12TH MIDDLE EASTERN SIMULATION & MODELLING MULTICONFERENCE

MESM'2011

2nd GAMEON-ARABIA CONFERENCE

GAMEON-ARABIA'2011

EDITED BY

Marwan Al-Akaidi

Ken Newman

NOVEMBER 14-16, 2011

Amman, Jordan

A Publication of

EUROSIS-ETI

Printed in Ghent, Belgium

12TH MIDDLE EASTERN SIMULATION & MODELLING MULTICONFERENCE

2ND GAMEON-ARABIA CONFERENCE

NOVEMBER 14-16, 2011

Organised by

ETI

Sponsored by

EUROSIS

AOU

BITE	IEEE UKRI - SPC	German Jordanian University
Ghent University	HTI	Orange
LITIS	Université du Havre	University of Skovde

Hosted by

AOU – Arab Open University

Amman, Jordan

INTERNATIONAL PROGRAMME COMMITTEE

Modelling Methodology and Modelling Tools
John Dalseng, Finnmark College, Alta, Norway
Helge Hagenauer, Universitaet Salzburg, Salzburg, Austria
Brian Hollocks, Bournemouth University, Bournemouth, United Kingdom
Bjoern Johansson, Chalmers University of Technology, Goeteborg, Sweden
Panajotis T. Katsaros, Aristotle University Thessaloniki, Thessaloniki, Greece
Yan Luo, NIST, Gaithersburg, USA
Michel Minoux, Universite P. et M. Curie-CNRS, Paris cedex 05, France
Tudor Niculiu, University Politecnica Bucharest, Bucharest, Romania
Zdravko Terze, University of Zagreb, Zagreb, Croatia
Dmitry Zaitsev, Odessa National Telecommunications Academy, Odessa, Ukraine

Analysis Methodology
Abdel Aitouche, Hautes Etudes Industrielles, Lille Cedex, France
Bruno Ciciani, Universita di Roma "La Sapienza", Rome, Italy
John Dalton, University of Newcastle, Newcastle-upon-Tyne, United Kingdom
Julie Dugdale, Universite Paul Sabatier, Narbonne, France
Mauro Iacono, SUN - II Universita di Napoli, Aversa (CE) Italy
Jens Kohlmeyer, Universitaet Ulm, Ulm, Germany
Francesco Moscato, Seconda Universita di Napoli, Aversa, Italy
Jairo R. Montoya-Torres, ENS des Mines de Saint-Etienne, Gardanne, France
Nadia Zanzouri, ENIT, Tunis, Tunisia

Simulation of Networks and Communication
Salvatore D'Antonio, Consorzio Interuniversitario Nazionale per l'Informatica, Naples, Italy
Mahmoud Al-Qutayri, Khalifa University of Science, Sharjah, UAE
Med Salim Bouhlel, Sfax University, Sfax, Tunisia
Eugen Dedu, LIFC, Montbelliard, France
Carlos Palau, UPV, Valencia, Spain
Djamel Sadok, The Federal University of Pernambuco, Brazil
Ahmad A. Sharieh, University of Jordan, Amman, Jordan
Veronique Veque, Universite de Paris Sud LRI, Orsay, France

Complex Systems Simulation
Marta Pla Castells, Universidad de Valencia, Valencia, Spain
Michele Colajanni, Universita di Modena, Modena, Italy
Krzysztof Fujarewicz, Silesian University of Technology, Gliwice, Poland
Tom Dhaene, Ghent University, Ghent, Belgium
Daniel Huber, Heinz Nixdorf Institute, University of Paderborn, Paderborn, Germany
Jan Lemeire, VUB, Brussels, Belgium
Jorge M. Perez, ITP, S.Fernando de Hernares, Spain
Panos Periorellis, University of Newcastle Upon Tyne, Newcastle, United Kingdom
Fatima Rateb, University College London, London, United Kingdom
Marco Remondino, University of Turin, Turin, Italy
Marc Rouff, University of Caen Basse-Normandie, Caen, France
Jan Studzinski, Systems Research Institute of Polish Academy of Science, Warsaw, Poland

Simulation of Signal and Image Processing
Bruno Ciciani, Universita di Roma "La Sapienza", Rome, Italy
Aboul Ella Hassanine, Kuwait University, Safat, Kuwait
Tania Jimenez, LIA, Universite d'Avignon, France
Viorel Nicolau, "Dunarea de Jos" University of Galati, Romania
Bogdan Raducanu, UAB, Cerdanyola, Spain
Renate Sitte, Griffith University, Gold Coast Mail Centre, Australia
Joao Manuel R. S. Tavares, University of Porto, Porto, Portugal
Hanene Trichili, University of Sfax, Sfax, Tunisia
Ahmed M. Zeki, University of Bahrain, Bahrain

Energy Systems Simulation
Paulo Maciel, UFPE, Recife, Brazil
Musleh B. AL-Otaibi, Alahmadi, Kuwait
Ahmad Jammal, Ministry of Education & Higher Education, Beirut, Lebanon
Janos Sebestyen Janosy, KFKI, Budapest, Hungary
Carlos Ramos, Polytechnic of Porto (ISEP/IPP), Porto, Portugal
Tzanko Tzanov, Polytechnical University of Catalunya, Barcelona, Spain
Patrick Waeger, EMPA, St. Gallen, Switzerland

Simulation in the Chemical and Petrochemical Industry
Ammar Al Khani, Espoo, Finland
Diganta Bhusan Das, Loughborough University, Loughborough, United Kingdom
Eugenio Ferreira, Universidade do Minho, Braga Portugal
Vahid Nassehi, Loughborough University, Loughborough, United Kingdom
Mohammad Al-Nawafleh, Tafila Applied University College, Al-Balqa' Applied University, Tafila, Jordan
Cornel Resteanu, Research Institute of Informatics, Bucharest, Romania
M R Riazi, Kuwait University, Safat, Kuwait

Multimedia and Virtual Reality Systems
Christos Bouras, University of Patras, Rion, Patras, Greece
Eckhard C. Bollow, Software Technology and IT-Project Management, Lueneburg, Germany
Maria Paule Ruiz del Puerto, University of Oviedo, Oviedo, Spain
Zhisheng Huang, Vrije University Amsterdam, The Netherlands
Christoph Laroque, University of Paderborn, Paderborn, Germany
Ahmed Mostefaoui , Laboratoire d'Informatique de Franche-Comte Montbeliard Cedex, France
Elena Pelican, Ovidius University of Constanta, Romania

INTERNATIONAL PROGRAMME COMMITTEE

Dorin-Mircea Popovici, Ovidius University of Constanta, Romania
Christian Reimann, C-LAB, Paderborn, Germany
Marco Roccetti, University of Bologna, Bologna, Italy
Zasriati Azla Sabot, University Tun Abdul Razak (UNITAR) Selanngor, Malaysia
Mike Sips, University of Konstanz, Konstanz, Germany

Decision Processing in Management
Hadhoum Boukachour, Le Havre University, Le Havre, France
Naoufel Cheikhrouhou, Ecole Polytechnique Federale de Lausanne, Lausanne, Switzerland
Joel Colloc, Le Havre University, Le Havre, France
Jeanne Schreurs, Universiteit Hasselt, Diepenbeek, Belgium
Christine Sybord, Lyon 2 Lumiere University, Lyon, France
Behrouz Zarei, Sharif University of Technology, Tehran, Iran

Modelling and Simulation in Biomedical Systems
Manuel Alfonseca, Escuela Politecnica Superior, Universidad Autonoma de Madrid, Madrid, Spain
Clemens Heitzinger, University of Vienna, Vienna, Austria
Manuel Gonzalez Hidalgo, Edf. Anselm Turmeda, Palma de Mallorca, Baleares, Spain
Olaf Tietje, Systaim GmbH, Zurich, Switzerland

Modelling and Simulation for Industrial Applications
Track Chair: Prof. David Stockton, de Montfort University, Leicester, United Kingdom
Luis Felipe Acebes Arconada, Universidad de Valladolid, Valladolid, Spain
Leopoldo Arnesto, UPV, Valencia, Spain
Michel Benne, University of La Reunion - LE2P, Saint-Andre, France
Pascal Berruet, Universite Bretagne Sud, Lorient cedex, France
PJ Byrne, DCU Business School, Dublin City University, Dublin, Ireland
A. Chatzinikolaou, Bosch Rexroth SA, Athens, Greece
Olaf Diegel, Massey University, Albany Campus, North Shore Mail Centre, Auckland, New Zealand
S Collart Dutilleul, LAGIS-EC-LILLE, Villenneuve d'Ascq, France
Georges Habchi, Polytech Savoie, Annecy-le-Vieux, France
Sergej Kalaschnikow, VA TECH ELIN EBG Elektronik GmbH & Co, Vienna, Austria
Markus Koch, C-LAB, Paderborn, Germany
Habtom Mebrahtu, Anglia Polytechnica University, Chelmsford, United Kingdom
Pascal Meyer, Forschungszentrum Karlsruhe, Karlsruhe, Germany
Lars Moench, TU Ilmenau, Ilmenau, Germany
Carlos E. Pereira, Universidad Federal do Rio Grande Do Sul, Porto Alegre, Brazil
Caterina Rizzi, Universita degli Studi di Bergama, Dalmine, Italy
Amir Shirkhodaie, Tennessee State University, Nashville, USA
Hans Veeke, TU Delft, Delft, The Netherlands
Ramiro Velazquez, Universite Paris 6 - CNRS, Fontenay aux Roses, France
Joao de Silva Wellington, UFPE, Recife, Brazil

Simulation in Logistics, Traffic, Transport and Harbour Simulation
Paul Davidsson, Blekinge Institute of Technology, Ronneby, Sweden
Isabel Garcia Gutierrez, Universidad Carlos III de Madrid, Leganes, Madrid, Spain
Thomas Hanne, Fraunhofer-Institut ITWM, Kaiserslautern, Germany
Daniel Krajzewicz, GermanAerospace Centre, Berlin, Germany
Seung Man Lee, NASA Ames Research Center, Moffett Field, USA
Herve Manier, Universite de Technologie de Belfort-Montbeliard (UTBM), Belfort Cedex, France
Vittorio Maniezzo, Universita di Bologna, Bologna, Italy
Rafael J. Martinez, Universitat de Valencia, Valencia, Spain
Philippe Mussi, INRIA Sophia Antipolis, France Frederic Serin, IUT du Havre, Le Havre, France

Web based Simulation
Victor Bassilious, University of Abertay Dundee, United Kingdom
Jorge Marx Gomez, Otto-von-Guericke-Universitaet Magdeburg, Magdeburg, Germany
Wolfgang Kuehn, University of Wuppertal, Wuppertal, Germany
Jose Machado, Universidade do Minho, Braga Codex, Portugal
Mara Nikolaidou, University of Athens, Ilissia, Athens, Greece
Francesco Quaglia, Universita' di Roma "La Sapienza", Roma, Italy
Stefan Sarstedt, University of Applied Sciences Hamburg, Hamburg, Germany
Alfonso Urquia, UNED, Madrid, Spain

Simulation in Information Processing
Julien Bourgeois, LIFC, Montbelliard, France
Christophe Claramunt, Naval Academy Research Institute, Brest Naval, France
Patrick De Causmaecker, KU Leuven, Camous Kortrijk, Kortrijk, Belgium
Khalil Drira, LAAS-CNRS, Toulouse Cedex 04, France
Laurent Guillon, Naval Academy Research Institute, Brest Naval, France
Jean-Claude Hochon, GFI Consulting, Toulouse, France
Peter Lawrence, Deakin University, Bentleigh, Australia
Christophe Roche, Universite de Savoie, Le Bourget du Lac cedex, France
Christian Siemers, TU Clausthal, Clausthal-Zellerfeld, Germany
Jim Torresen, University of Oslo, Oslo, Norway

Simulation in Fuzzy Systems, Neural Networks and Genetic Algorithms
Track Chair: Kifaya Qaddoum, Warwick University, Coventry, United Kingdom
Chrissanthi Angeli, Technological Education Institute of Piraeus, Athens, Greece

INTERNATIONAL PROGRAMME COMMITTEE

Piotr Bania, AGH - University of Science and Technology, Krakow, Poland
Mokhtar Beldjehem, St.Anne's University, Halifax, Canada
Paulo Alexandre Ribeiro Cortez, University of Minho, Guimaraes, Portugal
Rashidy Kanan, University of Amirkabir, Tehran, Iran
Antoni Ligeza, AGH - University of Science and Technology, Krakow, Poland
Marie-Ange Manier, Universite de Technologie de Belfort-Montbeliard (UTBM), Belfort Cedex, France
Viorel Nicolau, University of Galati, Galati, Romania
Paulo Moura Oliveira, UTAD-Universidade de Tres-os-Montes e Alto Douro, Vila Real, Portugal
Angus Wu, City University of Hong Kong, Hong Kong

