36, no. 1, pp. 41–50, Jan. 2003.

[4] J. Famaey, S. Latrea, J. Strassner, and F. De Turck, “A hierarchical approach to autonomic network management,” *2010 IEEE/IFIP Network Operations and Management Symposium Workshops*, pp. 225–232, 2010.

[5] M. Aldinucci, M. Danelutto, and P. Kilpatrick, “Towards hierarchical management of autonomic components: a case study,” in *Parallel, Distributed and Network-based Processing, 2009 17th Euromicro International Conference on.* IEEE, 2009, pp. 3–10.

[6] T. Mukherjee, A. Banerjee, G. Varsamopoulos, and S. K. Gupta, “Model-driven coordinated management of data centers,” *Computer Networks*, vol. 54, no. 16, pp. 2869–2886, Nov. 2010.

[7] J. O. Kephart, H. Chan, R. Das, D. W. Levine, G. Tesauro, and F. R. A. C. Lefurgy, “Coordinating multiple autonomic managers to achieve specified power-performance tradeoffs,” *in IEEE Intl. Conf. on Autonomic Computing, Jun*, pp. 145–154, 2006.

[8] M. Steinder, I. Whalley, J. E. Hanson, and J. O. Kephart, “Coordinated management of power usage and runtime performance,” in *NOMS 2008 - 2008 IEEE Network Operations and Management Symposium.* IEEE, 2008, pp. 387–394.

[9] A. Schaeffer-Filho, E. Lupu, N. Dulay, S. L. Keoh, K. Twidle, M. Sloman, S. Heeps, S. Strowes, and J. Sventek, “Towards Supporting Interactions between Self-Managed Cells,” in *First International Conference on Self-Adaptive and Self-Organizing Systems (SASO 2007)*, no. Saso. IEEE, Jul. 2007, pp. 224–236.

[10] A. Schaeffer-Filho, E. Lupu, and M. Sloman, “Realising management and composition of self-managed cells in pervasive healthcare,” in *Pervasive Computing Technologies for Healthcare, 2009. PervasiveHealth 2009. 3rd International Conference on.* IEEE, 2009, pp. 1–8.

[11] X. Zhu, D. Young, B. J. Watson, Z. Wang, J. Rolia, S. Singhal, B. McKee, C. Hyser, D. Gmach, R. Gardner, T. Christian, and L. Cherkasova, “1000 Islands: an Integrated Approach To Resource Management for Virtualized Data Centers,” *Cluster Computing*, vol. 12, no. 1, pp. 45–57, Nov. 2008.

[12] D. Nurmi, R. Wolski, C. Grzegorczyk, G. Obertelli, S. Soman, L. Youseff, and D. Zagorodnov, “The Eucalyptus Open-Source Cloud-Computing System,” *2009 9th IEEE/ACM International Symposium on Cluster Computing and the Grid*, pp. 124–131, 2009.

[13] O. Mola and M. Bauer, “Collaborative policy-based autonomic management: In a hierarchical model,” *Network and Service Management (CNSM), 2011 7th International Conference on*, pp. 1–5, 2011.

[14] O. Mola and Michael A. Bauer, “Towards Cloud Management by Autonomic Manager Collaboration,” *Int’l J. of Communications, Network and System Sciences*, vol. 04, no. 12, pp. 790–802, 2011.

Mobile Custom-made Handheld Chemical Detection Device

Iem Heng
Computer Engineering Technology Department
New York City College of Technology (NYCCT)
Brooklyn, NY, USA
iheng@citytech.cuny.edu

Andy S. Zhang
Mechanical Engineering Technology Department
New York City College of Technology (NYCCT)
Brooklyn, NY, USA
azhang@citytech.cuny.edu

Raymond Yap
Computer Engineering Technology Department
New York City College of Technology (NYCCT)
Brooklyn, NY, USA
raymond.yap@hotmail.com

***Abstract*— In today's society, there is a challenge to detect and to avoid exposure to harmful and lethal chemicals. This remains an issue to public health and has not been addressed adequately. To address this challenging issue, the objective of this paper is to provide a unique perspective in designing and building a custom-made mobile handheld chemical detection (CHCD) device that can detect harmful and lethal chemical gases (NO_2, N_2, CO, CO_2, LPG, CH_4, CNG, C_2H_5OH, NH_3, H_2 and others) in public and in private gathering places. This mobile handheld device can relay the information of the chemical concentration levels detected to smartphone or tablet or laptop in any place at any time. Applications of this useful mobile CHCD prototyping device include detection of harmful gases in public and in private gathering places such as subway stations, shopping malls, airports, and residential houses. Additionally, this mobile CHCD device provides an alternative and affordable resource for people to have and use it as an advance warning system within the proximity of dangerous areas.**

Keywords- chemical detection; mobile communication; remote sensing.

I. INTRODUCTION

Over the years, chemical gases in the U.S. and around the world have caused and taken many innocent lives that could have been prevented. Of primary concern are the human health effects of chemical gas, including premature mortality and chronic illnesses such as bronchitis and asthma. Despite the tremendous economic costs and pervasive negative health impacts of bad chemical gas – chemical gas often goes unnoticed because it is largely invisible. Much of what happens in our immediate environment passes without being noticed by the public despite the fact that there are recording and crowd-sourcing devices installed in some neighborhood that monitor the air quality. A mobile CHCD device captures a spectrum of that lost reality and returns it to the users in real-time as the events unfold. By making these specific environmental events available to participants in real time and location, the CHCD supplements the qualitative information reported by government agency with quantitative information obtained from handheld sensing device that observes and records aspects of the environments that are either impossible to perceive directly (e.g., pollutant gas concentrations) or difficult to quantify and communicate in a consistent manner. A mobile CHCD device allows individuals to broadcast what is happening with their environment, crowdsource their own information with that from other participants, and identify patterns and commonalities. Thus, this mobile CHCD device makes the detection of chemical gas possible by concern citizens, thereby empowering communities to advocate for healthy environments.

Unlike the current commercial chemical detection devices, the mobile CHCD device is a unique and novel device in term of miniature size, provides an advance communication warning system accessed by other smart devices (smartphone, tablet, and laptop), and provides an affordable low-cost detection device for consumers. For instance, comparing the commercial hazardous vapor warning LCD 3.3 [1] and Nose Gas Sensor [2] devices to the mobile CHCD device, the LCD 3.3 and Nose Gas Sensor devices are designed as one unit with LCD screen used for displaying the gas concentration levels and are not capable of communicating with other smart devices. Additionally, both devices are not small and not cost effective for general consumers. Furthermore, many other commercial detection devices have similar features as the LCD 3.3 and Nose Gas Sensor. This is why the mobile CHCD device is a unique and novel detection device for an advance warning system to the public.

In this paper, we are going to provide a unique perspective in designing and building a CHCD device that can detect harmful and lethal chemical gases in public and in private gathering places. First, we look into how the chemical gas sensors work. Then a schematic of CHCD device is developed. Based on the hardware schematic of CHCD device, the custom-made physical prototype of this device is created. The CHCD device is capable of communicating to portable devices such as smartphone or tablet or laptop through the use of Bluetooth technology. We then have done several tests (indoor and outdoor scenarios) of mobile CHCD device with those portable devices. From those tests, the raw analog signal data is acquired and is calibrated. This calibrated data is what provided for people to understand the useful benefit of mobile CHCD device.

II. Chemical gas Sensors

Currently, there are many different types of harmful chemicals and gases (NO_2, N_2, CO, CO_2, LPG, CH_4, CNG, C_2H_5OH, NH_3, and others) that can harm innocent people and could give serious negative environmental impact on the planet we live on. To prevent the loss of innocent human lives, effective detection and handheld monitoring systems need to be developed. The mobile CHCD is a state-of-the art device to detect many kinds of harmful chemical gases in the air and on the ground through the use of various chemical gas sensors. This detection device provides an advance warning system and alerts the general public through their smart devices. This could reduce human casualties, environmental destruction, and property loss.

How do the various chemical gas sensors work? Many gas sensors use a heater to detect certain gases. In general, many of these gas sensors have similar schematic diagram [3] as follows:

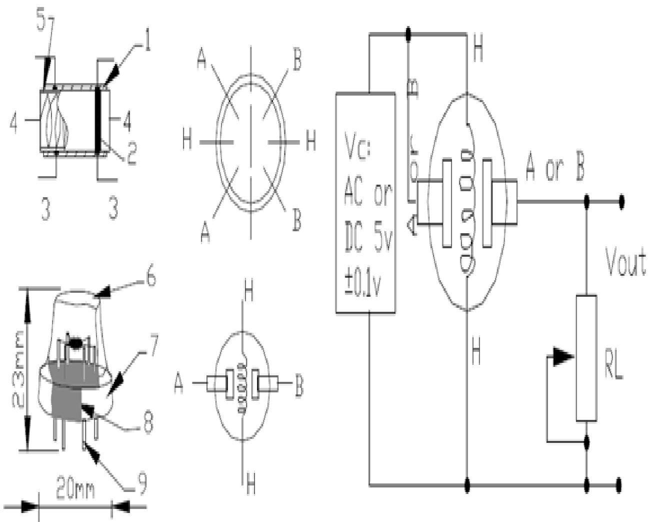


Figure 1. General Gas Schematic.

According to the chemical gas sensor schematic diagram shown in Figure 1, there are 6 pins coming out of the gas sensor itself. However, some chemical gas sensors have only 3 or 4 pins. 5 volts can supply to pins (A H A); whereas, both pins B can be used as an analog output signal. The other pin H, in between both pins B, can be used as a ground (GND) pin. This is illustrated in Figure 2 as follows:

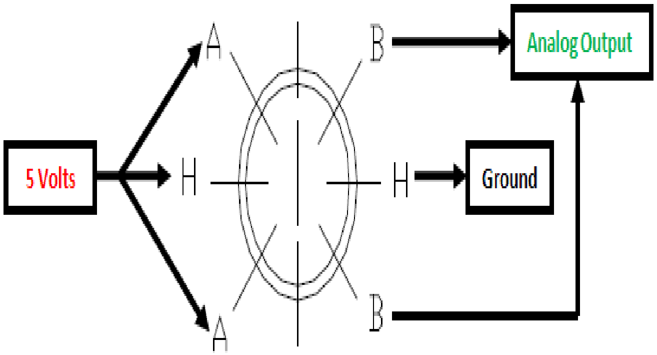


Figure 2. Gas Sensor.

Then the data from an analog output signal is incorporated into specially designed software codes capable of detecting a wide range of harmful and lethal chemical gases by using different types of gas sensors. The sensor testing and calibration of analog signal data acquisition will be fully discussed in later Section IV.

In term of the working principle for the gas sensor for CO_2, as an example, it takes on the solid electrolyte cell principle and is composed by the following solid cells: Air Au|NASICON|| carbonate|Au, air, CO_2 [4].

When the CO_2 sensor exposes to the CO_2 environment, it will have electrochemical reaction with the following equations [5]: Cathodic reaction equation:

$2Li + + CO_2 + 1/2O_2 + 2e - = Li_2CO_3$.

Anodic reaction equation: $2Na+ + 1/2O_2 + 2e- = Na_2O$.

The overall chemical reaction equation:

$Li_2CO_3 +2Na + = Na_2O +2Li + + CO_2$.

As a result of electrochemical reaction, according to Neste equation (Nernst), the process will produce the following electromotive force (EMF): $EMF = Ec - (R \times T) / (2F) \ln (P(CO_2))$, where PCO_2 is the partial pressure of CO_2; Ec is a constant; R is the gas constant; T is temperature in Kelvin; F is the Faraday constant.

From Figure 1, the sensor heating voltage supplied from other circuit. When its surface temperature is high enough, the sensor equals to a cell, its two sides would output voltage signal, and its result accord with Nernst's equation [5]. In sensor testing, the impedance of amplifier should be within 100-1000GΩ, Its testing current should be control below 1pA.

III. Mobile CHCD device

In this section, we will discuss and provide more details on how we design the prototype of Mobile Custom-made Chemical Detection (CHCD) Device.

A. Hardware Schematic of Mobile CHCD Device

The set up pins of a gas sensor in Figure 2 provide some ideas of how to hook up the hardware schematic diagram of mobile CHCD device. Using Fritzing [6] software program, the design of the overall hardware schematic for the device is shown in Figure 3.

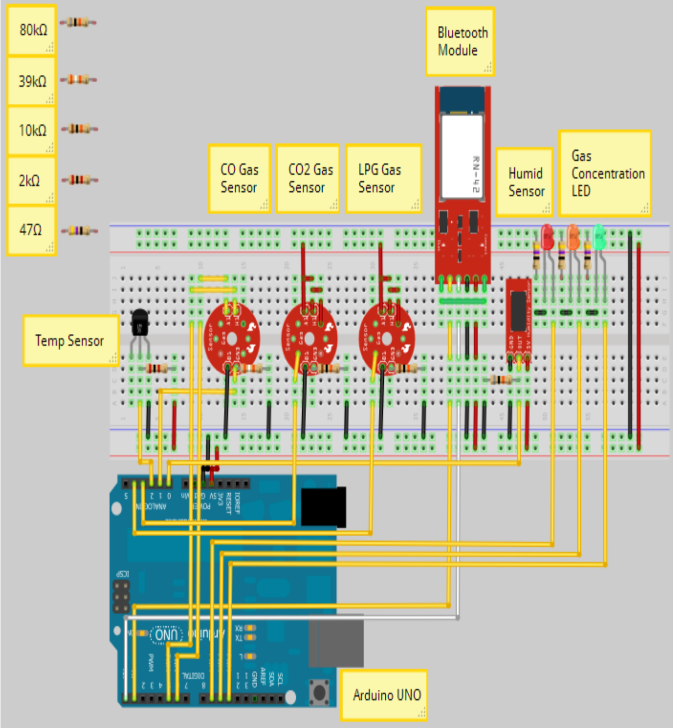


Figure 3. Schematic of CHCD Device.

From the schematic of Figure 3, the yellow, red, and black wires are used as data signal communication, voltage, and ground, respectively. The signal lines for the three sensors (CO gas, CO_2 gas and LPG gas) are connected to the Arduino microcontroller analog pins A1, A3 and A4. In addition to the three gas sensors in Figure 3, temperature and humidity sensors are added to the CHCD device for monitoring the effect of data acquisition in correlation to the three gas sensors. Whereas, the red and black lines across the breadboard in Figure 3 are connected to Arduino 5V and ground pins, respectively. The yellow and white wires from the Bluetooth module are used as the transceivers and are connected to Arduino pins TX (transmitter – yellow wire) and RX (receiver – white wire). The use of a Bluetooth module provides the wireless communication lines between CHCD device and smartphone (or tablet or laptop). Also, for instance, the three color LEDs indicate the CO gas concentration levels. Green, orange, and red color LEDs are indicated the least, medium, and highest PPM, respectively. PPM is for part per millions, and it is used to measure the concentrations of chemical gas. Hence, this makes CHCD device as a unique embedded mobile handheld device in monitoring the surrounding areas of one’s present.

B. *Prototyping of Mobile CHCD Device*

Based on the hardware schematic of CHCD device in Figure 2, the custom-made physical prototype of this device is created. The following Figure 3 shows the final prototype design of the CHCD device.

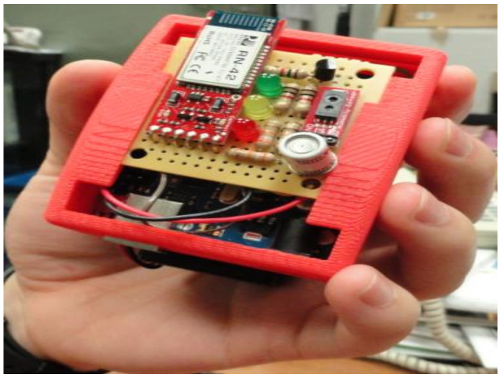


Figure 4. Prototyping of CHCD Device.

Currently, as seen in Figure 4 above, the custom-made prototype of chemical detection device has been used for testing the CO (Carbon Monoxide) concentration levels along with the temperature and humidity. This device will be expanded to include the testing of CO_2 (Carbon Dioxide), LPG (Liquefied Petroleum Gas), and among other known chemical gases.

Figure 5 shows the progress stages in designing the CHCD device leading to the final prototype.

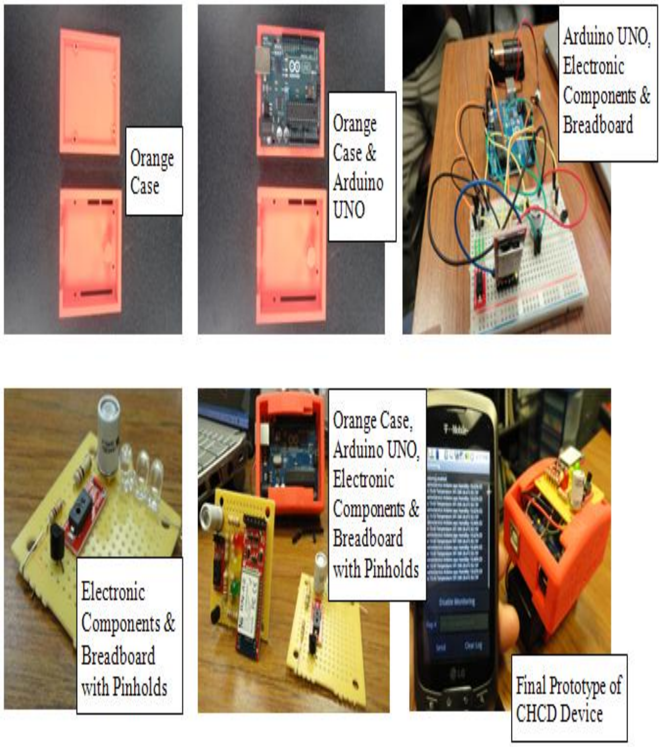

Orange Case
Orange Case & Arduino UNO
Arduino UNO, Electronic Components & Breadboard
Electronic Components & Breadboard with Pinholds
Orange Case, Arduino UNO, Electronic Components & Breadboard with Pinholds
Final Prototype of CHCD Device

Figure 5. Design Stages of CHCD Device.

The CHCD device cover (orange case) is for holding the Arduino UNO microcontroller and electronic components (sensors, resistors, LEDs, and Bluetooth) with breadboard. To accomphish this, a computer model of the CHCD device is created using Autodesk Invnentor [7] software. Figure 6 shows a computer model of the CHCD device assembly. Figure 7 is a computer rendering of the device. Then a phyical prototype was made using 3D rapid prototyping machine as shown in Figure 8. The electronic components are soldered on the back side of the breadboard. Upon completing all the electronic components soldering on the breadboard, we are able to slide the completed breadboard onto the orange case. This completes the design stages of the CHCD prototype as seen in Figure 4.

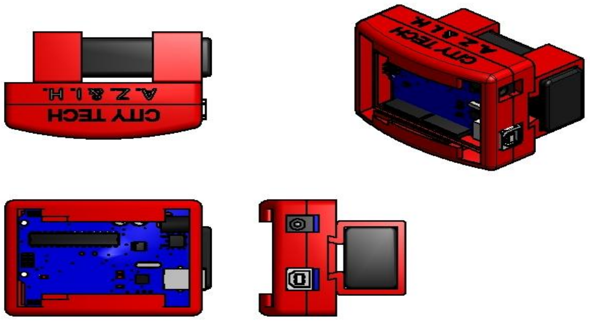


Figure 6. Assembly View of the CHCD Device.

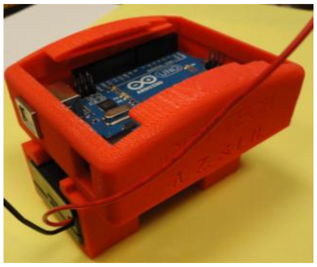


Figure 7. Computer Rendering of CHCD Device.

Figure 8. Physical Prototype of CHCD Device.

C. *Wireless Communication*

Figure 9 is a block diagram of Bluetooth wireless communication for the mobile CHCD device.

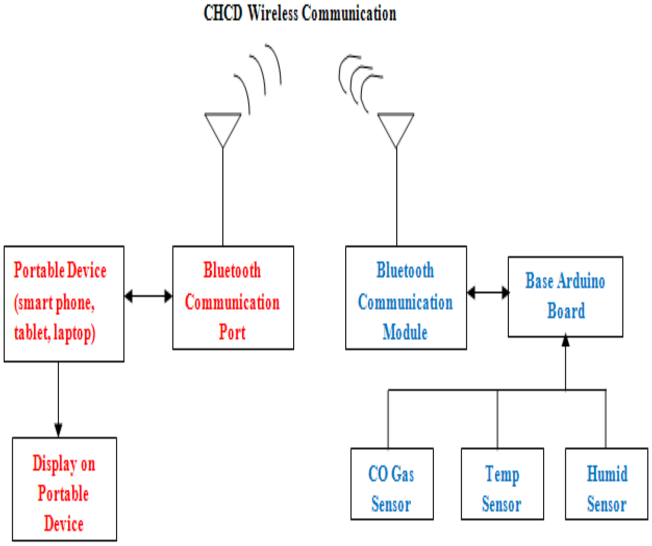

CHCD Wireless Communication
Portable Device (smart phone, tablet, laptop)
Bluetooth Communication Port
Bluetooth Communication Module
Base Arduino Board
Display on Portable Device
CO Gas Sensor
Temp Sensor
Humid Sensor

Figure 9. CHCD Wireless Communication.

Namely, the mobile CHCD device is capable of communicating to portable devices such as smartphone or tablet or laptop through the use of Bluetooth technology. Bluetooth wireless technology is based on the IEEE 802.15 standard. Bluetooth was developed to replace the cables that were connect to desktop and portable computers, mobile phones, handheld devices, computer accessories and peripheral electronic devices [8]. Thus, the use of Bluetooth wireless communication enables the users to retreat the data from the chemical gas sensor and, at the same time, display gas concentration levels on the portable devices.

To make the CHCD device to communicate and interface wirelessly with portable device (smartphone, tablet and laptop), as illustrated in Figure 9, the programming source codes must be introduced to provide access of communication and interface between devices. First, the source codes are written in Arduino sketch [9] to communicate and interface with the electronic components such as CO gas sensor, temperature sensor, humidity sensor, and LEDs, as shown in Figure 5. Upon the success of interfacing with the sensors and LEDs in Arduino sketch, the Android library MeetAndroid is import to the Arduino library folder, so that the data acquired from the Arduino Serial Monitor is sent to the Amarino Application (App) progam. The Android library MeetAndroid is part of Amarino driver device that is required to be imported in the library folder of Arduino sketch. The Amarino progam [10] is a freeware program that incorporates a plug-in mechanism which allows programmers and developers to integrate their events into Amarino. Then, this provides a gateway to communicate with smartphones and tablets based on the Android open source operating system. Figure 10 illustrates the details of communication between Arduino Sketch , Amarino App, and Android operating system.

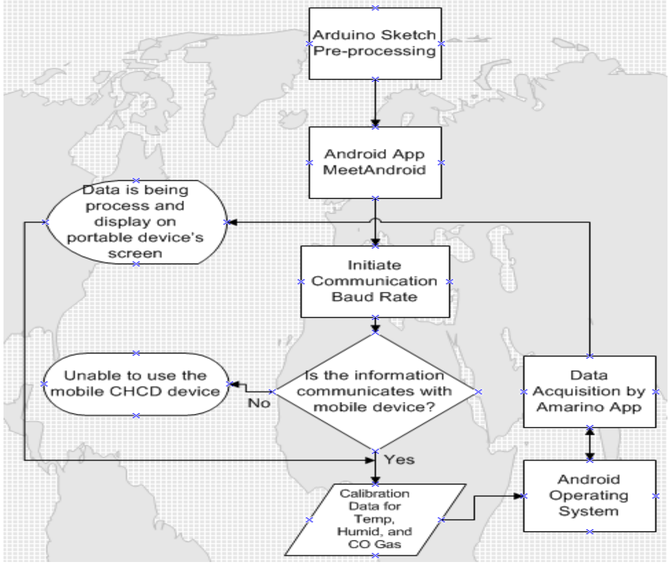

Arduino Sketch Pre-processing
Android App MeetAndroid
Data is being process and display on portable device's screen
Initiate Communication Baud Rate
Unable to use the mobile CHCD device
No
Is the information communicates with mobile device?
Yes
Data Acquisition by Amarino App
Android Operating System
Calibration Data for Temp, Humid, and CO Gas

Figure 10. Flowchart of Wireless Interface and Communication.

IV. CO GAS SENSOR CALIBRATION

In this section, the calibration of data acquisition from CO gas sensor is discussed in details. Using the CHCD device in Figure 4, the raw analog signal data is acquired from the CO gas sensor. Then the raw data must be calibrated with respect to that analog data. For instance, taking all factors such as the type of sensor and the conditions of the application into consideration, the proposed calibration procedure is based on the following block diagram:

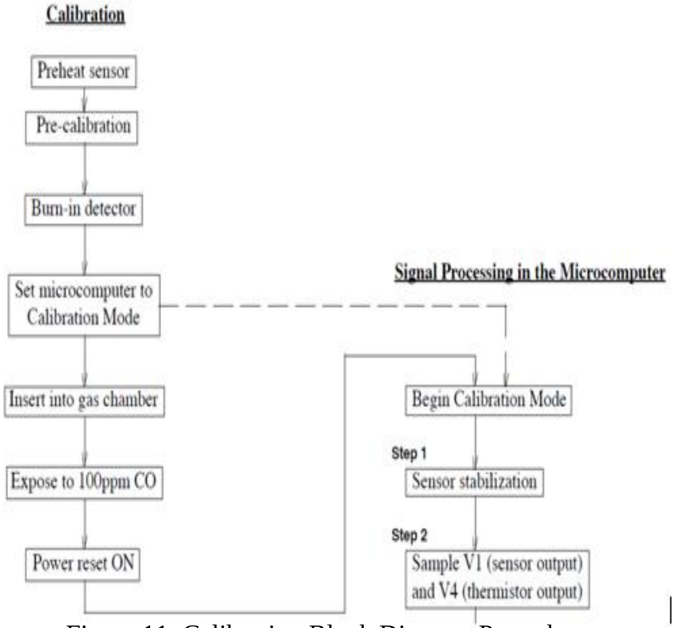
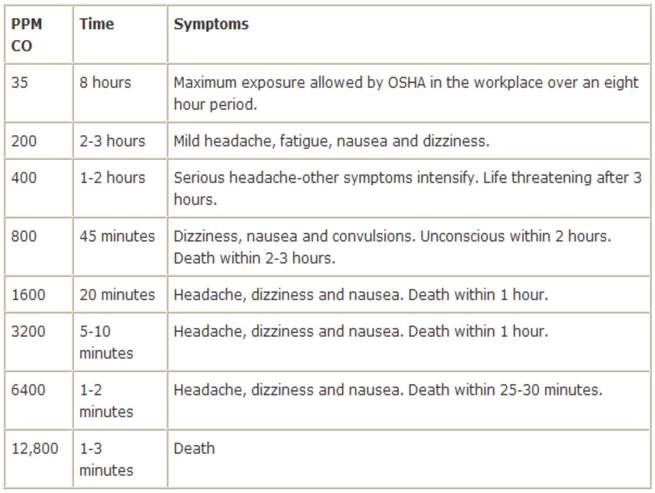


Figure 11. Calibration Block Diagram Procedures.

Based on the flowchart in Figure 11 above, 100 PPM is exposed and is used for calibration standard. Applications of how safe and unsafe of PPM for CO gas can be seen in Figure 12 below. For instance, 200 PPM would have a symptom of mild headache, fatigue, nausea and dizziness in two to three hours [11].



PPM CO	Time	Symptoms
35	8 hours	Maximum exposure allowed by OSHA in the workplace over an eight hour period.
200	2-3 hours	Mild headache, fatigue, nausea and dizziness.
400	1-2 hours	Serious headache-other symptoms intensify. Life threatening after 3 hours.
800	45 minutes	Dizziness, nausea and convulsions. Unconscious within 2 hours. Death within 2-3 hours.
1600	20 minutes	Headache, dizziness and nausea. Death within 1 hour.
3200	5-10 minutes	Headache, dizziness and nausea. Death within 1 hour.
6400	1-2 minutes	Headache, dizziness and nausea. Death within 25-30 minutes.
12,800	1-3 minutes	Death

Figure 12. Gas Concentration in PPM and Symptoms.

In addition to the block diagram in Figure 11, the gas concentration chart [12] in Figure 13 is used as part of the calibration for the CO concentration in PPM.

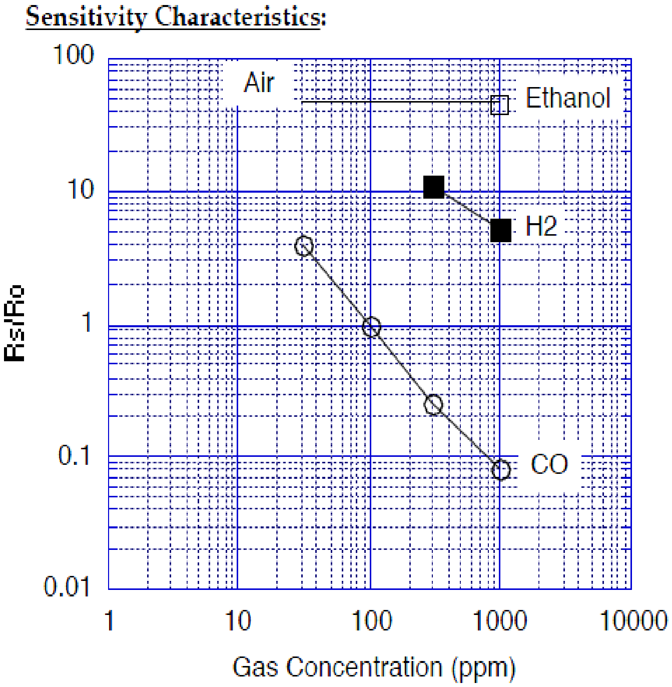


Figure 13. Gas Concentration in PPM.

Figure 13 represents typical sensitivity characteristics of CO concentration levels. The Y-axis is indicated as sensor resistance ratio (Rs/Ro) [12] which is defined as follows:

Rs = Sensor resistance of displayed gases at various concentrations and Ro = Sensor resistance in 100 PPM CO

In other way of looking at Ro is the level of expose gas to the sensor in clean air. For instance, if we pour 100 PPM gas in the container with confined space, what would Rs sensor read? It may read 98 PPM or 102 PPM.

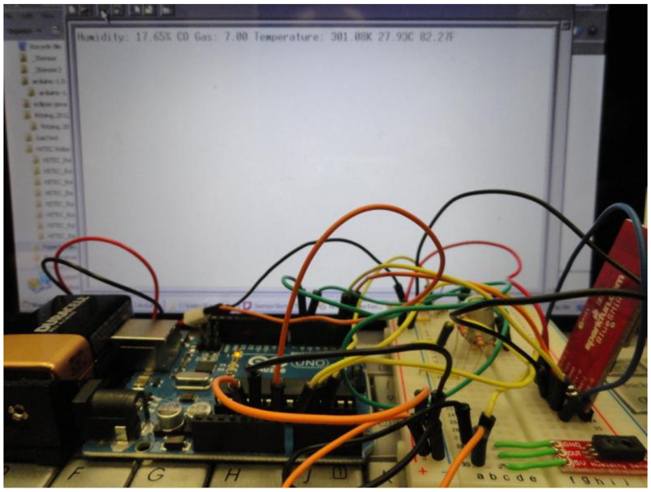
The formula [13] for defining the sensor resistance Rs is as follows:

$$Rs = \frac{Vc \ x \ R_L}{V_{out}} - R_L \qquad (1)$$

From equation (1), Vc is the voltage input, and it is 5 Volts from Arduino microcontroller embedded in CHCD device. R_L is the load resistance (in this case, we use 39kΩ) that is connected to CO gas sensor. Vout is a voltage signal from the CO gas sensor, and it varies depending on the amount of CO concentration within PPM (parts per million). Then the value of Rs in equation (1) changes according the amount of CO gas present, and as seen in Figure 13 above, the typical range for CO gas concentration is from 30 to 1000 PPM. If Rs resistance value is the same as Ro resistance value, it means 100/100 = 1, which correlates to 100 PPM in Figure 13. In theory, Ro represents the X axis in Figure 13 if conditions are perfect.