Simulation in the Built Environment
Bernadette O'Regan, University of Limerick, Limerick, Ireland
Agustin Yague, Universidad Politicnica de Madrid, Madrid, Spain

Simulation in Archaeology
Jean Yves Blaise, UMR CNRS/MCC 694 MAP, Marseille, France
Anders Hast, University of Gavle, Gavle, Sweden
Wolfgang Kreutzer, University of Canterbury, Christchurch, New Zealand
Marcos A Rodrigues, Sheffield Hallam University, Sheffield , United Kingdom

Simulation Projects in the Arab World
Saleh Al-Jufout, Tafila Technical University, Tafila, Jordan
Mohamed-Aly O. Louly, University of Technology, Troyes, France

TIE Workshop
Habib Abdulrab, INSA Rouen, Rouen, France
Aladdin Ayesh, De Montfort University, Leicester, United Kingdom
Eduard Babkin, Higher School of Economics (Nihzny Novgorod), Russia
Cyrille Bertelle, Le Havre University, Le Havre, France
Gérard H.E. Duchamp, Paris XIII University, Paris, France
Rawan Ghnemat, Princess Sumaya University of Technology, Amman, Jordan
Laszlo Gulyas, Eotvos University, Budapest, Hungary
Barna L. Iantovics, Petru Maior University of Tg. Mures, Romania

GAMEON-ARABIA
Marco Block-Berlitz, Mediadesign Hochschule in Berlin, Berlin, Germany
Eva Hudlicka, Psychometrix Associates, Blacksburg, VA, USA
Pieter Jorissen, Karel de Grote Hogeschool, Hoboken, Belgium
Ian M Marshall, Coventry University, Coventry, United Kingdom
Joseph M. Saur, GTRI, Joint Systems Integration Center, USA

MESM’2011

GAMEON-ARABIA’2011

All articles published in these Proceedings have been peer reviewed

EUROSIS-ETI Publications are ISI-Thomson and INSPEC referenced

A CIP Catalogue record for this book is available from the Royal Library of Belgium under nr.12620

For permission to publish a complete paper write EUROSIS, c/o Philippe Geril, ETI Executive Director, Greenbridge NV, Wetenschapspark 1, Plassendale 1, B-8400 Ostend Belgium.

EUROSIS is a Division of ETI Bvba, The European Technology Institute, Torhoutsesteenweg 162, Box 4, B-8400 Ostend, Belgium

Printed in Belgium by Reproduct NV, Ghent, Belgium
Cover Design by Grafisch Bedrijf Lammaing, Ostend, Belgium

EUROSIS-ETI Publication
ISBN: 978-90-77381-67-0
ISN: 978-90-77381-67-0

Preface MESM’2011

Dear conference delegate,

I have pleasure to welcome you to the 12th Middle East Simulation & Modelling Multiconference (MESM2011), organized by the European Technology Institute and EUROSIS, held in the wonderful city of Amman, Jordan. The MESM2011 is sponsored by the IEEE-UKRI SPC and the Arab Open University.

This year’s event also includes for the second time presentations for GAMEON-ARABIA, which we will hope will introduce the non-gaming participants of the MESM to the exciting new simulation and AI technology and methodology of computer game design.

The aim of this conference is bring people from various parts of the Middle East in contact with colleagues working in Modelling & Simulation from around the world.

A number of studies have found Simulation & Modelling to be one of the most practical and effective problem solving techniques. However, there is little practical study literature available to guide those interested in the field. Modelling and simulation make a powerful combination to improve the system and organisation of the future.

The conference highlights recent and significant advances in many research areas of modelling and simulation related to Methodology, Networks Communication, Signal & Image Processing, Biomedical Applications, Industrial Applications, Software Engineering, Neural Networks and Fuzzy System.

Modeling and Simulation is a discipline for developing a level of understanding of the interaction of the parts of a system, and of the system as a whole. The level of understanding which may be developed via this discipline is seldom achievable via any other discipline.

A system is understood to be an entity which maintains its existence through the interaction of its parts. A model is a simplified representation of the actual system intended to promote understanding. Whether a model is a good model or not depends on the extent to which it promotes understanding. Since all models are simplifications of reality there is always a trade-off as to what level of detail is included in the model. If too little detail is included in the model one runs the risk of missing relevant interactions and the resultant model does not promote understanding. If too much detail is included in the model the model may become overly complicated and actually preclude the development of understanding.

Modelling and Simulation in Engineering aims at providing a forum for the discussion of formalisms, methodologies and simulation tools that are intended to support the new, broader interpretation of Engineering. Competitive pressures of the Global Economy have had a profound effect on manufacturing in Europe, Japan and the USA with much of the production being outsourced. In this context the traditional interpretation of the engineering profession linked to the actual manufacturing needs to be broadened to include the integration of outsourced components and the consideration of logistic,

Preface MESM'2011

economical and human factors in the design of engineering products and services. Modelling and Simulation in Engineering intends to report leading-edge scientific contributions from mathematics, computer science, various sub-disciplines of engineering, management, psychology and cross-cultural communication, all of which focus on the modelling and simulation of human-centred engineering systems. A simulation generally refers to a computerized version of the model which is run over time to study the implications of the defined interactions. Simulations are generally iterative in their development. One develops a model, simulates it, learns from the simulation, revises the model, and continues the iterations until an adequate level of understanding is developed.

Your presence at this conference emphasizes an important fact: the challenge and opportunities that surrounding the Modelling & Simulation practice anywhere in the world. We in EUROSIS believe that Modelling & Simulation make a powerful combination to improve the systems and organisation of the 21st century. It will continue to increase public awareness of the quality of life, and growing need to improve this through better organizations and systems.

As a General Conference & Programme Chair, I would like to express my thanks to the Rector of Arab Open University Professor Musa Mihsein, for sponsoring this conference and giving me the time to Chair this conference and thanks also to the committee members for reviewing the papers. Thanks are also due to Professor Mohammad Hamdan, for supporting this conference and to our local chair Prof. Abu Qudais, and his colleagues in AOU, Amman, Jordan, in organizing this event.

Thanks to my colleague Philippe Geril, executive director of EUROSIS office for supporting the event and for his time. Last but not least thanks to all authors without whom the conference would not be a successful conference.

We are sure you will enjoy your stay at the conference and we hope we will able to let you savour the sights and sounds of Amman and its surroundings during the next couple of days.

Professor Dr Marwan Al-Akaidi
General Conference and Programme Chair
MESM Chair
Arab Open University
P.O.Box: 3322 Safat
13033 Kuwait
Email:m.al-akaidi@arabou.edu.kw

Preface GAMEON-ARABIA'2011

In a year of great political upheaval across the Middle East and North African region it is particularly gratifying to acknowledge the small but dedicated cohort of researchers who work in this region and whose work is assembled in this volume. As always the diverse range of topics reflects a broad spectrum of the big questions in Game and Simulation research.

Within the GameOn sessions we have Nashed and Hull's investigation of machine learning in first-person shooter games, and an insightful analysis by Christou of social dilemmas, unfair practices and sanctions which have evolved in Massively Multiplayer games. Williams contributes to the theory of game design process with his nine interactive components model.

Collected in the Modeling and Simulation sessions (MESM) are Al-Tahat's visual approach to teaching object-oriented programing, Shawar's report on system integration for the purposes of organizational rules, Al-Sadi's review of eLearning tools and functionality to support disabled students. This section concludes with my own small contribution to web-based heuristics, data collecting techniques and structuring in a discussion of my plausible ancestor generator.

The GameOnArabia/ MESM' 2011 is a remarkable event not for its size but for the context in which it occurs amid the two big stories of the year - the Arab Spring and the global economic crisis. Thanks go to the organizers at the Arab Open University in Jordan and to EUROSIS for making it happen.

Prof. Ken Newman
IT and Creative Industries
HTC, Abu Dhabi

CONTENTS

CONTENTS

SCIENTIFIC PROGRAMME

DATA SIMULATION AND MANAGEMENT

Self-Organizing Maps and Principal component analysis for tomato yield datasets

KEFAYA QADDOUM, EVOR HINES, DACIANA ILLIESCU, MUHD KHAIRULZAMAN ABDUL KADIR

School of Engineering, University of Warwick, UNITED KINGDOM

k.s.qaddoum@warwick.ac.uk

Keywords. Principal component analysis, self-organizing map, yield prediction.

ABSTRACT

A Self-Organizing Map (SOM) is an artificial neural network model that uses competitive and unsupervised training. A SOM network has two main characteristics: it makes obtaining a simplified model of the training data possible and it has the ability to project them on a two dimensional map that shows the existing relations among them. Kohonen [5] suggested the first model of SOM, where the complete network structure had to be specified in advance and which remained static during all of the training process. PCA is a projection-based dimensionality reduction method to reduce data dimensionality by linearly transforming the data to a new coordinate system In this paper we present a good visualisation ability, where simulations have been carried out on tomato yield data sets for classification and prediction tasks, as well as on some data analysis have been done.

1 INTRODUCTION

1.1 SOM

Data clustering and classification is a straightforward approach as well as K-means algorithm (MacQueen 1967), which calculates each cluster centre as the mean of data vectors within the cluster. This is sometimes difficult if we have no a priori knowledge of the data distribution, all reference vectors are adjusted depending on their proximity to the input vector. Kohonen's self-organizing feature map (SOM) (Kohonen 1982) is another model incorporating a soft-max adapting rule, using a Mexican hat neighbourhood function to modify map nodes. It is a good choice for vector quantification in applications such as speech and image coding, featuring topology preserving ability and approximation of data distribution. SOM has found great success with a large amount of applications in various fields (Kohonen 1997).SOM is basically an unsupervised algorithm and it is rarely used for tasks such as time series prediction and pattern classification. Kohonen (1997) suggests using LVQ in these cases. Vesanto (1997) incorporates a local linear regression model on the top of a SOM network in a time-series prediction problem and achieved very good results. He constructs local data sets for the prototype vectors and uses linear regression models on these local data sets.

Other variations of the SOM models try to introduce improvements into the computation process and the effectiveness of the feature mapping. Blackmore and Miikkulainen (1993) proposed an incremental grid growing algorithm, where nodes can be added to, or deleted from, the feature map, which is of fixed low dimensionality. Fritzke (1991,1994) proposed a growing cell structure (GCS), which uses a fixed topology dimension for reference vector space, but there is no pre-defined layout order for network nodes and the topology is much more flexible. The network creates new nodes whenever input data is not closely matched to existing reference vectors, and sets up connections between nodes. Bruske and Sommer (1995) presented another similar model, dynamic cell structure (DCS-GCS), differing from GNG slightly in the location of node insertion. GCS, GNG and DCS-GCS can be applied to supervised learning by adding an additional output weight layer which adopts a delta learning rule. A comparison on the performance of GNG, GCS and fuzzy ARTMAP is made in (Heinke and Hamker 1998).

The name Self-Organizing Map (SOM) signifies a class of neural-network algorithms in the unsupervised-learning category. In its original form the SOM was invented by the founder of the Neural Networks Research Centre, Professor Teuvo Kohonen in 1981-82, and numerous versions, generalizations, accelerated learning schemes, and applications of the SOM have been developed since then.

The central property of the SOM is that it forms a nonlinear projection of a high- dimensional data manifold on a regular, low-dimensional (usually 2D) grid. In the display, the clustering of the data space as well as the metric-topological relations of the data items is clearly visible. If the data items are vectors, the components of which are variables with a definite meaning such as the descriptors of statistical data, or measurements that describe a process, the SOM grid can be used as a groundwork on which each of the variables can be displayed separately using a grey-level or pseudo-color coding. This kind of combined display has been found to be very useful for the understanding of the mutual dependencies between the variables, as well as of the structures of the data set.

The SOM has spread into numerous fields of science and technology as an analysis method. We have compiled a list of over 4000 scientific articles that apply the SOM or otherwise benefit from it.

The most promising fields of application of the SOM seem to be

- data mining at large, in particular

visualization of statistical data and document collections,

- process analysis, diagnostics, monitoring, and control,
- biomedical applications, including diagnostic methods and data analysis in bioinformatics, and
- data analysis in commerce, industry, macroeconomics, and finance.

1.2 Principal Component Analysis

PCA is a projection-based dimensionality reduction method to reduce data dimensionality by linearly transforming the data to a new coordinate system such that the greatest variance by any projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on. The main advantage of using PCA in a study on the morphology of bones is that the patterns or the major variations within high dimensional 3D surface data can be captured and that the data can be compressed without losing patterns of the data. PCA is closely related to SSM. It decomposes the training set in Eigen space as well as restricts the shape variations of the model close to the known variations. Given a n × 3m data matrix x of n shapes consisting of m surface points (3m dimensional), the computation of PCA starts with a mean-centred matrix y obtained from the empirical mean x (x = P^ni=1xi/n) of the distribution being subtracted from the variables: y = x − x. Two approaches are mostly used to calculate the principal component.