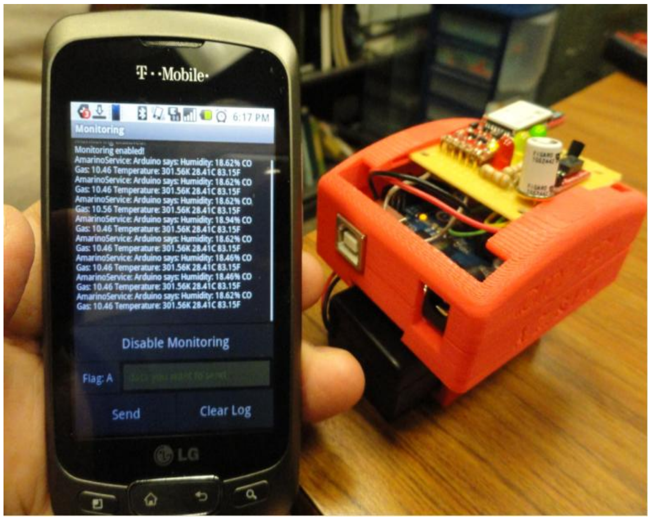
V. PRELIMINARY TESTING AND RESULTS

We have done several tests of mobile CHCD device with smartphone, laptop and tablet. The indoor tests were performed at the College as seen in Figure 14 and Figure 15.



```
Humidity: 17.65% CO Gas: 7.00 Temperature: 301.08K 27.93C 82.27F
```

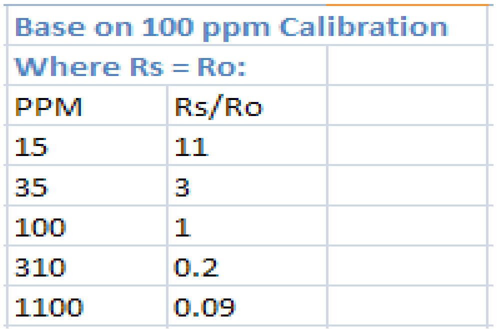
Figure 14. Indoor Testing Using Laptop.



```
Monitoring
Monitoring enabled!
AmarinoService: Arduino says: Humidity: 18.62% CO
Gas: 10.46 Temperature: 301.56K 28.41C 83.15F
AmarinoService: Arduino says: Humidity: 18.62% CO
Gas: 10.46 Temperature: 301.56K 28.41C 83.15F
AmarinoService: Arduino says: Humidity: 18.62% CO
Gas: 10.56 Temperature: 301.56K 28.41C 83.15F
AmarinoService: Arduino says: Humidity: 18.94% CO
Gas: 10.46 Temperature: 301.56K 28.41C 83.15F
AmarinoService: Arduino says: Humidity: 18.62% CO
Gas: 10.46 Temperature: 301.56K 28.41C 83.15F
AmarinoService: Arduino says: Humidity: 18.46% CO
Gas: 10.46 Temperature: 301.56K 28.41C 83.15F
AmarinoService: Arduino says: Humidity: 18.46% CO
Gas: 10.46 Temperature: 301.56K 28.41C 83.15F
AmarinoService: Arduino says: Humidity: 18.62% CO
Gas: 10.46 Temperature: 301.56K 28.41C 83.15F
```

Figure 15. Indoor Testing Using Smartphone.

And the preliminary data from the indoor tests can be seen in Figure 16. The data is then tabular in Excel spreadsheet.



Base on 100 ppm Calibration		
Where Rs = Ro:		
PPM	Rs/Ro	
15	11	
35	3	
100	1	
310	0.2	
1100	0.09	

Figure 16. Data from Preliminary Indoor Tests.

Then the data in Figure 16 is plotted in Excel chart. The chart can be seen in Figure 17. And it is used to tell the sensitivity characteristics of CO concentration levels.

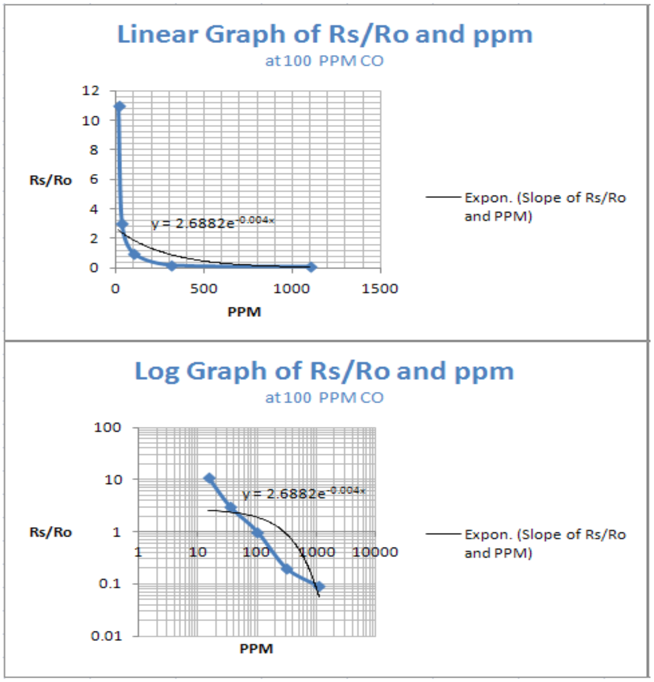
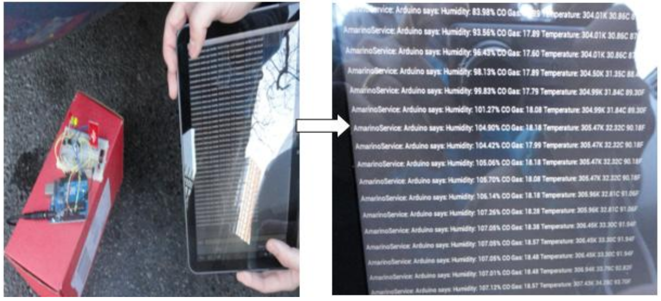


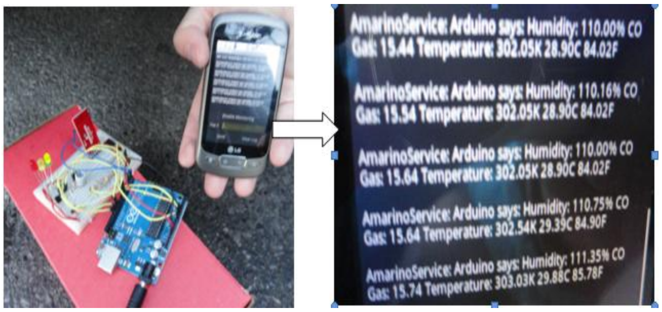
Figure 17. Sensor Resistance Chart for Indoor Tests.

Similarly, the outdoor tests of CHCD device (without the orange case) were performed with the tablet and smartphone. The test was done by placing the CHCD device behind the car's exhaust pipe, while the car engine is on, as seen in Figure 18 and Figure 19 below.



```
[illegible]rvice: Arduino says: Humidity: 83.98% CO Gas: [illegible] Temperature: 304.01K 30.86C [illegible]
AmarinoService: Arduino says: Humidity: 93.56% CO Gas: 17.89 Temperature: 304.01K 30.86C [illegible]
AmarinoService: Arduino says: Humidity: 96.43% CO Gas: 17.60 Temperature: 304.01K 30.86C [illegible]
AmarinoService: Arduino says: Humidity: 98.13% CO Gas: 17.89 Temperature: 304.50K 31.35C [illegible]
AmarinoService: Arduino says: Humidity: 99.83% CO Gas: 17.79 Temperature: 304.99K 31.84C 89.30F
AmarinoService: Arduino says: Humidity: 101.27% CO Gas: 18.08 Temperature: 304.99K 31.84C 89.30F
AmarinoService: Arduino says: Humidity: 104.90% CO Gas: 18.18 Temperature: 305.47K 32.32C 90.18F
AmarinoService: Arduino says: Humidity: 104.42% CO Gas: 17.99 Temperature: 305.47K 32.32C 90.18F
AmarinoService: Arduino says: Humidity: 105.06% CO Gas: 18.18 Temperature: 305.47K 32.32C 90.18F
AmarinoService: Arduino says: Humidity: 105.70% CO Gas: 18.08 Temperature: 305.47K 32.32C 90.18F
AmarinoService: Arduino says: Humidity: 106.14% CO Gas: 18.18 Temperature: 305.96K 32.81C 91.06F
AmarinoService: Arduino says: Humidity: 107.26% CO Gas: 18.28 Temperature: 305.96K 32.81C 91.06F
AmarinoService: Arduino says: Humidity: 107.05% CO Gas: 18.38 Temperature: 306.45K 33.30C 91.94F
AmarinoService: Arduino says: Humidity: 107.05% CO Gas: 18.57 Temperature: 306.45K 33.30C 91.94F
AmarinoService: Arduino says: Humidity: 107.05% CO Gas: 18.48 Temperature: 306.45K 33.30C 91.94F
AmarinoService: Arduino says: Humidity: 107.01% CO Gas: 18.48 Temperature: 306.94K 33.79C 92.82F
AmarinoService: Arduino says: Humidity: 107.12% CO Gas: 18.57 Temperature: 307.43K 34.28C 93.70F
```

Figure 18. Outdoor Testing Using Tablet.

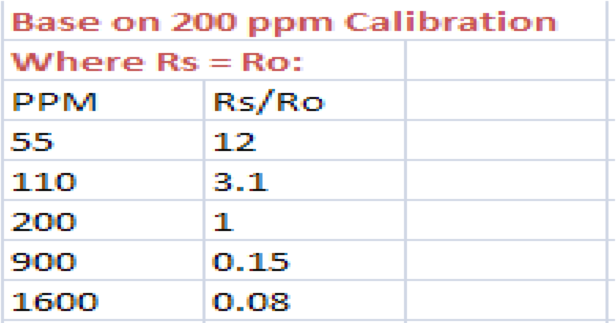


```
AmarinoService: Arduino says: Humidity: 110.00% CO
Gas: 15.44 Temperature: 302.05K 28.90C 84.02F
AmarinoService: Arduino says: Humidity: 110.16% CO
Gas: 15.54 Temperature: 302.05K 28.90C 84.02F
AmarinoService: Arduino says: Humidity: 110.00% CO
Gas: 15.64 Temperature: 302.05K 28.90C 84.02F
AmarinoService: Arduino says: Humidity: 110.75% CO
Gas: 15.64 Temperature: 302.54K 29.39C 84.90F
AmarinoService: Arduino says: Humidity: 111.35% CO
Gas: 15.74 Temperature: 303.03K 29.88C 85.78F
```

Figure 19. Outdoor Testing Using Smartphone.

ISBN: 978-1-61208-202-8

The preliminary data from the outdoor testing scenario can be seen in the following Figure 20:



Base on 200 ppm Calibration		
Where Rs = Ro:		
PPM	Rs/Ro	
55	12	
110	3.1	
200	1	
900	0.15	
1600	0.08	

Figure 20. Data from Preliminary Outdoor Tests.

Then the data in Figure 20 is plotted in Excel chart and can be seen in the following Figure 21:

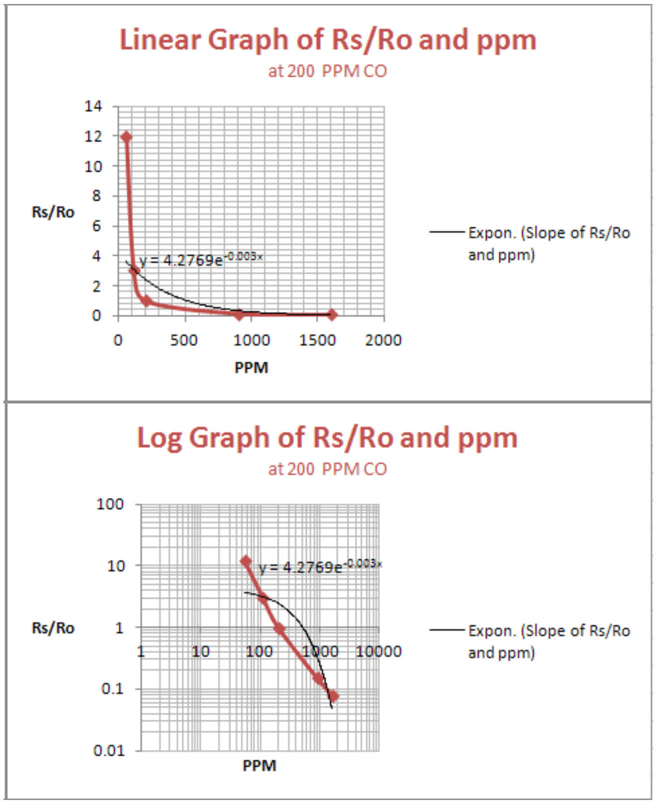


Figure 21. Sensor Resistance Chart for Outdoor Tests.

Two scenarios had been conducted on the mobile CHCD device using smartphone, tablet and laptop. Of the two scenarios, the worst case was found when the CHCD device was placed behind the car's exhaust pipe while the car engine was running. These results are encouraging and show that the mobile CHCD device provides a reliable data to determine the chemical concentration PPM levels. Note that when performing outdoor tests of CHCD device, it is taken into considerations of wind, temperature, and humidity. Clearly, this is a small preliminary sample test for the mobile prototype CHCD device. More testing of this device will be performed in the near future.

VI. SUMMARY AND CONCLUSION

"Detecting explosives is not an easy thing," said David Danley, a retired Army colonel and head of defense programs at Combimatrix Corp., a small biotechnology company near Seattle [14].

David Danley is right, and the goal for this research project is to develop a mobile and portable chemical detection device that will assist and provide significant impact to society in term of reducing the potential loss of human life through detection and prevention. The mobile CHCD device provides early warning system if there is a possible expose of harmful and lethal chemical concentration levels within distance. This early detection and prevention could save many human lives from harmful chemical gases. Thus, the mobile CHCD device becomes an analytics engine capable of picking out emergent patterns in human environments and biology.

The mobile CHCD device is a low-cost miniature detection device that could provide crucial instant information of chemical detection and preventing the loss of human life. This crucial information of sensing and detecting the quality of the air become possible with the aid of modern technologies (smartphone, tablet, and laptop). Hence, the mobile CHCD device, along with modern technologies, provides an alternative affordable resource for the people to have access and use it to identify the invisible harmful chemicals at early warning stage and could possibly lead to save many human lives.

The current cost for producing the prototype of this mobile CHCD device is approximately $60 to $80.

ACKNOWLEDGMENT

This research is partially supported by a grant from the National Science Foundation (NSF ATE No 1003712). The authors appreciate greatly the supports from NSF. In addition, the authors would like to thank Michael Heimbinder who is involved in making mobile CHCD device.

REFERENCES

[1] Smith Detection Group, [retrieved: March, 2012] http://www.smithsdetection.com/1025_4601.php

[2] University of Illinois, [retrieved: March, 2012] http://www.futurity.org/science-technology/sensor-sniffs-out-shoe-bombs/

[3] Parallax Inc., [retrieved: February, 2012] http://www.parallax.com/Portals/0/Downloads/docs/prod/sens/MQ-7.pdf

ISBN: 978-1-61208-202-8

[4] Parallax Inc., [retrieved: February, 2012] http://www.parallax.com/Store/Sensors/GasSensors/tabid/843/CategoryID/91/List/0/SortField/0/Level/a/ProductID/598/Default.aspx

[5] 8085 Projects. Info, [retrieved: February, 2012] http://www.8085projects.info/default.aspx

[6] Fritzing Inc., [retrieved: January, 2012] http://www10.fritzing.com/

[7] Autodesk, Inc., [retrieved: January, 2012] http://www.autodesk.com

[8] I. Heng, F. Zia, and A. Zhang, "*Wired* and Wireless Port Communication." In proceedings of The 118th Annual ASEE Conference & Exposition, June 26 -29, 2011. Vancouver, British Columbia, Canada.

[9] Arduino, [retrieved: January, 2012] http://arduino.cc/en/

[10] Bonifaz Kaufmann, [retrieved: March, 2012] http://www.amarino-toolkit.net/index.php/home.html

[11] About.com Biology, [retrieved: March, 2012] http://biology.about.com/od/molecularbiology/a/carbon_monoxide.htm

[12] Figaro USA Inc., [retrieved: April, 2012] http://www.figarosensor.com/products/2442pdf.pdf

[13] A. Sri-on, S. Sanongraj, and M. Pusayatanont, "*A Simple Microcontroller Circuit for Carbon Monoxide Sensor.*" The 8th Asian-Pacific Regional Conference on Practical Environmental Technologies, Ubon Ratchathani University, Ubonratchathani, Thailand, March 24-27, 2010.

[14] MSN, [retrieved: April, 2012] http://www.msnbc.msn.com/id/8552323/ns/technology_and_science-tech_and_gadgets/t/future-technology-could-help-thwart-terrorism/

[15] M. Goldstein, "Carbon monoxide poisoning." Journal of Emergency Nursing: JEN: Official Publication of the Emergency Department Nurses Association, Vol. 34, pp. 538 – 542, 2008.

A Behavior-Based Method for Rationalizing the Amount of IDS Alert Data

Teemu Alapaholuoma, Jussi Nieminen, Jorma Ylinen, Timo Seppälä, Pekka Loula
Telecommunication Research Center
Tampere University of Technology, Pori Unit
Pori, Finland
teemu.alapaholuoma@tut.fi, jussi.nieminen@tut.fi, jorma.ylinen@tut.fi, timo.a.seppala@tut.fi, pekka.loula@tut.fi

***Abstract*—Intrusion detection systems typically rely on signatures. A signature describes a rule, which is realized as an alert whenever an IP packet matching the rule is observed in the network by an intrusion detection system. In the configuration phase of a signature based intrusion detection system, the operator usually activates the signatures considered interesting. Interesting typically refers to aberrant traffic and behavior in the network. The classification of signatures as interesting or uninteresting is typically based on prior knowledge about the characteristics of the monitored network. In this paper, we introduce a method based on network behavior for identifying, which alerts and signatures could be considered interesting. Based on the identification, only the signatures labeled as interesting should be activated, in order to rationalize the amount of alert data produced. The method is based on the K-means clustering of intrusion detection system alert data.**

Keywords-Alert; Detection; Intrusion; Clustering; Snort

I. Introduction

The unstoppable increase in data transferred on the Internet is reflected directly in the amount of measurement and maintenance data produced in operator systems. This is also the case with IDS (Intrusion Detection System), providing data about threats and intrusion attempts in the network. It is impossible for the operator to analyze every alert manually, and make conclusions about the severity or relevancy of the alerts. Usually the operator classifies the signatures as interesting or uninteresting. The classification can be based on prior knowledge about the target network characteristics, or simply on some specific point of interest. The operator might be interested only in a small amount of signatures, and activate only those. On the other hand, the operator might know that a certain service is never used in the network and the signatures related to it can be deactivated. This paper describes a behavior-based method for classifying signatures as interesting or uninteresting. The method is based on passive data analysis of alert data generated by an IDS system. In the data generation phase, all available signatures are activated. K-Means [1] clustering is utilized as a classification method. The method learns the normal behavior from the alert data, meaning that those alerts can be labeled as uninteresting.

As a data provider we have used Snort [2], an open-source signature-based IDS system. Snort was chosen because it is the leading open-source IDS system in the market. It is actively maintained, and the signature set is updated frequently. Snort is also known to cope with a large packet throughput rate, which is an essential feature in this particular analysis.

The target network was a campus area network. It is known that the information security policy in the target network is fairly free. From the data point of view, this is a benefit. If there were many restrictions in the network, the alert data would not be as generally applicable as it is in this case. The monitoring and analysis execution rely on anonymity. The IP address details of the alerting IP packet are anonymized in the packet capture phase, so the IDS system does not know the real IP addresses communicating in the network.

The data is pre-processed before the actual analysis. The arguments for clustering are computed from the raw alert data, and the data set construction is changed from a time-sorted list into a one-hour time series format. In the analysis phase, a data set of two days' total length is used.

This paper is divided into six sections as follows. In the next section, the research work related to reducing the intrusion detection alert data is presented. Section three presents the environment and processes of obtaining the data. The methods utilized are briefly presented in Section four. In Section five, data set pre-processing is presented together with the execution and results of the analysis. The conclusions and future plans are presented in Section six.

II. Related Work

The rationalizing of the amount of intrusion detection alerts has been studied by many research centers and communities. Typically, the main goal in these studies has been the reduction of the amount of alerts, and presenting only the essential information to the network administrator or operator. In many studies, DARPA data sets have been used. They have been found suitable for information security related analysis, and also for evaluation of IDS systems. The data sets consist of both normal and aberrant network traffic, offering a convenient opportunity to compare observations under different circumstances. Another connective element between the various studies is the selected IDS software. The Snort IDS system is the primary element in many studies, when the intrusion detection is based on signatures. The main reason why the Snort IDS system is so popular among researchers is its good level of performance, affordability,

ISBN: 978-1-61208-202-8

and suitability for different environments. The Snort IDS system is freely downloadable from the Internet offering an effective solution for monitoring IP networks.

Alharby et al. have done related work in [3], where the reduction of false positive alerts is studied using continuous and discontinuous patterns. They achieved significant results where the amount of false positive alerts could be decreased by as much as 90%. Perdisci et al. [4], have collected alerts from different IDS systems in their study. They have used the clustering technique to form upper level alert classes. The creation of the alarm classes was not based on predefined definitions, as they were formed dynamically.

The above-mentioned journals are focused only on reducing the number of alerts. This study has also concentrated on the optimization of an active rule set. In some cases when a new rule set is adopted in the IDS system, some signatures might cause false positive alerts, because they are related to normal traffic and normal behavior in the network.

III. Obtaining The Data

The data set was collected in February 2011. The collection of the data set was carried out using the Snort IDS system (v2.8.6.1). Before starting the monitoring sequence, the Snort IDS was installed and configured. Since it offers only basic elements for network monitoring, the latest rule set was downloaded from the Snort support pages and extracted to the appropriate Snort subdirectory. The HOME_NET and EXTERNAL_NET variables were defined in a Snort configuration file. The first variable specified the IP address space of the campus area network and the second variable encompassed all IP address spaces, excluding the home network. In normal situations, the definition of the HOME_NET variable is a straightforward operation, but in this study an anonymization procedure was used to masquerade as the original IP address space. Before monitoring the campus area network with the Snort IDS system, network traffic is anonymized using the Libtrace library provided by the WAND Network Research Group from the University of Waikato (New Zealand) [5]. The anonymization process itself is performed real-time where original network traffic is captured from a physical interface, anonymized, and forwarded to a virtual interface. Instead of listening to the physical interface, the Snort IDS system monitors network traffic from the virtual interface.

The network traffic is monitored at a backbone link between the home network and the Internet. Unfortunately, the network traffic between different network segments could not be monitored, because it would have complicated the monitoring setup too much. Moreover, we did not want to disrupt the function of network switches by overloading them, so the monitoring phase carried out from a single point.

The home network under observation consisted of 4092 hosts. In order to temper the load of the Snort under high speed network traffic, two simultaneous Snort processes were used. Both processes analyzed a network segment of their own. Instead of storing observations in separate log files, both Snort processes stored the alerts in the same database. This simplified the analysis of the alerts, because we were able to use built-in SQL query clauses, instead of implementing a tool for parsing the log files.

IV. Utilized Methods

K-means clustering is utilized to be able to extract normal behavior from the alert data. K-means was selected as a clustering algorithm because of its tendency as a centroid based clustering algorithm to form round clusters, and for its applicability for clustering large data sets. Round clusters are desirable in this case since we operate in two dimensional argument space, where the clustering arguments are nearly on the same scale. The clustering process in general requires decision-making before execution. The main questions to be answered are:

1. What arguments do we choose for clustering?
2. Which distance metric should we use with this type of data?
3. What is the optimal amount of clusters for this data set?

In this paper, arguments were selected that described the amount of alerts. The time series format effects the selection of arguments, so counters have been used in the data pre-processing phase. The first argument indicates the amount of alerts per signature. This was an obvious choice because the aim in this study was to reduce the amount of alerts. The second argument clarifies how many hosts are causing an alert. This simplifies interpretation of the alerts in a situation where a single host produces a large amount of erroneous traffic that is triggered as alerts. The arguments that are used in the clustering are "*Total number of alerts per signature*" and "*Total number of alerting hosts per signature*", respectively. Both arguments are calculated in the data pre-processing phase.

It is an advantage if the clustering can be executed in a two- or three-dimensional space. Obviously, the execution of the clustering is less expensive from the computation point of view, and secondly, the visualization of the results is straightforward and simple. The visualization point is essential in our case, and therefore the target was to execute the clustering with a maximum of three arguments.

The Euclidean distance metric was chosen as it is typically applied in data analysis in general. It is suitable for centroid-based clustering algorithms operating in low-dimensional data spaces such as the data set in this study.

The optimal cluster amount for the data set is determined by calculating the Davies-Bouldin index [6] for cluster amount values of 2 to 10. The maximum index value indicates the best cluster amount. Additionally, visual interpretation was used in estimating the amount of clusters. Visual interpretation was easy to carry out because there were only two arguments used in the clustering analysis. If the number of arguments had exceeded the two-dimensional space, the visual interpretation would have been a much tougher task.

V. Analysis Execution And Results

A. Pre-processing the data set

There were over 117 million alerts during the four-week period, so it is clear that a method for rationalizing the amount of alerts is needed. A two-day period was selected from the four-week time period for the analysis. The total number of alerts was around 7 million. Approximately 4 million alerts originated from the home network, and the remaining 3 million or so alerts were from the external network. Although there was a huge amount of alerts, only 57 different home network bound alarm types were observed. Correspondingly, 79 alarm types caused alerts that originated from the external network.

Instead of using millions of rows in the data analysis, a one-hour time series was formed from the alert data. The time series data consist of eight arguments. Two of the eight arguments were selected as clustering arguments, and the remaining six were left aside at this point. They will be used in the result analysis phase, when the clustering results are back-traced to the original data set.

When the one-hour time series data was formed from the Snort alerts, the number of rows decreased drastically in comparison with the amount of raw data. After formation of the time series data, there were 1123 rows related to the home network and 1783 rows related to the external network, i.e., a total of 2906 rows.

A signature identification number was the key argument in the formation of the time series data. The identification numbers of the Snort alerts are system specific. A certain signature might be associated with a different identification number in different systems. This naturally complicates the comparison of the results between the various Snort systems. Instead of parsing the time series data from text-based log files, the data was formed from the database using SQL query clauses. Obtaining the alert distribution between the home and external networks was a relatively easy task, because the database supported query clauses, where an IP address range can be expressed by dotted-quad representation (IPv4). The data sets were stored in files of their own, where the values of the eight arguments were separated by a comma (CSV). Finally, we were able to proceed to the mathematical analysis phase.

B. Reduction of home network related alerts

To simplify and accelerate mathematical analysis, in this study the MATLAB program (2011b2) was used, which offered a comprehensive selection of tools for our purposes. Before beginning the statistical analysis, the removal of outliers has to be addressed. If the data set contains data points that are considered measurement errors, or otherwise differ significantly from the rest of the data, outlier removal is necessary. One possible means for removing outliers is a procedure where values that are three times larger than standard deviation are removed from the data set. In this study, outlier removal is not required because there is no method for distinguishing either the normal alert distribution or an aberrant one. We simply assume that all the data produced by the IDS system is valid.

The two arguments used as clustering arguments are the key factors in cluster analysis: the Home Count argument indicates the "*Number of alerts per signature*" and correspondingly, the Home Hit Count argument indicates the "*Number of alerting hosts per signature*". To minimize the effect of the different value ranges of the parameters, the values of the arguments were scaled in two phases, first with logarithmic scaling, and secondly with the MATLAB Zscore function. Logarithmic scaling simply takes a logarithm from the parameter values. Zscore subtracts the mean from every parameter value and divides the values by a standard deviation.

The alerts reduction of the IDS system is carried out in three phases, using K-means as a clustering algorithm, and the Davies-Bouldin index for determining the optimal cluster amount. Based on our earlier experience with the K-Means algorithm, we decided to use the algorithm in this study, too [7]. The decision was also favored by the performance of the algorithm. In the first step, the amount of clusters was estimated using the Davies-Bouldin index. This information was used in the clustering analysis, which was accomplished by using the K-means algorithm. This is a very straightforward process. In the second step, the outcome of the clustering analysis was interpreted. If an alert belongs to the same cluster over the time period, and it appears in every hour over the time period, it can be interpreted as normal behavior. Normal behavior is not interesting and it could be ignored from further analysis, or corresponding alerts could be removed from the Snort rule set. This decision requires human interpretation. In the third step, the alerts related to the normal behavior in the network were removed from the data set. This sequence was repeated until there were no alerts that matched our definitions, i.e., existed constantly and in all periods of the time series. In the final iteration round, alerts were scattered in different clusters and there were no alerts that appeared in every hour over the time period. Steps one to three were carried out separately for alerts, which originated from the home or external networks. This division simplifies the interpretation of the results, because the administrator or network operator clearly sees the initiator of a given alert, and can react to it in the appropriate way. In the following chapters, the mechanism for reducing and rationalizing alerts and signatures is described.

C. Results of the reduction

The results from the first iteration round are presented in a two-dimensional scatter-plot in Figure 1 and in the bar chart in Figure 2.

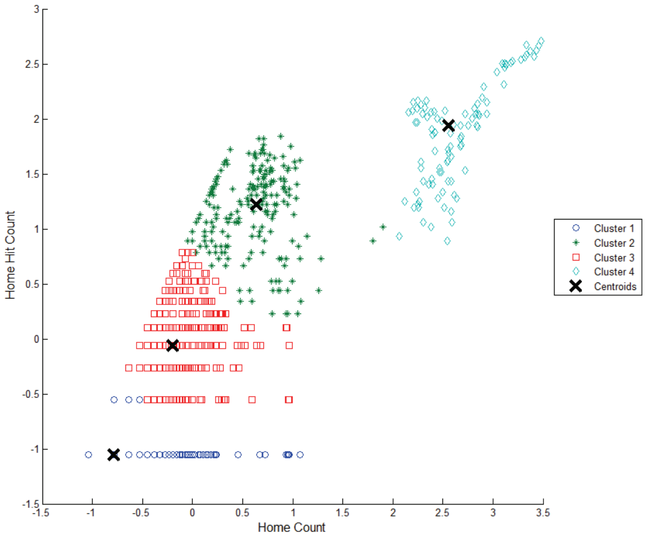


Fig. 1. Snort alerts clustering, home network initiated alerts, round 1

On the X-axis, the values of the argument Home Count are presented. The values of the argument Home Hit Count are presented on the Y-axis. In both cases the values of the arguments have been scaled. In the figure, four different clusters can be seen. The centroids of each cluster are marked by a bolded '*X*'. Clusters 1, 2, and 3 formed a super group on the left side, whereas cluster 4 is clearly separated from the other clusters. At this point we can present the hypothesis that cluster 4 consists of signatures that cause many of the uninteresting alerts, i.e., are related to the normal behavior in the network. In the previous paragraph, there was discussion about the handling of outliers. If the outliers had been removed from the data set, significant information would have been lost about the behavior of the signatures. Most likely the whole of cluster 4 would have been interpreted as outliers. The signature distribution between the clusters is illustrated in Figure 2.

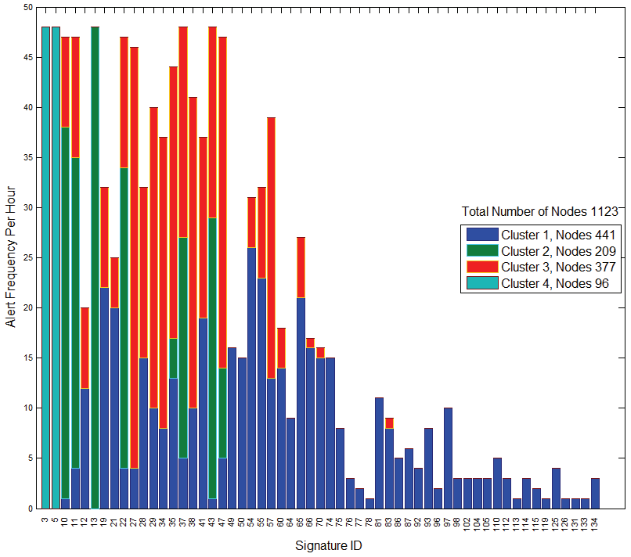


Fig. 2. Alert distribution into clusters, home network initiated alerts, round 1

Signature ID numbers are presented on the X-axis and one-hour time periods are presented on the Y-axis. The maximum value of the Y-axis is 48 hours, which refers to the two-day sampling period. When examining the graph more closely, it can be observed that only a few signatures appeared in every hour and existed in the same cluster over the time period. Signatures 3 and 5 in cluster 4, and signature 13 in cluster 2, fulfilled our requirements of existing in the same cluster in each one-hour period and over the whole 48 hours. Other signatures were distributed into the different clusters or they would not have appeared in every hour over the time period. It might be thought that those three signatures caused many uninteresting alerts. According to our assumption, these signatures relate to normal behavior in the network, and the alerts they caused can be ignored or removed from the rule set. To confirm this conclusion, we studied the nature of the alarms in question, and we found that the alarms in question did not violate the information security policy, so they can be omitted from further analysis. For reasons of privacy protection, only signature ID numbers are used in this paper, rather than using verbal signature identifiers. However, we will open the curtain a little to justify why some signatures can be ignored. Cluster 4 consisted of two signatures that caused many uninteresting alerts. The signatures were triggered from applications that use the IPv6 or BitTorrent protocol, so the hypothesis proved right. Correspondingly, cluster 2 consisted of eight signatures but only one signature can be ignored. The signature triggered form applications that use BitTorrent.

In two days the three signatures caused a total of 3877369 alerts, covering 99 percent of the total amount of alerts, which is a very high proportion. At this point it is clear that by using the method presented in this paper, the amount of alerts can be rationalized drastically.

In the second iteration round, signatures 3, 5, and 13 were removed from the data set in the preprocessing phase and after this procedure, the cluster estimation and analysis were performed again, as in the first iteration round. The results from the second clustering round are depicted in Figure 3.

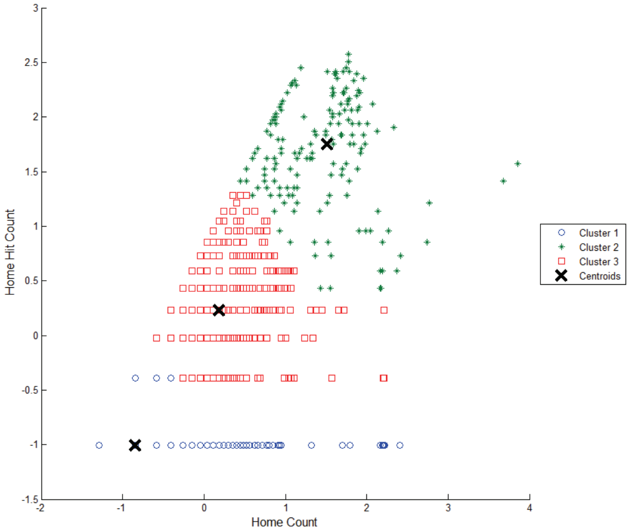


Fig. 3. Snort alerts clustering, home network initiated alerts, round 2

The number of clusters has been decreased from four to three, and there are no more clusters clearly separated from other clusters. Clusters 2 and 3 are strictly connected together, but most of the nodes of cluster 1 are separated from the other clusters. The signature distribution into different clusters over the time series is illustrated in Figure 4.

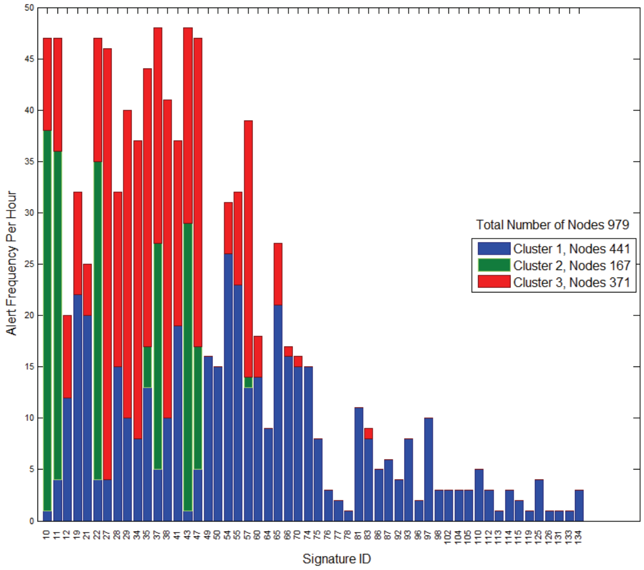


Fig. 4. Alert distribution into clusters, home network initiated alerts, round 2

If we examine the distribution of the signatures into different clusters, we observe that there are no more signatures appearing in every hour of the time series and belonging to the same cluster over the two-day time period. In this case, it can be concluded that signatures cannot be removed after this round, so there is no need to perform the third iteration round of clustering. Signatures 37 and 43 appear in every hour over the two days, but their behavior is inconsistent, which can be concluded from their distribution into different clusters. In contrast, signature 27 belongs mostly to cluster 3 but does not appear in every hour during the two days.

At this point the reader might wonder why clustering the time series data is used to rationalize the amount of alerts. Why not take the five biggest alerting signatures and simply ignore them? We will defend our approach using the following example. In some cases there can be a sudden event that causes an enormous amount of erroneous traffic, which the IDS system interprets as alerts. If in this case the five biggest alerting signatures are simply ignored, we could draw the wrong conclusions about the network behavior. The time series format and the clustering methods ensure that this misinterpretation cannot be made.

D. Reduction of external network related alerts

During the two-day period in question, IP traffic from the external to home network caused about 3 million alerts. In the time series format, the data set consisted of 1783 rows. As in the case of the home network, the data set consisted of eight arguments, but only two variables were used in mathematical analysis. The values of the arguments were scaled using the same MATLAB functions as in the case of the home network. The cluster estimation and analysis were carried out using the Davies-Bouldin index and the K-Means algorithm, respectively.

There were three different iteration rounds related to the external network. After those iteration rounds, a total of 13 alerting signatures could be ignored from further analysis and furthermore, removed from the Snort rule set. The signatures that were removed from future analysis were related to the ICMP, IPv6, and BitTorrent protocols. Generally it can be said that the results from the first iteration round were the most significant. In the first iteration round there were 79 alerting signatures under analysis and 69 in the second round analysis. Finally, in the third round, three more alerting signatures could be labeled as uninteresting.

During the two days, the 13 signatures formed a total of 2986773 alerts, covering 99 percent of the total amount of alerts, which is a very large proportion. When we analyzed the alerting signatures related to the external network, we noticed that 10 of the signatures were related to ICMP traffic. The share was so high that we suspected it to be port scanning related. Table I illustrates how the amount of significant alerts was reduced after each iteration round in the case of alerts originating in the home and external networks.

TABLE I. SUMMARY OF DIFFERENT ITERATION ROUNDS

	Home Network			**External Network**		
Round	***Signatures***	***Alerts***	***%***	***Signatures***	***Alerts***	***%***
1	57	3907585	100	79	3011376	100
2	54	3877369	99	69	2692465	89
3	-	-	-	66	294308	10
Result	**54**	**30216**	**1**	**66**	**24603**	**1**

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-202-8

VI. Conclusion And Future Work

The goal of this study was to obtain a method for rationalizing the amount of intrusion detection alerts by identifying the alerts related to the normal traffic in the target network. Removing the uninteresting alerts related to normal traffic resulted in a set of alerts that potentially held all the interesting alerts related to aberrant traffic. Our experience from the Snort IDS reveals that the number of alerts can easily exceed the limit of one million per day in a large-scale IP network. A manual examination of millions of alarms would drive a system administrator or network operator to desperation! We defined two arguments for mathematical analysis by post-processing the raw alert data. In total there were eight arguments. The other six variables offered valuable information during the analysis phase. Instead of using millions of rows in the mathematical analysis, we formed time series data from the original data set. To simplify the analysis phase, the alerts were divided into home and external data sets, based on the source IP address of the alert.

When the two data sets had been obtained, the reduction of the alerts was started. The amount of clusters was estimated with the aid of the Davies-Bouldin index, and the cluster analysis was carried out with the K-means algorithm. We analyzed signature behavior over the whole of the 48-hour time period. If a signature belonged to the same cluster over the whole time period and it appeared in every hour over the time period, it was associated with the normal traffic in the network, and thus ignored from further analysis and removed from the Snort rule set. When the inappropriate signatures had been removed from the data set, cluster estimation and analysis were processed again, until there were no signatures suitable for our definitions.

The method fulfilled our requirements surprisingly well, as seen in Table I. By using well-known mathematical functions and appropriate arguments, the number of alerts can be significantly reduced. This simplifies the further analysis of the alerts and optimization of the Snort rule set. When the optimized rule set is activated, a network operator can react faster to critical threats. On the other hand, the number of alerts is still fairly high after the iteration rounds, so alternative methods have to be investigated in order to further reduce the amount of alerts. In addition, we are planning to extend our study in the future to compare our method with other existing methods. Unfortunately, this could not be carried out in time for this paper because the time window of the study was coming to an end.

Another question is how well does the signature-based IDS system apply to the monitoring of large-scale IP networks, where the number of alerts might exceed the limit of one million per day. In addition to this, many malware applications use strong encryption algorithms to protect the control traffic, which, in turn, is difficult to detect from the normal network traffic. Consequently, there are still many open questions related to intrusion detection left to be addressed.

References

[1] J. Macqueen, “Some methods for classification and analysis of multivariate observations,” Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, University of California Press, 1967, pp. 281-297.

[2] M. Roesch, “Snort - Lightweight intrusion detection for networks,” Proceedings of the 13th Conference on Systems Administration (LISA-99), USENIX Association, Nov. 1999, pp. 229-238.

[3] A. Alharby and H. Imai, “IDS False Alarm Reduction Using Continuous and Discontinuous Patterns,” Applied Cryptography and Network Security, vol. 3531, 2005, pp. 423-442, doi:10.1007/11496137_14.

[4] R. Perdisci, G. Giacinto, and F. Roli, “Alarm clustering for intrusion detection systems in computer networks,” Engineering Applications of Artificial Intelligence, vol. 19 (4), Jun. 2006, pp. 429-438, doi:10.1016/j.engappai.2006.01.003.

[5] Libtrace, version 3.0.8, http://research.wand.net.nz/software/libtrace.php, retrieved Dec. 2011.

[6] D. L. Davies and D. W. Bouldin, “A cluster separation measure,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. PAMI-1 (2), Apr. 1979, pp. 224-227, doi:10.1109/TPAMI.1979.4766909.

[7] O. Knuuti, T. Seppälä, T. Alapaholuoma, J. Ylinen, P. Loula, P. Kumpulainen, and K. Hätönen, “Constructing communication profiles by clustering selected network traffic attributes,” Proceedings of the Fifth International Conference on Internet Monitoring and Protection ICIMP 2010, May 2010, pp. 105-109, doi:10.1109/ICIMP.2010.21.

The Application of Fuzzy Clustering in Life Quality Assessment of Lung Cancer Patients

Peiyu Chen
College of Electronic Information and Control Engineering
Beijing University of Technology
Beijing, China
530795037@qq.com

Liying Fang, Pu Wang, Shuang Li
College of Electronic Information and Control Engineering
Beijing University of Technology
Beijing, China
{fangliying, wangpu}@bjut.edu.cn,
690736804@qq.com

***Abstract*—In order to research the relationship between the FACT score and lung cancer patients' tumor progression with the patients' death situation, we manage to identify the variation tendency rules of the FACT score. This paper proposes a new, simple and efficient representation for the one-dimensional numerical time series. The representation uses the improved fuzzy C-means clustering algorithm to cluster all the sequence segments, realizing the symbolization of the numerical data sequence. At last,we find the frequent patterns in the new symbolic FACT score sequence. The relation between the FACT score and the patients' death situation can be found through the experiments.**

Keywords-fuzzy C-means clustering; time series; symbolization; GSP.

I. INTRODUCTION

This part contains a review of the data object, i.e., FACT score, and expounds the research necessity and significance.

A. FACT score

Life quality assessment of lung cancer patients has been widely focused, as the traditional disease evaluation and prognosis are not adapting the patients' needs, the health concepts and nowadays medical mode any more.

FACT is a Functional Assessment of Cancer Therapy which was developed by Cella etc. of the Chicago Rush-Presbyterian-St. Luke Medical Center at United States. It contains a basic module, FACT-G, measuring the life quality of cancer patients and some specific cancer subscale. The FACT-G has 27 indexes, divided into four parts: the somatic condition (7 indexes), the social/family situation (7 indexes), the emotional status (6 indexes) and the functional status (7 indexes). Specific cancer assessment contains the basic module and a specific module [1].

The FACT-L, a specific cancer assessment, that is a self-assessment of the lung cancer patients which containing the FACT-G and the lung cancer specific module (9 indexes). The lung cancer specific module is mainly used to the patients who receive chemotherapy, radiation therapy and have been well-educated. Each index is the grade entry, the forward entry direct count from zero to four points and reverse entries reverse count when computing the score, i.e., four points to fill the first one grade within three points on two levels, etc.

B. Overview

The basic task of data mining is to find frequent patterns from the data set. This kind of problem is to decide which mode is frequent in a class of data set mode possible existence. Frequent pattern is found based on the symbolic sequence that including several basic item sets, however, the large number of the item sets will inevitably reduce the support and typicality of frequent pattern. In this way, finding frequent pattern from the numerical data sequence will be a great challenge.

There are certain kinds of data types in the medical data source, such as the symbol type, the Boolean type, the numeric type, etc. Due to the limited number of basic item sets, the frequent patterns using the frequent pattern algorithm can be directly mined from the first two data types. The number of the numerical data in the basic set is hundreds of thousands that is not conducive to frequent pattern mining, so mining frequent patterns need to transforming dataset from numerical to symbol.

This paper proposed a new representation of time series, and then mined data pattern from FACT score in lung cancer cases for above problems. The experiments show that the representation can re-paint the original sequence in a certain accuracy conditions and convenient for their pattern mining.

In the second part, introducing the whole algorithm and detailed explain the each step of the algorithm. In the third part, the experiment is accomplished based on the algorithm and analysis. At last, the conclusion and future work is presented.

II. PATTERN FIND BASED ON THE FUZZY CLUSTERIN

During the process of finding the pattern, the first step is classifying the FACT score with tumor progression and death situation. Then symbolizes the original sequence to find the pattern. At last verifies the result, the sequential pattern, which is compliant with reservation requirements through the experimental results. If the results do not reach reservation requirements, it needs to re-select the dimensions

ISBN: 978-1-61208-202-8

and other parameters, until the condition is met. Figure 1 shows the flowchart:

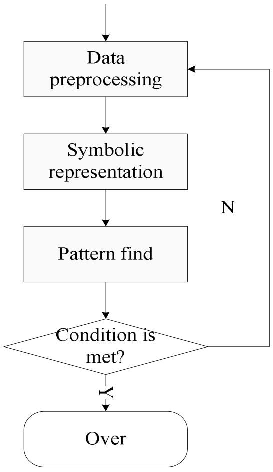


Figure 1. The flowchart of Pattern find

A. Data preprocessing

Each patient has at most 4 treatment periods and 14 follow-up periods, and the collected data content of each treatment and follow-up are the same. Preprocess the patients' FACT score. Take out the ID number of patients and the corresponding FACT score while ignoring the missing point directly from the database which stores the each treatment and follow-up data of patients.

B. Symbolic representation

First handle the patient's data thus we could segment an N-dimensionality into an N-1 one by making every treatment and follow-up data a break-point. The slope of the segment represent the trend of the curve, which has a better result compared with the Euclidean distance, Pearson correlation coefficient etc. Extract the slope of the segment as the standard of clustering, and symbolize the segment that each line in the corresponding class.

In this paper, we proposed an improved fuzzy C-means clustering algorithm used to cluster FACT score. Fuzzy clustering algorithm is based on the best practices function and it use calculus computing technology to get the optimal cost function. The fuzzy clustering algorithm defines the neighbor function between the vector and the cluster, and the membership function set provides the membership degree of the clustering vector. In the fuzzy method, the vector membership degree in the different cluster are interrelated. Simultaneously using the distance of slope of the segment replacing the traditional Euclidean distance could pay more attention to the trend of the FACT score in the clustering algorithm.

The fuzzy C-means clustering algorithm's input parameter is C and then divide N data objects into C clustering. Clustering results to meet high similarity in one cluster, and meet smaller similarity in the different clustering results [2].

Fuzzy C-means clustering algorithm:

Input: the number of clusters C, and the data sets with N data objects.

Output: the matrix of cluster centers and partition matrix.

Algorithm process:

Step 1: Initialize the partition matrix Uik with values in a random number between 0 and 1.

Step 2: Calculate the cluster center.

Step 3: Calculate the value function J. If it is less than a certain threshold or the value change comparing with last value is less than a threshold, the algorithm stop.

Step 4: Recalculate the partition matrix Uik, then return to step 2.

Figure 2 shows the flowchart:

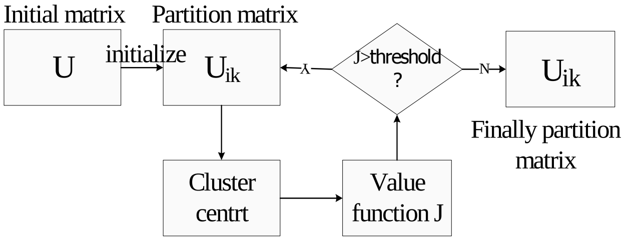


Figure 2. The flowchart of Fuzzy Cluster

The pseudo code of the algorithm:

```
Initialize the number of clusters = 3;
Initialize the partition matrix Uik;
While J> threshold
    For each cluster:
        Calculate the cluster center;
        Calculate the value function J;
Return Uik;
```

In the algorithm, the data object is a sub-sequence after segment; the distance between the segments is measured by the change value of the two points of the segment [3]. The distance between segment l_i and l_j is:

$$D_{ij} = \sqrt{(s_i - s_j)^2} \qquad (1)$$

C. Pattern find algorithm

Frequent pattern find algorithm uses pruning strategies of redundant candidate pattern and special data structure-a hash tree realizing the fast memory access of candidate patterns that using the GSP algorithm, which is similar to Apriori algorithm [4][5].

Algorithm process:

Step 1: Scan sequence database, getting sequential pattern L1 whose length is 1 as the initial seed set.

Step 2: According to the seed set Li, candidate sequential patterns Ci+1 whose length is i+1 by connecting operations and pruning operations; then scan the sequence database, calculating the support of each candidate sequence patterns and generate the sequence pattern Li+1 whose length is i+1 which is used to be a new seed set;

Step 3: Repeat the second step, until no new sequence mode or a new candidate sequence patterns.

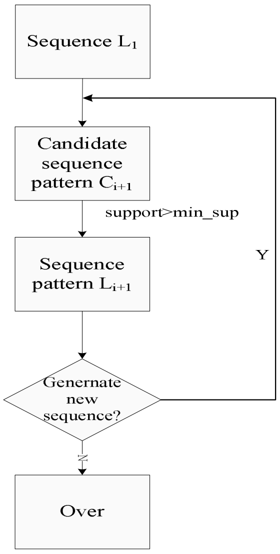

Sequence L_1
Candidate sequence pattern C_{i+1}
support>min_sup
Y
Sequence pattern L_{i+1}
Genernate new sequence?
N
Over

Figure 3. The flowchart of GSP

The pseudo code of the algorithm:

```
Initialize the sequence L1;
While generate new sequence
    Generate the candidate pattern Ci+1
    For each Ci+1
       If the support of Ck>min_sup
            Ck is a FP;
       Get the new sequence pattern Li+1;
   Return Li+1;
```

In above flowchart, the initial sequence is the result of the fuzzy clustering after restored base on the original numeric sequence by the patients.

III. EXPERIMENTAL RESULTS AND ANALYSIS

Experimental data is FACT score which is in the treatment and follow-up cases of non-small cell lung cancer provided by the Beijing Chinese Medicine Hospital. All patients' data were right-align (reversal the sequence), and only intercepting the first 4 data of the sequence. All cases have been deposited in the Oracle 10g database.

There are 4 class patients which classified by tumor progression and death situation: first class is no tumor progression but death; second class is tumor progressed and death; third class is no tumor progression and live; forth class is tumor progressed but live. Every patient have a sequence whose length is between 4 and 17; then cluster the FACT segment by the fuzzy C-means clustering algorithm, while C is 3 that is increment, decrement and constant which are represented by a,b,c, respectively; at last, use GSP algorithm to finding the frequent pattern from the sub-sequence with four data.

We use above algorithm to find the frequent pattern from FACT score. TABLE I shows the result:

TABLE I. THE RESULT OF THE PATTERN FIND

class (patients' number)	Support (percentage)	Frequent sub-sequence
I(22)	15(68.2%)	<a c c>
II(74)	52(70.3%)	<a c c>
III (59)	55(93.2%)	<c c>
IV (75)	66(88%)	<c c>

From the table, it can be seen that a clear upward trend in the last follow-up data in Class I and II patient, and lead to death; however, Classes III and IV patient population in the last several follow-up FACT score value is always maintained a flat trend, and survived.

IV. CONCLUSION AND FUTURE WORK

Transformed the numeric sequence to symbolic sequence by the improved fuzzy C-means clustering algorithm, and find the frequent pattern from the new sequence. The method shows the relationship between the FACT score and the patients' death situation. But there are some disadvantages, such as the support of the frequent sequence <a c c> in class I and II is not enough. Further improvements of the algorithm we expect better results.

ACKNOWLEDGMENT

This paper is supported by 2010 Program for Excellent Talents in Beijing Municipal Organization Department (2010D005015000001), the New Centaury National Hundred, Thousand and Ten Thousand Talent Project, and got the cooperation with Beijing Hospital of Traditional Chinese Medicine Affiliated to CPUMS. Special thanks have been given there.

REFERENCES

[1] C. Wan and C. Zhang, Development and evaluation of the Chinese version of the FACT-L (V1.0) for patients with lung cancer, Quality of Life Newsletter, pp. 19.

[2] H. Chu and B. Chao, Novel Optimization Method for Fuzzy C-Means Algorithms, Journal of Information Engineering University, vol.12, No.3, Jun.2011. (in Chinese)

[3] C. S. Mller-Levet, F. Klawonn, K. H. Cho, and O. Wolkenhauer. Fuzzy Clustering of Short Time-Series and Unevenly Distributed Sampling Points. in Advances in Intelligent Data Analysis V. 5th International Symposium on Intelligent Data Analysis, IDA 2003, 28-30 Aug. 2003. 2003. Berlin, Germany: Springer Verlag.

[4] M. Xia and X. Wang, Research on sequential pattern mining algorithms, Computer Technology and Development, 16th ed., vol. 4. (in Chinese)

[5] L. Liu and J. Cui, Comparison and Analysis between Algorithm of GSP and PrefixSpan, Journal of Liaoning Institute of Technology, vol. 26, No.5, Oct. 2006. (in Chinese)

Mobility platform for video content with Augmented Reality

Ángel García Crespo, Israel González Carrasco, José Luis López Cuadrado
Computer Science Department
Universidad Carlos III de Madrid
Leganes, Spain
acrespo@ia.uc3m.es, igcarras@inf.uc3m.es, jllopez@inf.uc3m.es

Carlos Celorrio, Álvaro González
Planet Media Advanced IT Solution
C/ Torrelaguna, 77
Madrid, Spain
carlos.celorrio@planetmedia.es,
alvaro.gonzalez@planetmedia.es

Abstract— **Augmented reality is to include multimedia elements in real images captured from the real world. Nowadays mobile devices allow visualizing of high quality multimedia elements and processing video, images and information like a personal computer. The inclusion of augmented reality applications in the mobile devices is growing but so far it has not been massively exploded. Developing games is one of the major applications of this technology. Gymkhanas, guided tours or treasure hunts are funny games that can be improved by the application of AR. However there are not tools that allow the final users for developing their own games. This paper presents an ongoing work that aims to develop an online framework for creating augmented reality games for mobile devices which allows the user for creating their own games sharing them with other people. The application of this tool will not be only focused to gaming: training programs, fire prevention or guided tours could be developed in an easy way.**

Keywords- Augmented Reality, Developing Tool, Mobile Devices, Games

I. INTRODUCTION

The Augmented Reality (AR) technologies are being integrated gradually into mobile devices, enabling the creation of increasingly sophisticated games and other useful applications for day to day. From the point of view of video games different proposals aimed at introducing real game virtual items has been developed. However, the authors didn't find an approach that allows the user to define their own game, in their environment and their own goals, and simultaneously seek interaction with users of different devices.

In addition, consumers of digital media content in the form of games are usually accustomed to playing on scenarios and pre-defined objectives. AR allows the use of related real spaces for users to define different stages of a game. Although there are games that take advantage of these real spaces, there is not a tool that facilitates the user to create their own game without having deep technical knowledge. The approach of this technology to the user opens the ability to benefit from the advantages of AR in everyday situations, either as a game or, as already mentioned, for the definition of simulation and guides. Besides the creation of an online tool enables collaboration of multiple users from different locations and facilitates the publication and access to content through a web server. On the other hand, a tool for development this type of games will to facilitate to companies the creation of content for mobile devices and the possibility of quick development of AR-based games.

II. LITERATURE REVIEW

Augmented reality consists of expanding the reality captured by image information generated by computer. The areas of application of this technology are very different but the content creation tools for these technologies are usually single-user desktop applications.

Originally virtual reality was used as a technique for rehabilitation in different areas. Thus, Wilson and others assert that the main benefits of applying these techniques involve the ability to create activities in a simulator, without the limitations imposed by their disability, to allow the people with disabilities to perform them safely [2]. For example, such exercises can be simulated in a kitchen without the dangers associated with it (fire, etc.). There have been various approaches to therapy based on virtual reality, e.g. Jack et al. propose a virtual reality system in which the patient can perform the exercises from two gloves that allow interaction with the system [3].

Furthermore, AR presents a promising future, being an emerging technology with interesting potential applications in many domains. Among them, the entertainment, and more specifically the videogame, takes an important place. This alternative is conceived as computer software created for entertainment in general and based on the interaction between one or more persons with an electronic device that runs the game. The video game market is one of the emerging sectors of the communications industry and entertainment, exceeding its sales to the cinema in some countries [4]. Taking these data into account it is understandable that companies are betting in a clear way for AR in this field, since this technology offers very attractive possibilities for end users

Despite the global crisis the fall in sales in Spain is below than the rest of Europe [4]. In this sense in recent years, AR has made its way into the game technology, thanks in large part to the great evolution of gaming technology and the great competition among the companies. The inclusion of reality factor in the computer image makes the exercise gain in credibility for the end user compared to the previous solutions based solely on virtual reality. Today there are increasing examples of commercial video games based on this technology. Global dimension companies such as Sony Computer Entertainment with Invizimals [1] have made contributions to this technology for different consoles platforms. Other attractive AR solution developed for mobile devices is Can You See Me Now? of Blast Theory [6]. This is an online game of persecution on real streets of different cities based on locations obtained from GPS devices. Another important project was the Project ARQuake [7], adaptation of the popular video game to be played in real locations. Even newly emerging platforms such as Android or iPhone begin to take advantage of AR to provide the end user a different experience and advantageous. An example is Wikitude World Browser which allows from an Android device plan trips and find content of interest about the user's environment through information from Wikipedia and Qype [8]. Finally, Augmentizer for iPhone provides information of reality through labels that identify points of interest (POI) via the mobile phone [9].

Regarding the generation of AR-based effects, there are several tools that will facilitate the creation of applications and tools with these features. With these toolkits, developers will be able to comprehensively implement AR experiences, from initial concept or idea to final product delivery. Atomic [10] is a tool that allows you to define patterns that will be recognized from the image of a webcam to be replaced by virtual images. Similar tools asDart [11] or BuildAR [12] or programming libraries like OSGART [13] facilitate the creation of AR-based effects that can be integrated in other projects and applications, enabling integration of software architectures under the same solution. However, none of these tools allows online development with the collaboration of multiple users.

With respect to existing devices, an essential component of AR is the camera. Currently, over 30% of the installed PC has a webcam. In 2008, the web camera market was valued at 1,500 million euros, and forecasts for 2015 are to achieve a growth of over 2,500 million. Laptops manufactured in recent years, incorporated in most cases a web camera. Moreover, all 3G-based mobile phones have video cameras. The latest generation of consoles (Xbox 360 with Kinect, Play Station 3 with the move motion controller, Wii ...) allows the acquisition and use of webcams. The new generation of consoles, including portable (e.g. Nintendo 3DS, PSPgo, PSP-3000, and iPhone4 and Sony Ericsson Xperia Play), also included standard video cameras. The functionality and the use of these webcams is very limited, sticking exclusively to the videoconference and to take pictures, but the number of webcams installed is very significant, with expected growth in coming years. This is a great opportunity for the application of AR to enhance the appearance of new applications and services based on these devices, since this technology is based precisely on the capture and image processing for generating real-time image synthesis seamlessly integrated with reality.

III. PROPOSED SOLUTION

A. INTRODUCTION

The aim of this research is to develop an online platform to for creating AR games which requires the interaction between mobile devices. It is a complex system that requires the development of complex components that must be integrated together.

AR is to introduce virtual reality elements into real-world images, providing additional information. This technology has been applied in the video-games area offering advanced video-games. The evolution of mobile devices has opened a new world of possibilities in the creation of AR applications: from guides of museums to driving assistants. The bandwidth of mobile communications to the Internet grows with the technology of the new devices, allowing the reception of data in reasonable time and providing with more possibilities of communication and processing. However, the technologies of the mobile devices are heterogeneous and most of the developments are device dependents.

The development of the components required for the execution of games in mobile devices is direct consequence of the requirement of probe that the proposed development framework works properly.

The proposed technology will allow users without technical knowledge for the development of personalized games. The games developed are not restricted to leisure aims: it can also be used as professional training in role-playing or simulation (especially simulation of situations of risk), or as an assistant for mobility. Thus a simulation of the evacuation of a building in a fire emergency could be defined. Each player with a mobile device with a camera could visualize the images of the place in which the game is played. From the central server, each player would receive the information about virtual obstacles or hints about the best way to go.

Achieving the aims of this research implies the validation of the idea of an operative framework as is, as well as the establishment of the basis for the development of a generic technology for defining AR video-games that allows the integration of several players with different devices in the same activity guided by the rules of the game. It allows the user for taking part in a real-time collective game without technological restrictions.

B. FRAMEWORK ARCHITECTURE

This project is based on creating a platform for developing and publishing online games based on AR. The aim of the game defined could be different than recreational purposes: as mentioned, it can be used for simulation exercises or tests emulating risk situations difficult to obtain in the real world.

As shown in the literature review, there are games based on the movement in a defined environment. However, these games are invariables, meaning that either the goals or the field must always be the same. While there have been found development frameworks for AR, they are not intended for non-technical users or developments in an online platform where multiple developers can participate. The other feature of the game is that a director, in real-time, is interacting with the players to lead the achievement of the objectives of the game.

The main components of this framework are depicted in Figure 1. The main feature of the games is that they will be directed from a central computer via a game server component from which the information is sent to various players involved in the game. Players, with heterogeneous mobile technology, will capture images of where they are. Through cameras that incorporate such devices. The elements that are defined for the game, and other elements the director of game want to send, will be included in these images by means of the mobile client component.

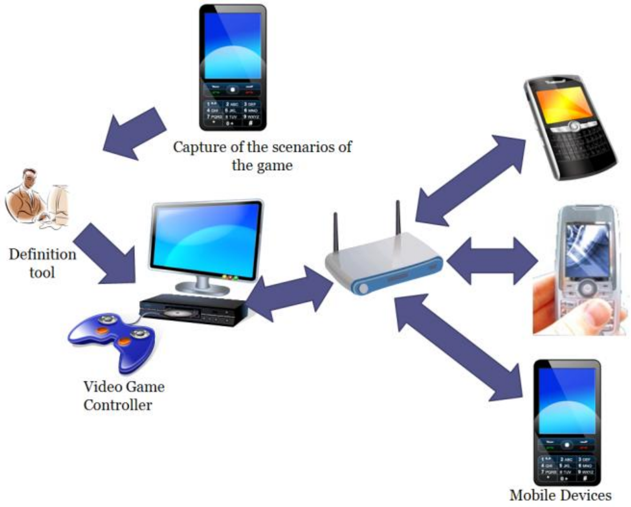


Figure 1. General architecture: components of the complex system.

The demonstration of the validity of the proposed framework will be made on a game server component, which will include an element of control for the game director, and a client component of the game for Android platforms. Figure 2. depicts the sequence of game definition from a set of base components.

The development framework will be based on a Model Driven Architecture (MDA). The foundation of this architecture lies in the definition of the games in a device-independent way, so that performance on specific devices will be based on the interpretation of the generic definition of the game. It will allow the execution of games in several existing platforms as well as future platforms, by means of specific components for each architecture. Furthermore, modelling each game in a device-independent way facilitates the creation of different generic components that can be used for different games on different platforms. It will also ease the process of setting up each game without device-dependent considerations.

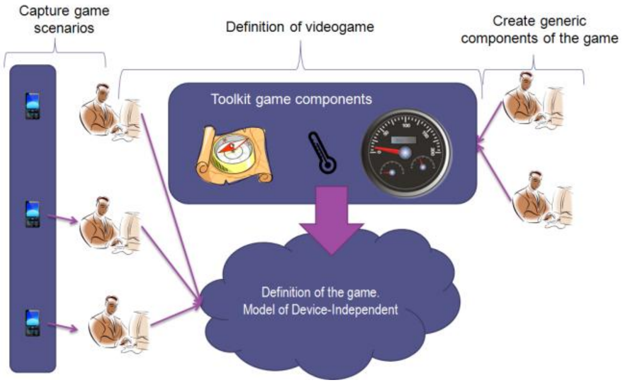


Figure 2. Process for defining the game (device-independent)

C. OPERATION STEPS

1. The first step is to define the scenario in which the game will take place. Since the game can be defined in the place that the user decides, it is necessary to establish a set references that can be used later to define virtual elements on them.
2. After obtaining the images and / or geographic positions in which the elements of the game will be introduced, the user will enter this information on the game development tool. With this tool, the user defines different virtual items to be entered into the game and which ones are the goals of the game (get to a certain point, find a certain number of elements, etc.).
3. Once the game has been defined, the execution module will register the devices that will participate in the game and establish the communication between them. The defined game is loaded in the game control server. Each device read the model by means of the execution component. This component adapts the independent model to the specific device in which it will be executed. From that moment, the devices will send their GPS position and will receive the elements defined based on their position.
4. The master of the game, from the control tool, could monitories the position of the players.
5. In real time, the master of the game can sent messages and other non previously defined elements to the players. In this way the game is more dynamic and can be adapted by the master to unexpected situations.

The communication between devices is done through a network connection, either wireless or mobile phone connections, depending on the characteristics of the device. Initially there are no restrictions, except that technical analysis or further field tests determine significant constraints on the connection requirements.

IV. CONCLUSIONS

This paper has introduced the work in progress on a novel framework for developing games based on mobile technologies and AR. The impact of the proposed framework in the market is twofold: on the one hand it can become a professional development tool for providers of mobile services; on the other hand a restricted version aimed at home users can created their own games.

Moreover, the component for defining the games will bring the creation of AR-based content to users without much knowledge of computers, so they can create their own games in the environment they prefer without having deep knowledge about 3D technology, AR and, of course, without any programming.

The nature of this research is clearly horizontal because this technology is not only focused on games based on AR but also it can be applied as discussed to create simulations that allow training evacuation plans, defining guided tours including POI information, or including advertising elements. In short, the platform opens a new world of possibilities for creating AR content and for its distribution in mobile devices with heterogeneous technologies.

REFERENCES

[1] Animalz, http://petz.es.ubi.com/ (Last visited 31/05/2010).

[2] Wilson PN, Foreman N, Stanton D., Virtual reality, disability and rehabilitation. Disability and Rehabilitation, 19(6):213-20. 1997.

[3] Jack, D., Roian, R. Merians, A.S., Tremaine M., Burdea, G.C., Adamovich, S.V., Recce, M. and Poizner H. Virtual Reality-Enhanced Stroke Rehabilitation. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 9(3): 308-318. 2001.