By arranging the eigenvalues in descending order and their corresponding eigenvectors so that λi > λi+1 makes λj nearly zero at a certain point j. The first corresponding j eigenvectors form the principal axes describing the most important modes of variables of the data in eigen space. The proportion of the total variability f i explained by each principal mode is equal to its eigenvalue divided by the sum of all n eigenvalues: fi = λi/ P i λi. The eigenvectors φ provide the so-called modes of shape variations, and the eigenvector corresponding to the largest eigenvalue accounts for the largest variation. The magnitude of eigenvalue shrinkage is a function of the type of shrinkage, sample size, the number of variables in the correlation matrix, the ordinal root position, the population eigen structure, and the choice of PCA or principal factors analysis [BS84].

The n (when 3m > n) largest eigenvalues and their corresponding eigenvectors of the covariance matrix C3mx3m can be determined from the n eigenvalues and eigenvectors of Cnxn = ynx3myT nx3m. The accumulation of the first few eigenvalues has such a large proportion of the accumulation of all eigenvalues that the rest of the eigenvalues approximate a zero proportion. Let λi(i = 1,2, ..., n) be the eigenvalues of eigenvectors, the first j eigenvalues λj and corresponding eigenvectors φj of C3mx3m can be calculated as be the generally mean when they say factor analysis). The question of sample size in factor analysis and PCA has been studied for decades: some statisticians have argued that a large sample size N improves the precision of estimates from PCA [Cli70, Bag83]; others have shown that the ratio of samples to variables is more important in interpreting outcomes of PCA [GV88]. The commonly accepted minimal number of samples for the analysis should be larger than 100 or five times the number of variables being analysed. However, it is fair to say that there are no absolute rules for deciding sample size.

MacCallum et al suggested that the minimum sample size depended on the nature of the data, most notably its strength [MHW01]. Strong data are data in which variable communalities are consistently high and many variables are expected to have high load (or weight) on at least three or four components. Communality represents the proportion of the variance of a particular variable that is shared with other variables. The values are calculated as the sum of squared component weights for a given variable [Try98]. The meaning of load (or weight) of PCA is the same as for point distribution models, which will be explained in the following section.

The reliability of applying PCA to the shoulder bone shapes is not validated in this paper. Instead, all applications based on PCA results in the present work are tested using leave-one-out validations.

2. SOM Vs PCA EXPERIMENTS AND DISCUSSION

The Self-Organizing Map (SOM) algorithm has attracted a great deal of interest among researchers and practitioners in a wide variety of fields. The SOM has been analyzed extensively, a number of variants have been developed and, perhaps most notably, it has been applied extensively within fields ranging from engineering sciences to medicine, biology, and economics. We have collected a comprehensive list of 4465 scientific papers that use the algorithms, which have benefited from them, or contain analyses of them. The list is intended to serve as a source for literature surveys. A set of 13 topical categories was selected by combining classes used by the INSPEC (tm) database. The number of articles in those categories was then plotted as a function of the year of publication. The plots only contain the articles available in the INSPEC collection

.

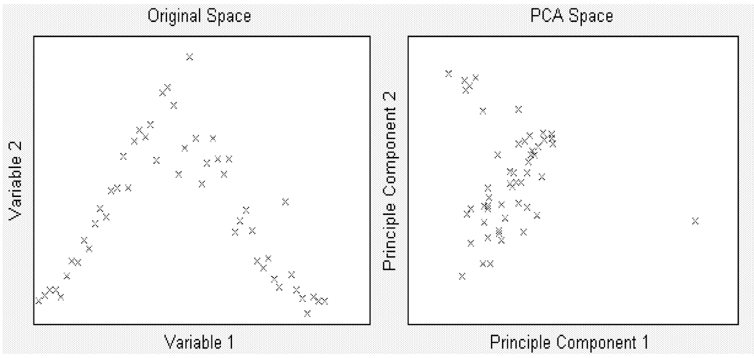


Figure 1. PCA components analysis

The results shown in Figure 1 reveal, for instance, that the transformed data is the linear combinations of the original data. In the new coordinate system, the number of coordinates is generally equal to the original coordinate system and all coordinates are orthogonal to each other. The new coordinates are ordered so that the first coordinate accounts for the most variations in the original data; the second coordinate explains the maximum variances for the residual data; the third coordinate explains the most variation for the next residual data and so on. Hence, the first coordinate is considered to be the most important coordinate, known as the first principal component (PC); the second coordinate is called the second PC, and so on. The SOM references were organized onto a document map to study the relationships between the topic categories, and to provide an interface for browsing and searching the collection. A WEBSOM [2] was computed using the titles of the documents. For some documents also an abstract was available and it was used in the computation.

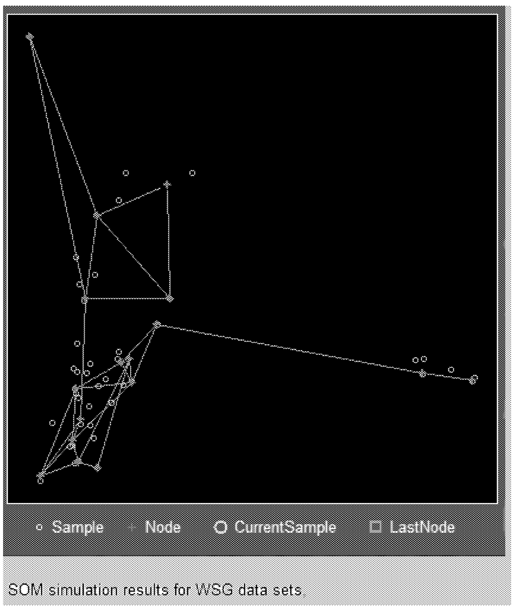


Figure 2. SOM components analysis

One of the drawbacks of the SOM is that the user must select the map size in advance. This may lead to many experiments with different sized maps, trying to obtain the optimal result. This is very time consuming. A problem with large SOMs Figure 2 is that when the map size is increased, the time it takes to do any operations on the map increases linearly. Training and using these large maps may again be quite slow. Several classical approaches have been suggested to circumvent these problems. One approach is to start with a small SOM and add nodes to it during training. Blackmore and Miikkulainen [3] propose *incremental grid growing*. In their system a rectangular grid is used. Neighbours that are close to each other in the data space are connected, whereas very distant ones are not. New map nodes are added at locations that poorly represent the data space. Fritzke's *Growing Cell Structure* [4] abandons the symmetric grid structure altogether. Instead his system has a group of hyper tetrahedrons. During training new hyper tetrahedrons are grown from the old ones. The system also removes harmful hyper tetrahedrons for greater flexibility. Related approaches have been presented by Martinez and Schulten [5] and Bruske and Sommer [6].

Growing the size of the grid dynamically is an intuitive and good principle. However it does not solve the other problem mentioned above. By the grid size becomes very large, finding the best matching unit takes longer and longer. Larger maps require some sort of additional data structures that make the search faster.

PCA as one of the most famous mathematical techniques widely used in data dimensionality reduction performs vector space transformation on the given dataset to rearrange the data into a new coordinate system (Dunteman, 1989). The transformed data is the linear combinations of the original data. In the new coordinate system, the number of coordinates is generally equal to the original coordinate system and all coordinates are orthogonal to each other. The new coordinates are ordered so that the first coordinate accounts for the most variations in the original data; the second coordinate explains the maximum variances for the residual data; the third coordinate explains the most variation for the next residual data and so on. Hence, the first coordinate is considered to be the most important coordinate, known as the first principal component (PC); the second coordinate is called the second PC, and so on. Generally speaking, a certain number of PCs, less than the number of coordinates in the original data, are enough to account for the most variance in the original data. This is to say that PCA can transform a high dimensional dataset to a lower dimensional space without losing significant amounts of information when compared with the original dataset (Dunteman, 1989).

In general, a PCA transformation can be performed using the following 4 steps (assuming X is an n-by-p matrix, where n is the number of variables and p is number of records in each variable):

Step 1: subtract the mean value

$$\mathbf{X' = X - X}$$

Where X is an n-by-1 vector containing the mean of each variable in dataset X, and X' is the new dataset.

Step 2: calculate the covariance matrix

$$\mathbf{A = cov(X') = (X'X'T)/(p - 1)} \qquad \mathbf{(3.12)}$$

Where S represents the covariance matrix, X'T is the transpose of matrix X' and p is the size of each variable.

Step 3: calculate the eigenvectors and the eigenvalues of the covariance matrix

$$\mathbf{Ax = Alx}$$
$$\mathbf{(A - Al)x = 0}$$
$$\mathbf{det(A - Al) = 0}$$

Where A is the square covariance matrix derive in step 2,I is the identity matrix, A is a vector of eigenvalues and x is the matrix of eigenvectors.

Step 4: reorder the eigenvectors based on their associated eigenvalues, from the highest to the lowest, which represent the explained variance of the eigenvectors

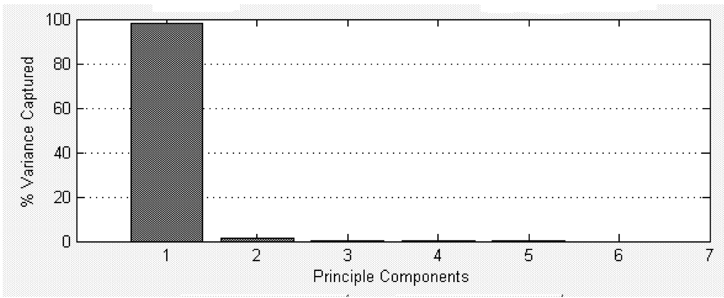


Figure 3. PCA components feature analysis for Temp, CO2, Light and week

3. RESULTS

The classification results, averaged over the 10-fold cross-validation experiments, can be seen in Figure 3. The figure shows the classification results averaged over the five features. The overall results of the PCA are noticeably better than the SOM percentages. This kind of good performance on the hard portions of the data space is very desirable. Figure 4 gives a better overall view of the classification of four variables

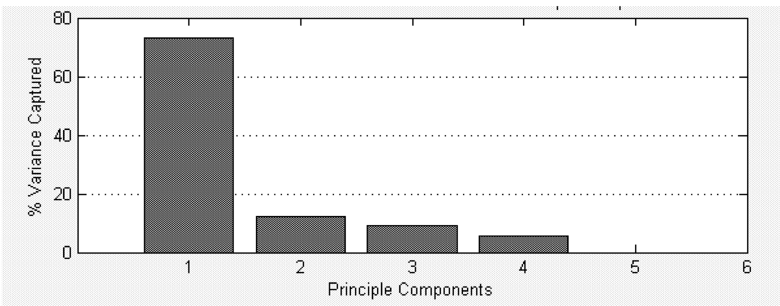


Figure 4. PCA components feature analysis for Temp, CO2, Light and week, and yield variables

4. CONCLUSION

Self Organizing Maps (SOMs) are used to organize unsupervised data. While the Kohonen SOM has proven to be very popular and quite effective, there is no theoretical or empirical research to suggest that it is the optimal strategy for SOMs. So, we tried to evolve the rules of the SOM. We represented the rules of a SOM for self-organization on unsupervised data. Our data was arranged into definite clusters. While the samples that make up these clusters were randomly generated, the centres of the clusters were fixed. We changed the network for minimizing the distance of the vectors in the map to actual cluster centres. We used two different sets of cluster centres. We found that even with random starting vectors and random sample points (from fixed cluster centres), SOMs achieved very good reduction in distance to cluster centres when the cluster centres were used one set at a time. By this measure, the classification accuracy of the SOMs was perfect when possible and near optimal otherwise. However, when both of the sets were used to determine fitness, we find that the distance reduction and classification accuracy suffered.

In comparison to the PCA, we found that SOMs consistently performed better in all the experiments much better than PCA. While this shows great promise, the results cannot be generalized over practical applications of unsupervised learning without further experimentation.

5. REFERENCES

[1] Teuvo Kohonen, Self-Organizing Maps, Springer Series in Information Sciences, Vol. 30, Springer, Berlin, Heidelberg, New York, 1995, 1997, 2001, 3rd edition.

[1] J. M. Baldwin. A new factor in evolution. The American Naturalist, 30(354):441–451, June 1896.

[2] R. Cattral, F. Oppacher, and D. Deugo. Principles of Data Mining and Knowledge Discovery, chapter Using Genetic Algorithms to Evolve a Rule Hierarchy, pages 289–294. Springer, Berlin / Heidelberg, 1999.