[4] Asociación Española de Distribuidores y Editores de Software de Entretenimiento. Balance económico de la industria del videojuego. http://www.adese.es/pdf/dossier_prensa%20_balance_economico_2009.pdf (Last visited 31/05/2010).

[5] INVIZIMALS: http://www.invizimals.com/ (Last Visited 31/05/2010)

[6] Can You See Me Now?: http://www.blasttheory.co.uk/bt/work_cysmn.html (Last visited 31/05/2010).

[7] ARQuake: http://wearables.unisa.edu.au/projects/arquake/ (Last visited 31/05/2010).

[8] WIKITUDE WORLD BROWSER: http://www.wikitude.org/wikitude-ar-guide (Last visited 31/05/2010)

[9] AUGMENTIZER: http://augmentizer.net (Last visited 31/05/2010).

[10] ATOMIC: http://www.sologicolibre.org/projects/atomic/es/

[11] DART: http://www.cc.gatech.edu/dart/ (Last visited 31/05/2010)

[12] BUILDAR: http://www.hitlabnz.org/wiki/BuildAR (Last visited 31/05/2010)

[13] OSGART: http://www.artoolworks.com/community/osgart/ (Last visited 31/05/2010).

The Design and Implementation of Bare PC Graphics

Alexander Peter
Computer Science
Towson University
Towson, MD 21252
apeter9@students.towson.edu

Ramesh K. Karne
Computer Science
Towson University
Towson, MD 21252
rkarne@towson.edu

Alexander L. Wijesinha
Computer Science
Towson University
Towson, MD 21252
awijesinha@towson.edu

Patrick Appiah-kubi
Computer Science
Towson University
Towson, MD 21252
appiahkubi@towson.edu

Abstract—**Most multimedia applications today run with the support of an operating system, a graphics driver and related libraries. We present a lean graphics architecture for a bare PC that has no operating system or kernel running in the machine. The architecture enables a multimedia application to be independent of any computing environment and avoids dependencies on other software. To maintain simplicity, the graphics implementation uses the basic primitives to display a pixel, line, circle and a bitmap image. It can be used to implement complex graphics in spite of its simplicity. The bare PC graphics implementation is small in size, extensible and easy to maintain. This design allows graphics programmers to achieve higher performance by eliminating operating system overhead and using direct interfaces to the hardware.**

Keywords-Bare Machine Computing (BMC); Multimedia Graphics; Bare PC Graphics; Graphics Design; Application Graphics Object (AGO).

I. MOTIVATION

The bare machine computing (BMC) paradigm has been demonstrated using applications such as Web servers, Web mail servers, Email servers and clients, SIP servers [1], secure applications [13], and VoIP soft-phones [2, 14]. These applications used text-only interfaces for user interactions. The availability of a bare graphics interface will enrich these applications and make them more convenient for users. BMC applications are self-contained and independent of any operating system (OS), kernel, or execution environment. The work presented is a first step towards building bare PC graphics interfaces and lays a foundation for future multimedia applications that can run on bare devices.

II. INTRODUCTION

Current multimedia and graphics applications are built on graphics, video and audio drivers that are accessible through some platform such as Windows or Linux. In handheld devices, multimedia software is embedded in the devices to allow graphics capabilities. These embedded systems rely on some lean OS or kernel. In most cases, multimedia or graphics applications are dependent on the device platform. Modern multimedia applications use a graphics processing unit (GPU) along with video memory to provide parallel processing power before rendering to screen or storage. The video card's processing power and technological advancements in hardware pave the way for new software architectures that exploit the capabilities of modern systems. For example, since video cards provide gigabytes of low cost memory, paging and virtual memory are unnecessary, and multiple address spaces can be avoided by using a single monolithic executable code with real memory [15]. Today's high definition video cards can stream well over 60 frames per second, and are even over-clocked to higher speeds for 3D simulation using the Wiggle effect [4].

However, the above technological trends and techniques are platform-dependent and are not easily ported from one environment to another. The paper considers the design and implementation of a bare graphics architecture that is self-contained and does not require any operating system, kernel or environment to run. It is written in C/C++ and accesses video memory directly from its application program. This paper describes our approach in detail with some preliminary data.

The remainder of the paper is organized as follows. Section III presents related work; Section IV describes the architecture of bare PC graphics; Section V discusses its design and implementation; Section VI presents the results; and Section VII contains the conclusion.

III. RELATED WORK

The BMC concept also known as dispersed operating system computing (DOSC) [10] enables computer applications to run on a bare machine or a bare PC. Eliminating operating system abstractions [6] has been studied by many authors and the benefits include significant performance improvements as shown in Exokernel [7], Micro-kernel [9], lean kernel [18] and OS-Kit [19]. The BMC approach in contrast completely avoids any centralized OS or kernel. This results in the BMC paradigm wherein an application programmer has sole control of the application and its execution environment. Multimedia applications can also be built using the BMC paradigm by extending the Application Object (AO) [12] concept to develop an application graphics object (AGO) model comprising graphics, voice and video. The AGO run on a bare PC without the support of any OS or kernel. The AGO provides direct hardware communication interfaces to AO programmers thus eliminating all the abstraction layers introduced by OSs and their environments. Direct BMC hardware interfaces for C/C++ applications are described in [11]. These interfaces enable program load, screen display, mouse and keyboard access, process management, and network and audio card control. This paper describes new BMC interfaces constituting a hardware API for graphics applications.

There has been considerable research and significant advances in the areas of graphics and multimedia. In [5], an OpenGL-based scalable parallel rending framework that provides a graphics API was discussed. In [20], two user interfaces for interactive control of dynamically-simulated character using embedded system platforms were demonstrated. A lean mapping graphics interfaces that uses a method for real-time filtering of specular highlights in bump and normal maps was described in [16]. All such approaches require conventional OS-based platform support. A comprehensive low-level graphics design and implementation was described in [17]. However, these graphics interfaces use DOS (Microsoft Disk Operating System) primitives and interrupt 21h, which require DOS environment. In bare PC applications, only required interrupts are used and included with the application. At present, there appears to be no direct hardware API for multimedia applications that can run on a bare PC with no OS support.

IV. Architecture

A. Architectural Decription

The BMC Graphics architecture differs in many aspects from that of conventional or embedded graphics systems as the interfaces are directly accessible by the AO programmer. A given interface executes without any interruption as a single thread of execution. The AO programmer can control activation, suspension and resumption of this thread at program time. Figure 1 illustrates our graphics architecture that is suitable for any IBM PC based system. Currently, this architecture has been implemented using only Intel IA32 processor-based PCs.

An application programmer writes a graphics application (e.g., animation, visualization) application object (AO) in C++ or C using the direct hardware graphics interfaces (API). These interfaces are provided by the application graphics object (AGO). The AGO implements high level application logic if needed and sets up parameter passing for shared memory. The AGO invokes “C” language interfaces (using extern “C” {}) to invoke C calls from C++. The C calls in turn will invoke assembly calls for a given graphics interface. The assembly call then calls a graphics API software interrupt (int 0xfa). The AO, AGO, C, assembly calls have full access through a memory interface to read or write data in shared memory. This is accomplished by using a MEMDataSel selector that allows access to shared memory in real and protected modes using zero base select value. All of this code is executed in protected mode.

The software interrupt above is an interrupt gate that takes the call to real mode. The graphics interfaces are implemented in assembly code that run in real mode. These interfaces in real mode have access to video memory as shown in Figure 1. PC BIOS interrupts are also used to control video memory and graphics modes. Interrupt descriptor table (IDT), global descriptor table (GDT), local descriptor table (LDT), task state segment (TSS), boot, loader, interrupt service routines are all part of an AO. The AO is a self-contained, self-managed and self-executable module. The AO programmer has sole control of the facilities that are needed to run a given application.

The graphics architecture views the screen as just a multi-dimensional array of pixels that can be represented using vector-based mathematical models. This differs from conventional approaches that deal with each pixel every time the graphic changes.

The BMC graphics interfaces reduce complexity by providing a pointer to video memory that dynamically binds the interfaces at the hardware level to the video memory buffer. In this approach, there is no need to synchronize the GPU with the CPU functions [4] as done in a conventional system.

The AGO architectural design is classified into eight broader categories:

- *Application Program (AO) Protected Mode:* actual user program main() executes here.
- *Application Graphics Object (AGO):* software and hardware API used by AO.
- *C-Programming Interface:* used as a gateway between C++ and ASM language abstracts.
- *PC Assembly Interface:* used to interact with real-mode shared memory.
- *Software Interrupt*: Interrupt to bridge between real and protected mode dynamically.
- *Interrupt Gate to Real Mode:* used in real-mode to interact directly with the Hardware Video Memory.
- *Graphics Operations Real Mode:* used to invoke low-level graphics primitives including screen access, screen framework, font/symbol, and other attributes as outlined in section V.

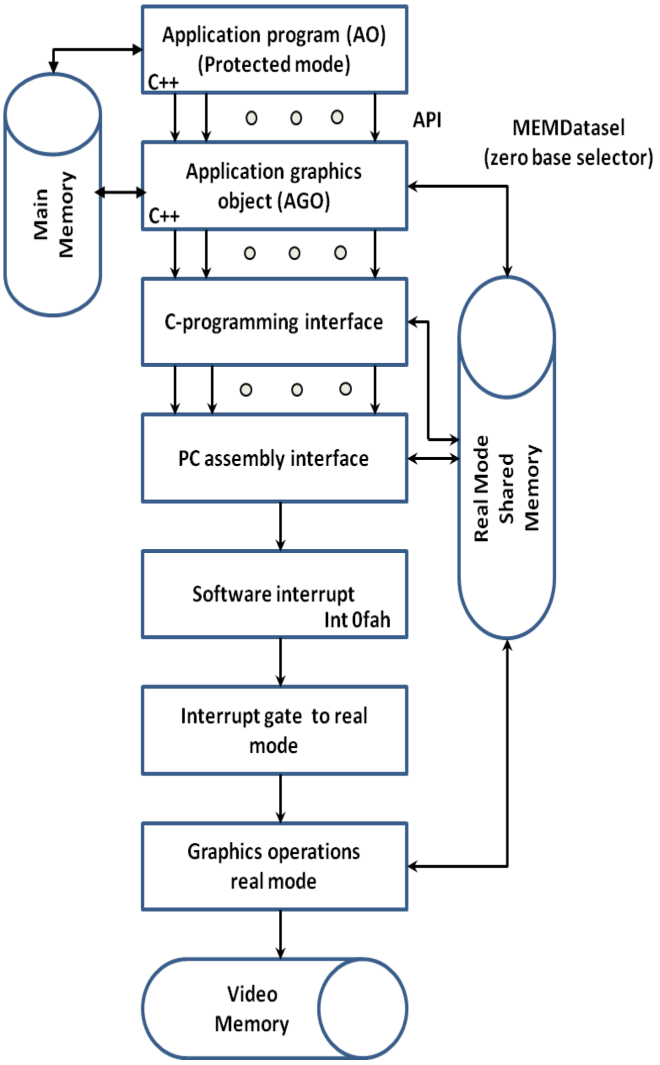


Figure 1. BMC Graphics Architecture

B. Architecural Novelties

The lean graphics architecture has many novel characteristics. The interfaces developed here can be directly invoked from standard C or C++ programs and are fully controlled by the programmer. The architecture is very generic and can be implemented on any pervasive device. As this approach uses only video memory instead of graphics card interfaces, a given graphics application can be ported easily to many types of devices. The current AGO does not use any graphics accelerators and hardware support, bare interfaces to audio, video and graphics cards can be written if necessary to resolve any performance issues. When more sophisticated bare graphics and multimedia becomes feasible, users can carry their own USB with a Web client and browse the Web on any bare device without carrying any hardware and with no reliance on any OS, kernel or environment. In addition to being convenient for users, this graphics architecture eliminates overhead and may be easier to secure due to its simplicity.

V. Design and Implementation

A Bare PC graphics system is designed to perform graphics functions, such as drawing geometric figures with fraction (fractal) primitives, displaying text characters, and performing other attributes such as color, pixel, line, circle and displaying a bitmap image.

The following graphics functions are implemented using standard C and Intel assembly language. They are based on the design and implementation principles in [17] and modified to work with a bare PC system as described below:

- *Screen access:* clear screen, set the entire screen to a color or attribute, save the screen image in memory, and restore a saved screen image.
- *Screen framework:* set a shape screen area to a given color or attribute, save and retrieve a screen area in video memory.
- *Font/Symbols:* based on Vector / ASCII
- *Images/video:* based on pixel and compression
- *Shapes:* based on Vector/Pixel Mathematical Algorithms.
- *Attributes:* set the current drawing color, set the current fill color, set the current shading attribute, set the current text color, set the current text font, set the current line type (continuous, dotted, dashed, etc.), and set the current drawing thickness.
- *Image transformation:* scale, rotate, translate, and clip image.
- *Bit operations for performance:* BIT Shifters; XOR, OR, NOT and AND bitwise operations.

In order to illustrate our implementation, we describe five basic direct graphics APIs:

(1) **draw_pixel():** This interface takes x and y coordinates of a given pixel and computes its video memory location. The video memory address, location of a pixel, color of a pixel and its "opcode" are stored in shared memory. As illustrated in Figure 1, this interface goes from protected mode to real mode to the graphics operations code. It then obtains the pixel parameters from shared memory and places the pixel in the video memory. After displaying the pixel on the screen, it will return to its AO. The entire API process is executed as a single thread of execution. More optimal approaches to drawing a pixel on the screen will be considered later.

(2**) draw_line():** This is simply drawing many pixels to draw a line using Bresenham's algorithm [3].

(3) **draw_box():** This API uses draw_line() API repeatedly to plot a box.

(4) **draw_circle():** The circle is implemented without using the sine and/or cosine functions. It uses the algorithm described in [8] and uses draw_pixel() API.

(5) **draw_bitmap():** First, the bitmap file is read from the removable device (USB) during the program load and stored in main memory. Second, the bitmap file header is parsed for size, color and image data offset location parameters as shown in Figure 3. Third, the display mode is setup to match minimum color palette requirements for the image as shown on Figure 2. Fourth, the video memory address, pointer to the bitmap image data and its opcodes are stored in shared memory as illustrated in Figure 1. Finally, the AGO will copy the image data from shared memory to video memory for display as shown in Figure 3. After placing the bitmap on the screen, it will return to its AO.

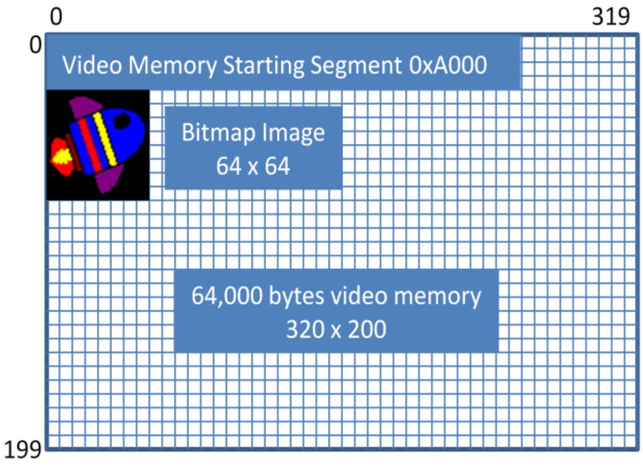

0
319
0
199
Video Memory Starting Segment 0xA000
Bitmap Image
64 x 64
64,000 bytes video memory
320 x 200

Figure 2. Video Memory Layout

Video memory is a contiguous linear addressing model, which differs from the x and y coordinates of the computer screen. To plot a pixel, the offset is calculated from the beginning of the video memory as follows: y coordinate multiplied by the total width of the screen and the x coordinate added to it.

In the example shown in Figure 2, we use the VGA Mode 0x13 with screen dimension of 320 pixels in width and 200 pixels in height. This translates to 0 to 319 on the x axis (width) and 0 to 199 on the y axis (height). The top left corner starts at coordinate (0, 0). Each pixel represents 8 bits (1 byte). Thus, the memory needed to store images of this size (320x200) is 64,000 bytes.

Figure 3 shows the Bitmap Image structure. It consists of several components that are described below.

The File Header in Figure 3 contains the FileType which starts with 4D42h ("BM"). FileSize is the Size of the image file in bytes. Reserved fields are used for future enhancements, with default values set to 0. The BitmapOffset stores the starting position of image data in bytes. The total size of the File Header is 14-bytes.

"Size" is the size of this header in bytes, and Width and Height are the Image width and height in pixels. A plane is the number of color planes and BitsPerPixel is the number of bits per pixel. The total size of the Bitmap Header is 40-bytes.

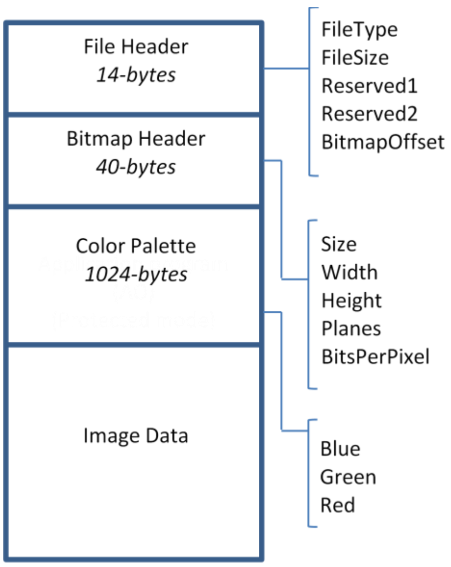


Figure 3. Bitmap Image Format

The BMC Color Palette specifies the red, green, and blue values of each pixel in the bitmap data by storing a single value used as an index into the color palette. In the newer versions of the BMP standard, the Color Palette and Image Data are merged together. In our example, we are using the Image Data directly, since we have pre-defined our palette to 256-colors. The total size of the Color Palette is 1024-bytes.

The pseudo-code in Figure 4 is used to display a 64x64 bitmap to video memory; since video memory is linear, a simple computation based on the x and y values can be used.

```
void draw_bitmap(AO, AGO Object Reference)
Initialize Memory Locations;
Initialize Variables;

Loop While screen_height < 200
{
 For (y=0; y < image_width; y++) {
    For (x=0; x < image_height; x++){
     //AGO Implementation, copy to Video Memory

        Video_Memory_Pointer [x+y*320]=
             Shared_Memory_Pointer[x+y*64] } }
}
```

Figure 4. Bitmap to Video Memory Pseudo-Code

VI. Testing

The testing was conducted on a standard VGA graphics card and VESA enabled BIOS on VGA Mode 13, with 320-by-200 pixel resolution in 256-Colors. The graphics was tested on a Dell Optiplex GX260 PC with 512 MB memory.

The preliminary response time in Table I was conducted using the UNIX system time; it is an end-to-end measurement, which includes other components such as the VGA hardware device and other display intermediaries.

Below, we illustrate five basic primitives as described in Section V. Each API is a direct hardware interface available to the AO programmer that can be invoked directly from C or C++ code.

TABLE I. BMC Graphics – Preliminary Response Time

AGO Object	Response Time (Microseconds)
draw_pixel() graph Figure 5	2.25
draw_line() graph Figure 6	83.25
draw_ circle() graph Figure 7	250
draw_bitmap() graph Figure 8	5

A. Pixels

Using the draw_pixel() AGO API, 5000 pixels were plotted as shown in Figure 5. The x, y coordinates and color were chosen randomly. Preliminary performance tests shows a response time of 2.25 microseconds.

Figure 5. BMC Graphics - 5,000 Random Pixels and Colors

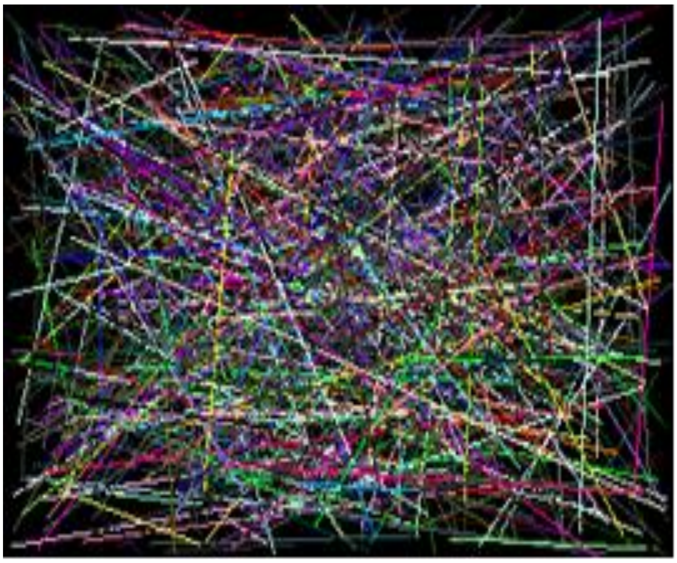


Figure 6. BMC Graphics - 5,000 Random Lines and Colors

B.Line

Using the draw_line() AGO API, 5000 lines were rendered as shown in Figure 6. The x1, x2, y1, y2 coordinates and color where chosen randomly. The draw_circle API is a direct hardware interface inherited from the draw_pixel() AGO. Preliminary performance tests shows a response time of 83.25 microseconds.

C.Circle

Using the draw_circle() AGO API, 5,000 circles were rendered as shown in Figure 7. The x, y coordinates, radius size and color were chosen randomly. The draw_circle API is a direct hardware interface inherited from the draw_pixel() AGO. Preliminary performance tests shows a response time of 250 microseconds.

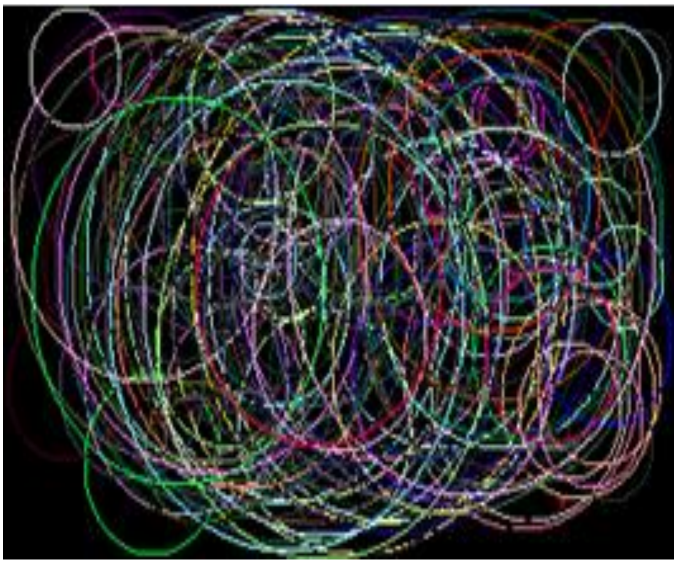


Figure 7. BMC Graphics - 5,000 Circle with Random Size and Color

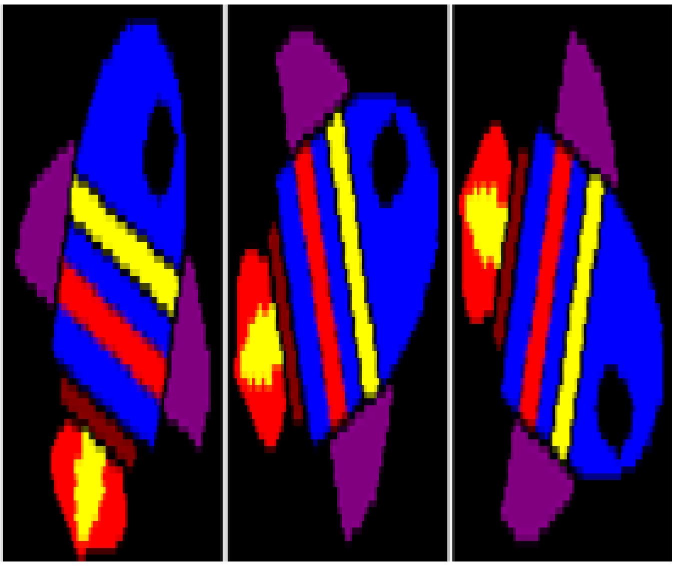


Figure 8. BMC Graphics – Bitmap

D. Bitmap

The bitmap used in this example is a 64 by 64 256-color bitmap with 8 bits per pixel, the file format is Windows RGB-encoded BMP format uncompressed. For a 256-color bitmap, there is a 54-byte header and a 1024-byte palette table in addition to the actual bitmap data.

Using the draw_bitmap() AGO API, the bitmap image was loaded into video memory directly for display. The bitmap was loaded three times with different orientation to show rotation and animation. This can be achieved by changing the pixel loading order while copying to memory as shown on Figure 8. Preliminary performance tests shows a response time of 5 microseconds.

VII. CONCLUSION

We presented the architectural design for building graphics applications that can run on bare devices with no operating system, kernel or environment support. We also gave details of its implementation model using C/C++. The AGO API illustrates some of the fundamental graphics elements and their functionality. The preliminary performance data indicates applicability of bare machine graphics for complex graphics applications. The direct hardware graphics API can be used in a variety of pervasive devices to achieve common graphics operations. We have also presented the benefits of running graphic applications on a bare PC including simplicity, elimination of abstraction layers, and self-containment. With bare PC graphics, the programmer has direct access to the video graphics device and complete control of all hardware resources enabling autonomy with performance advantages due to elimination of system overhead.

ACKNOWLEDGMENT

We sincerely thank NSF and in particular the late Dr. Frank Anger, who initially supported this work by funding through SGER grant CCR-0120155. Without his encouragement, bare machine computing concept and explorations could not have been possible.

REFERENCES

[1] A. Alexander, A. L. Wijesinha, and R. Karne, "A Study of Bare PC SIP Server Performance," The Fifth International Conference on Systems and Networks Communications, ICSNC, Nice, France, pp. 392 – 397, August 2010.

[2] A. Alexander, A. L. Wijesinha, and R. Karne, "Implementing a VOIP Server and a User Agent on a Bare PC," The Second International Conference on Future Computational Technologies and Applications, Future Computing, Portugal, Lisbon, pp. 8 – 13, November 2010.

[3] D. Brackeen, Developing Games in Java. Berkeley, CA: New Riders Games, pp. 63-70, 2003.

[4] D. Salomon, The Computer Graphics Manual, Ithaca, NY: Springer-Verlag Publisher, pp. 200 – 240, 2011.

[5] S. Eilemann, M. Makhinya and R. Pajarola. “Equalizer: A Scalable parallel rending framework,” IEEE Transactions on Visualization and Computer Graphics, pp. 436 – 452, June 2008.

[6] R. Engler and M.F. Kaashoek, "Exterminate all operating system abstractions," In Fifth Workshop on Hot Topics in Operating Systems, pp. 78, May 1995.

[7] D. Engler, "The Exokernel Operating System Architecture," Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Ph.D. Thesis, 1998.

[8] J. E. Frith. “Fast Circle Algorithm,” http://www.tutego.de/aufgaben/j/insel/additives/base/fcircle.txt, Copyright (c) 1996 James E. Frith, Email: jfrith@compumedia.com [Retrieved: March, 2011].

[9] B. Ford, M. Hibler, J. Lepreau, R. McGrath, and P. Tullman, “Interface and execution models in the Fluke Kernel,” Proceedings of the Third Symposium on Operating Systems Design and Implementation, USENIX Technical Program, New Orleans, LA, pp. 101-115, February 1999.

[10] R. K. Karne, K.V. Jaganathan, T. Ahmed, and N. Rosa, “DOSC: Dispersed Operating System Computing,” OOPSLA, 20th Annual ACM Conference on Object-Oriented Programming, Systems, Languages, and Applications, Onward Track, Sandiego, CA, pp. 55-61, October 2005.

[11] R. K. Karne, K. Venkatasamy and T. Ahmed, "How to run C++ applications on a bare PC," In proceeding of 6th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel / Distributed Computing (SNPD), pp. 50 – 55, May 2005.

[12] R. K. Karne, "Application-oriented Object Architecture: A Revolutionary Approach," In 6th International Conference, HPC Asia, Poster presentation, December 2002.

[13] N. Kazemi, A. L. Wijesinha, and R. Karne, "Evaluation of IPsec Overhead for VoIP using a Bare PC," 2nd International Conference on Computer Engineering and Technology (TCCET), vol. 2, pp. 586 – 589, April 2010.

[14] G. Khaksari, A. L. Wijesinha, R. K. Karne, L. He, and S. Girumala, “A Peer-to-Peer Bare PC VoIP Application,” IEEE Consumer Communications and Networking Conference, Seamless Consumer Connectivity, CCNC, Las Vegas, Nevada, pp. 803 – 807, January 2007.

[15] P. Kovach and J. Richter, Inside Direct3D, Redmond, WA: Microsoft Press, 2000.

[16] M. Olana, and D. Baker, “Lean Mapping,” Proceedings of 2010 ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games, pp. 181-188, February 2010.

[17] J. Sanchez and C. Maria, Computer Animation Programming Methods & Techniques, McGraw-Hill, 1995.

[18] “Tiny OS,” Tiny OS Open Technology Alliance, University of California, Berkeley, CA, 2004, http://www.tinyos.net/ [Retrieved: June, 2007].

[19] "The OS Kit Project," School of Computing, University of Utah, Salt Lake City, UT, June 2002, http://www.cs.utah.edu/flux/oskit [Retrieved: May, 2009].

[20] P. Zhao and M. Van de Panne, “User interfaces for interactive control of physics-based 3D characters,” Proceeding of the 2005 symposium on Interactive 3D graphics and games, pp. 87 – 94, April 2005.

Formal Performance Measures for Asymmetric Communication

Paulius Tervydis
Department of Telecommunications
Kaunas University of Technology
Kaunas, Lithuania
paulius.tervydis@ktu.lt

Ramutis Rindzevicius
Department of Telecommunications
Kaunas University of Technology
Kaunas, Lithuania
ramutis.rindzevicius@ktu.lt

Jonas Valantinas
Department of Applied Mathematics
Kaunas University of Technology
Kaunas, Lithuania
jonas.valantinas@ktu.lt

Abstract **— Typically, any asymmetric network is characterized either by non-uniform link transmission bit rate or by uneven traffic intensity. Through monitoring and asymmetric network status control one can improve performance of the service-sector. In this paper, some exact analysis methods, oriented to achieve comparatively high utilization of the data packet transmission link and satisfy the service quality of the asymmetric loss and queueing systems, are proposed. The developed mathematical approach turns out to be extremely useful for the analysis of the network node with asymmetric transmission links. Some comments concerning the application of a few different strategies to the selection of an unoccupied data packet transmission link are discussed. The straight process analysis in the asymmetric loss and queueing systems is carried out using continuous-time Markov chains. Performance of the asymmetric system is measured using accurate expressions. Finally, in the case of Poisson arrivals and exponential transmission times, an exact analytical model is applied to the system.**

Keywords - queueing system; quality of service; asymmetric system; Markov process.

I. INTRODUCTION

Computation of the performance measures is simple enough if all data packet transmission links have the same parameters and packet transmission times are distributed exponentially. Though, in the data transmission network, heterogeneous transmission links appear often and often. Performance of the data transmission over an asymmetric system with multiple links depends on a particular link that transmits a data packet. Therefore, it is highly expedient and useful to compute accurately the performance measures of the said asymmetric systems. The more detailed description of the latter systems is presented in [1, 2]. In [3], it is shown that a well-known approach to investigating symmetric M/M/m and GI/G/m systems (Kendall's notation is used) can be applied also to the suchlike asymmetric systems, provided the service rates of data packet transmission links differ only slightly (by a ratio <10) and the system utilization is high enough. Some interesting methods, used to compute several characteristics of the asymmetric queueing systems, are described in [4] wherein computations reduce to solving equations associated with continuous-time Markov chains. In such a way, some accurate results are obtained for elementary asymmetric lossless M/M/m queueing systems. In [5], the asymmetric finite capacity queueing systems are analyzed with the use of analytical and simulation models based on Moore and Mealy automata. In [6], performance measures for an asymmetric node, with a priority flow and two data transmission links, are estimated applying analytical and simulation models. Advanced resources sharing methods, in asymmetric networks, are proposed in [7]. In [8], Lakshman et al. produce the network control mechanism facilitating and supporting TCP/IP data transmission over the asymmetric networks. Authors determine the throughput as a function of buffering, and state conditions under which the transmission link is fully utilized. In [9], Krithikaivasan et al., employing control and routing, outline rigorously how to improve performance in congested parts of the asymmetric network.

Not going into minor details, we here emphasize that asymmetric systems, as well as their performance analysis, are far from being investigated thoroughly. Further research is necessary.

The rest of the paper is organized as follows. Section II introduces an asymmetric loss system. Section III evaluates data packets arrival rate impact on the performance measures of the asymmetric loss system. Analysis of the data packet rate impact on the performance measures of the asymmetric queueing system is presented in Section IV. Some conclusive commentary is presented in Section V.

II. ANALYSIS OF THE ASYMMETRIC LOSS SYSTEM

An asymmetric loss system with different rates of the data packet transmission links is modelled. The functional diagram of the system is presented in Fig. 1.

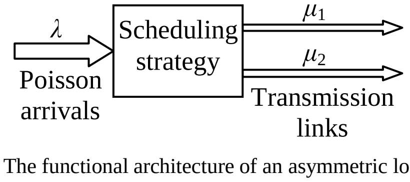


Figure 1. The functional architecture of an asymmetric loss system.

The best performance measures of the system, such as the link utilization and the data packet loss, are obtained by switching between transmission links.

Let us denote the data packet transmission rate over a link i (i=1, 2) by μ_i. We shall take an exact analytical model of the asymmetric loss system with Poisson data packet arrivals (intensity λ) and exponential data packet transmission time over each link (intensities μ_1 and μ_2, $\mu_1 > \mu_2$). Such an asymmetric system can be represented as the continuous-time Markov chain (the system itself being in a stable state). Two parameters (components of the vector *XY*) are attached to each state of the system, where *X* represents the state of the data packet transmission link 1, and *Y* represents that of the data packet transmission link 2. If *X* or *Y* equals zero the respective link is free (unoccupied), otherwise (*X* or *Y* equals 1), the respective link is busy. Let us consider a loss system, provided a few different strategies for selecting transmission link are applied (Fig. 2 – Fig. 4).

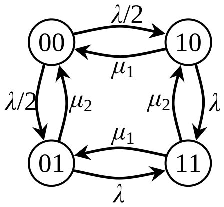


Figure 2. Markov process for the asymmetric loss system (the data packet transmission links are occupied, with the same intensities; *Case* 1).

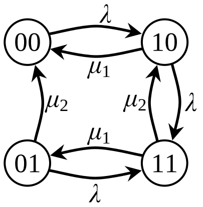


Figure 3. Markov process for the asymmetric loss system (the faster transmission link 1 is occupied first; *Case* 2).

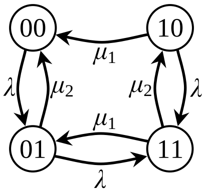


Figure 4. Markov process for the asymmetric loss system (the slower transmission link 2 is occupied first; *Case* 3).

The usage of the global balance concept for the Markov chains enables us to put down the following equations (for evaluation of the system state probabilities P_{XY}):

In *Case* 1 (Fig. 2),

$$\begin{cases} P_{00} + P_{10} + P_{01} + P_{11} = 1, \\ \lambda P_{00} - \mu_1 P_{10} - \mu_2 P_{01} = 0, \\ (\lambda + \mu_1) P_{10} - \dfrac{\lambda}{2} P_{00} - \mu_2 P_{11} = 0, \\ (\lambda + \mu_2) P_{01} - \dfrac{\lambda}{2} P_{00} - \mu_1 P_{11} = 0, \\ (\mu_1 + \mu_2) P_{11} - \lambda P_{10} - \lambda P_{01} = 0. \end{cases} \tag{1}$$

In *Case* 2 (Fig. 3),

$$\begin{cases} P_{00} + P_{10} + P_{01} + P_{11} = 1, \\ \lambda P_{00} - \mu_1 P_{10} - \mu_2 P_{01} = 0, \\ (\lambda + \mu_1) P_{10} - \lambda P_{00} - \mu_2 P_{11} = 0, \\ (\lambda + \mu_2) P_{01} - \mu_1 P_{11} = 0, \\ (\mu_1 + \mu_2) P_{11} - \lambda P_{10} - \lambda P_{01} = 0. \end{cases} \tag{2}$$

In *Case* 3 (Fig. 4),

$$\begin{cases} P_{00} + P_{10} + P_{01} + P_{11} = 1, \\ \lambda P_{00} - \mu_1 P_{10} - \mu_2 P_{01} = 0, \\ (\lambda + \mu_1) P_{10} - \mu_2 P_{11} = 0, \\ (\lambda + \mu_2) P_{01} - \lambda P_{00} - \mu_1 P_{11} = 0, \\ (\mu_1 + \mu_2) P_{11} - \lambda P_{10} - \lambda P_{01} = 0. \end{cases} \tag{3}$$

The asymmetric system state probabilities P_{XY} are obtained by solving the above linear systems. In particular, one can easily find some other system performance measures, such as:

- the data packet transmission link utilization

$$\begin{aligned} \rho_1 &= P_{11} + P_{10}, \\ \rho_2 &= P_{11} + P_{01}; \end{aligned} \tag{4}$$

- the data packet loss probability

$$P_{loss} = P_{11}. \tag{5}$$

III. Data packets arrival rate impact on the performance measures of the asymmetric loss system

Performance measures of the asymmetric loss system, represented in the form of a function of the data packets arrival rate λ, are shown in Fig. 5 and Fig. 6.

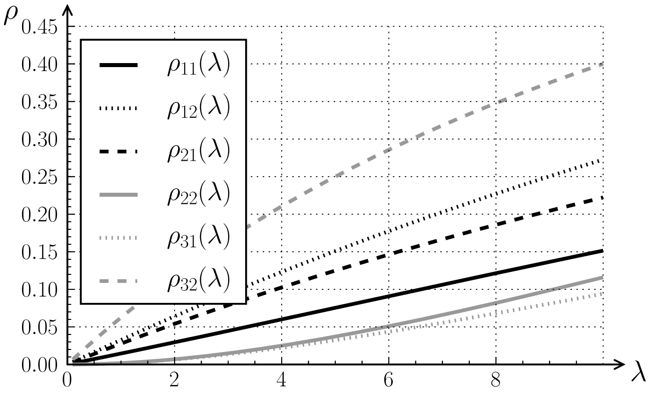


Figure 5. The data packet transmission link utilizations as a function of λ (*Cases* 1,2,3; μ_1=35 and μ_2=15).

In Fig. 5, the dependence of the data packet transmission link utilizations ρ_{i1} and ρ_{i2} on the strategy used to select an unoccupied data transmission link i $(i=1,2,3)$ facilitates selection of the data transmission link (*Cases* $1,2,3$).

The data packet loss probability P_{loss} attains its maximal value in *Case* 3 and minimal value in *Case* 2 (the faster transmission link is occupied first; Fig. 6).

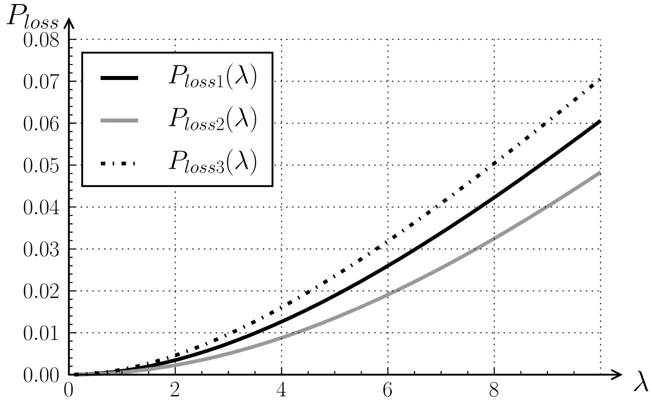


Figure 6. The data packet loss probabilities as a function of the data packet arrival rate λ, assuming different data packet transmission link scheduling strategies are applied (*Cases* 1,2,3; μ_1=35, μ_2=15).

We here observe that the analytical model of the queueing system is accurate only in the case of Poisson arrivals and exponential data packet transmission times (in the links).

IV. Data Packets Arrival Rate Impact on the Performance Measures Of The Asymmetric Queueing System

In this section, an asymmetric queueing system with two data packet transmission links is analysed. To estimate the queueing system performance measures, an exact analytical model has been developed. The queueing system itself is characterized by different data packet transmission rates μ_1, μ_2 and a finite buffer of size K (the lower part; Fig. 7).

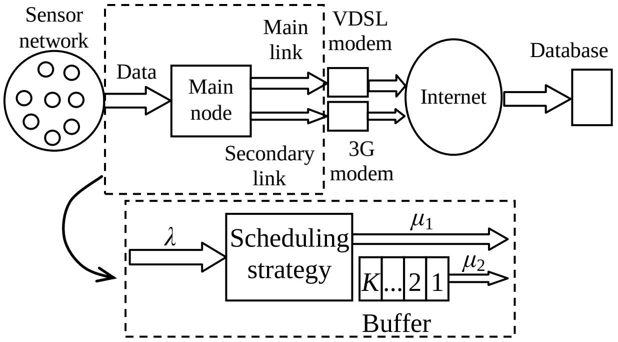


Figure 7. Functional architecture of the asymmetric queueing system.

For instance, the above model can be applied to evaluating performance measures of the main node of a sensor network. As the basis for calculations, the situation shown in the upper part of Fig. 7 is chosen. The data flow, from the network of sensors, is directed to the main node which sends the data over the Internet to the remote database. The main node has two data transmission links: the primary link that is connected to the Internet over VDSL modem, and the secondary link that is connected to the Internet over 3G modem. The mean data transmission rates over the primary and the secondary links are equal to C_1=4Mb/s=500000B/s and C_2=2Mb/s=250000B/s, respectively. The secondary link is used if and only if the primary link is busy. The mean length of the data packet equals L=1000B. Therefore, the buffer of B=16KB can store up to K=B/L=16 data packets. The data packet transmission intensities over the first and the second links equal μ_1=C_1/L=500 packets/s and μ_2=C_2/L=250 packets/s, respectively.

Consider the Poisson data packet arrival flow, with intensity λ, and the data packet transmission time (over each link) distributed exponentially. The data packet transmission link scheduling strategy is such that the link with data packet transmission rate μ_1 is occupied first. The data packet from the buffer (finite capacity) is transmitted only over the second transmission link with transmission rate μ_2. A birth-and-death Markov model of the suchlike queueing system is shown in Fig. 8.

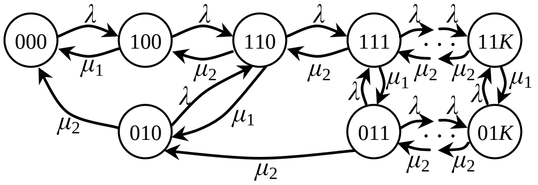


Figure 8. The continuous-time Markov chain for the asymmetric queueing system.

Each steady state of the system is described using three parameters (components of the vector XYZ), where X represents the state of the first link (0 – unoccupied, 1 –busy) Y represents that of the second link and Z represents the number of data packets in the buffer (in our case, from 0 to K).

For finding the system state probabilities, the following system of algebraic equations is used:

$$\begin{cases} \lambda P_{000}-\mu_1 P_{100}-\mu_2 P_{010}=0; \\ (\lambda+\mu_1)P_{100}-\lambda P_{000}-\mu_2 P_{110}=0; \\ (\lambda+\mu_2)P_{010}-\mu_1 P_{110}-\mu_2 P_{011}=0; \\ (\lambda+\mu_1+\mu_2)P_{110}-\lambda P_{100}-\lambda P_{010}-\mu_2 P_{111}=0; \\ (\lambda+\mu_2)P_{011}-\mu_1 P_{111}-\mu_2 P_{012}=0; \\ (\lambda+\mu_1+\mu_2)P_{111}-\lambda P_{011}-\lambda P_{110}-\mu_2 P_{112}=0; \\ (\lambda+\mu_2)P_{012}-\mu_1 P_{112}-\mu_2 P_{013}=0; \\ (\lambda+\mu_1+\mu_2)P_{112}-\lambda P_{111}-\lambda P_{012}-\mu_2 P_{113}=0; \\ ------------------ \\ (\lambda+\mu_2)P_{01K}-\mu_1 P_{11K}=0; \\ (\mu_1+\mu_2)P_{01K}-\lambda P_{11(K-1)}-\lambda P_{01K}=0. \end{cases} \qquad (6)$$

The obtained state probabilities P_{XYZ} can be applied to finding performance measures of the above asymmetric system, such as:

- data packet loss probability

$$P_{loss} = P_{11K}\,; \tag{7}$$

- data packet transmission link utilizations

$$\rho_1 = P_{100} + \sum_{i=0}^{K} P_{11i}\,; \tag{8}$$

$$\rho_2 = \sum_{i=0}^{K} P_{11i} + \sum_{i=0}^{K} P_{01i}. \tag{9}$$

Let us denote the data packet arrival (to the first and the second links) intensities by λ_1 and λ_2, respectively. Then the link utilization can be alternatively computed this way:

$$\rho_1 = \lambda_1 / \mu_1\,;\ \rho_2 = \lambda_2 / \mu_2\,; \tag{10}$$

here

$$\lambda_1 = \lambda(P_{000} + \sum_{i=0}^{K} P_{01i})\,; \tag{11}$$

$$\lambda_2 = \lambda(P_{100} + \sum_{i=0}^{K-1} P_{11i})\,. \tag{12}$$

The average number of the data packets in the buffer equals

$$\overline{N_q} = \sum_{i=1}^{K} i \cdot P_{11i} + \sum_{i=1}^{K} i \cdot P_{01i}\,. \tag{13}$$

The mean waiting time value (for the data packet) in the queue is obtained in accordance with Little's theorem, i.e.

$$\overline{W} = \frac{\overline{N_q}}{\lambda_2} = \frac{\overline{N_q}}{\mu_2 \rho_2}. \tag{14}$$

The probability that a new data packet will enter the queue is given by

$$P_{wait} = P(W > 0) = \sum_{i=0}^{K} P_{11i}\,. \tag{15}$$

The average number of data packets in the asymmetric queueing system

$$\overline{N_s} = \overline{N_q} + \rho_1 + \rho_2\,. \tag{16}$$

The average time, spent by the data packet in the asymmetric queueing system, equals

$$\overline{T_s} = \frac{\lambda_1}{\mu_1(\lambda_1+\lambda_2)} + \frac{\lambda_2}{\mu_2(\lambda_1+\lambda_2)} + \overline{W}. \tag{17}$$

Performance measures of the queueing system, expressed in the form of a function of the queueing system parameters λ and *K*, are shown in Fig. 9 - Fig. 15.

The data packet loss probability increases considerably when the data packet transmission link utilization achieves 0.5 (λ>400) (Fig. 9).

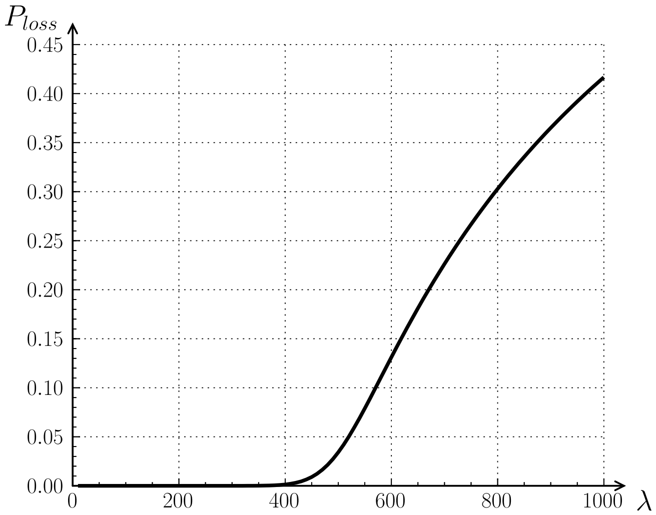


Figure 9. The data packet loss probability P_{loss} as a function of λ, given μ_1=500, μ_2=250, *K*=16.

The data packet loss in the system occurs if and only if the buffer is full. The data packet loss probability can be lessened in several ways: by increasing the buffer capacity, by increasing the data transmission rate over the links or by decreasing the data packet arrival rate. In the given example, the decrease of the data packet arrival rate is achieved by limiting the number of data collection sensors. For instance, if one of the data sensors produces 10 data packets per second, then the main node can serve 40 sensors with minimal risk of data packet loss.

The link utilization level turns out to be another important concern. The right estimation of the link utilization level is used to guarantee that the packet loss will not occur. Also, the estimation results can be used to evaluate economic aspects of the link usage. For instance, the cost of the data transmission over the secondary link over the 3G Internet connection can be higher. Thus, the given model can be explored to estimate how intensively the links will be used, calculate the usage price or make a decision concerning data transmission rates operable in the links.

In Fig. 10, dependence of the link utilization on the data packet arrival intensity λ, is demonstrated. It can be seen that the links are used according to the selected scheduling strategy: first of all, the primary link (with greater data transmission rate) is occupied, the secondary link is used if and only if the first one is busy.

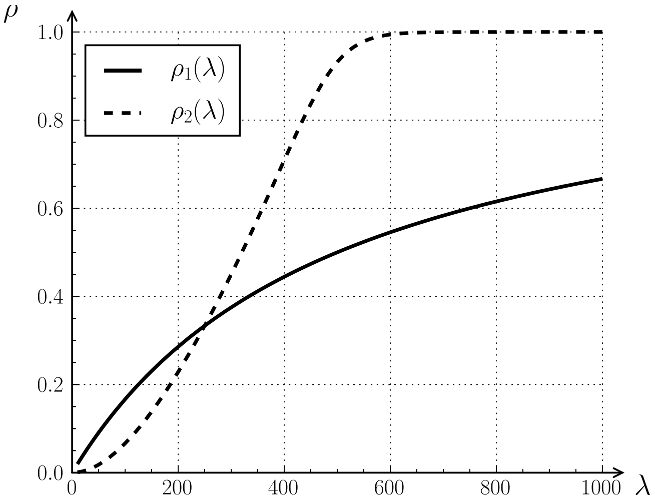


Figure 10. The data packet transmission link utilization ρ_1 and ρ_2 as a function of λ, given μ_1=500, μ_2=250, K=16.

The selected link usage (scheduled) strategy also affects other performance measures.

The average number of data packets in the queue (Fig. 11) rapidly increases when the primary link is busy and the loading of the secondary link goes up.

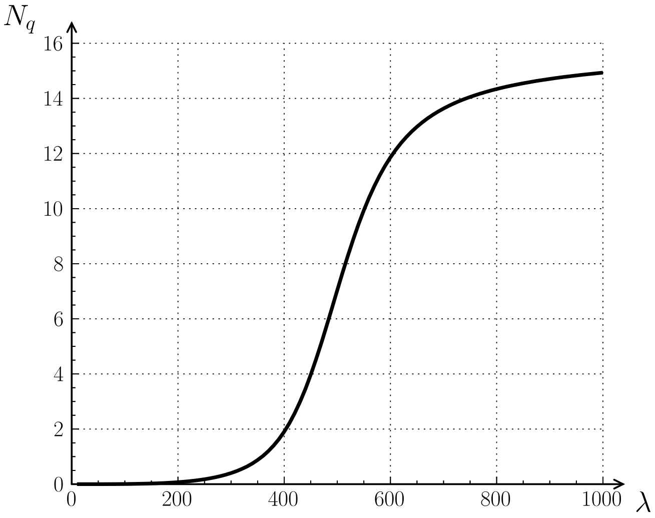


Figure 11. The mean value of data packets in the buffer N_q as a function of λ, given μ_1=500, μ_2=250, K=16.

The probability that an arriving data packet will enter the queue (Fig. 12) also increases when the intensity λ of data packet arrival is increased. Although, it is clear that it should be in this way, but the proposed model gives the exact values, which have an interesting nonlinear fashion.

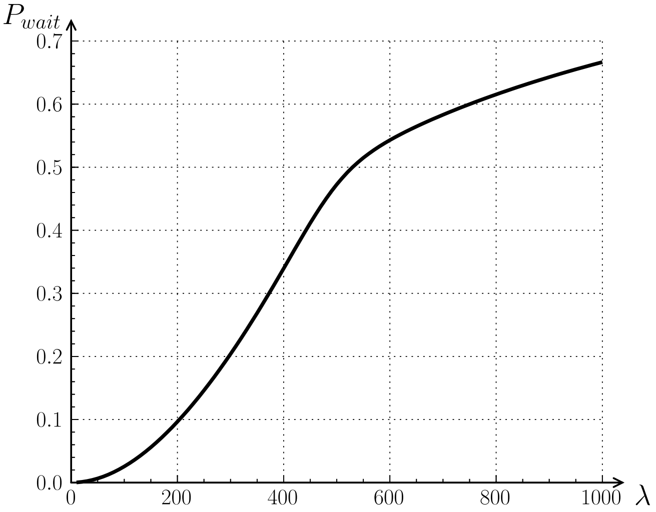


Figure 12. The probability that an arriving data packet will enter the queue $P(W>0)$ as a function of λ, given μ_1=500, μ_2=250, K=16.

The mean values of the time, spent by a data packet in the queue (W) and in the system (T_s), are very important parameters (Fig. 13). Those values facilitate evaluation of the data packet transmission delay or the processing rate.

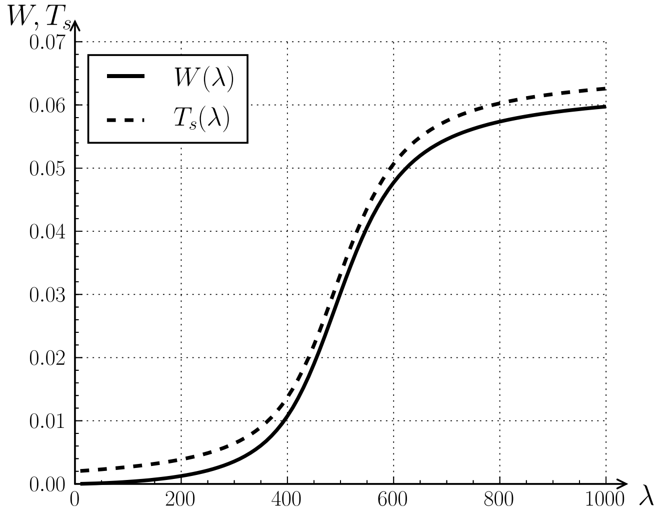


Figure 13. The mean values of the time, spent by a data packet in the queue W [seconds] and in the system T_s [seconds], as a function of λ, given μ_1=500, μ_2=250, K=16.

It is recommended to transmit data packets, which are sensitive to delay, via the link 1 (in the presented asymmetric system).

The size of the buffer also influences the system performance parameters. The influence degree can be estimated using the proposed model.

The data packet loss probabilities P_{loss}, expressed in terms of λ and K, are presented in Fig. 14. As it can be seen, the greater the buffer K, the lesser the probabilities P_{loss}. On the other hand, the difference is negligible, as the packet transmission link utilization approaches 1 (λ>750).

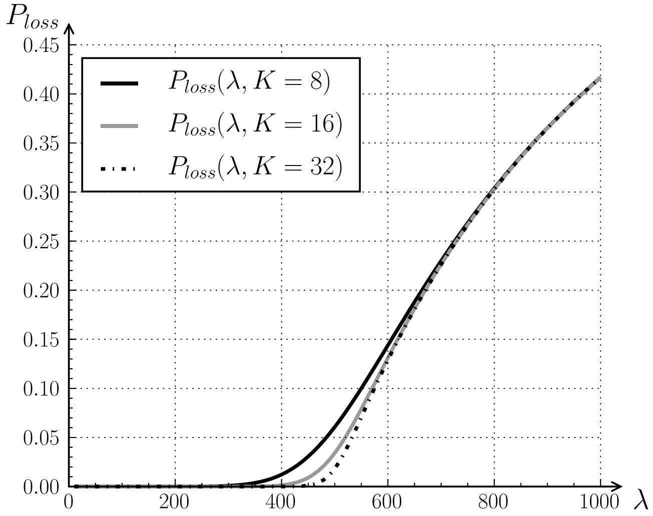


Figure 14. The data packet loss probability P_{loss} as a function of λ and K, given μ_1=500, μ_2=250, K=8,16,32.

The mean values of time spent by a data packet in the system T_s, expressed in terms of λ and K, are presented in Fig. 15.

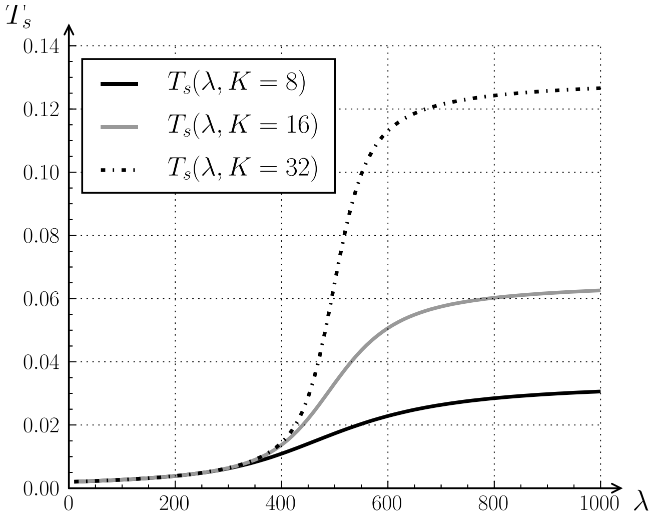


Figure 15. The mean values of time spent by a data packet in the system T_s [seconds] as a function of λ and K, given μ_1=500, μ_2=250, K=8,16,32.

The values of W appear to be greater for greater values of K. It can also be seen that the values of T_s grow apart, as the data packet transmission link utilization approaches 1 (λ>750).

V. Conclusion And Future Work

In the paper, the queueing performance measures, such as the probability of the data packet loss in a finite buffer, the mean queue length, the mean waiting time, the arrival rate impact on the performance measures, are investigated using appropriate analytical models. In the general case (say, non-Poisson data packet flow, non-exponential service time distribution), an exact analytical model turns out to be very complicated. So, simulation is recommended to achieve task-oriented investigation results.

Obviously, the proposed formal approach to the analysis of asymmetric systems is nothing but the starting point for those who are interested in the processes associated with asymmetric data packet transmission systems, i.e. for those specialists who wish to identify new research trends in the area of asymmetric transmission systems for better resource sharing and increasing transmission link performance measures.

Undoubtedly, accurate modelling of the data packet transmission processes in asymmetric systems is truly an important step in optimizing any data transmission network.

Acknowledgment

We would like to express our sincere thanks to a number of our friends and researchers at Kaunas University of Technology who encouraged us to submit the paper. Also, we wish to thank Dr. Egle Rindzeviciute (Sweden) and Dr. Martin Dodsworth (UK) for their assistance in preparing English version of the paper.

References

[1] R. Geist and K. Trivedi, "The Integration of User Perception in the Heterogeneous M/M/2 Queue," in A. Agrawala and S. Tripathi, editors, Proc. Performance, pp. 203-216, Amsterdam, 1983. North-Holland.

[2] K. Trivedi, "Probability and Statistics with Reliability, Queueing, and Computer Science Applications," Prentice-Hall, Englewood Cliffs, N.J., 1982.

[3] G. Bolsh and A. Scheuerer, "Analytische Untersuchungen Asymmetrischer Prioritatsgesteuerter Wartesysteme," in W. Gaul and A. Bachem editors, Operations Research Proc. 1991, pp. 514-521, Stuttgart, Berlin, September 1991, Springer.

[4] M. Baer, "Verlustsysteme mit untersschiedlichen mittleren Bedienungszeiten der Kanale," Wissenschaftliche Zeitschrift der Hoschschule fur Verkehrswesen-Friedrich List, 1985.

[5] A. Zvironiene, Z. Navickas and R. Rindzevicius, "Performance Analysis of an Asymmetric Internet Node"// ITI2007: Proc. of the 29th International Conference Information Technology Interfaces, June 25-28, Dubrovnic, Croatia/ University of Zagreb, 2007. ISBN 9789537138097. pp. 663-670.

[6] R. Rindzevicius, V. Pilkauskas and K. Gvergzdys, "Analysis of an Asymmetric Data Network Node with Priority Call Flows"// ITI2005: Proc. of the 27th International Conference Information Technology Interfaces, June 25-28, Dubrovnic, Croatia/ University of Zagreb, ISBN 953-7138-02, 2005 pp. 513-519.

[7] P. Cheng Fu and C. Soung Lierw, "A Remedy for Performance Degradation of TCP Vegas in Asymmetric Networks" //IEEE Communication Letters, vol.7, No1, January, 2003 pp. 42–44.

[8] T. V. Lakshman, U. Madhow and B.Suter, "TCP/IP Performance with Random Loss and Bidirectional Congestion"//IEEE/ACM Transactions on Networking, Vol 8, No5, October, 2000 pp. 541-555.

[9] B. Krithikaivasan, S. Srivastava, C. Beard, A. Van de Liefvoort and D. Medhi, "Controlling Performance in the Congested Parts of an Aymmetric Network Using Controls and Routing"//Proc. Of the Eight IEEE conference ISCC'03, 2003 pp. 1530-1346.

PMIPv6-based Inter-Domain Handover using Efficient Buffering Scheme

Daeseon Park
Department Of Computer Science and Engineering,
Korea University,
Seoul, South Korea
daesun88@korea.ac.kr

Minsoo Woo / Sung-Gi Min*
Department Of Computer Science and Engineering,
Korea University,
Seoul, South Korea
{geniiwoo, sgmin*}@korea.ac.kr

***Abstract*—Global Handover (HO) requirement between domains is increasingly in demand to provide home network accessibility in remote places. This necessity will lead continuous services between different domains by the Proxy Mobile IPv6 (PMIPv6)-based mobility. In this paper, we introduce the enhanced inter-domain HO mechanism using the efficient buffering scheme to provide continuous services and buffered data backup mechanism for the data delivery to reduce the inter-domain HO delay between different PMIPv6 domains. The simulation result shows that the efficient buffering scheme and the buffered data backup mechanism deliver buffered data among the different domains right after completion of inter-domain handover.**

Keywords-PMIPv6; handover; inter-domain; buffering scheme; backup mechanism.

I. INTRODUCTION

MIPv6 [1] was proposed to meet the global mobility demand by the development of wireless communication technology and the commodity of hand-held mobile devices. IETF introduced Proxy Mobile IPv6 (PMIPv6) for the network-based mobility protocol since MIPv6 has the long Layer 2 and Layer 3 handover latency due to the Mobile IP signaling message involving Mobile Node (MN).

Localized mobility without requiring the MN to participate in any mobility related signaling messages is PMIPv6's [2] purpose, which provides a network-based mobility. Mobile Access Gateway (MAG) typically residing in the access router detects the attachment of the MN to the access link when the MN enters the PMIPv6 domain. MAG verifies if the MN is eligible to the network-based mobility management service by RADIUS [3] or Diameter [4] protocol when a MN attachment is detected. MAG sends a Proxy Binding Update (PBU) including MN information and Proxy-CoA to the Local Mobility Anchor (LMA) for the registration of MN after an authentication & authorization procedure with AAA server. When LMA responds the Proxy Binding Acknowledgement (PBA) message including home network prefix to MAG, bi-direction tunnel is finally established for the data delivery from or to the MN between LMA and MAG. Moreover, Router Advertisement (RA) with home network prefix of MN is sent to MN after home network prefix registration for MN is done in MAG. If the address analysis for the home network prefix received from MAG is successful in MN, MN can eventually communicate with Correspondent Node (CN).

A PMIPv6-based HO and data forwarding in the inter-domain was proposed in [5] for the inter-domain HO solution. Draft NetLMM-Neumann [5] suggests the home LMA as Session Mobility Anchor (SMA) to handle all incoming and outgoing packets for MN during of mobility session. LMA of new PMIPV6 domain initializes a tunnel for SMA in the home domain to continue serving as an anchor point for MN after inter-domain HO procedures. That is, SMA in the home domain acts as the anchoring end point for the LMA of visited domain when the inter-domain HO occurs.

In [6], there was a PMIPv6-based inter-domain roaming scenario just from the home domain to visited domain with two concatenated tunnels between visited MAG and home LMA via visited LMA.

Our proposal will extend [6] to the PMIPv6-based inter-domain HO scenarios including between visited domains and returning to the home domain from the visited domain and the efficient buffering scheme and buffered data backup mechanism will be applied to reduce the inter-domain HO delay.

The remainder of this paper is organized as follows: In Section II, the related works are explained. Section III will describe the proposed schemes. In Section VI, performance of analysis of our scheme is studied. Concluding remarks will be given in Section VII.

II. RELATED WORKS

Soochang [6] introduced a PMIPv6-based inter-domain roaming mechanism using concatenated tunnel construction from visited MAG to visited LMA and from home LMA to visited LMA when MN moves to a visited domain. Visited MAG sends PBU message including home LMA address of MN received from the home AAA response message via visited AAA server when MN is attached to the visited MAG and MN is recognized as roaming user after interacting with the visited AAA server. Inter-domain tunnel is additionally established between visited LMA and home LMA when visited LMA and home LMA exchanges the PBU/PBA message. In [6], two concatenated tunnels are required to roam from the home domain to the visited domain. Two tunnels were composed of one intra-domain tunnel from visited MAG to visited LMA and one inter-domain tunnel from visited LMA to home LMA.

Hyo-beom [7] proposed efficient buffering scheme to prevent the HO packet loss within the same PMIPv6 intra-domain. This scheme introduced the ARQ handling function

ISBN: 978-1-61208-202-8

similar to Go-Back-N to implement the buffered data delivery without duplicated packets between MAG and LMA. Home MAG (MAGh) and home LMA (LMAh) exchange the PBU/PBA message to construct the bi-directional tunnel when a MN is attached in home MAG. ARQ anchor handler in LMAh makes buffer per MN, temporarily stores every packet in each buffer and sends buffered packets to MAG including the sequence number located in IP-in-IP encapsulation header with destination option header according to [5]. On receiving a packet from LMA, The ARQ handler of MAGh sends packets to MN and responds the delivered sequence number to LMA. Then, ARQ anchor handler of LMAh will remove the delivered packets from the buffer by the reported sequence number. In PMIPv6 [2], efficient buffering scheme using the sequence number to check duplicated packets is required to provide the seamless mobility in ARQ handler of MAG and ARQ anchor handler of LMA since the HO start time of MN is not predictable in the network side.

III. PROPOSED SCHEMES

Roaming scenario was only proposed for a PMIPv6-based MN roaming scenario from the home domain to the visited domain in [5][6]. Our proposal extends [6] to the PMIPv6-based inter-domain HO scenarios including HO case between visited domains and HO case for the returning to the home domain using the sequence number-based efficient buffering scheme and buffered data backup mechanism to prevent the packet loss between domains. To distinguish PBU message from the LMA of another PMIPv6 domain and PBU message from MAG within the same PMIPv6 domain, flag S will be set to indicate the inter-domain PBU message, which is originated from another LMA in the different PMIPv6 domain according to [5].