[3] Samuel Kaski, Jari Kangas, and Teuvo Kohonen. Bibliography of self-organizing map (SOM) papers: 1981–1997. Neural Computing Surveys, 1(3&4):1–176, 1998. Available in electronic form at http://www.icsi.berkeley.edu/'jagota/NCS/: Vol 1, pp. 102–350.

[4] Teuvo Kohonen, Samuel Kaski, Krista Lagus, Jarkko Salojärvi, Jukka Honkela, Vesa Paatero, and Antti Saarela. Self organization of a massive document collection. IEEE Transactions on Neural Networks, 11:574–585, 20

[5] Blackmore J., Miikkulainen R. (1993) Incremental Grid Growing: Encoding High-Dimensional Structure into a Two-Dimensional Feature Map, Proc. ICNN'93, Int. Conf. on Neural Networks, Vol. I, 450-455, IEEE Service Center.

[6] Deboeck, G.(1999) Investment maps for emerging markets. Neuro-fuzzy techniques for intelligent information systems (N.Kasabov and R.Kozma Eds.). Physica Verlag, 373-395.

[7] Deng, D., Kasabov, N. (1999) Evolving Self-organizing Map and its Application in Generating A World Macroeconomic Map. Proceedings of the ICONIP/ANZIIS/ANNES'99 International Workshop, 7-12, University of Otago, Dunedin.

[8] European Central Bank, Montly Bulletin (1999).

[9] Fritzke, B. (1994) Growing cell structures - a self-organizing network for unsupervised and supervised learning. Neural Networks, 7, 1441-60.

[10] Heskes, T.M., Kappen B. (1991) Learning processes in neural networks. Physical Review A, 44, 2718-26.

[11] Kasabov, N. (1998) The ECOS framework and the ECO learning method for evolving connectionist systems. Jour. of Advanced Computational Intelligence, 2, 1-8.

[12] Kasabov, N. (1998) Evolving fuzzy neural networks - algorithms, applications and biological motivation. in: Yamakawa and Matsumoto (Eds.) Methodologies for the Conception, Design, and Application of Soft Computing, World Scientific, 271-274.

[13] Kasabov, N., Erzegoveri, L. et. al. (2000) Hybrid intelligent decision support systems and applications for risk analysis and prediction of evolving economic clusters in Europe. To appear in N.Kasabov (ed), Future Directions for Intelligent Systems and Information Sciences, Physica Verlag (Springer Verlag).

[14] Kaski, S. (1997) Data exploration using self-organizing maps. Acta Polytechnica Scandinavica, Mathematics, Computing and Management in Engineering Series No. 82.

[15] Kohonen T. (1982). Self-organizing formation of topologically correct feature maps, Biological Cybernetics, v. 43, 59-69.

[16] Kohonen, T. (1997) Self-Organizing Maps, second edition. Springer.

[17] Mao J., Jain, A.K. (1995) Artificial neural networks for feature extraction and multivariate data projection. IEEE Transaction on Neural Networks, 6, 296-317.

[18] Sammon, Jr., J. (1969) A nonlinear mapping for data structure analysis. IEEE Transaction on Computers, 18, 401-09.

[19] Sarle, W. (1999) Newsgroup FAQ of comp.ai.neural-nets, part 2. URL ftp://ftp.sas.com/pub/neural/ FAQ.html.

[20] Serrano-Cinca, C. (1996) Self organizing neural networks for financial diagnosis. Decision Support Systems, 17, 227-38

SMART SEARCH: AN INTELLIGENT SEARCH SYSTEM FOR THE OMAN NATIONAL TRANSPORT COMPANY USING WEB ONTOLOGY LANGUAGE

Sherimon P.C., Arab Open University, Muscat

Raja Waseem Anwar, Arab Open University, Muscat

Youssef Saad Takroni, Arab Open University, Muscat

Reshmy Krishnan, Muscat College, Muscat

Vinu P.V., HCT, Muscat

KEYWORDS
Ontology, OWL, Protégé, ONTC, Semantic Web, Knowledge Base, SQWRL.

ABSTRACT

This paper presents an ontology development for a knowledge base based intelligent searching system for the Oman National Transport Company (ONTC). The Company operates intercity and long distance bus services. A traditional database information system is implemented currently in the ONTC for users to retrieve necessary information such as schedule, routes, ticket charges etc. Here the information retrieval is in syntactic nature. It means that data is retrieved according to the values passed to the database. The search is not done in an intelligent way. Semantic Web concepts enable information systems to perform search in an intelligent way. It empowers machines to understand and response the human queries based on their meaning and making searching in an effective way. Ontology is the back bone of semantic web. It is an explicit specification of a conceptualization. Without ontology there cannot be vocabulary for representing knowledge. In this paper we propose an intelligent searching system using the concept of Semantic Web for the ONTC which will enable users to perform an efficient search and can create more accurate results. We use Protégé, the ontology development editor for developing and inferring the ONTC ontology and SQWRL (Semantic Query-Enhanced Web Rule Language) for querying the ontology.

INTRODUCTION

The Oman National Transport Company is the largest public bus and coach operators in the Sultanate of Oman. The company operates to different destinations from its headquarters. A person who wishes to travel in the ONTC can book his/her credentials in advance. A new form of web enabled system will empower computers to function in an intelligent way. With the development of information technology and wide use of the Internet in the recent years, the existing knowledge system available is not sufficient to satisfy the great deal of knowledge in the Web. So, there is need for the scholar to currently resolve the problem to organize knowledge validly, and to make it apply efficiently to the system. Ontology is a discipline that is part of the knowledge representation field and could resolve the bottleneck problem within the knowledge to obtain (Berners-Lee, 2001). Ontology describes the system at the level of semantic knowledge. The increasing importance of semantic web is based on the ontology technology. Now ontology has been widely used in any domains like medicine, military, cooking, e-learning, jobs, enterprise, and transportation and so on.

THE ONTC ONTOLOGY DOMAIN

Knowledge representation is the symbolization or formalization of the knowledge of the world (Alexander, 2002). Various Knowledge representation methods are predicate logics, production rules, framework, and ontology and so on. In this paper, we use the ontology and the rules to represent the knowledge of ONTC (Thomas R. Gruber, 1995). Ontology is an engineering artifact that is constituted by a specific vocabulary used to describe a certain reality (domain), plus a set of explicit assumptions regarding the intended meaning of the vocabulary (Neches et.el, 1991). Insufficiency in expressivity and reasoning features are the problems affect the perfect ness of ontology. Rules can lay the foundation for expression and reasoning capabilities. Adding rules with the ontology will solve the expressivity and reasoning problems.

The Role of Ontology

Ontology is defined as: "An explicit specification of a conceptualization." (Maria 2010). It usually contains modeling primitives such as concepts, relations between concepts, and axioms. The basic element in the design of ontology is a 'category', meaning its domain. Each category should have 'individual' and 'instance'. 'Inheritance' and 'implement' is their basic relation. 'Slot' could be used in order to depict them. At the same time, slot is divided into attribute and relationship and limited with facet. Ontology can be used in an integration task to describe the semantics of the information sources and to make the contents explicit (S.Chandrasekharan et.el.,1999). Ontology has shown to be the right answer to the structuring and modeling problems arising in Knowledge Management (Gruber, 1993). They provide a formal conceptualization of a particular domain that can be shared by a group of people (in and between organizations).

Role of Ontology in the Transportation Domain

Retrieval of data and information is the vital part of any system. An intelligent approach could improve the efficiency and accuracy of data retrieval (Ohno, 1998). Semantic Web concepts enable information systems to perform intelligent search. It empowers machines to understand and response the human queries based on their meaning and it makes searching more effective. Ontology is the backbone of semantic web (Brachman, R.J.et.el.,1991).

Using ontology as a knowledge base improves the efficiency of data retrieval. Customers can get data in meaning full way. (An intelligent approach) .Semantic web approach is vital in any kind of transport system because an intelligent approach will improve the query result.

Components of Ontology

Ontology consists of four main components to represent a domain (Berners-Lee et.al 2001). They are: classes (concepts of the domain), their attributes (slots), restrictions on slots (facets), relations (different types of binary associations between concepts or data values), and instances (real world individuals). Formally, ontology often boils down to an object model represented by a set of concepts or classes C, which are taxonomically related by the transitive IS-A relation $H \in C \times C$ and non-taxonomically related by named object relations $R^* \in C \times C \times String$. Axioms denote a statement that is always true (Hyvonen, 2004).

Methodology to Create an Ontology

Determine the domain and scope of the ontology. We suggest starting the development of ontology by defining its domain and scope (R.Subhashini and Dr. J. Akilandeswari,2001). That is, answer several basic questions such as what is the domain that the ontology will cover, for what we are going to use the ontology, for what types of questions the information in the ontology should provide answers, who will use and maintain the ontology, what is the level of formality of the intended ontology? etc.

Enumerate important terms in the ontology. Developers should write down a list of all terms in the domain, either to make statements about or to explain to a user (Natalya F. Noy and Deborah L. McGuinness, 2001)
.

Define the classes and the class hierarchy. Classes have to be defined with appropriate names. The hierarchies have to be defined in three different ways, viz. top-down, bottom-up, and hybrid. A top-down approach starts with the most general classes in the domain and subsequent specialization of the classes. A bottom-up approach starts with the most specific classes, the leaves of the hierarchy, with subsequent grouping of these classes into more general classes. A hybrid approach is a combination of the top-down and bottom-up approaches.

Define the properties of classes. Properties describe the internal structure of classes. The object property relates one class to another class (Das, 2002). The data type property assigns the value type to the class. The annotation property describes the generic comments of the class.

Define the facets of the properties. Properties can have different facets describing the value type, allowed values, the number of the values (cardinality), and other features of the values, the properties can have.

Create instances. At last create the individuals or instances of the classes.

Ontology Tools

There are several editors available to create ontology. Ontology editors are tools that enable the users for inspecting, browsing, codifying, and modifying ontology and support in this way the ontology development and maintenance task. Prot´eg´e, OntoEdi, OilE, WebOD, and Ontolingua are some examples of ontology editors.

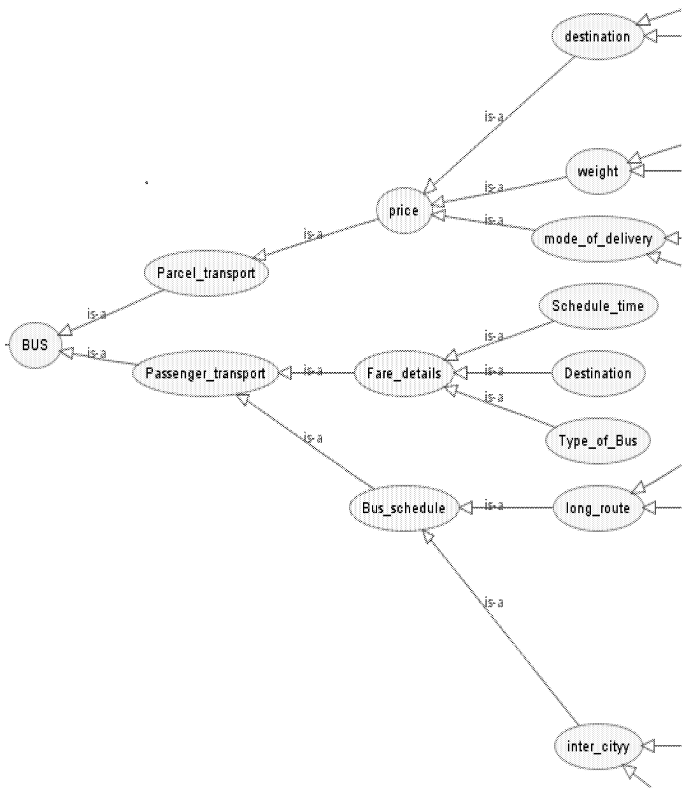
Protégé
Protégé is a free, open-source platform that provides a growing user community with a suite of tools to construct domain models and knowledge-based applications with ontology. Further, Protégé can be extended by way of a plug-in Architecture and a Java based Application Programming Interface for building knowledge-based tools and applications.

SWRL
Semantic Web Rule Language, an acronym for SWRL, is a rule language. The SWRLTab is a development environment for working with SWRL rules in Protege-OWL. It supports the editing and execution of SWRL rules. It provides a set of libraries that can be used in rules, including libraries to interoperate with XML documents, and spreadsheets, and libraries with mathematical, string, RDFS, and temporal operators. The SWRLTab has several software components, like, SWRL Editor which supports editing and saving of SWRL rules in OWL ontology, SWRL Built-in Libraries which includes the core SWRL built-ins defined in the SWRL Submission and built-ins for querying OWL ontology.