Sequence number-based efficient buffering scheme in [5] can be extended to inter-domain HO using buffered data backup mechanism. Sequence number is allocated to the packet, which is sent to MAG through tunnel from the LMA to identify the packet. MAG reports to LMA whether packet is delivered to MN or not by ACK message using two flags; flag RR (Receiver Ready) is set for the periodic report by the expiry of timer in MAG and the other flag REJ (Reject) is set to inform LMA of the fact that there is the first missing packets after successful delivery for the immediate report. Namely, setting flag REJ in ACK message may indicate the HO preparation.

Figure. 1 shows the example of the sequence number-based buffering management in the inter-domain HO. When LMAh receives the data 1 destined to MN, LMAh allocates the sequence number 1 to the received packet to identify the packet, which is sent to the MAG through the tunnel and LMAh temporarily stores the data D(1) and sequence number S(1) per MN. Sequence number S(1) is sent to the MAG included in the IP-in-IP encapsulation header with new destination option header. S(2)/D(2) and S(3)/D(3) are subsequently sent to MAG and stored in LMAh. If MAGh detects the packet D(3) is not sent to MN, delivery failure is immediately reported to LMAh using the flag REJ(3) in the ACK message. In other words, it means that D(1) and D(2) are appropriately sent to MN and needs to remove the data D(1) and D(2) in the buffer. To prevent buffer overflow, flag RR in the ACK message is used for the periodic packet delivery status report to the LMAh by the expiry of the RR timer in the MAGh; RR(6) indicates that MAGh already delivered all packets to MN and buffered data less than those indexed with sequence number 6 should be cleared in the buffer of LMAh. Since data are not delivered to MN and MAGh returns the flag REJ(6) and REJ(7) to LMAh during the inter-domain HO, all the buffered data are delivered to new LMAv via the inter-domain tunnel after MN attachment to new domain and the establishment of inter-domain tunnel between LMAh and LMAv. All buffered data are eventually flushed to MN through the new MAGv. Sequence number-based buffer management is restarted in the new LMAv according to the packet delivery status report from the new MAGv after the inter-domain HO.

A. PMIPv6-based HO from Home to Visited Domain

Figure. 2 shows that MAGv1 performs the authentication step with visited 1 AAA when MN moves to visited domain 1. It redirects authentication messages to home AAA server according to the service level agreement between two AAA servers since the visited 1 AAA server knows that MN is involved in home AAA by the realm portion of user NAI (username@realm). Response messages including LMAh address information from home AAA server are sent to MAGv through the visited 1 AAA. For the construction of inter-domain tunnel between LMAv and LMAh, MAGv1 sends the PBU including LMAh address information to LMAv. LMAv sends the PBU message with flag S set to LMAh address received from MAGv for inter-domain tunnel establishment. Hence, after responding the PBA message to LMAv and transmitting the RA message to MN, LMAh can finally send the sequence number-based buffered data to LMAv for flushing the buffered data through MAGv according to the efficient buffering scheme. On flushing the data, LMAv newly acts as ARQ anchor handler and MAGv serves as ARQ handler.

B. PMIPv6-based HO between Visited Domains

In Figure. 3, buffered data in LMAv1 should be firstly returned to LMAh in HO scenario from visited domain 1 to visited domain 2 in order to prevent the inter-domain tunnel from extending the repeated ARQ handler from the home domain to another visited domain via visited domain 1. LMAv1 returns the buffered data to LMAh after receiving the PBU with flag S set from the home domain according to buffered data backup scheme. When MN moves to visited domain 2 and is attached to MAGv2, MAGv2 will send the PBU to LMAv2 for the establishment of inter-domain tunnel between LMAh and LMAv2 and then, LMAv2 will notify LMAh that the inter-domain HO occurs by sending the PBU with flag S set to LMAh, LMAh requests the LMAv1 to return the sequence number-based buffered data to LMAh by sending PBU with flag S set. After inter-domain tunnel establishment between LMAv2 and LMAh, LMAh can eventually transfer the sequence number-based buffered data to the LMAv2, which requested the PBU for the buffered data flushing through the MAGv2. Old intra-domain tunnel between MAGv1 and LMAv1 is terminated by predefined timer. According to this backup mechanism from the previous LMAv2 to LMAh for the buffered data, LMAv2, which acts as new ARQ anchor handler does not need to know the information of previously anchored LMA information and it is not required to trigger additional signaling message between AAAh and LMAh for the buffered data delivery.

C. PMIPv6-based HO for Returning to Home Domain

Figure. 4 indicates that MN can directly receive the buffered data from LMAh when MN comes back from the visited domain 2 to the home domain and is attached to MAGh because buffered data in previous LMA is already transferred to LMAh according to the buffered data backup mechanism. Predefined timer also terminates the old intra-domain tunnel between MAGv and LMAv.

IV. EXPERIMENTS

The simulation using NS-3 network simulator was performed in IEEE 802.11 wireless environments. Simulation network topology is shown by Figure. 5.

There are three inter-domain HO scenarios and each domain has one LMA connected with each different domain and two MAGs are linked each other within domain. The link delay of all wired links is 10msec and the link capacity is 50Mbps for the wired links and 11Mbps for the wireless links according to the 802.11a standard. As for the wireless delay, propagation delay model is applied to the simulation. MN moves with velocity 10m/sec across the PMIPv6 domains from the home domain to the visited domain 2 via visited domain 1 and then finally returns to the home domain. CN is attached with link capacity 50Mbps & link delay 100msec and communicates with MN through CBR over UDP with rates 1 Mbps. Sequence number will be measured to check the packet delivery in the simulation because ARQ anchor handler in LMA makes buffer per MN, temporarily stores every packet in each buffer and sends buffered packets to MAG including the sequence number located in IP-in-IP encapsulation header with destination option header according to [5]. In the simulation, inter-domain HO delay will be compared to the test case without efficient buffering scheme and buffered data backup mechanism to measure the sequence number for the packet delivery performance in the following cases: inter-domain HO to visited domain, inter-domain HO between visited domains, and inter-domain HO returning to the home domain.

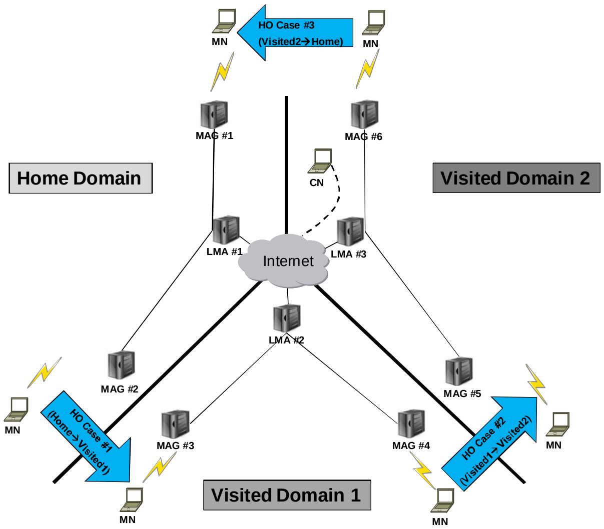


Figure 5. A network topology for simulation.

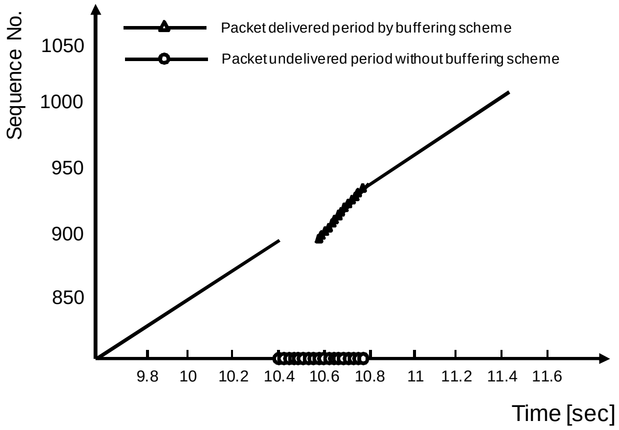


Figure 6. Received sequence number after inter-domain HO from home domain to visited domain 1 with buffering scheme when CN sends packets to MN.

Figure. 6 shows that inter-domain HO occurs around 10.39s from the home domain to the visited domain 1. Buffered packets are delivered from LMAh to LMAv1 through the inter-domain tunnel for the data flushing in 10.59s before MAG in the visited domain 1 can deliver packets to MN without the efficient buffering and buffered data backup scheme in 10.77s.

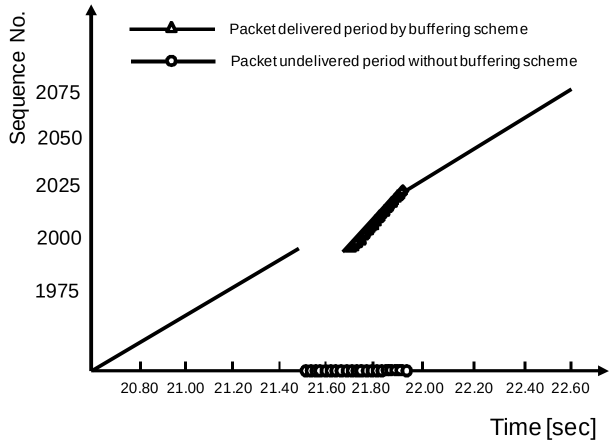


Figure 7. Received sequence number after inter-domain HO from visited domain 1 to visited domain 2 with buffering scheme when CN sends packets to MN.

Figure. 7 shows that inter-domain HO occurs around 21.50s from visited domain 1 to visited domain 2. Buffered packets are delivered from LMAv1 to LMAv2 through the LMAh using the backup mechanism for the buffered data flushing in 21.69s before MAG in the visited domain 2 can deliver packets to MN without the efficient buffering and buffered data backup scheme in 21.87s.

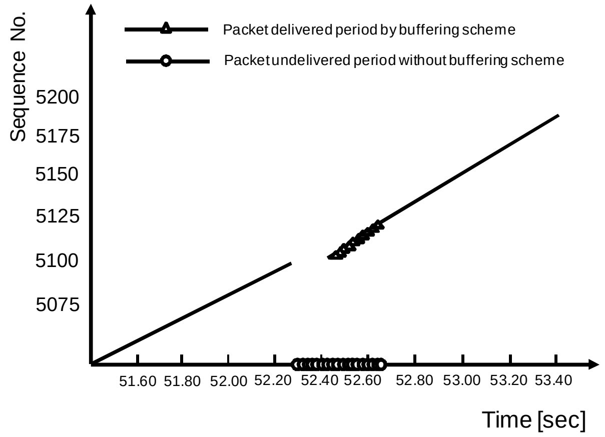


Figure 8. Received sequence number after inter-domain HO from visited domain 2 to home domain with buffering scheme when CN sends packets to MN.

Figure. 8 shows that inter-domain HO occurs around 51.12s from visited domain 2 to the home domain. Buffered packets are delivered from LMAv2 to LMAh for the data flushing in 52.42s before MAG in the visited domain 1 can deliver packets to MN without efficient buffering and buffered data backup scheme in 52.65s.

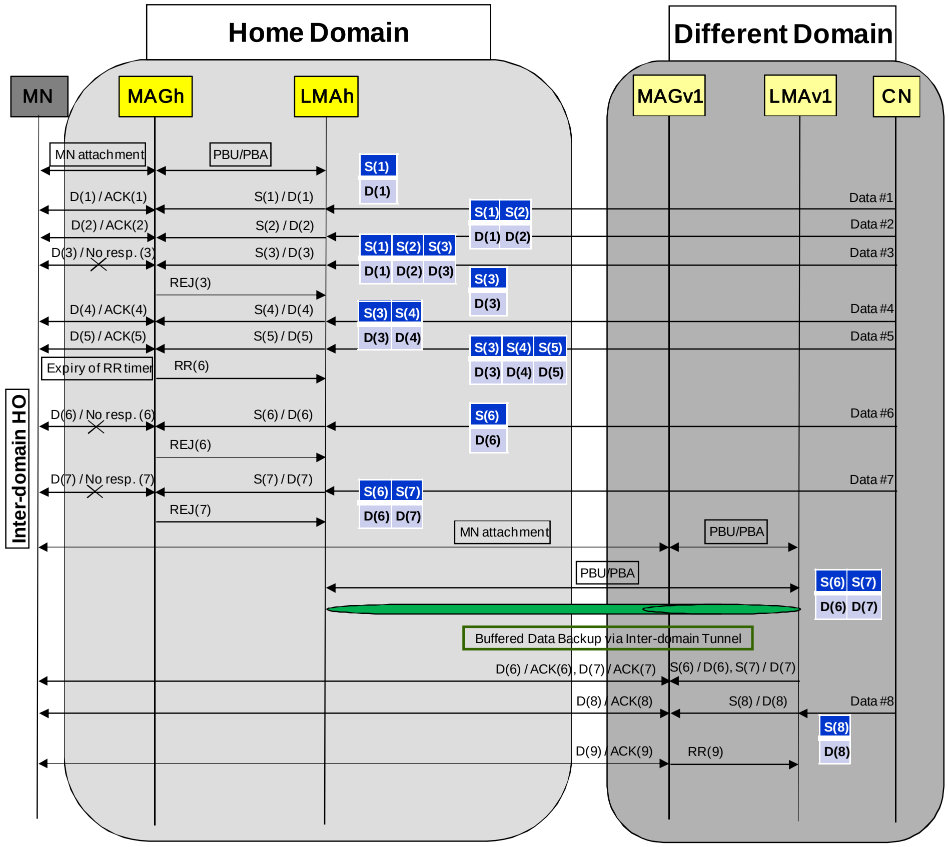


Figure 1. A sequence number and buffered data management in the inter-domain HO

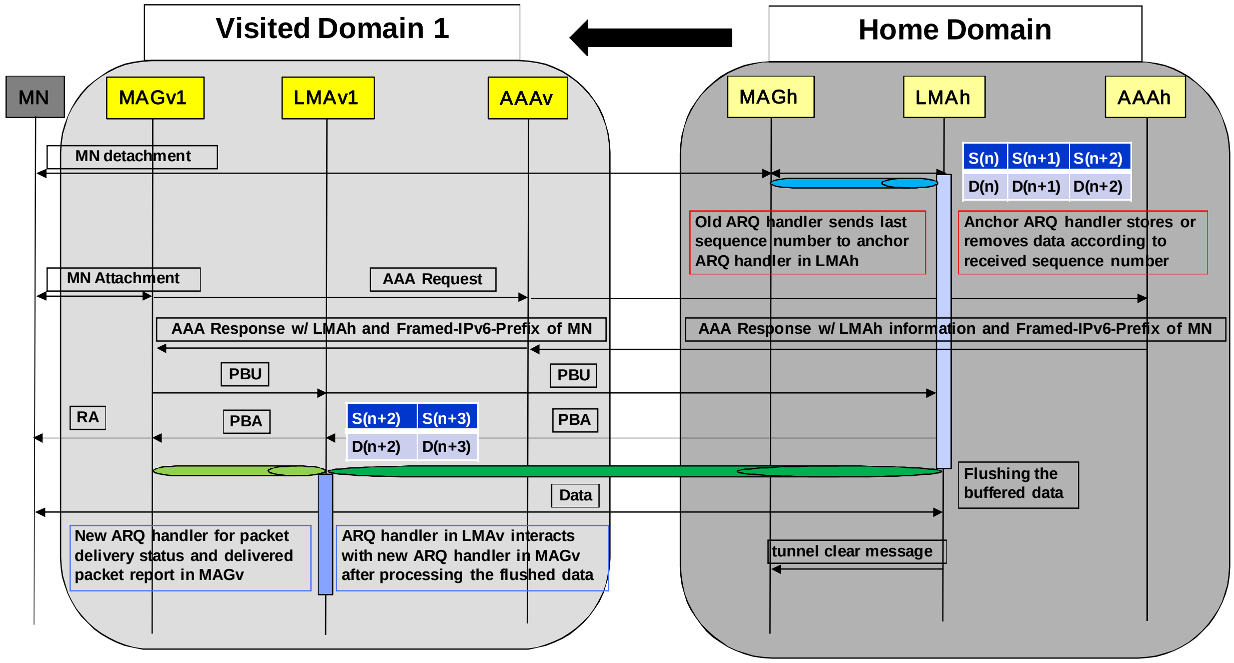

Visited Domain 1
Home Domain
MN
MAGv1
LMAv1
AAAv
MAGh
LMAh
AAAh
MN detachment
S(n)
S(n+1)
S(n+2)
D(n)
D(n+1)
D(n+2)
Old ARQ handler sends last sequence number to anchor ARQ handler in LMAh
Anchor ARQ handler stores or removes data according to received sequence number
MN Attachment
AAA Request
AAA Response w/ LMAh and Framed-IPv6-Prefix of MN
AAA Response w/ LMAh information and Framed-IPv6-Prefix of MN
PBU
PBU
RA
PBA
S(n+2)
S(n+3)
D(n+2)
D(n+3)
PBA
Flushing the buffered data
Data
New ARQ handler for packet delivery status and delivered packet report in MAGv
ARQ handler in LMAv interacts with new ARQ handler in MAGv after processing the flushed data
tunnel clear message

Figure 2. A call flow for HO from home to visited domain.

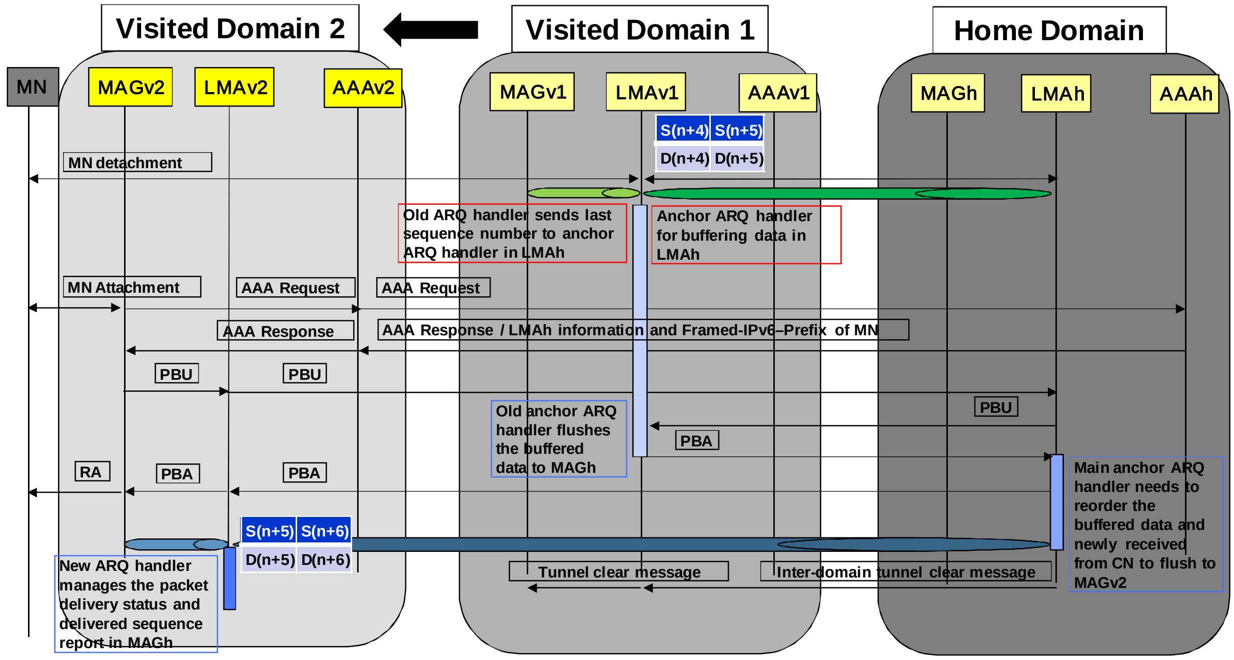

Visited Domain 2
Visited Domain 1
Home Domain
MN
MAGv2
LMAv2
AAAv2
MAGv1
LMAv1
AAAv1
MAGh
LMAh
AAAh
S(n+4)
S(n+5)
D(n+4)
D(n+5)
MN detachment
Old ARQ handler sends last sequence number to anchor ARQ handler in LMAh
Anchor ARQ handler for buffering data in LMAh
MN Attachment
AAA Request
AAA Request
AAA Response
AAA Response / LMAh information and Framed-IPv6–Prefix of MN
PBU
PBU
PBU
Old anchor ARQ handler flushes the buffered data to MAGh
PBA
RA
PBA
PBA
Main anchor ARQ handler needs to reorder the buffered data and newly received from CN to flush to MAGv2
S(n+5)
S(n+6)
D(n+5)
D(n+6)
New ARQ handler manages the packet delivery status and delivered sequence report in MAGh
Tunnel clear message
Inter-domain tunnel clear message

Figure 3. A call flow for HO from visited domain 1 to visited domain 2.

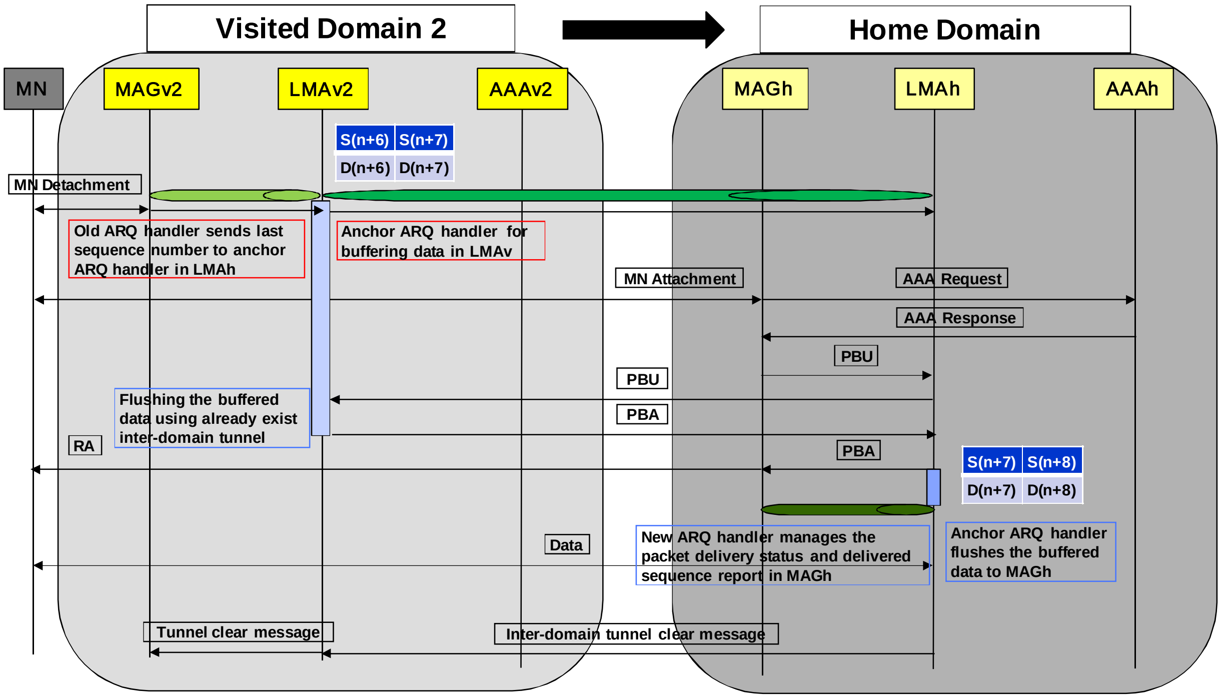

Visited Domain 2
Home Domain
MN
MAGv2
LMAv2
AAAv2
MAGh
LMAh
AAAh
S(n+6)
S(n+7)
D(n+6)
D(n+7)
MN Detachment
Old ARQ handler sends last sequence number to anchor ARQ handler in LMAh
Anchor ARQ handler for buffering data in LMAv
MN Attachment
AAA Request
AAA Response
PBU
PBU
Flushing the buffered data using already exist inter-domain tunnel
PBA
RA
PBA
S(n+7)
S(n+8)
D(n+7)
D(n+8)
Data
New ARQ handler manages the packet delivery status and delivered sequence report in MAGh
Anchor ARQ handler flushes the buffered data to MAGh
Tunnel clear message
Inter-domain tunnel clear message

Figure 4. A call flow for HO from visited domain 2 to home domain.

V. Conclusions

Our proposal expands Draft NetLMM-Neumann [5] and Soochang [6] to inter-domain HO scenarios including the case between visited domains and the case for the returning to the home domain by using the buffered data backup scheme. In inter-domain HO between visited domains, buffered data backup mechanism is proposed to prevent inter-domain tunnel from extending the sequential buffered data relay through the multiple LMA to reach the target LMA. For the seamless mobility and service continuity in the inter-domain HO, efficient buffering scheme [7] in the intra-domain HO is extended to inter-domain HO scenario.

The simulation results show that efficient buffering scheme provides the reduction of inter-domain HO delay and the enhancement of seamless services by supporting the sequence number-based data management between MAG and LMA and the buffered data delivery from the old LMA to new target LMA through the inter-domain tunnel.

References

[1] C. Perkins, Ed., D. Johnson, and J. Arkko, “Mobility Support in IPv6,” RFC6275, July 2011.

[2] S. Gundavelli, Ed., K. Leung, V. Devarapalli, K. Chowdhury, and B. Patil ”Proxy Mobile IPv6,” RFC5213, August 2008.

[3] J.Salowey and R.Droms, “RADIUS delegated-IPv6-Prefix attribute,” RFC4818, April 2007.

[4] P. Calhoun, J. Loughney, E. Guttman, G. Zorn, and J. Arkko, “Diameter base protocol,” RFC3588, September 2003.

[5] N. Neumann, X. Fu, and J. Lei, “Inter-Domain Handover and Data Forwarding between Proxy Mobile IPv6 Domains,” draft-neumann-netlmm-inter-domain-02, March 2009.

[6] Soochang P. and Euisin L., “Inter-Domain Roaming Mechanism Transparent to Mobile Nodes among PMIPv6 Networks,” Vehicular Technology Conference (VTC 2010-Spring), pp. 1608-1611, May 2010.

[7] Hyo-Beom L., Sung-Gi M., Kyoung-Hee L., and Hyun-Woo L. “PMIPv6-Based NEMO Protocol with Efficient Buffering Scheme,” Ubiquitous Information Technologies and Applications (CUTE), pp. 1-6, December 2010.

ISBN: 978-1-61208-202-8

Upper bounds and optimal solutions for a Deterministic and Stochastic linear Bilevel Problem

Pablo Adasme
Universidad de Santiago de Chile
departamento de Ingenieria Eléctrica
pablo.adasme@usach.cl
Laboratoire de Recherche en Informatique,
Universite Paris-Sud XI,
Batiment 650, 91405, Orsay Cedex France
pablo.adasme@lri.fr

Abdel Lisser
Laboratoire de Recherche en Informatique,
Universite Paris-Sud XI,
Batiment 650, 91405, Orsay Cedex France
abdel.lisser@lri.fr

***Abstract*—In this paper, we compute upper bounds and optimal solutions for a deterministic linear bilevel programming problem and then, for a stochastic version of this problem. The latter is formulated while adding probabilistic knapsack constraints in the upper level problem of the initial deterministic model. The upper bounds are computed using a Lagrangian iterative minmax algorithm and linear programming relaxations. To this purpose, we first transform both problems into the so called Global Linear Complementarity problems. We then, use these models to derive equivalent mixed integer programming formulations. This allows comparing the iterative minmax algorithm and the linear programming upper bounds with the optimal solution of the problem for the deterministic and stochastic instances as well. Our numerical results show tight near optimal bounds for both, the stochastic and deterministic linear programming relaxations and larger gaps for the iterative minmax algorithm.**

***Keywords*-Linear bilevel programming; stochastic programming; mixed integer programming.**

I. Introduction

In mathematical programming, the bilevel programming problem (BPP) is a hierarchical optimization problem. It consists in optimizing an objective function subject to a constrained set in which another optimization problem is embedded. The first level optimization problem (upper-level problem) is known as the leader's problem while the lower-level is known as the follower's problem. Formally, it can be written as follows

$$\begin{aligned} \min_{\{x\in X, y\}} \quad & F(x,y) \\ \text{s.t.} \quad & G(x,y) \leq 0 \\ & \min_{\{y\}} f(x,y) \\ & \text{s.t.} \quad g(x,y) \leq 0 \end{aligned}$$

where $x \in R^{n_1}$, $y \in R^{n_2}$, $F : R^{n_1} \times R^{n_2} \to R$ and $f : R^{n_1} \times R^{n_2} \to R$ are the decision variables and the objective valued functions for the upper and lower level problems, respectively. Similarly, the functions $G : R^{n_1} \times R^{n_2} \to R^{m_1}$ and $g : R^{n_1} \times R^{n_2} \to R^{m_2}$ denote upper and lower level constraints. Bilevel programming is commonly used to model situations in which two or more decision makers control part of the variables within a particular decision process [1]. The main goal is thus, to find an optimal point such that the leader and the follower minimizes their respective objective valued functions $F(x,y)$, $f(x,y)$ subject to their respective linking constraints $G(x,y)$ and $g(x,y)$. Notice that either the leader (or the follower) might also have their own particular constraints such as the set X in the above leader problem. Applications concerning BPP include transportation, networks design, management and planning among others (for different domains of applications see for instance [6]).

It has been shown that BPPs are strongly NP-hard even for the simplest case in which all the involved functions are affine [8]. Hereafter, we only consider the case in which all the above functions $F(x,y), f(x,y), G(x,y), g(x,y)$ are linear. Besides, if a particular constrained set exists in the leader or in the follower problem, we assume that it is a polyhedral affine space.

Stochastic programming (SP), on the other side, is an optimization technique which deals with the uncertainty of the input parameters of a mathematical program [16]. The underlying idea of SP is that the input parameters can be modeled as random variables to which the theory of probabilities can be applied. The probability distributions governing the data are usually assumed to be known in advance or that they can be estimated. The probability space is also usually assumed to be discrete and as such, one can consider finite sets of scenarios for the input parameters. There are two well known scenario based approaches in SP. The first one is known as the recourse model approach [5], [7] while the second one is known as probabilistic constrained approach [7]. The literature related to SP has grown considerably in last decades. A general survey can be found for instance in [14] and the reader is also referred to [3], [9], [15] or to a more recent book in [16] for a deeper comprehension.

In this paper, we consider the probabilistic knapsack constrained approach proposed in [7] when embedded into the upper level problem. Under this approach, it is imposed a threshold risk on the probability of occurrence for some (or all) of the constraints within a particular mathematical model. This means that some of the constraints should be satisfied, at least for a given percentage, while the rest of them are discarded.

The paper is organized as follows. In Section II, we provide a brief state of the art concerning joint aspects of bilevel and stochastic programming. Then, in Section III, we state the linear bilevel programming problem (LBPP) and briefly explain the probabilistic constrained approach considered. In Section IV, we derive the Global Linear Complementarity problem (GLCP) and also explain how the iterative minmax (IMM) algorithm works in order to compute the upper bounds. In Section V, we derive from the GLCPs, mixed integer and linear programming formulations (Resp. MIP and LP) according to [1]. Numerical results are given for the LBPP and for the stochastic LBPP (SLBPP) in Section VI. Finally, in Section VII we give the main conclusions of the paper.

II. Related Work

Although there exist many application domains in which bilevel programming can be suitably applied, joint stochastic and bilevel programming aspects have not yet widely been explored so far. Some preliminary works are the following [2], [4], [11]–[13], [17].

In [11], Luh et al. study a deterministic pricing problem and propose a stochastic counterpart for it by assuming that the inducible region is subject to uncertainty. Here, the inducible region is defined as the feasible set of the follower problem induced by the decision of the leader problem. Next, Patriksson et al. also incorporates uncertainty in the input data of hierarchical mathematical Programming problems [13]. In both papers [11], [13], the authors discuss theoretical aspects such as necessary and sufficient conditions for optimality, existence of solutions, convexity, and propose algorithms to deal with the problem at hand. Subsequently, Christiansen et al. [4], consider a stochastic bilevel programming problem which corresponds to an application in structural optimization where again, theoretical aspects such as existence of optimal solutions, Lipschitz continuity and differentiability aspects are discussed. More recently, applications concerning telecommunication network problems have been studied in [2], [17]. Therein, the analysis covers both theoretical and also practically oriented issues. In particular, special attention is given to different formulations of one and two stage stochastic bilevel programming problems where necessary optimality conditions for each of these problem instances are stated. Additionally, in [17], it is also proposed an algorithm which uses a stochastic quasi-gradient method to solve the problem.

Finally in [12], Özaltin et al. consider a stochastic bilevel knapsack problem with uncertain right-hand sides, and derive necessary and sufficient conditions for the existence of an optimal solution. In particular, they provide an equivalent two stage stochastic formulation when the leader problem take only integer values for the decision variables, although at the cost of having binary decision variables in the follower problem. Branching based algorithms are proposed to solve large scale instances of the problem.

In this paper, we focus more on computational numerical experiments rather than on theoretical aspects. Hence, we proceed as follows. We first compute upper bounds and optimal solutions for a generic linear bilevel programming problem (LBPP). We then, extend this generic LBPP by introducing knapsack probabilistic constraints in the upper level problem [7]. Hence, we compute upper bounds and optimal solutions for this stochastic LBPP (SLBPP) as well. The upper bounds are computed using a Lagrangian iterative minmax (IMM) algorithm proposed in [10] and also using linear programming (LP) relaxations we formulate from the so called Global Linear Complementarity Problem (GLCP) according to [1]. In [10], Kosuch et al. neither provide optimal solutions for deterministic or stochastic problems nor calculate gaps to measure IMM efficiency. Furthermore, even when Audet et al. propose links to derive an equivalent MIP formulation from a linear bilevel programming problem [1], they do not provide numerical comparisons to measure the tightness of its LP relaxation. Therefore, this paper can be seen as an extension of the works presented in [10] and [1] in the sense that now, we do provide optimal solutions and upper bounds for the IMM and for the LP relaxations as well as numerical comparisons between them, for deterministic and stochastic instances. In particular, we compute the optimal solutions using the MIP equivalent formulations [1].