SQWRL
Semantic Query enhanced Web Rule Language is an extension of SWRL rule language. It takes rules' antecedent as a pattern specification for a query and takes rules' consequent as a retrieval specification. Any valid SWRL antecedent is a valid SQWRL pattern specification. That means, SQWRL places no restriction on the left side of a query. It uses SWRL's built-in libraries as an extension point. SWRL editor can be used to generate and edit the queries. To execute the query SQTab has to be added, by adding the jar file jess 7.1p2 in the protégé plug-in directory. The core operator is aqwrl:select. The select operator takes one or more arguments, which are variables in the pattern specification of the query, and builds a table using the arguments as the columns of the table.

STRUCTURE OF THE ONTC ONTOLOGY

Building blocks of Ontology can be taken as classes (Fr. Rubin, 2011). They describe the concepts of the domain. The concept of this domain is the types of the buses such as parcel coaches and passenger buses. The price of the parcel coaches will be calculated depending upon the destination, mode of delivery and weight of the parcel. Price will vary depending on the destination. If it is within the city, the price will be less and outside the city, the price will be more. If the customer wants an emergency mode of sending the parcel, more charge will be taken than the normal mode. Price varies according to the weight of the parcel. If weight is more, price also will be more. All these specifications can be mentioned using the data property of the classes (Junli Wang, 2005). The system can retrieve the appropriate result according to the query input. For example for an input of destination Dubai, the system can retrieve the best fare.



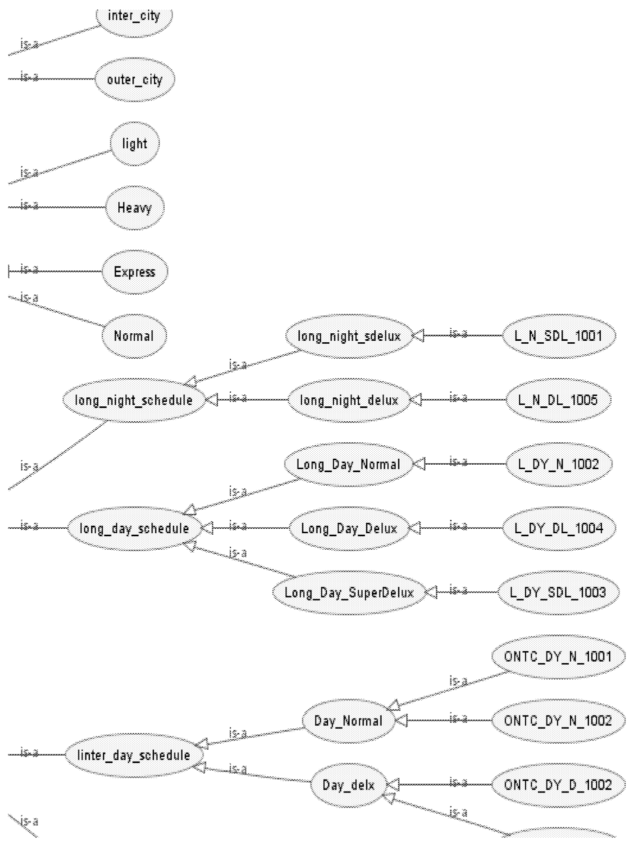


Figure 1. Class – Subclass "is-a" mapping in the ONTC Ontology

Ontology Structure in Protégé

In the ONTC ontology, BUS is a subclass of owl:Thing. Parcel_Transport and Passenger_Transport are subclasses of BUS.

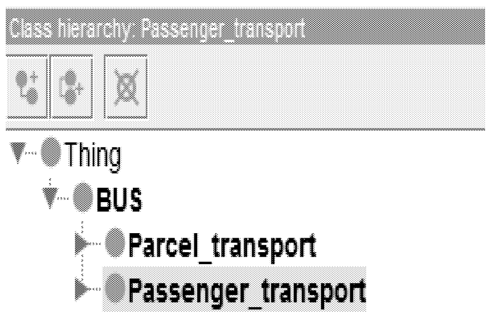


Figure 2. Class Hierarchy - Top Level

In the case of Passenger bus, there are two type of information are available. Bus_schedule and Fare_details of the Bus travel.

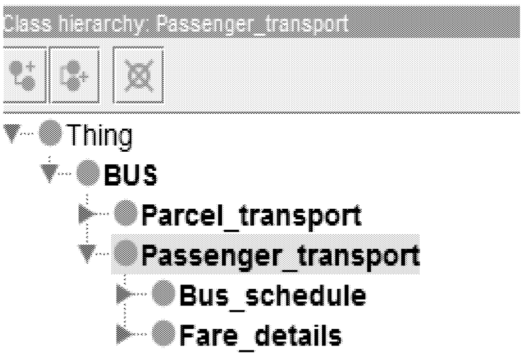


Figure 3. Class Hierarchy – Passenger_transport sub classes

In the case of Parcel_transport, price is the subclass. Destination, mode_of_delivery and weight are the subclasses of price class.

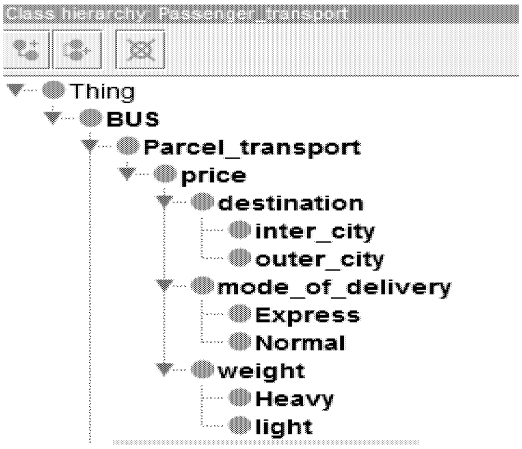


Figure 4. Class Hierarchy – Parcel_transport sub classes

inter_ city and long_ route are the subclasses of Bus_schedule. In each route buses are being scheduled in day and night to provide maximum facility to the public. During the day time and night time, ticket rates are different. Generally three types of buses are allotted, Super Delux, Deluxe, and Ordinary buses. Depending on the distance and timing, only certain types of the buses are allotted each time. For e.g. during the day time, for the long distance, all three types of buses are available. Passengers can utilize the service of three types of buses according to their convenience. But during night and long distance travel, only delux and super delux buses are chartered since ordinary buses are not advisable due to security reasons.

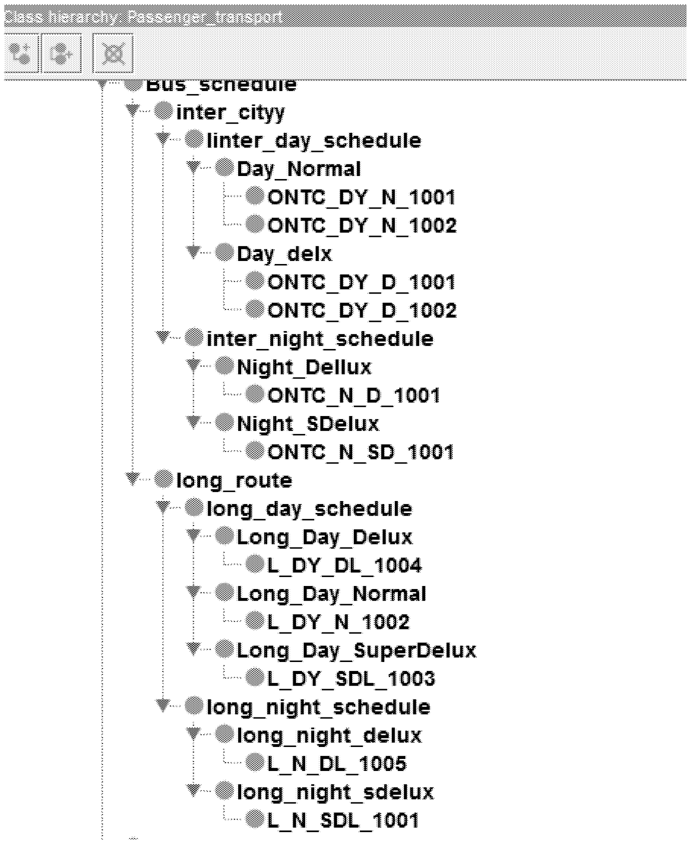


Figure 4. Class Hierarchy - Bottom Level Sub Classes

Similarly for the inter_city service, during the day time only ordinary and deluxe buses are permitted since the distance is less. During the night time only delux and super delux are available because of the shortage of the drivers.

For each service the bus numbers are also given so that passengers can get the result if the appropriate query is given. The properties can have different facets describing the value type, allowed values and the number of values that property can take. The datatype properties of the ONTC ontology are hasdestination, hasarrivaltime, hasday, hasdeparturetime, hasdestination station, hasdistance, hasrouteno, hassourcestation, hasBusno. The Object property has domain and range. The classes to which property is described is called domain of the property. The allowed classes or type instances for the property are often called the range of the property. The Object Property of the ONTC ontology are hasFirstclassFare, hasBusinessfair, haseconomyfare, has Expressfare, hasOrdinaryfare, hassuperdeluxfare, hasdeluxfare, hasheavyparcelfare, haslight parcelfare. The last step in the development of the ontology is creating individuals or instances of classes in the hierarchy. Defining an individual of a class requires is 1) choosing the class 2) creating an individual of that class and 3)filling the property values. The typical individuals are ONTC_DY_N_1001, ONTC_DY_N_1002, INTC_DY_D_1001, ONTC_DY_D_1002, ONTC_N_D_1001, ONTC_N_SD_1001, L_DY_DL_1004,L_DY_N_1002, L_DY_SDL_1003, L_N_DL_1005, and L_N_SDL_1001.

The ONTC Ontology Based Reasoning

In our work, Fact++ reasoner[ ] is used to check the consistency of the ontology, classifying the classes and performing the individual detection (Chimaera, 2000). Checking the consistency is used to identify the semantic contradiction which makes ambiguity in the description of the domain. Classifying the classes is used to test the subsumption relationship between classes and classes are classified into different level in the class hierarchy. Reclassification is necessary if there is any inconsistency.

Rule Based Reasoning

In this work, SWRL is used for creating rules and SQWRL is used for supporting OWL queries. For example, To display the fare of the given bus route and destination and source station, we can write the query like this :
hasDestination(?x,"SALALAH")^hasRouteno(?x,"L_N_DL_1005") ^hasSourcestation (?y,"MUSCAT")→ sqwrl(?a) ^sqrl:column names("delux fare")

CONCLUSION & FUTURE WORK

The owl based ontology for the ONTC was constructed with Protégé 4.1.beta after following the listed steps for the ontology development process. We have used SWRL for inferring the constructed Ontology. Consistency checking of ontology and classification of classes has been done with the support of FACT++ reasoner. Ontology design is a creative process and two ontologies designed by different people would never be the same. We hope that this ontology will help the public very much in providing sufficient information. In future we will enhance the ontology with the local names of the places with natural language support and automatic ontology development by the system itself. Also we will integrate this ontology with the ontology for airlines in Sultanate of Oman.

REFERENCES

Alexander Maedche, Ontology Learning For The Semantic Web, Kluwer Academic Publishers, USA, 2002.

Berners-Lee, T., Hendler, J., Lassila, O, "The Semantic Web", Scientific American, May and Grid, 2005. SKG '05. First International Conference on 2001.

Brachman, R.J.et.el., 1991 Living with CLASSIC: When and how to use KL-ONE-like language. *Principles of Semantic Networks*. J. F. Sowa, editor, Morgan Kaufmann: 401-456.

Chimaera (2000). Chimaera Ontology Environment. www.ksl.stanford.edu/software/chimaera.

Das, A., Wu, W., McGuinness, D, "Industrial Strength Ontology Management. The Emerging Semantic Web", IOS Press, 2002.

Fr. Rubin Thottupuram and Sumathy Easwaran, Farm-Agro Ontology formation: A black pepper model, International Journal of Research and Reviews in Software Engineering, Vol. 1, No. 1, March 2011

Gruber, T.R. (1993). A Translation Approach to Portable Ontology Specification. *Knowledge Acquisition* 5: 199-220.

Hyvonen, E., Saarela, S., Viljanen, K, "Application of Ontology Techniques To View-Based Semantic Search And Browsing, The Semantic Web: Research and Applications",Proceedings of the First European Semantic Web Symposium(ESWS), 2004.

Junli Wang; Zhijun Ding; Changjun Jiang; Dept. of Comput. Sci. & Eng., Tongji Univ., Shanghai **"** An Ontology –based Public Trnasport Query sytem**"** Semantics, Knowledge 2005,page 62,ISBN 0-7695-2534-2**.**

Maria Keet **,** "Introduction to Ontology Engineering with emphasis on Semantic Web Technologies" KRDB Research Centre, Free University of Bozen-Bolzano, Italy.as part of Masters Ontology Winter School 2010.

Natalya F. Noy and Deborah L. McGuinness. ``Ontology Development 101: A Guide to Creating Your First Ontology''. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880, March 2001.