III. Problem Formulation

In this section, we first present the generic LBPP under study. Then, we extend this generic model by adding knapsack probabilistic constraints in the upper level problem according to [7]. Since the probabilistic constrained approach introduces binary variables in the problem, we then obtain a mixed integer linear bilevel programming problem (MILBP) which we transform back into a LBPP [1]. We consider the following LBPP:

$$\begin{aligned} \text{LBP1:} \quad \max_{\{x\}} \quad & c_1^T x + d_1^T y && (1)\\ \text{s.t.} \quad & A^1 x + B^1 y \leq b^1 && (2)\\ & 0 \leq x \leq \mathbf{1}_{n_1} && (3)\\ & y \in \arg\max_{\{y\}} \{c_2^T x + d_2^T y\} && (4)\\ \text{s.t.} \quad & A^2 x + B^2 y \leq b^2 && (5) \end{aligned}$$

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-202-8

$$0 \leq y \leq \mathbf{1}_{n_2} \tag{6}$$

where $x \in R^{n_1}$ and $y \in R^{n_2}$ are decision variables. Vectors $\mathbf{1}_{n_1}$ and $\mathbf{1}_{n_2}$ are vectors of size n_1 and n_2 with entries equal to one. Matrices A^1, B^1, A^2, B^2 and vectors $c_1, c_2, d_1, d_2, b_1 \in R^{m_1}, b_2 \in R^{m_2}$ are input real matrices/vectors defined accordingly. In LBP1, (1)-(3) correspond to the leader's problem while (4)-(6) represent the follower's problem. Knapsack probabilistic constraints can be added to the upper-level problem of LBP1 as follows. Let $w = w(\omega) \in R_+^{n_1}$ and $S = S(\omega) \in R_+$ be two random variables distributed according to a discrete probability distribution Ω. We consider the following knapsack probabilistic constraints in the upper level problem

$$P\left\{w^T(\omega)x \leq S(\omega)\right\} \geq (1-\alpha) \tag{7}$$

where α represents the risk we take while not satisfying some of the constraints. Since Ω is discrete, one may suppose that $w = w(\omega)$ and $S = S(\omega)$ are concentrated in a finite set of scenarios such as $w(\omega) = \{w_1, .., w_K\}$ and $S(\omega) = \{s_1, ..s_K\}$, respectively with probability vector $p^T = (p_1, .., p_K)$ for all k such that $\sum_{k=1}^K p_k = 1$ and $p_k \geq 0$. According to [7], constraints in (7) can be transformed into the following pair of deterministic constraints

$$w_k^T x \leq s_k + M_k z_k \quad k = 1:K \tag{8}$$
$$p^T z \leq \alpha \tag{9}$$

where vector $z^T = (z_1, .., z_K,)$ is composed of binary variables. This means, if $z_k = 0$ then the constraint is included, otherwise it is not activated. M_k for each $k = 1 : K$ is defined as

$$M_k = \sum_{i=1}^{n_1} w_k^i - s_k$$

where w_k^i denotes the ith component of vector w_k. Putting it altogether yields the following deterministic mixed integer linear bilevel program

$$\begin{array}{lll}
\text{MILBP1:} & \max\limits_{\{x,z\}} & c_1^T x + d_1^T y \\
& \text{s.t.} & A^1x + B^1y \leq b^1 \\
& & 0 \leq x \leq \mathbf{1}_{n_1} \\
& & w_k^T x \leq s_k + M_k z_k \quad k = 1:K \\
& & p^T z \leq \alpha \\
& & z_k \in \{0,1\}^K \\
& & y \in \arg\max\limits_{\{y\}}\{c_2^T x + d_2^T y\} \\
& \text{s.t.} & \qquad A^2x + B^2y \leq b^2 \\
& & \qquad 0 \leq y \leq \mathbf{1}_{n_2}
\end{array}$$

Although MILBP1 contains binary variables, it can be converted back into an equivalent continuous LBPP [1] as follows

$$\begin{array}{lll}
\text{LBP2:} & \max\limits_{\{x,z\}} & c_1^T x + d_1^T y \\
& \text{s.t.} & A^1x + B^1y \leq b^1 \\
& & 0 \leq x \leq \mathbf{1}_{n_1} \\
& & w_k^T x \leq s_k + M_k z_k \quad k = 1:K \\
& & p^T z \leq \alpha \\
& & 0 \leq z_k \leq 1, \quad \forall k \\
& & v = 0_K \\
& & (y,v) \in \arg\max\limits_{\{y,v\}}\{c_2^T x + d_2^T y + \mathbf{1}_K^T v\} \\
& \text{s.t.} & \qquad A^2x + B^2y \leq b^2 \\
& & \qquad 0 \leq y \leq \mathbf{1}_{n_2} \\
& & \qquad v \leq z \qquad (10) \\
& & \qquad v \leq \mathbf{1}_K - z \qquad (11)
\end{array}$$

In LPB2, we denote by $\mathbf{1}_K$ and 0_K, the vector of all ones and the vector of all zeros of dimension K. As explained in [1], the transformation from MILBP1 into LBP2 can be done by performing the following steps. First the binary variables $z \in \{0,1\}^K$ for each $k = 1 : K$ in the upper level problem should be relaxed inside the interval [0,1]. In parallel, a new continuous variable vector $v = 0_K$ should be placed in the leader's problem imposing that all its entries be equal to zero. In fact, vector v is introduced in the follower's problem when adding the term $\mathbf{1}_K^T v$ in its objective function together with the new constraints (10)-(11). The term added in the objective function together with the latter constraints will enforce all the entries in vector z to be either equal to zero or one. We then, have derived an equivalent LBPP formulation for MILBP1. Notice that v is a variable vector in the follower's problem while vector z is a variable vector in the leader's problem.

In the next section, we derive the so called Global Linear Complementarity Counterparts for LBP1 and LBP2. Subsequently, we briefly present and explain the Lagrangian iterative minmax algorithm proposed in [10].

IV. The GLCP and IMM Algorithm

In this section, we explain all the necessary transformation steps until reaching the GLCP counterparts for LBP1 and LBP2. Then, we present IMM algorithm and describe how it works in order to compute the upper bounds. Finally, we derive from the GLCP problems equivalent MIP formulations according to [1] together with their LP relaxations.

A. The Global Linear Complementarity Problem

The GLCP is a single level quadratic optimization problem. The main idea of deriving the GLCP consists of replacing the original follower's problem with its initial constraints, dual constraints and complementary slackness conditions. The decision variables of GLCP are thus: the leader, the follower and the follower's dual variables as well. In order to derive a GLCP model for LBP1, we first write the dual of the follower problem as follows

$$\begin{array}{lll} \text{LBPD1:} & \min\limits_{\{\lambda,\mu\}} & \lambda^T(b^2 - A^2x) + \mathbf{1}_K^T\mu \qquad (12) \\ & \text{s.t.} & (B^2)^T\lambda + I_{n_2}\mu \geq d_2 \qquad (13) \\ & & \lambda \geq 0, \mu \geq 0 \qquad (14) \end{array}$$

where λ and μ are Lagrangian multipliers vectors of appropriate size. I_{n_2} represents the identity matrix of order n_2. Now, we add the complementary slackness conditions we construct by using LBP1 and LBPD1 together with its respective dual constraints (13)-(14). We may obtain the so called GLCP counterpart for LBP1 as follows

$$\begin{array}{lll} \text{LBPG1:} & \max\limits_{\{x,y,\lambda,\mu\}} & c_1^Tx + d_1^Ty \\ \text{s.t.} & A^1x + B^1y \leq b^1 & \\ & 0 \leq x \leq \mathbf{1}_{n_1} & \\ & A^2x + B^2y \leq b^2 & \\ & 0 \leq y \leq \mathbf{1}_{n_2} & \\ & (B^2)^T\lambda + I_{n_2}\mu \geq d_2 & \\ & \lambda \geq 0, \mu \geq 0 & \\ & (b^2 - A^2x - B^2y)^T\lambda = 0 & (15) \\ & (\mathbf{1}_{n_2} - I_{n_2}y)^T\mu = 0 & (16) \\ & ((B^2)^T\lambda + I_{n_2}\mu - d_2)^Ty = 0 & (17) \end{array}$$

where (15)-(17) are the complementary slacknes conditions. To derive the GLCP counterpart for LBP2, we proceed similarly as for LBP1. In this case, the dual formulation for the follower problem can be written as

$$\begin{array}{lll} \text{LBPD2:} & \min\limits_{\{\lambda_1,\mu_1,\mu_2,\mu_3\}} & \lambda_1^T(b^2 - A^2x) + \mu_1^Tz + \\ & +\mu_2^T(\mathbf{1}_K - z) + \mu_3^T\mathbf{1}_{n_2} & (18) \\ \text{s.t.} & (B^2)^T\lambda_1 + I_{n_2}\mu_3 \geq d_2 & (19) \\ & I_K\mu_1 + I_K\mu_2 \geq \mathbf{1}_K & (20) \\ & \lambda_1 \geq 0, \mu_1 \geq 0, \mu_2 \geq 0, \mu_3 \geq 0 & (21) \end{array}$$

where λ_1, μ_1, μ_2 and μ_3 are Lagrangian multiplier vectors respectively. Subsequently, the GLCP in this case reads

$$\begin{array}{lll} \text{LBPG2:} & \max\limits_{\{x,y,z,\mu_1,\mu_2,\mu_3,\lambda_1\}} & c_1^Tx + d_1^Ty \\ \text{s.t.} & A^1x + B^1y \leq b^1 & \\ & 0 \leq x \leq \mathbf{1}_{n_1} & \\ & A^2x + B^2y \leq b^2 & \\ & 0 \leq y \leq \mathbf{1}_{n_2} & \\ & w_k^Tx \leq s_k + M_kz_k \quad k = 1:K & \\ & p^Tz \leq \alpha, \quad 0 \leq z_k \leq 1 \quad \forall k = 1:K & \\ & (B^2)^T\lambda_1 + I_{n_2}\mu_3 \geq d_2 & \\ & I_K\mu_1 + I_K\mu_2 \geq \mathbf{1}_K & \\ & \lambda_1 \geq 0, \mu_1 \geq 0, \mu_2 \geq 0, \mu_3 \geq 0 & \end{array}$$

$$\begin{array}{ll} \lambda_1^T(b^2 - A^2x - B^2y) = 0 & (22) \\ \mu_1^Tz = 0 & (23) \\ \mu_2^T(\mathbf{1}_K - z) = 0 & (24) \\ \mu_3^T(\mathbf{1}_{n_2} - y) = 0 & (25) \\ y^T((B^2)^T\lambda_1 + I_{n_2}\mu_3 - d_2) = 0 & (26) \end{array}$$

In LBPG2, the last constraints (22)-(26) are due to the complementary slackness condition.

In the next subsection, we briefly illustrate how IMM algorithm works when solving a minmax relaxation derived from the GLCP [10].

B. The IMM Algorithm

To show how the IMM algorithm works, we take for illustration purposes, the GLCP we have already derived from the previous subsection denoted by LBPG2. Notice that this model is a quadratic optimization problem since their complementary constraints (22)-(26) are quadratic, and thus it is hard to solve directly. The first step of IMM consists in relaxing these quadratic constraints into the following Lagrangian function

$$\begin{array}{ll} \mathcal{L}(x, y, z, \lambda_1, \mu_1, \mu_2, \mu_3) = & \\ = c_1^Tx + d_1^Ty + & \\ +\lambda_1^T(b^2 - A^2x - B^2y) + & \\ +\mu_1^Tz + \mu_2^T(\mathbf{1}_K - z) + & \\ +\mu_3^T(\mathbf{1}_{n_2} - z) + & \\ +y^T((B^2)^T\lambda_1 + I_{n_2}\mu_3 - d_2) & (27) \end{array}$$

This allows writing a minmax relaxation for LBPG2 as follows

$$\begin{array}{lll} \text{LGN2:} & \min\limits_{\{\mu_1,\mu_2,\mu_3,\lambda_1\}} \max\limits_{\{x,y,z\}} & \mathcal{L}(x, y, z, \lambda_1, \mu_1, \mu_2, \mu_3) \\ \text{s.t.} & A^1x + B^1y \leq b^1 & \\ & 0 \leq x \leq \mathbf{1}_{n_1} & \\ & A^2x + B^2y \leq b^2 & \\ & 0 \leq y \leq \mathbf{1}_{n_2} & \\ & w_k^Tx \leq s_k + M_kz_k \quad k = 1:K & \\ & p^Tz \leq \alpha, \quad 0 \leq z_k \leq 1 \quad \forall k = 1:K & \\ & (B^2)^T\lambda_1 + I_{n_2}\mu_3 \geq d_2 & \\ & I_K\mu_1 + I_K\mu_2 \geq \mathbf{1}_K & \\ & \lambda_1 \geq 0, \mu_1 \geq 0, \mu_2 \geq 0, \mu_3 \geq 0 & \end{array}$$

The second step of IMM consists of decomposing LGN into two linear programming subproblems: LGNs and LGNd as

$$\begin{array}{lll} & \text{LGNs:} \max\limits_{\{x,y,z,\varphi\}} & \varphi \\ & \varphi \leq \mathcal{L}(x, y, z, \lambda_1^q, \mu_1^q, \mu_2^q, \mu_3^q), & \\ & \forall q = 0, 1, ..., N-1 & (28) \\ \text{s.t.} & A^1x + B^1y \leq b^1 & \\ & 0 \leq x \leq \mathbf{1}_{n_1} & \end{array}$$

$$
\begin{aligned}
& A^2x + B^2y \leq b^2 \\
& 0 \leq y \leq \mathbf{1}_{n_2} \\
& w_k^T x \leq s_k + M_k z_k \quad k = 1:K \\
& p^T z \leq \alpha, \quad 0 \leq z_k \leq 1 \quad \forall k = 1:K
\end{aligned}
$$

and

$$
\begin{aligned}
\text{LGNd:} \quad & \min_{\{\mu_1,\mu_2,\mu_3,\lambda_1,\beta\}} \quad \beta \\
& \beta \geq \mathcal{L}(x^q, y^q, z^q, \lambda_1, \mu_1, \mu_2, \mu_3), \\
& \forall q = 1, ..., N \qquad (29) \\
\text{s.t.} \quad & (B^2)^T \lambda_1 + I_{n_2}\mu_3 \geq d_2 \\
& I_K \mu_1 + I_K \mu_2 \geq \mathbf{1}_K \\
& \lambda_1 \geq 0, \mu_1 \geq 0, \mu_2 \geq 0, \mu_3 \geq 0
\end{aligned}
$$

where φ and β are defined as free real variables. Finally, the third step of the algorithm consists in solving iteratively both LGNs and LGNd. At iteration q, the auxiliary constraint (28) (resp. (29)) is added to LGNs (resp. LGNd) in order to enforce the convergence of their optimal solution values towards the optimal solution value of LGN. The iteration process stops when either $\beta - \varphi < \delta$ or $(\beta - \varphi)/\beta < \varepsilon$ for small $\delta > 0$ and $\varepsilon > 0$. The convergence of IMM is proven in [10]. Notice that even when IMM does not converge to a stationary point, it provides, at least, an upper bound for the GLCP. Hereafter, we denote by LGN1 and LGN2 the minmax relaxations we formulate starting from LBP1 and LBP2 respectively. In this paper, we compute upper bounds for LGN1 and LGN2 using IMM algorithm. Afterward, we compare these upper bounds with LP relaxations we derived from equivalent MIP formulations according to [1].

V. MIP and LP Formulations

In this subsection, we present for each GLCP problems (LBPG1 and LBPG2 respectively) an equivalent MIP formulation. The method basically consists of replacing each quadratic constraint of the GLCP by two linear constraints that include a new binary variable. According to [1], a MIP formulation for LBPG1 can be written as follows

$$
\begin{aligned}
\text{MIP1:} \quad & \max_{\{x,y,\lambda,\mu,\nu^1,\nu^2,\nu^3\}} c_1^T x + d_1^T y \\
\text{s.t.} \quad & A^1x + B^1y \leq b^1 \\
& 0 \leq x \leq \mathbf{1}_{n_1} \\
& A^2x + B^2y \leq b^2 \\
& 0 \leq y \leq \mathbf{1}_{n_2} \\
& (B^2)^T\lambda + I_{n_2}\mu \geq d_2 \\
& \lambda \geq 0, \mu \geq 0 \\
& b^2 - A^2x - B^2y + L\nu^1 \leq L\mathbf{1}_{m_2} \qquad (30) \\
& \lambda \leq L\nu^1, \quad \nu^1 \in \{0,1\}^{m_2} \qquad (31) \\
& \mathbf{1}_{n_2} - I_{n_2}y + L\nu^2 \leq L\mathbf{1}_{n_2} \qquad (32) \\
& \mu \leq L\nu^2, \quad \nu^2 \in \{0,1\}^{n_2} \qquad (33) \\
& (B^2)^T\lambda + I_{n_2}\mu - d_2 + L\nu^3 \leq L\mathbf{1}_{n_2} \qquad (34) \\
& y \leq L\nu^3, \quad \nu^3 \in \{0,1\}^{n_2} \qquad (35)
\end{aligned}
$$

In this model, constraints in (30)-(31),(32)-(33),(34)-(35) are equivalent to constraints (15),(16),(17) in LBPG1 respectively. These constraints force at least one of the terms within each product term to be equal to zero. To this end, a large constant L is needed [1]. Similarly, we can derive a MIP formulation for LBPG2 as follows

$$
\begin{aligned}
\text{MIP2:} \quad & \max_{\{x,y,z,\mu_1,\mu_2,\mu_3,\lambda_1,\theta_1,\theta_2,\theta_3,\theta_4,\theta_5\}} c_1^T x + d_1^T y \\
\text{s.t.} \quad & A^1x + B^1y \leq b^1 \\
& 0 \leq x \leq \mathbf{1}_{n_1} \\
& A^2x + B^2y \leq b^2 \\
& 0 \leq y \leq \mathbf{1}_{n_2} \\
& w_k^T x \leq s_k + M_k z_k \quad k = 1:K \\
& p^T z \leq \alpha, \quad 0 \leq z_k \leq 1 \quad \forall k = 1:K \\
& (B^2)^T\lambda_1 + I_{n_2}\mu_3 \geq d_2 \\
& I_K\mu_1 + I_K\mu_2 \geq \mathbf{1}_K \\
& \lambda_1 \geq 0, \mu_1 \geq 0, \mu_2 \geq 0, \mu_3 \geq 0 \\
& b^2 - A^2x - B^2y + L\theta_1 \leq L\mathbf{1}_{m_2} \qquad (36) \\
& \lambda \leq L\theta_1, \quad \theta_1 \in \{0,1\}^{m_2} \qquad (37) \\
& z + L\theta_2 \leq L\mathbf{1}_K \qquad (38) \\
& \mu_1 \leq L\theta_2, \quad \theta_2 \in \{0,1\}^K \qquad (39) \\
& \mathbf{1}_K - z + L\theta_3 \leq L\mathbf{1}_K \qquad (40) \\
& \mu_2 \leq L\theta_3, \theta_3 \in \{0,1\}^K \qquad (41) \\
& \mathbf{1}_{n_2} - y + L\theta_4 \leq L\mathbf{1}_{n_2} \qquad (42) \\
& \mu_3 \leq L\theta_4, \quad \theta_4 \in \{0,1\}^{n_2} \qquad (43) \\
& (B^2)^T\lambda_1 + I_{n_2}\mu_3 - d_2 + L\theta_5 \leq L\mathbf{1}_{n_2} \qquad (44) \\
& y \leq L\theta_5, \quad \theta_5 \in \{0,1\}^{n_2} \qquad (45)
\end{aligned}
$$

Analogously, in this model constraints (36)-(45) replace constraints (22)-(26) in LPBG2. We denote by LP1 and LP2 the corresponding linear programming relaxations derived from MIP1 and MIP2, respectively.

VI. Numerical Results

In this section, we present numerical results for MIP1, MIP2, LP1, LP2, LGN1 and LGN2. The input data is generated as follows. The entries in matrices A^1, A^2, B^1, B^2 are filled with random values uniformly picked from [-1,1] except for the last row which is uniformly filled with values in [0,1]. The entries of b^1, b^2 are generated in the following way:

$$
b_i^1 = \sum_{j=1}^{n_1} A_{ij}^1 + \sum_{j=1}^{n_2} B_{ij}^1 + \rho_i^1, \quad i = \{1, .., m_1\} \qquad (46)
$$

$$
b_i^2 = \sum_{j=1}^{n_1} A_{ij}^2 + \sum_{j=1}^{n_2} B_{ij}^2 + \rho_i^2, \quad i = \{1, .., m_2\} \qquad (47)
$$

Table I
UPPER BOUNDS AND OPTIMAL SOLUTIONS FOR THE DETERMINISTIC PROBLEM (LBPP)

#	Instance Size				MIP1	LGN1		Time LGN1	# LPs	LP1		Time LP1	Gaps	
	m_1	m_2	n_1	n_2		Ubs	Std			Ubs	Std		LGN1	LP1
1	25	25	50	50	318.6297	400.5155	55.2680	0.4578	19.9000	324.0572	53.2167	0.1703	20.5718	1.5381
2	25	25	50	100	570.9695	754.9484	45.3639	0.5563	15.7000	579.6330	45.8222	0.1641	24.2832	1.3892
3	25	25	50	150	800.2270	1078.6	38.0506	0.7156	14	806.3612	39.1128	0.1844	25.7922	0.7522
4	25	25	50	250	1319.1	1758.4	69.2061	1.1703	12.4000	1324	55.1661	0.2797	24.9855	0.3751
5	25	25	100	50	534.4669	616.3301	46.8627	0.9031	30.7000	541.6181	49.7854	0.2375	13.3941	1.3551
6	25	25	100	100	823.3725	993.2053	43.7868	0.9297	22.1000	830.9123	50.2394	0.2422	17.1145	0.8795
7	25	25	100	150	1061	1323.7	75.8933	1.0781	18.3000	1062.8	94.5218	0.2562	19.9749	0.1677
8	25	25	100	250	1501	1975.4	80.0849	1.5844	15.7000	1512.4	85.2606	0.2719	24.0231	0.7090
9	25	25	150	50	796.3497	879.2877	71.7902	1.1422	31.2000	799.2495	71.0114	0.2391	9.4376	0.3403
10	25	25	150	100	1050	1232	47.2699	1.4094	27.6000	1057.9	56.2905	0.2391	14.8390	0.7229
11	25	25	150	150	1288.4	1567.1	66.5875	1.4109	20.6000	1303.8	61.2962	0.2609	17.8121	1.1917
12	25	25	150	250	1763.7	2213.5	96.8940	2.3250	20.1000	1768.8	83.8630	0.3000	20.3168	0.2782
13	25	25	250	50	1348.7	1436.5	68.2226	1.4578	29.6000	1349.4	59.5212	0.2656	6.0902	0.0493
14	25	25	250	100	1541.7	1736.5	61.1188	2.0219	31.1000	1552	55.1324	0.2797	11.2055	0.6480
15	25	25	250	150	1777	2047.9	59.2596	1.9344	23.5000	1782.3	44.7907	0.2781	13.2086	0.2986
16	25	25	250	250	2292.5	2723	48.3565	2.5031	20.1000	2297.4	61.8148	0.3234	15.8171	0.2108
17	50	50	50	50	181.8254	256.2218	48.6907	0.8703	17.6000	184.6658	45.4258	0.2516	29.3389	1.4090
18	50	50	50	100	399.8696	570.6325	83.5419	2.3438	24	404.9089	63.2648	0.2562	29.8847	1.2420
19	50	50	50	250	1116.7	1581.4	60.9783	2.9844	14.8000	1117.8	57.5520	0.3422	29.3974	0.1003
20	50	50	50	500	2413.4	3281.7	68.6972	3.3344	10.6000	2415.2	56.4318	0.4969	26.4553	0.0747
21	50	50	100	50	338.3935	401.1113	80.9225	2.6719	34.6000	338.8181	68.0222	0.2562	15.5971	0.1357
22	50	50	100	100	639.0975	804.6371	58.2480	5.3641	39	642.3280	62.9545	0.2813	20.7142	0.5395
23	50	50	100	250	1403.2	1836.7	54.4123	4.9594	21.6000	1408.8	76.2232	0.3516	23.6332	0.4281
24	50	50	100	500	2595.3	3471.6	89.8953	5.4188	14.6000	2596.1	102.6473	0.4656	25.2606	0.0311
25	50	50	250	50	1146.4	1223.7	79.9139	6.2484	54.6000	1156.5	75.4866	0.3266	6.2685	0.8414
26	50	50	250	100	1374.8	1544.5	66.4153	7.4703	50.8000	1381.5	64.8566	0.3563	10.9772	0.4713
27	50	50	250	250	2136.1	2551.2	96.8040	10.4750	37	2137.3	69.5131	0.4203	16.2371	0.0592
28	50	50	250	500	3282.1	4180.2	79.3646	11.9203	23.7000	3287	73.9953	0.5359	21.4839	0.1534
29	50	50	500	50	2392.7	2472.2	94.0571	12.0469	60.4000	2394.9	88.9616	0.4484	3.2086	0.0864
30	50	50	500	100	2586.4	2750	43.8231	11.7516	52	2590.4	47.6176	0.4609	5.9523	0.1548
31	50	50	500	250	3386.5	3828.9	63.0136	16.3172	45.7000	3390.8	57.1894	0.5453	11.5474	0.1240
32	50	50	500	500	4574	5499.7	92.8812	18.2703	29.9000	4574.8	82.5642	0.6703	16.8290	0.0173
33	100	100	150	150	764.0421	999.6294	89.3524	32.7984	49.5000	767.1260	91.5415	0.4359	23.7201	0.3816
34	100	100	150	200	1010.3	1330.6	56.0125	38.4875	45.4000	1010.5	56.4980	0.4531	24.0978	0.0212
35	100	100	150	300	1483.4	2006.4	61.0033	33.0938	43.8000	1485.2	42.4945	0.5594	26.0603	0.1240
36	100	100	150	500	2518.9	3400.1	73.6681	39.9469	30.4000	2520.9	92.7737	0.7281	25.9167	0.0807
37	100	100	200	150	1107.4	1358.6	102.1064	46.0844	60.4000	1108.5	77.3447	0.4437	18.4364	0.1010
38	100	100	200	200	1362.7	1703.6	161.0181	49.2656	51.6000	1363.7	137.4408	0.4719	20.0013	0.0709
39	100	100	200	300	1774.1	2296.3	52.6927	38.4375	49.2000	1776	44.1224	0.5906	22.7256	0.1049
40	100	100	200	500	2742	3602.8	55.1760	45.6688	33.4000	2744.2	41.0465	0.7656	23.8907	0.0809
41	100	100	300	150	1495.9	1758.3	102.0865	65.8469	84.6000	1497.8	97.7465	0.5406	14.9401	0.1207
42	100	100	300	200	1782.6	2157.1	56.6420	45.8000	70.4000	1783.5	66.7431	0.5938	17.3630	0.0469
43	100	100	300	300	2196.5	2718.8	30.9680	50.4688	58.2000	2197.8	31.8063	0.7000	19.2058	0.0609
44	100	100	300	500	3259.4	4111.9	67.2305	73.8406	47	3261	91.4262	0.8625	20.7402	0.0461
45	100	100	500	150	2525	2768.5	140.0770	61.5875	89.4000	2525.2	130.6626	0.7438	8.7985	0.0105
46	100	100	500	200	2782.7	3146.5	73.4873	53.1187	67.6000	2786.3	105.9573	0.7813	11.5874	0.1255
47	100	100	500	300	3246.4	3765.5	128.5285	80.2375	76	3249.1	120.3366	0.8500	13.7950	0.0848
48	100	100	500	500	4159.1	5026.3	71.2547	90.6281	51.6000	4160.3	61.3897	1	17.2430	0.0282

where ρ_i^1 and ρ_i^2 for each i, are random numbers picked from the interval $[0, 2]$. This procedure ensures that the inducible region generated by the upper level and lower level constraints be non-empty and bounded. Each input data vector w_k, for each probabilistic constraint in LBP2, is chosen uniformly distributed from [0,1] while s_k are picked from the interval $[\frac{1}{2}\widetilde{W}_k, \widetilde{W}_k]$. Here, $\widetilde{W}_k$ is computed as $\widetilde{W}_k = w_k^T \mathbf{1}_{n_1}$ for $k = 1 : K$. Finally, vectors c_1, c_2, d_1, d_2 are randomly chosen from $(0, 10]$ and $\alpha = 0.05$. Again, this procedure guarantees boundedness for the feasible region of the bilevel instances, although it does not guarantee non-emptiness anymore [10].

Without loss of generality we set the large value L needed for the resolution of the MIP and LP formulations be equal to $L = 10^5$. The IMM algorithm as well as the MIP and LP formulations are implemented using Matlab 7.8 and Cplex 12.2. The simulations are run in a 2100 MHz computer with 2 Gb Ram under windows XP.

Table I shows numerical results for MIP1, LGN1 and LP1 while table II shows the same information for MIP2, LGN2 and LP2, respectively. These numerical results correspond to averages computed over 50 sample runs for each instance, except for the instances 33 to 48 in tables I and II. For these instances, we only compute the average over 10 runs since solving the MIP models become prohibitive for larger instances. The two tables provide similar information. In table I, columns 2 to 5 give the instance sizes. Column 6 provides the optimal solution of MIP1. Columns 7 and 8 give the upper bounds and the standard deviation obtained while using IMM to solve LGN1. Columns 9 and 10 give the cpu time in seconds and the number of LPs IMM needs to converge. Similarly, columns 11 to 13 provide the upper bounds we obtain with the LP1 relaxation, its standard deviation and the cpu time in seconds. Finally, relative gaps are given in columns 14 and 15 for LGN1 and LP1, respectively. The gaps are computed as $\left(\frac{Ubs - MIP1}{Ubs}\right) \cdot 100$ in each case.