Neches, R., Fikes, R.E., Finin, T., Gruber, T.R., Swartout,W.R, "Enabling Technology For Knowledge Sharing", AI Magazine 12(3), 1991, pp. 36-56.

Ohno Machado, L., Gennari, J.H., Murphy, S.N., Jain, N.L., et al., "The Guideline Interchange Format", J. American Medical Informatics Association (5), 1998, pp. 357-372.

R.Subhashini,(2001) Dr. J. Akilandeswari(2001). A Survey on ontology construction methodologies., International Journal of Enterprise Computing and Business systems(online),Vol1.Issue1.January 2001.

S.Chandrasekharan et.el.,1999,What are Ontology and Why do we need them ? IEEE Intelligent Systems Volume 14 Issue 1, January 1999.

Thomas R. Gruber, (1995). Towards principles for the design of ontology used for knowledge sharing. International Journal of Human-Computer Studies vol.43: 907-928.

AUTHOR BIOGRAPHIES

Sherimon P.C. is a faculty of Computer Studies at Arab Open University (AOU), Muscat, Sultanate of Oman. He holds a Master's Degree in Computer Applications from Madras University, India. His research interests lies in the area of semantic web and web programming. He has presented research papers in several National and International Conferences. He has publications in Intrenational journals.

Raja Waseem Anwar is a faculty of Computer Studies at Arab Open University (AOU), Muscat, Sultanate of Oman. He holds the Master's Degree in Computer Science from

Karachi University. His research interests lies in the area of Information Security and Computer Networks. He has presented research papers in several National and International Conferences. He has publications in Intrenational journals.

Dr. Youssef Saad Takroni Dr. Reshmy Krishnan is an Assistant Professor at Arab Open University,Muscat, Sultanate of Oman. He holds a PhD degree in computer science from CUNY -USA. His research interests lies in the area of AI and Neural Network. He has presented several research papers in International Conferences. He has publications in Intrenational journals.

Dr. Reshmy Krishnan is an Assistant Professor at Muscat College, Muscat, Sultanate of Oman. She holds a PhD degree in computer science and engineering from Sathyabhama University and masters degree in computer science and engineering from Madras University India. She is a reviewer in many journals and chaired conference sessions. Her research interests lies in the area of Ontology and Semantic web. She has presented several research papers in International Conferences. She has publications in Intrenational journals.

Vinu P.V. is a lecturer in Information Technology at Higher College of Technology (HCT), Muscat, Sultanate of Oman. She holds a Master's Degree in Computer Science from Mahatma Gandhi University, India. Her research interests lies in the area of pervasive computing, web technologies and programming. She has presented research papers in International Conferences. She has publications in Intrenational journals.

A PLAUSIBLE ANCESTOR GENERATOR

Prof Ken Newman PhD
Center for Research and Training
Higher Colleges of Technology
Abu Dhabi
United Arab Emirates
E-mail: Ken.Newman@hct.ac.ae

KEYWORDS
Geneology, Historical Geographic Information Systems, Semantic Web, Screen-scraping, Wikibots

ABSTRACT

This paper demonstrates methods for creating a server application with the task of generating a plausible family tree based on the user. The application does not try to be factual, merely plausible. In creating this app a number of server based techniques are employed to solve the various problems. 1 - Heuristics were defined to simulate plausible birth and death dates, and also the parent's age at the birth of the each generation. 2 - Screen scraping techniques were employed to harvest names from name databases which did not offer an api. 3 – Google places api was used to select a plausible birthplace for each ancestors. 4 – Wikibots were used to retrieve a plausible occupation plausible for each ancestor. The purpose of this study is to investigate how a variety of servers-based methods can be employed within a single application to provide a complex result which appears intelligent, has a certain level of semantic credibility, is fun, engaging for the user and appears entirely plausible.

INTRODUCTION

An ancestor generator may appear a surprising choice for an investigation of the semantic web a term first coined a decade ago (Berners-Lee et al, 2001). The authors point towards a web capable of a level of semantic intelligence – interconnection of meanings and relationships. Consider then, that the output of the ancestor generator is a set of ancestor objects with a similar data structure (dates, names, gender, locations, occupations, socio-political affiliations). This type of information is just the sort which can be simulated with heuristics and data collection methods. This investigation not only demonstrates emerging techniques for data collection from web sources and I, hope, contributes in some small way to the on-going question of semantic intelligence on the Web. But it is also quite satisfying for a user to see the output of their ancestory request.

HOW MANY ANCESTORS

From informal hand counts in auditoriums (not exactly scientific method but it has confirmed my initial expectations) … most people know the name and birthplace of both their parents, and can work out their birth year given enough time. Less than half of the people can do the same for all four of their grandparents , and very few can supply so much as a name of even one of their 8 great-grandparents.

As the number of ancestors for a single individual double every generation the numbers very quickly become very large. Each previous generation x will give the modern desendant 2x ancestors. If we make some broad assumptions, such as making an average age of a parent at the birth of any child is say 25 years old that gives us a not unreasonable average of 4 generations every century. In this scenario sometime in the 1300s a modern descendant had around 280 million ancestors, which begs the question, how many people were actually on earth at that time? Expert opinion tends to agree that the world population in the 1300s was around 240 million (Kremer, 1993). Even if we assume the largest estimates to be more accurate, it can be safely assumed that by the mid 1300s everyone alive today had more ancestors alive at that time, than there were people on the planet at that time. What becomes obvious is that a single ancestor connects to a modern individual by possibly many separate ancesteral lines.

So if a modern tourist visits an 11th century cathedral and wonders if any of her ancestors worked on the building of this cathedral the numbers would suggest that it is almost impossible that not just one but many of the original workers are among her billions of ancestors at that time, and some of them will be her ancestor via thousands of different ancesteral lines. What we will never be able to map with clarity is the unbroken lineage to all of those billions of ancestoral lines in the 1100s and beyond (as the numbers keep doubling).

SInce the historical records do not exist for the vast majority of these lineages, then how plausible can a fictitious and randomly generated lineage be? Specifically the questions considered for this study were:

How can a plausible gender, birth and death dates, age of ancestor at the birth of their child be generated?

How can a plausible name for each ancestor appropriate to their birth region be generated?

How can a plausible place of birth for each ancestor, given that the user is trying to find an ancestor in a random country/region?

How can a plausible occupation for the ancestor given the period they lived and the culture of their place and time?

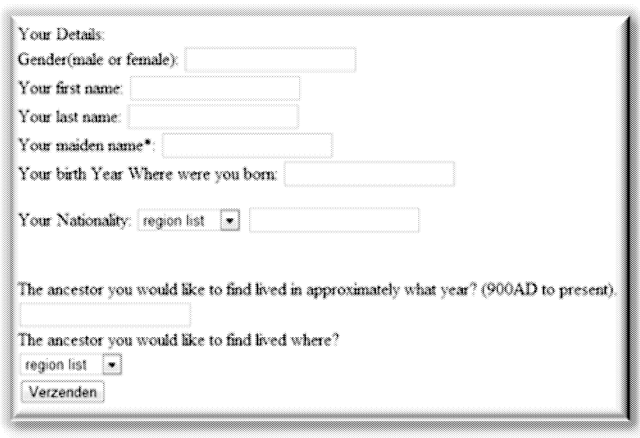
The methods used for these questions were for gender and dates - a set of heuristics; for names - screen scraping techniques; for the place of birth - the new Google Places api; and for the occupation - wikibots. The following sections describe these methods. The scope was limited to an ancestor no earlier than 900AD, and also to only European ancestors – since the name website used in this study contained only European names.

HEURISTICS FOR GENDER AND DATES

In answering the question "How can a plausible gender, birth and death dates, age of ancestor at the birth of their child be generated?" the following heuristics were employed to produce a set of plausible data for each ancestor.

1. An ancestor has a 50% chance of being male or female.
2. If an ancestor is male he has the same family name as his child.
3. If an ancestor is female she has the same married family name as her child, but a different maiden name.
4. The birth year of each ancestor is the child's birth-year minus the ancestors age at the child's birth (random in range 15-35).
5. The death year of each ancestor is the child's birth year plus the number of years they lived after that (random in range 0-50).
6. Before 1900 the chance that a woman died in childbirth rises to 1 in 5.

The interface asked the user to fill in basic information about themselves and the ancestor they were trying to find. A limit was set at 900AD, as place and individual naming prior to that becomes less plausible with the existing data.



Figures 1: The Ancestor Generator Interface

Using these simple rules it was possible to generate an ancesteral lineage which was entirely plausible. The server application was coded in php using an ancestor class, for which instances were kept in an array. Initially this first version used a simple list of English names -10 boys names, 10 girls names and 10 common surnames, as places holders for the investigation of the second question – How can a plausible name for each ancestor appropriate to their birth country be generated?

SCREEN-SCRAPING NAMES

'Behind the name' is probably the most comprehensive online site for firstnames (male and female) and last names. There are currently around 17,000 first names and 5000 family names. It also has the a úsage'fetures which lists regions/countries where each name is used. Unfortunately the site does not offer an API to handle server requests – the only way to access the vast list of names is via their own web interface, hence – screen-scraping.
The term Screen-scraping has come to mean the process of collecting data from a web service which does not offer an api – i.e. the only access is a dynamic html output intended to be displayed in a web browser. By taking the html output and studying it's formatting it is often possible to write server functions to collect large amounts of the data of the original web service. The tools for the would-be screen-scraper are:

- a robust server-side http application which can emulate browser requests and handle returns, and
- a system for pattern matching html text with regular expressions.

These tools are available in the php architecture in various ways. The method used in this study was the set of cURL classes which offer server-side http emulation. In figure 2 the php code example is using the cURL class to make a http request to the "behind the name"web service, requesting the second of four pages of last names beginning with the letter "a".

```
$btn_url ='http://surnames.behindthename.com/
    php/browse.php?letter=a2';

$ch = curl_init();
curl_setopt($ch, CURLOPT_URL, $btn_url);
curl_setopt($ch, CURLOPT_HEADER, 1);
curl_setopt($ch, CURLOPT_FOLLOWLOCATION, true);
curl_setopt($ch, CURLOPT_RETURNTRANSFER, true);
curl_setopt($ch, CURLOPT_FILETIME, true);
$data = curl_exec($ch);
curl_close($ch);
```

Figures 2: Using cURL to request lastnames

The target web server does not distinguish between this request and a browser request and it faithfully returns the page as a html/text string which is placed in the $data variable for processing. The resulting raw html contains a lot of peripheral information but eventually it beomes a list of names with a common format as shown in figure 3.

Pattern matching can now be done using regular expressions to extract the name, the usage and any notes for each name. Regular expressions contain a set of symbols with meanings for finding and matching patterns in text.

```
<table><tbody><tr><td class="normalcell">
<b>Angelov</b>
</td></tr>
<tr><td class="minicell">
<b>Usage:</b> <font class="info">Bulgarian</font>
<br>
</td></tr>
<tr><td class="normalcell">
Means "son of <a href="http://www.behindthename.com/name/find/angel"
>Angel</a>".</td></tr>
</tbody></table>
```

Figures 3: A snippet of retrieved html text

The code snippet in figure 4 demonstrates regular expressions being used to clean up the html text and then assemble all the html text containing the name data into an array for further processing. This was possible since studying the retrieved html revealed that the only tables used on the page were those containing the name data.

```
//remove line returns and feeds
$data = preg_replace('/\r|\n/','',$data);
//find the names from the data
$re = '/(<table>)[\s\S]+?(<\/table>)/';
preg_match_all($re, $data, $matches);
```

Figures 4: Cleaning and Assembling html

This allows for further processing of the html text to extract the usable name data. The code snippet in figure 5 demonstrates how the <tr> tag and the word 'usage' were used to identify the name, usage and notes for each name. These three values are then are stored in $lineData, which in turn is added to the container array $nameData.

```
foreach ($matches[0] AS $type => $nameData) {
    //get the name
    $i = 0;
    $re2 = '/(<tr>).+?(<\/tr>)/';
    preg_match_all($re2, $nameData, $matches2);
    foreach($matches2[0] AS $type => $lineData){
        $linetype = array('name','usage','notes');
        if($i==0){
            //remove the <b> from the front of the name
            $name = preg_replace('/<.+?>/', '', $lineData);
        }elseif($i==1){
            //remove the <b> from the front of the name
            $usage = preg_replace('/<.+?>/', '', $lineData);
            $usage = preg_replace('/Usage:/', '', $usage);

        }elseif($i==2){
            $notes = preg_replace('/<.+?>/', '', $lineData);
        }
        $i++;
    }
```

This simple and elegant function scrapes the data from the html output of the web service and allows it to be placed in a useable form. In the case of the ancestor application random names (first and last names) can now be generated according to region and gender. This discussion raises the question of the ethics of screen-scraping, and this is a delightfully thorny question best left for another paper. The experiments done in this study have been done without the knowledge or consent of Behind the Name – which is the whole point of Screen-scraping – the target server does not know it is being scraped.

FINDING A PLAUSIBLE BIRTH LOCATION

The Google Places API, released in May 2011, provides an adequate, though problematic solution to the question "How can a plausible place of birth for each ancestor be generated, given that the user is trying to find an ancestor in a random European country/region?"

Starting with a region/country as defined by our user and within the constraints of the usage regions available in the names data (e.g. Spain France, England), a request (again using cURL) to the Google Maps API will return the region/country's bounding rectangle as a set of two geolocs (a geoloc contains a latitude and longitude value). This time of course there is no need for screen scraping as the Google Maps and Places API's contain a comprehensive set of interrogating methods. The output can be returned in xml or Json formats as the requesting application chooses.

The two geolocs returned to define a region/country's bounding rectangle are called 'north-east'and 'south-west'. A random location can then be generated from within the bounding rect and this random geoloc can be sent to Google places API with a request to return any 'political'entities within a radius of, for example, 5 kms. A political entity in Google places API includes "localities, villages, towns, districts, states, nations". Although the Google places API is the best available online service for solving the problem of a plausible birth location there are two problems that are to date unsolved.

- Some countries/regions are not well defined by a bounding rect.
- Google Places does not contain historic data.

The first problem has a number of different sub-problems. While countries like France and Spain, simply because of their rectangular geographic shape, return about 90% successful locations, England, Scotland and Russia do not. Scotland, whose territories include the Shetland Islands, has a bounding rect that consists of about 80% ocean. England has the problem that it's bounding rect completely includes Wales, which should be excluded, since it is treated as a separate region for name generation. Russia returns an error that it's north-east geoloc cannot be greater than it's south-west geo-loc – further investigation needed. Some of these problems could be solved if Google Maps offered a bounding polygon rather than a bounding rect.

Alternatively, the ancestor app could keep it's own set of region polygon's – further investigation is needed for this.

The second problem is even more interesting. If the ancestor app is looking for a birthplace for an ancestor in the 1500s in Spain, Google Places will return a town in Spain but it has no information on how long that town has been there, and in fact whether Spain even existed at that time. Systems that capture this type of information are generically known as Historic Geographic Information Systems (HGIS) and there have been a number of notable attempts to create a workable, universal standard HGIS. To date none of these have been developed to a level of becoming an accepted standard. Interestingly all of the attempts to develop HGISs pre-date both Google Maps and Google Places (Knowles & Hiller, 2008) suggesting this is a topic ripe for an innovative research and development project using Google Places as the underlying architecture.

Swiss company EuroAtlas has produced a series of maps showing the cities, towns and political and tribal regions of Europe every 100 years from 100AD to 2000AD. This is the sort of information that could be captured in an HGIS, and not just for Europe. A list of current HGIS activities and resources is listed on the website of the Historical Research Network's website (www.hgis.org.uk). Despite these limitations, the ancestor generator generates birth location for each ancestor that would probably satisfy a lay-person's knowledge of history.

WHAT DID EACH ANCESTOR DO

The question "How can a plausible occupation for the ancestor given the period they lived and the culture of their place and time?" is a very interesting one and is the topic of on-going investigation at the time of writing this paper. The focus for this problem has been to develop a cultural object model and to populate data tables of this model with data from Wikipaedia by the use of so called 'wikibots' to conduct page searches. Wikibots are an area of server innovation that is often seen as exposing systems to malicious intentions(Geiger & Ribes, 2010).

Initial experiments searching for synonyms of 'occupation' (jobs, trades, professions) combined for searches of dates on the page using regular expressions yielded only limited useful results. The current path of investigation is to create a model of culture, and again populating the cultural data objects via the use of Wikbots. Each culture contains key elements of Pots, Poetry, Philosophy and Public Institutions – my four P's of culture. (I note that others including Dean(2011) use a different set of 4 Ps of Marketing). In my framework Pots represents the technologies of a culture; poetry the various forms of creative expression, arts and literature; philosophy the world view, morality and religion; and public institutions the administration, legal system, land ownership, military, market etc.

Plausible occupations are assigned to cultures based on the 4 Ps framework, and cultures can be assigned regions and timelines of influence. This question has exposed another topic ripe for further investigation – modeling Cultures. A comprehensive Culture model could also be the key to ever more detailed information on each ancestor to be generated, such as what language/languages did they speak, what were their family, tribal, political, and racial affiliations. Perhaps even significant events in their life could be plausibly described – their upbringing, marriage circumstances, travel, and death.

SUMMARY

This investigation has attempted to use existing server based heuristics to create plausible human stories. Within the limits of the study it has been successful in demonstrating the usefulness of the methods available, but it has also exposed some interesting limitations, and these, as have been noted through the paper are topics that invite some exciting possibilities for future investigation. These include a workable Historic Geographic Information System built as an overlay on the Google Places API, and a 'model of culture' from which a plausible occupation might be generated.

REFERENCES

Berners-Lee, T. Hendler, J. Lassila O. 2001. "The Semantic Web." *Scientific American*, May 2001.

Knowles A, Hillier A eds. "Placing History: How Maps, Spatial Data, and GIS Are Changing Historical Scholarship", ESRI Press, New York, 2008 ISBN 978-1589480131

Kremer, M. 1993. "Population Growth and Technological Change: One Million B.C. to 1990", Quarterly Journal of Economics, Vol 108, Issue 3, Pp 681-716. doi: 10.2307/2118405

Dean, R. 2011, "Designing and Building Culture at a Startup", Presentation, published at, http://www.slideshare.net/rishi12/designing-and-building-culture-at-a-startup-using-the-four-ps-framework

Geiger R, Ribes D 2010 "The work of sustaining order in wikipedia: the banning of a vandal", Proceeding CSCW '10 Proceedings of the 2010 ACM conference on Computer supported cooperative work ACM New York, ISBN: 978-1-60558-795-0 doi:10.1145/1718918.1718941

AUTHOR BIOGRAPHY

KEN NEWMAN was born in Queensland, Australia and graduated in Applied Film & TV studies in 1982. Through a career in animation, computer games, and educational software development, he morphed from a lackluster animator to a mediocre programmer to a better than average project manager. He took his first academic post in 2002 at Griffith University in Australia where he completed his PhD. Since then he has held posts in the UK, Netherlands, and currently holds the Chair for Cultural and Creative Industries at HCT in the UAE.

ELECTRONICS SIMULATION

Fuzzy Modeling of Dynamic Electronic Devices

Basil Sh. Mahmood
Department of Computer Engineering & IT
University of Mosul
Iraq
Email: basil_mahmood@yahoo.com

Omar Abd Al Razzak Ibrahim
Department of Electrical Engineering
University of Mosul
Iraq
Email: omaralrazzak@yahoo.com

KEYWORDS
Fuzzy Systems, Takagi-Sugeno Model, Fuzzy Modeling, Dynamic Circuits, Sector Nonlinearity, Chaotic Systems, LorenzOscillator.

ABSTRACT

This paper presents a novel view to model electronic devices employing Takagi-Sugeno Fuzzy structure. A systematic method is proposed to generate the rules and select the antecedent and consequent membership functions directly from the mathematical expressions of the electronic circuits. A zero order MISO discrete Sugeno structure is used to construct models for RC and RLC circuit as examples of linear dynamic electronic circuits. In nonlinear dynamic systems, a Takagi-Sugeno fuzzy model for chaotic Lorenz oscillator has been driven using sector nonlinearity approach.Finally, fuzzy model representing chaotic behavior for RLD circuit is considered.

INTRODUCTION

Fuzzy logic was develop as a way to imitate the capability of the human mind operating with vague concepts and approximated reasoning. Therefore, the significant advantage of using fuzzy logic in practical applications is to describe a system in a manner that tolerates imprecision and uncertainty. Employ fuzzy logic, a real-world systems can be model in terms of linguistic variables, fuzzy sets, and membership functions. This adds the expert knowledge from human operators to the modeling process. “Intelligent” model resulting is an alternative representation of the original system, which usually is much simpler in structure than a model obtained using a conventional approach (Zhong et al. 2006).

Fuzzy logic provides a simple and uncomplicated way to decompose the task of modeling and control design into a group of local tasks. These local tasks tend to be easier to handle. In the end, fuzzy logic provides the mechanism that blends these local tasks together to deliver the overall model and control design. In general, there are two approaches for constructing fuzzy models: Identification (fuzzy modeling) using input-output data and derivation from given linear or nonlinear system equations.

The procedure of fuzzy modeling using input-output data is suitable for plants that are unable or too difficult to be represented by analytical and/or physical models. On the other hand, if nonlinear dynamic model for systems can be readily obtained, then the second approach, which derives a fuzzy model from given nonlinear dynamical models, is more appropriate (Kazuo and Wang 2001).

TAKAGI-SUGENO FUZZY MODEL

TS models have recently become a powerful practical engineering tool for modeling and control of complex systems (Leandro and Araujo 2007). A general TS model employs an affine model with a constant term in the consequent part for each rule. This is often referred as an *affine TS model*. When the consequent part for each rule is represented by a linear model (without a constant term), this is the special type of TS fuzzy model and referred to it as a TS model with *linear rule consequence*, or simply *a linear TS model*(Kazuo and Wang 2001). Another special type of TS fuzzy model, when the consequents of the rules do not depend on the input variables, referred to this type of TS fuzzy model as a *zero-order Takagi-Sugeno model*(Jacek 2009).

A typical fuzzy rule in a zero-order Sugeno fuzzy model has the form:

IF x is AANDy is BTHEN$z = k$

Where A and$B$ are fuzzy sets in the antecedent, while k is a crisply defined constant in the consequent.All output membership functions are singleton spikes, and the implication method is simply multiplication, and the aggregation operator just includes all of the singletons (R. Jang and N. Gulley 2001).
The discrete TS fuzzy structure for modeling dynamic systems use functions of input variables as the rule consequent as the following example:

IF $y(n)$ is M_1 AND $y(n$-1) is M_2 AND $y(n$-2) is M_3 AND $u(n)$ is M_4 AND $u(n$-1) is M_5 THEN $y(n+1)=F(y(n),y(n-1),(y(n-2),u(n),u(n-1))$

Where M_1, M_2, M_3, M_4, M_5, and M_6 are fuzzy sets, $y(n)$, $y(n-1)$, and $y(n-2)$ are the output of the system to be modeled at sampling time (n), $(n-1)$ and $(n-2)$, respectively. And, $u(n)$ and $u(n-1)$ are system input at time $(n)$and $(n-1)$, respectively; $y(n-1)$ is system output at the next sampling time, $(n+1)$, $F()$ is an arbitrary function (H. Ying. 2000).

PERFECT TRANSFORMATION OF NONLINEAR SYSTEMS USING THE SIMPLEST TAKAGI-SUGENO MODEL

Here, the approximation of multiple inputs/single output (MISO) system is considered. The main goal is to translate the transfer functions or the state equations of the system to fuzzy rule-based formula according to Takagi- Sugeno model.

linear membership functions are used for input variables because, they have a simple conceptual, a clear interpretation and play a crucial role in many applications in the fuzzy modeling and control. They are sufficient for modeling complex highly nonlinear static or dynamic, continuous or discrete-time systems.