Table II provides exactly the same information for MIP2, LP2 and LGN2. The only difference now, is that the second column gives the number of scenarios $k = \{1, .., K\}$ we add in the leader's problem. From the numerical results, we mainly observe in table I, that the gaps decrease with

Table II
UPPER BOUNDS AND OPTIMAL SOLUTIONS FOR THE STOCHASTIC PROBLEM (SLBPP)

#	Instance Size					MIP2	LGN2		Time LGN2	# LPs	LP2		Time LP2	Gaps	
	K	m_1	m_2	n_1	n_2		Ubs	Std			Ubs	Std		LGN2	LP2
1	25	25	25	100	100	784.3203	989.2347	33.8779	1.3563	21.4000	820.4212	34.9967	0.2172	20.6631	4.3371
2		25	25	100	250	1513.4	2019.9	43.3322	1.7859	14	1564.5	69.4705	0.2281	25.0775	3.2478
3		25	25	250	100	1433.4	1742.2	48.9163	3.5781	35.5000	1583	45.0512	0.2406	17.7093	9.4391
4		25	25	250	250	2082.1	2700.1	57.2095	4.0766	23.4000	2267.3	76.3264	0.3516	22.8960	8.1650
5	50	25	25	100	100	760	979.1	42.86	1.7828	23.6999	795.01	49.1429	0.2	22.3878	4.3532
6		25	25	100	250	1487.31	1990.43	53.6890	1.9218	14	1555.41	51.2573	0.2250	25.2638	4.3767
7		25	25	250	100	1399.5	1707	56.8898	3.6875	29.9000	1553.6	53.2098	0.2656	17.9859	9.8821
8		25	25	250	250	2127.5	2719.5	55.6479	3.9328	20.8000	2288.7	62.2031	0.3094	21.7517	7.0396
9	75	25	25	100	100	760.4551	986.0295	31.6953	2.5156	25.9000	807.3817	36.6528	0.2172	22.8722	5.7832
10		25	25	100	250	1497.7	1998.7	54.0413	2.6828	17.4000	1573.3	62.6726	0.2578	25.0623	4.7602
11		25	25	250	100	1388.2	1743.4	60.6976	4.1031	27.3000	1572.8	59.4721	0.2938	20.3690	11.7179
12		25	25	250	250	2131.8	2761.1	60.2416	4.9328	22.5000	2311.9	73.1240	0.3250	22.7843	7.7919
13	100	25	25	100	100	772.6953	1001.2	84.1361	11.4016	37	826.4422	65.5359	0.2500	22.6190	6.3321
14		25	25	100	250	1472.6	1966.2	71.4463	2.7703	16.1000	1535.6	56.7641	0.2641	25.0811	4.0855
15		25	25	250	100	1377.4	1726.4	41.9370	4.9156	26.5000	1552.3	35.1868	0.3234	20.2008	11.2598
16		25	25	250	250	2102.2	2708.9	76.7820	5.5844	22.1000	2283.2	84.8015	0.3625	22.3871	7.8996
17	25	50	50	100	100	604.5257	781.5844	58.1403	6.0828	33.6000	612.4399	55.7202	0.2172	22.6608	1.2270
18		50	50	100	500	2540.1	3455.1	71.2097	6.6953	14.9000	2585.6	65.6119	0.4031	26.4760	1.7418
19		50	50	500	100	2376.6	2822.8	58.5527	14.5063	45.9000	2645.4	62.4349	0.4484	15.7803	10.1482
20		50	50	500	500	4275.7	5460.6	89.9955	26.5719	33.7000	4594.6	122.9921	0.6375	21.6995	6.9387
21	50	50	50	100	100	626.8187	814.5498	67.9897	6.5563	31.9000	638.3389	69.7229	0.2250	23.1182	1.6959
22		50	50	100	500	2583.6	3489.2	73.0974	7.5297	15.6000	2631.8	88.4050	0.4188	25.9651	1.8243
23		50	50	500	100	2272.6	2747.3	68.7120	20.6469	54.5000	2581.8	66.2905	0.5047	17.2523	11.9420
24		50	50	500	500	4229.5	5455.9	68.9315	28.6516	34.5000	4577	80.0359	0.6859	22.4812	7.5911
25	75	50	50	100	100	622.8990	794.1815	78.9943	3.9438	12.6000	637.6912	70.7837	0.2422	21.3467	2.1283
26		50	50	100	500	2573.2	3484.7	90.6039	8.0594	15.8000	2632	121.2170	0.4266	26.1898	2.2405
27		50	50	500	100	2289	2757.2	54.4159	27.1641	57.5000	2588.4	70.9270	0.5641	16.9958	11.5822
28		50	50	500	500	4284.8	5507.7	143.5341	28.1391	31.8000	4641.4	126.0542	0.7375	22.1955	7.6747
29	100	50	50	100	100	663.1588	865.8819	88.8559	8.8281	30.7000	684.2872	77.8045	0.2531	23.3672	2.9595
30		50	50	100	500	2532.7	3456.1	59.4832	9.0188	16.4000	2592.8	71.4109	0.4422	26.7103	2.2995
31		50	50	500	100	2305.5	2821.9	74.0348	25.1313	49.4000	2649.8	66.3914	0.6234	18.2569	12.9559
32		50	50	500	500	4182	5405.5	109.3389	30.8578	32	4537.1	131.9653	0.8078	22.6406	7.8144
33	25	100	100	200	200	1246.5	1583	62.0970	54.3094	50.6000	1252.3	47.2326	0.4969	21.2346	0.4482
34		100	100	200	500	2634.2	3560.8	56.6990	54.5812	35	2688.9	35.4166	0.7688	26.0167	2.0287
35		100	100	500	200	2709	3153.9	130.3294	84.5500	77	2838.2	139.4445	0.8281	14.1019	4.4996
36		100	100	500	500	4074.9	5194.1	140.2969	118.9000	57.2000	4306	124.6958	1.0531	21.5359	5.3538
37	50	100	100	200	200	1260.4	1579.9	77.4642	52.2313	44.4000	1270.7	67.9033	0.5281	20.1980	0.7712
38		100	100	200	500	2697.9	3641.6	125.9619	42.6219	28.4000	2756	153.8828	0.7813	25.9101	2.0966
39		100	100	500	200	2562.8	3044.8	73.0194	148.5656	101.4000	2696	102.6904	0.8500	15.8426	4.9312
40		100	100	500	500	3978.4	5096.3	105.5789	125.0719	62.4000	4209.8	82.0509	1.1375	21.9291	5.4933
41	75	100	100	200	200	1255.6	1612.9	42.5691	85.2406	63.8000	1275.4	67.2191	0.5375	22.1537	1.5483
42		100	100	200	500	2711.9	3683.5	29.0749	51.3500	32.6000	2775.8	99.9843	0.8000	26.3931	2.3101
43		100	100	500	200	2586.1	3039.6	131.0485	108.6281	72.8000	2769.2	95.2660	0.9031	14.8680	6.6020
44		100	100	500	500	3983.2	5108.9	96.7893	113.1469	55.2000	4223.8	111.6474	1.1406	22.0329	5.6904
45	100	100	100	200	200	1318.8	1649.6	70.3429	90.2156	60.4000	1340.7	63.9094	0.5594	20.0475	1.6257
46		100	100	200	500	2681.1	3622.4	76.0711	57.9453	35.2500	2733.7	108.2080	0.8164	26.0016	1.8915
47		100	100	500	200	2549.2	3121.6	128.1110	154.1375	93.2000	2778	119.7769	0.9344	18.2931	8.1927
48		100	100	500	500	3992.3	5062.4	116.0704	122.7125	56.2000	4200.6	149.1135	1.1844	21.1434	4.9305

the size of the instances and that they are very tight when compared to the optimal solution of the problem. On the other hand, the cpu times show that the LP relaxations are faster than IMM algorithm. For the latter, we observe a rapid growth which is directly related to the size of the instances. Concerning the average number of LPs IMM needs to converge, we notice a slightly increasing trend. Then, the growth in cpu time can be explained by the size of the LPs it solves within each iteration. Finally, we can see that the standard deviations show a constant behavior when compared to the average upper bounds in both cases, for the IMM and for the LP relaxation. The numerical results in table II, are a little bit different. Here, we observe that the relative gaps are not as tight as in table I for the LP relaxations, but still better than those obtained with IMM algorithm. Although, they become tighter as the size of the instances increase which is an interesting result. We can also see that the effect of increasing the number of scenarios in the probabilistic constraints does not have a significant impact on the numerical results. It is easy to note that these gaps are tighter when $n_1 < n_2$. Concerning the cpu times, we observe an increasing trend for the Lagrangian approach while for the LP relaxation they almost remain unchanged. The average number of LPs solved by IMM shows a slight increasing trend. Finally, we observe that the standard deviation behaviors are similar.

VII. CONCLUSION AND FUTURE WORK

In this paper, we computed upper bounds and optimal solutions for a deterministic linear bilevel programming problem and a probabilistic constrained linear bilevel counterpart due to [7]. The upper bounds were computed using the iterative minmax algorithm proposed in [10] and also using linear programming relaxations we derived according to the approach proposed in [1].

To this end, we transformed all the linear bilevel models into the so called Global Linear Complementarity problems from which we derived equivalent MIP and LP formulations. Our numerical results showed tight relative gaps for the upper bounds obtained with the LP relaxations. On the opposite, those obtained with IMM algorithm were considerably larger in all the instances we tested. In particular, we obtained better gaps on deterministic instances rather than for the stochastic ones, which means that probabilistic constraints decrease the effectiveness of the LP relaxations.

Finally, we argue that even when the LP relaxations give tighter bounds on these specific problems, IMM algorithm still provides a more general framework as it can be used to handle any type of non-linear constraints. Therefore, future research should also be devoted to strengthen IMM while testing it on different types of problems.

ACKNOWLEDGMENT

The first author is grateful for the financial support given by Conicyt Chilean government through the Insertion project number: 79100020.

REFERENCES

[1] Audet C., P. Hansen, B. Jaumard, and G. Savard, *Links Between Linear Bilevel and Mixed 0–1 Programming Problems*, Journal of Optimization Theory and Applications, Vol. **93**, pp. 273-300, 1997.

[2] Audestad J., A. Gaivoronski, and A. Werner, *Extending the stochastic programming framework for the modeling of several decision makers: Pricing and competition in the telecommunication sector*, Annals of Operations Research, Vol. **142**, pp. 19-39, 2006.

[3] Birge J. and F. Louveaux, *Introduction to stochastic programming*, Springer-Verlag, New York, 1997.

[4] Christiansen S., M. Patriksson, and L. Wynter, *Stochastic bilevel programming in structural optimization*, Structural and Multidisciplinary Optimization, Vol. **21**, pp. 361-371, 2001.

[5] Dimitris B. and Vineet Goyal, *On the Power of Robust Solutions in Two-Stage Stochastic and Adaptive Optimization Problems*, Mathematics of Operations Research, Vol. **35**, pp. 284-305, 2010.

[6] Floudas C. and P. Pardalos (Eds.), *Encyclopedia of Optimization*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 2001.

[7] Gaivoronski A., A. Lisser, and R. Lopez, *Knapsack problem with probability constraints*, Journal of Global Optimization, Vol. **49**, pp. 397-413, 2011.

[8] Hansen P., B. Jaumard, and G. Savard, *New Branch and Bound Rules for Linear Bilevel Programming*, SIAM Journal on Scientific and Statistical Computing, Vol. **13**, pp. 1194-1217, 1992.

[9] Klein W. and M. van der Vlerk, *Stochastic integer programming: general models and algorithms*, Annals of Operations Research, Vol. **85**, pp. 39-57, 1999.

[10] Kosuch S., P. Lebodic, J. Leung, and A. Lisser, *On Stochastic Bilevel Programming Problem*, Networks, Vol. **59**, pp. 107-116, 2012.

[11] Luh P., T. Chang, and T. Ning, *Pricing problems with a continum of customers as stochastic Stackelberg games*, Journal of Optimization Theory and Applications, Vol. **55**, pp. 119-131, 1987.

[12] Özaltin O., O. Prokopyev, and A. Schaefer, *The Bilevel Knapsack Problem with Stochastic Right-Hand Sides*, Operations Research Letters, Vol. **38**, pp. 328-333, 2010.

[13] Patriksson M. and L. Wynter, *Stochastic mathematical programs with equilibrium constraints*, Operations Research Letters, Vol. **25**, pp. 159-167, 1999.

[14] Sahinidis N., *Optimization under uncertainty: State-of-the-art and opportunities*, Computers and Chemical Engineering, Vol. **28**, pp. 971-983, 2004.

[15] Schultz R., S. Leen and M. Van Der Vlerk, *Two-stage stochastic integer programming: a survey*, Statistica Neerlandica, Vol. **50**, pp. 404-416, 1996.

[16] Shapiro A., D. Dentcheva, and A. Ruszczyński, *Lectures on Stochastic Programming: Modeling and Theory*, Volume 436. SIAM Philadelphia, Series on Optimization, Vol. **9** of MPS/SIAM, Philadelphia, 2009.

[17] Werner A., *Bilevel stochastic programming problems, analysis and application to telecommunications*, PhD. thesis, Department of Industrial Economics and Technology Management, Norwegian University of Science and Technology, 2004.

Measuring the Objective Complexity of Assembly Workstations

Definition and Analysis of Production Complexity

Luiza Zeltzer, Veronique Limère, El-Houssaine Aghezzaf, Hendrik Van Landeghem
Department of Industrial Management
Faculty of Engineering and Architecture, Ghent University
Ghent, Belgium
{Luiza.Zelter, Veronique.Limere, ElHoussaine.Aghezzaf Hendrik.VanLandeghem}@UGent.be

Abstract — **The large number of product variants, produced to satisfy customers, increases significantly the complexity of manufacturing systems. As a consequence, new approaches to deal with production processes are required. Because of the impact of complexity on productivity, it is in the first place important to understand what complexity is and what are its main drivers. Based on real data, a model is suggested to characterize workstations complexity. The model is presented and its validity and accuracy are discussed. This paper defines complexity in a production environment and it proposes an identifier for complex assembly workstations. This definition is able to characterize different manufacturing systems and to define a system as high complex or low complex.**

Keywords: Complexity; Complexity Definition; Mixed-model Assembly Line

I. Introduction

Over the last couple of years, in automotive industry, the number of introductions of new and different car models has increased drastically. These new models are mainly introduced to answer consumers' needs [1], but also because of the new generation of electrical driven products.

The increase of product variety is necessary to answer the market and sustainability demands, however this high variety makes the mixed-model assembly lines become rather complex. The introduction of new models increases the complexity of (re)designing factory processes and workstations, and consequently increases significantly the overall complexity of the production system.

Currently, one can conclude that the elements presented above, increase the manufacturing complexity but the exact causes and impacts on manufacturing processes are still unknown. This paper proposes a clear and objective definition of production complexity and attempts to determine its main drivers. To this end, the drivers of complexity are determined and a model is proposed enabling to define the complexity of a workstation.

The approach used is based both on theoretical and practical information, i.e. as a result of literature study and interactions with manufacturers. Section II presents a brief literature review and describes complexity in the manufacturing domain. In Section III, a complexity definition is proposed and the methodology of the study is presented. Section IV presents the results obtained and discusses some perspectives. The conclusions and future work are presented in Section V.

II. Literature Review

Lately, there has been a growing interest in the study of complexity of manufacturing processes and systems [2] [3]. One can distinguish three types of complexity: product complexity, process complexity and operational complexity. According to literature, one of the influencing factors of complexity is the manipulation of information. Complexity is directly related to the quantity of information, diversity of information and the information content [4]. The human element is also important [5] since it influences system performance. Moreover, complexity increases with the number of different product variants to be produced and the number of tasks within the production process [4].

In an attempt to understand complexity, its main drivers are determined and a taxonomy is proposed. The drivers of complexity are identified as: uncertainty, dynamics, multiplicity, variety, interactions and interdependencies [6] and a combination of such proprieties can render a system complex or not complex. A specific taxonomy is proposed where complexity is split in static and dynamic [7]. Static complexity is associated with the product, whereas dynamic complexity is linked to the process.

Although until now different approaches have been developed associating manufacturing complexity to product, process and human operator, it should be emphasized that there is to the best of our knowledge no existing model that quantifies the relationship between complexity as perceived by the operators and its drivers. This is the focus of this paper. In the following section a definition of complexity within the domain of this work is introduced.

III. Production Complexity Definition and Methodology

The research on which this paper is based was carried out within the vehicle industry of Belgium and Sweden, including both OEM's and their suppliers. The focus is therefore on workstations along a driven assembly line, which work on different vehicle models in a mixed model

ISBN: 978-1-61208-202-8

fashion. The research was subsequently carried out in three steps.

A. Complexity Definition

A good definition of complexity has to be generic enough to be applicable to different manufacturing systems and at the same time specific enough to guide the decision whether a system is complex or not. Although the literature review provided useful insights about manufacturing complexity, there still existed a need for a clear, simple and generic complexity definition. After extensive communication with the project partners the following definition is proposed:

"*Complexity is the sum of all aspects and elements that makes a task or a set of tasks mentally difficult, error-prone, requiring thinking and vigilance and inducing stress*".

This definition recognizes the fact that complexity of tasks is determined to a large extent by the person that executes them, hence termed subjective complexity. In many cases the same set of tasks can be judged differently by different people under different circumstances. This makes quantifying complexity in an unambiguous manner, the objective complexity, a real challenge. This paper focuses on how to measure complexity in an objective, repeatable manner.

B. Model Building Workshops

A series of workshops was done in collaboration with a group of automotive manufacturers, to gather knowledge about complexity in industry. Components of complexity were identified and classified as drivers or impacts and used to build a model.

Those workshops were a great opportunity to study and explore real manufacturing situations where complexity is present. In order to be able to gather as much useful information as possible, the participants included shop floor employees, production engineers, quality controllers and line management, who deal with complexity in their daily activities. All workshops were organized in a similar way:

Initially, the project objectives were explained to all participants. Next, the participants were asked to identify some low and some high complex workstations. Afterwards, three sets of open questions were presented to them and during a limited period of time they could reflect individually how these questions applied to the low and high complex cases respectively.

The goal of presenting the questions was to situate and identify how complexity is experienced. The questions were structured in three different sets. The first set focused on characterizing complexity. The second set of questions concentrated on revealing which consequences complexity has. The questions focus on the areas that are affected by complexity and on the influence of complexity on manufacturing work and teams. The third set of questions aimed at detecting the direct drivers of complexity, i.e. the variables that are directly linked to the complexity elements as causal factors.

Finally, after the participants' individual analysis, a brainstorming session was done where a list of ideas were discussed and gathered.

As the result of these workshops a high amount of important information was produced. In the next subsection, a causal model is presented as an outcome of the investigation of this information.

C. Causal model

As a result of the workshops, a causal model is defined with the goal to obtain a generic complexity model.

The model clusters the variables related with complexity

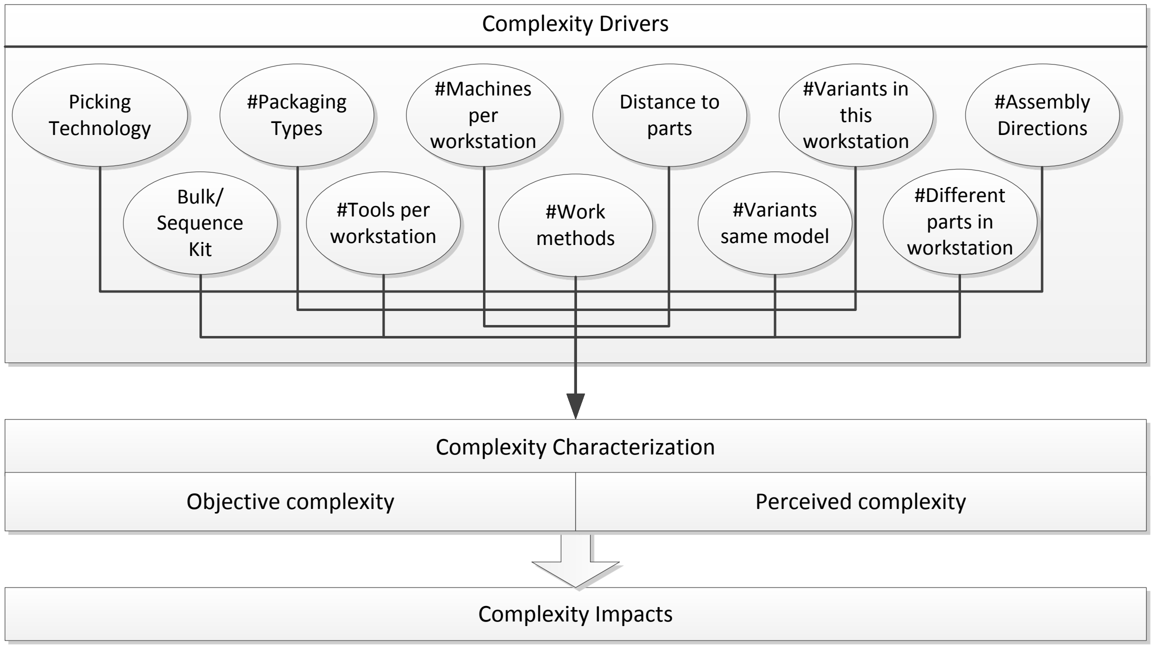


Figure 1. Complexity Causal Model

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-202-8

characterization (first set of questions), complexity impacts (second set of questions), and complexity drivers (third set of questions) into groups. Fig. 1 shows how the three categories of variables are linked together. As a result of the workshops, a causal model is defined with the goal to obtain a generic complexity model.

The model clusters the variables related with complexity characterization, complexity impacts, and complexity drivers into groups. Fig. 1 shows how the three categories of variables are linked together.

Complexity drivers are the variables that are linked with the source of complexity and are therefore represented at the top of the figure.

Next, complexity characteristics describe complexity. These characteristics are clustered into 2 main groups: objective complexity and perceived complexity. The main difference between both groups is that objective complexity can be analyzed quantitatively on external values, while perceived complexity can only be studied through the cognitive behavior of the operators. The authors decided to focus their research on the objective measurement of complexity. The subjective complexity was studied by a separate team in Sweden.

Finally, complexity impacts were derived from objective and perceived complexity.

IV. Results

The information obtained through the workshops was thus structured in a highly detailed causal model. Fig. 1 only shows an extract of the model, detailing the 11 direct drivers of complexity that were identified. These were further investigated.

A. *Complexity drivers*

The list of complexity-driving variables is presented in Table I together with a concise explanation of each variable. The next question to tackle was to characterize the relation between these variables and complexity, in an attempt to build a model. This set of variables is crucial to recognize what increases or decreases complexity. In order to now develop an objective complexity identifier for work stations, a dataset was created with the collaboration of the participants of the workshops.

TABLE I. Complexity-Driving Variables

Complexity-driving variables	**Description**
Picking technology	Fixed (F) : Operator takes part always on the same location from bulk storage. Signal (S) : Operator picks part from location indicated by a signal (light, display) Comparing (C) :Operator must compare simple information (symbols, colors) Manual (M) : Operator must read extensive information from manifest
Bulk/Sequence Kit	Sequenced (S) : Every part is in its package in correct assembly sequence Kit (K) : Parts are delivered in kits with exact set for one assembly operation Bulk (B) :Parts are by type in their own package
# Packaging types	The total number of different packaging types, a type having a specific layout. So 2 identical boxes with different inserts are 2 different types.
#Tools per workstation	The number of tools that the operator(s) need to handle to perform all possible assembly variants in this station, automatic tools (servants) excluded.
# Machines per workstation	Machines that perform automated tasks without operator assistance, with automatic or manual start.
# Work methods	Every unique set of work methods the operator must master in this workstation. A method contains several small steps.
Distance to parts	The farthest distance between the normal operator position (or the center of the workstations) and the parts at the border of line.
# Variants same model	The highest number of variants belonging to one model, among all models of which parts are assembled in this workstation.
# Variants in this workstation	Total number of variant parts, summed over all models that are assembled in this workstation.
# Different parts in workstation	Total number of unique part references that are assembled in this workstation, including all variants and models that typically occur in one year.
# Assembly directions	The number of different positions the operator must take to complete his workstation cycle, including repositionings of the upper body or the feet, but not small repositionings of the hands.

B. ***Experimental Dataset***

Using the list presented in Table I, the manufacturing collaborators were asked to select five workstations and define the value for each variable (driver). Moreover they were asked to classify each workstation as low complex or high complex. The result is a dataset composed of 76 workstations, 41 classified as low complex and 35 classified high complex, and the respective driver values.

In order to have more control over the scaling of each variable, we set up a Likert scale for each variable, dividing the data range over 4 levels on the scale. The result is shown in Table II.

TABLE II. DIRECT DRIVERS SCALE

Complexity-driving variables	**Likert scale**			
Picking technology	F	S	C	M
	1	2	3	4
Bulk/Sequence Kit	S	K	B	
	1	2	3	
# Packaging types	1	2-4	5-8	>8
	1	2	3	4
#Tools per workstation	1	2-4	5-8	>8
	1	2	3	4
# Machines per workstation	0	1	2	>2
	0	1	2	3
# Work methods	0-2	3-5	6-8	>8
	1	2	3	4
Distance to parts	0-1	1, 1-2	2, 1-4	>4
	1	2	3	4
# Variants same model	1	2-3	4-5	>5
	1	2	3	4
# Variants in this workstation	1	2-4	5-10	>10
	1	2	3	4
# Different parts in workstation	1-4	5-10	11-20	>20
	1	2	3	4
# Assembly directions	1	2-3	4-5	>5
	1	2	3	4

C. *Initial Model*

A complexity measurement is developed based on a weighted sum of the 11 variables. This measure determines if workstations have a low or high complexity according to equations 1 and 2:

where:

- basic complexity(w) is the complexity score of a workstation *w*,
- Score(i) is the value of the variable i according to the Likert scale,
- Weight(i) is the weight of the variable *i*,

$$\text{basic complexity}(w)=\sum_{\mathrm{i=1}}^{\mathrm{n}}\frac{\text{score(i)*weight(i)}}{\sum_{i=1}^{n} weight(i)} \tag{1}$$

$$\text{complexity(w)}=\frac{\text{basic complexity(w)} - \sum_{\mathrm{i=1}}^{\mathrm{n}} \text{min i}}{\sum_{\mathrm{i=1}}^{\mathrm{n}} \text{max i} - \sum_{\mathrm{i=1}}^{\mathrm{n}} \text{min i}}\ .10 \tag{2}$$

- max i is the maximum score value for variable i,
- min i is the minimum score value for variable i,
- complexity(w) is the complexity score of a workstation normalized into a scale from 0 to 10.

Fig. 2 shows the result of the calculated complexity measure compared with the subjective labels of LOW and HIGH complexity for each of the 76 workstations. The score for the LOW complexity workstations averages 4,8 and HIGH averages 7,2. The calculated score seems to distinguish HIGH from LOW complexity workstations, so the variables it is based on do seem to have a relation with the subjective complexity level. However, there is quite some fluctuation in the complexity scores. This suggests that not all variables have the same explanatory power, or even that some variables contradict others. Therefore, the next step is to adjust the weights or reduce the number of variables. In the following subsection a statistical model is developed to achieve just that.

D. *Statistical model*

The objective is to determine a good model for the prediction of the complexity of a workstation (high or low), based on the data gathered for the 76 work stations and their characterizing values for each of the 11 variables shown in Table I. Since the independent variable 'complexity' is a binary variable – it is either high or low – Logistic Regression is chosen for the analysis. In the analysis a prediction 0 corresponds to a high complex station, whereas

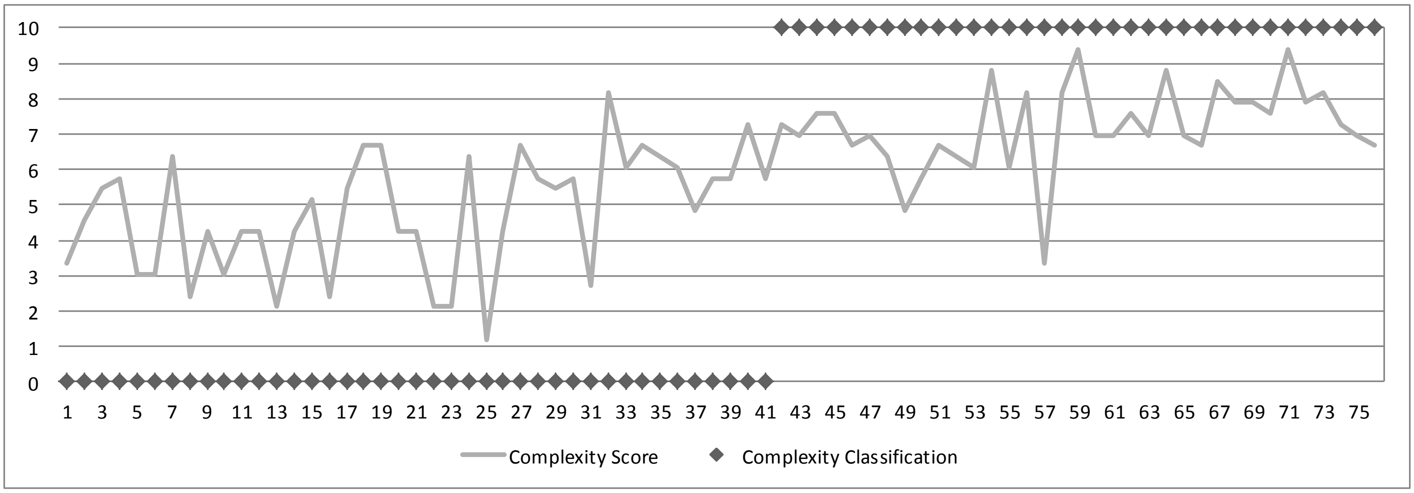


Figure 2. Workstations identification – Initial Model Versus Initial Classification

a prediction 1 corresponds to a low complex station.

Logistic regression will calculate the probability that the workstation's complexity is high or low from a combination of variable values in the following way:

$$Odds = \frac{P}{1-P} = e^{A+BX}$$
$$Ln(Odds) = A + BX = a + b_1 x_1 + b_2 x_2 + \dots \quad (3)$$

Where,

- P the chance to have a low complex station
- 1-P the chance to have a high complex station
- a a constant
- b_n coefficient for variable n
- x_n value for variable n

Based on all 76 cases, a model could be found with only 4 of the variables, able to classify 84% of all 76 cases correctly. The output is presented in TABLE III.

In the output, it can be seen that the predictive variables are:

- the Likert scale value for the number of packaging types,
- the number of assembly directions as measured directly
- the Likert scale value for the number of different parts in the workstation
- the number of work methods as measured directly

Of the 35 workstations identified as high complex, 31 could correctly be predicted as high complex by the model. Of the 41 workstations identified as low complex, 33 could correctly be predicted as low complex by the model. Fig. 3 shows the results.

TABLE III. STATISTICAL MODEL

Classification Table[a]

Observed			Predicted		
			HIGHLOW		Percentage Correct
			High	Low	
Step 1	HIGHLOW	High	31	4	88,6
		Low	8	33	80,5
	Overall Percentage				84,2

a. The cut value is ,500

Variables in the Equation

		B	S.E.	Wald	df
Step 1[a]	@#Packagingtypes	-1,127	,592	3,622	1
	Assemblydirections	-,243	,193	1,591	1
	@#Differentpartsinworkstation	-,874	,348	6,300	1
	Workmethods	-,058	,028	4,491	1
	Constant	6,676	1,837	13,200	1

V. CONCLUSION AND FUTURE WORK

This paper proposes a definition of production complexity wide enough to characterize different manufacturing systems and at the same time specific enough to define a system as high complex or low complex. A set of complexity direct drivers is extracted from real production data and interactions with manufacturers. Based on this set of complexity direct drivers two different complexity models are developed with the goal to measure and determine if workstations have a low and high complexity.

An initial model is proposed based on a complexity measure score. Then a statistical model is proposed based on logistic regression. To validate the proposed models, a set of experiments were carried out based in a set of 76 workstations which were classified as low or high complex. Initially this set contained 41 workstations classified as low complex and 35 workstations classified as high complex. The initial model was able to classify 82% of the

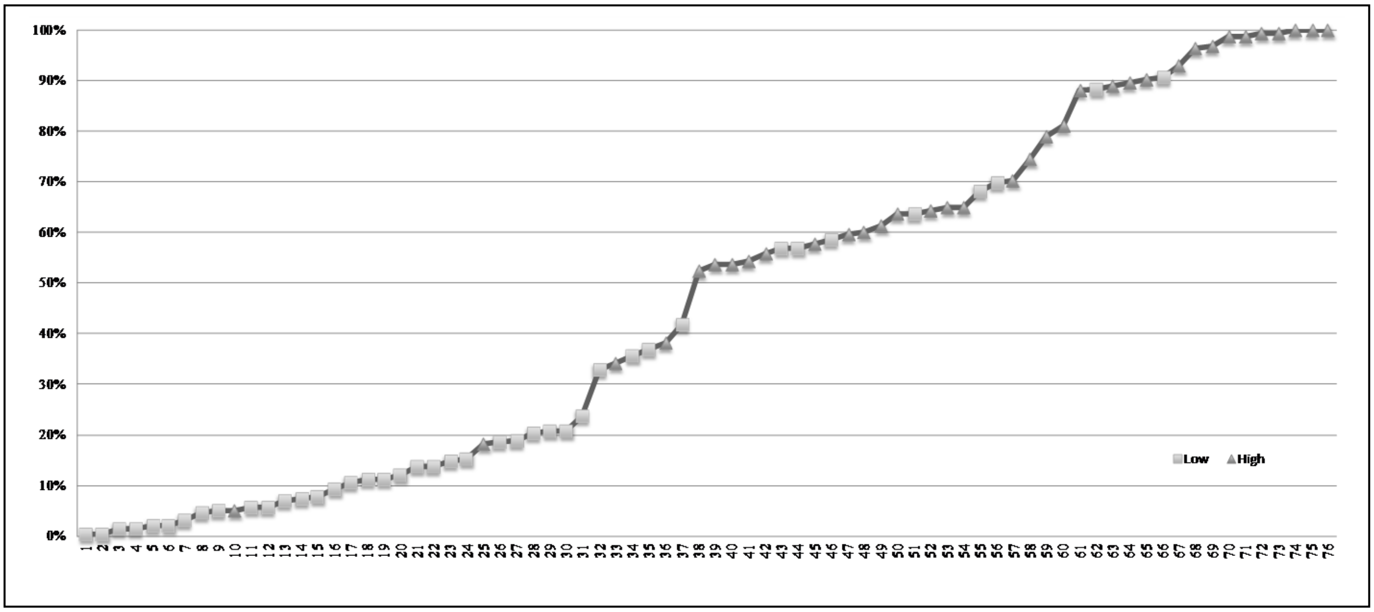


Figure 3. Workstations identification – Logit Model Versus Initial Classification

workstations correctly and the statistical model 84% of the workstations correctly.

The results obtained by the two models, provide some insight into the complexity-driving variables and their related scores. The results could also be used to measure the impact of complexity on both direct and indirect costs. They can also be useful for the subjective interpretation of complexity by the operator in the workstation.

The models give important insights in the impact of certain complexity drivers. Using the information from the models one should look into the extreme cases with wrong subjective labels, to assess whether the subjective label is wrong or not. In the former case this will enhance the value and validity of the models, and yield information about the subjective reasoning that led to the wrong classification. In the future the validity of the models should be further studied.

ACKNOWLEDGMENT

The authors would like to thank the industrial collaborators: Ford Genk, SAS, SML Genk, VCST, Van Hool Industrial Vehicles, Volvo Cars Ghent and Volvo Trucks Ghent. The authors also gratefully acknowledge the support of Flanders' Drive, the innovation center for the vehicle industry in Flanders.

REFERENCES

[1] J.P. MacDuffie, K. Sethuraman, and M.L. Fisher, "Product Variety and Manufacturing Performance: Evidence from the International Automotive Assembly Plant Study", Management Science, 1996, vol. 42 no. 3, pp. 350-369.

[2] Y.S. Kim, "A System Complexity Approach for the Integration of Product Development and Production System Design", MSc at MIT, USA, 1999.

[3] H.P. Wiendahl, and P. Scholtissek, "Management and Control of complexity in Manufacturing", CIRP Annals, 1996, vol. 43, issue 2, pp.: 533–540.

[4] W.H. ElMaraghy, and R.J. Urbanic, "Modelling of Manufacturing Systems Complexity", CIRP Annals, 2003, vol. 52, issue 1, pp.363-366.

[5] W.H. ElMaraghy, and R.J. Urbanic, "Assessment of Manufacturing Operational Complexity", CIRP Annals, 2004, vol. 53, issue 1, pp. 401–406.

[6] G. Schuh, L. Monostorib, B.Cs. Csájib, and S. Döringa. "Complexity-based Modeling of Reconfigurable Collaborations in Production Industry" , CIRP Annals, 2008, vol. 57, issue 1, pp. 445–450.

[7] C. Rodríguez-Toro, G.Jared, and K. Swift, "Product-development Complexity Metrics: a Framework for Proactive-DFA Implementation", 8th International Design Conference, Dubrovnik, Croatia. 2004, pp.:483-490.