Then for a MISO TS system, as figure (1) shows, with the inputs z_1, z_2, ..., z_n and the output S, every input $z_k \in [-\alpha_k, \beta_k]$ required that there is no interval degenerated to a single point, i.e. we assume $\alpha_k+\beta_k \neq 0$ for k =1, 2,...,n. For any z_k, there are two fuzzy sets with linear membership functions $N_k(z_k)$, and $P_k(z_k)$, where P_k is an algebraic complement of N_k as in figure (2).

$$N_k(z_k) = \frac{\beta_k - z_k}{\alpha_k + \beta_k}$$

$$P_k(z_k)=1 - N_k(z_k) \qquad k = 1, 2,...,n \qquad (1)$$

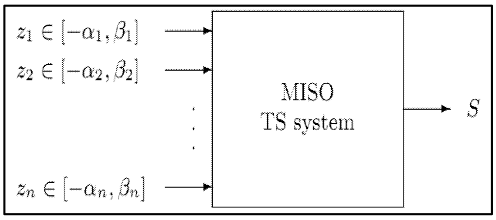


Figure 1: MISO TS fuzzy system

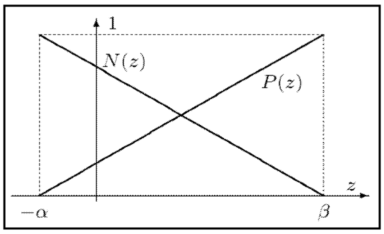


Figure 2: Linear membership functions of two fuzzy sets

This system is defined by 2^n rules in the form of implications:

$$\text{IF } p_{i_1,\ldots,i_n}, \text{ THEN } S = q_{i_1,\ldots,i_n} \qquad (2)$$

Where $(i_1, \ldots, i_n) \in \{0, 1\}^n$ and each antecedent $p_{(i_1,\ldots,i_n)}$ of an implication is the statement of the form "z_1 is A_{i_1} and ... and z_n is A_{i_n}". The A_{i_k} are as follows:

$$A_{i_k} = \begin{cases} N_k, \text{for } i_k = 0 \\ P_k, \text{for } i_k = 1 \end{cases} \quad , k = 1 \ldots n$$

If it is not stated differently, the consequents $q_{(i_1,\ldots,i_n)}$ of the rules in (2) assumed not depend on the input variables, i.e.a zero-order Takagi-Sugeno model will be considered. The rules (2) can be rewritten as follows:

$$\text{IF } P_v, \text{ THEN} S = q_v$$

Where $v \leftrightarrow (i_1, \ldots, i_n)$. For the inputs $z_1, \ldots, z_n$, the output is S and it is defined by the formula

$$S(z_1, \ldots, z_n) = \frac{\sum_{v=1}^{2^n} q_v h_v(z_1,\ldots,z_n)}{\sum_{v=1}^{2^n} h_v(z_1,\ldots,z_n)} \qquad (3)$$

Where

$$h_v(z_1, \ldots, z_n) = \mathsf{T}(A_{i_1}(z_1), \ldots, A_{i_n}(z_n))_v$$

The operator T denotes an algebraic t-norm $\mathsf{T}(x, y)= xy$, $A_{i_k}(z)$ are membership functions of the fuzzy sets, i.e. $A_{i_k} \in \{N_k, P_k\}$ for $i_k \in \{0,1\}$ and $k = 1, \ldots, n$. If the complementary property (1) is satisfied, then equation (3) reduces to(Jacek 2009):

$$S = \sum_{v=1}^{2^n} q_v h_v(z_1, \ldots, z_n)$$

FUZZY MODELING OF ELECTRONIC CIRCUIT

The design procedure described in this paper considers representing a given dynamic linear or nonlinear plant by Takagi-Sugeno fuzzy model.

Fuzzy Modeling of RC Circuit

The Discrete transfer function of RC circuit shown in figure (3) is as follows:

$$H(z) = \frac{1 - e^{-aT_s}}{z - e^{-aT_s}}$$

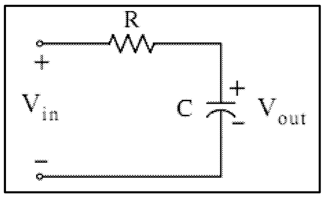


Figure 3: Simple RC circuit

Where a=1/RC, and T_s is the sample time. By assuming $\alpha_1 = e^{-aT_s}$, $\alpha_2 = (1 - e^{-aT_s})$, then the sequence representing this circuit is as follows:

$$y_k = \alpha_1 y_{k-1} + \alpha_2 x_{k-1} \qquad (4)$$

Assuming normalized bounded input/output, which means $y_{k-1} \in [0, 1]$, $x_{k-1} \in [0, 1]$, the membership functions then will be as shown in figure (4).

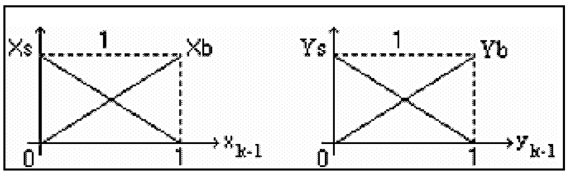


Figure 4: Antecedent MFs for RC circuit

Modeling this system explains in the following procedures:

Antecedent Membership Functions

The input universe of discourse (UD) is defined in the range [0, 1], and the antecedent MF sets according to figure (5) are as follows:

$$\begin{cases} \text{Xs represent } x_{k-1} - \text{small} \\ \text{Xb represent } x_{k-1} - \text{big} \\ \text{Ys represent } y_{k-1} - \text{small} \\ \text{Yb represent } y_{k-1} - \text{big} \end{cases}$$

Consequent Membership Functions

The following constants values will be set as crisp consequent MFs (singletons) $\{0, \alpha_1, \alpha_2, 1\}$, and hence there are four rules:

IF y_{k-1} is YsANDx_{k-1} is XsTHENy_k is 0
IF y_{k-1} is YsANDx_{k-1} is XbTHEN y_k is α_2
IF y_{k-1} is YbANDx_{k-1} is XsTHEN y_k isα_1
IF y_{k-1} is YbANDx_{k-1} is XbTHEN y_k is 1

Numerical Results

For R=C=1, and $T_s = 0.4$, equation (4) becomes:

$$y_k = 0.67\, y_{k-1} + 0.33\, x_{k-1} \qquad (5)$$

This linear system which is characterized by table (1) shown below, can be represented by a three-dimensional plane as shown in figure (5).

Table 1: characterized the system represented in equation(5).

y_{k-1}	x_{k-1}	y_k
0	0	0
0	1	α_2
1	0	α_1
1	1	1

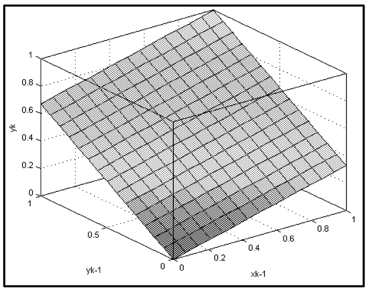


Figure 5: Fitted plans for equation (5)

Figure (6) shows the Simulink Implementation of the RC Circuit under consideration, and figure (7) represents the time responses of fuzzy model, as well as the real model of this circuit. It is obvious that both responses are identical.

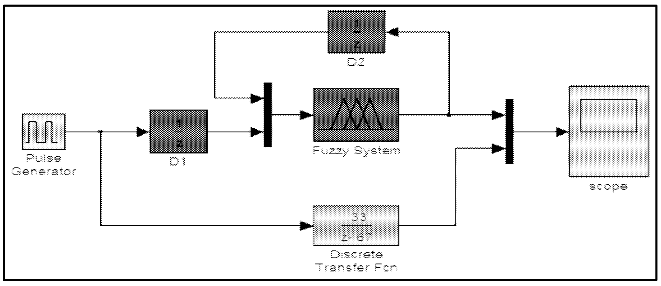


Figure 6: Simulink implementation of RC fuzzy circuit and real circuit

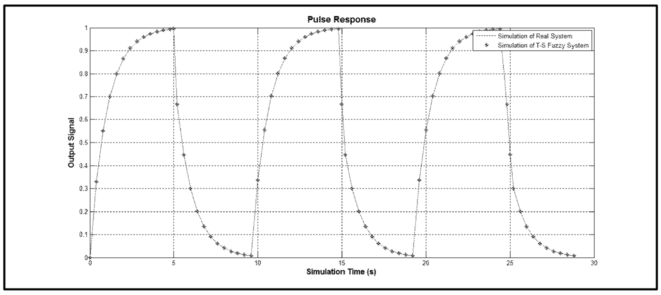


Figure 7: Time responses of the fuzzy RC model and the real circuit for square wave input

Fuzzy Modeling of RLC Circuit

The transfer function $H(s)$ of RLC circuit shown in figure (8) is as follows:

$$H(s) = \frac{\frac{R}{L}s}{s^2 + \frac{R}{L}s + \frac{1}{LC}}$$

For simplicity, the componentsof the circuit and the sampling time will be chosenasR =1, C = 1.3, L = 0.33, and$T$s=0.1 respectively.Then the sequence representing the output for this circuit is as follows:

$$y_k = 1.72y_{k-1} - 0.74y_{k-2} + 0.25x_{k-1} - 0.25x_{k-2} \quad (6)$$

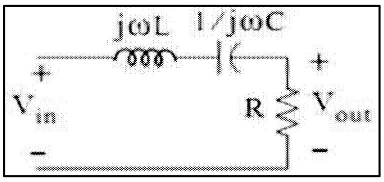


Figure 8: RLC circuit

Antecedent Membership Functions

For simplicity [0, 1] defined as a universe of discourse (UD) for input variablex_k , then the UD of variablesx_{k-1} ,x_{k-2}, are the same of x_k. The UD of other input variables i.e. (y_{k-1},y_{k-2}), can be found through a numerical simulation of equation (6), where the maximum and minimum of y_k represents the UD of y_{k-1} andy_{k-2}. So:X_{k-1},$X_{k-2} \in$ [0, 1],Y_{k-1}, $Y_{k-2} \in [-0.753, 0.753]$.The set of antecedent MFs as shown in figure (9) and Table (2), can characterize this linear system.

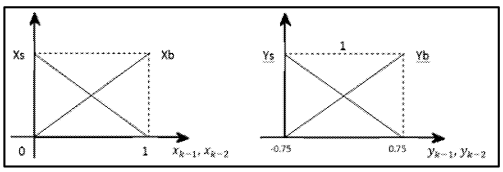


Figure 9: Antecedent membership functions for RLC model

Consequent Membership Functions

The UD of the output will be defining in the range [0, 1]. The following constants values must be set as crisp consequent MFs (singletons):(-0.74, -0.99, -0.48, -0.74, -1.85, -2.11, -1.59, -1.85, 1.85, 1.59, 2.113, 1.85, 0.74, 0.48, 0.99, and 0.74).

Table 2: characterized the system in equation (12)

y_{k-1}	y_{k-2}	x_{k-1}	x_{k-2}	y_k
-0.753	-0.753	0	0	-0.740
-0.753	-0.753	0	1	-0.998
-0.753	-0.753	1	0	-0.482
-0.753	-0.753	1	1	-0.740
-0.753	0.753	0	0	-1.856
-0.753	0.753	0	1	-2.114
-0.753	0.753	1	0	-1.597
-0.753	0.753	1	1	-1.856
0.753	-0.753	0	0	1.856
0.753	-0.753	0	1	1.597
0.753	-0.753	1	0	2.114
0.753	-0.753	1	1	1.856
0.753	0.753	0	0	0.740
0.753	0.753	0	1	0.482
0.753	0.753	1	0	0.998
0.753	0.753	1	1	0.740

The Rules

Since the fuzzy system has four inputs, so it is defined by 24 rules. These rules can be concluded from figure (9) as well as table (2). For example the first rule will be:

IF$y_{k-1}$is Ys ANDy_{k-2} is Ys AND x_{k-1} is Xs AND x_{k-2} is Xs THENy_k is -0.74

Simulations and Results

The Simulink implementation of the RLC circuit using TS model as well as the real system, are as show in figure (10). The time responses for the two models are as show in figure (11). It is obvious that both responses are identical.

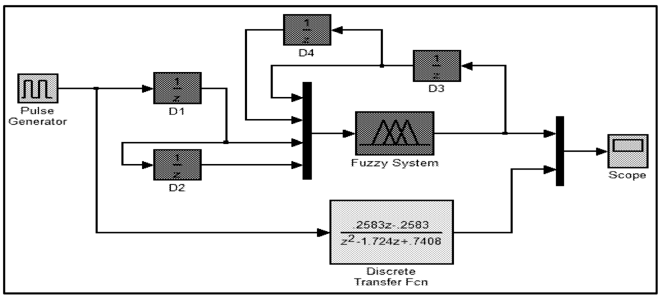


Figure 10: Simulink implementation of the RLC fuzzy circuit and real circuit

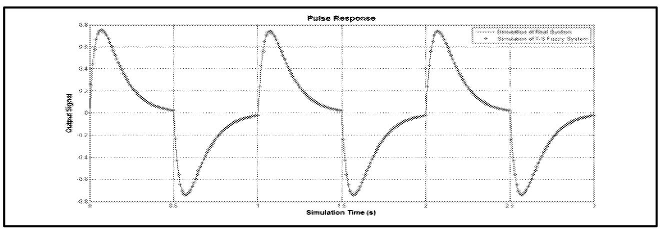


Figure 11: Time responses of the fuzzy RLC model and the real circuit for square wave input

SECTOR NONLINEARITY

Sector nonlinearity is based on the following idea: consider a simple nonlinear system $\dot{x}(t) = f(x(t))$, where $f(0) = 0$. The aim is to find the global sector such that the function $f(x(t))$ is bounded by these two lines $a_1 x\,(t)$ and $a_2 x\,(t)$, which means $\dot{x}(t) = f(x(t)) \in [a_1 a_2]x(t)$. Figure (12a) illustrates the sector nonlinearity approach. This approach guarantees an exact fuzzy model construction. However, it is sometimes difficult to find global sectors for general nonlinear systems. In this case, we can consider local sector nonlinearity. This is reasonable as variables of physical systems are always bounded. Figure (12b) shows the local sector nonlinearity, where two lines become the local sectors under condition $-d < x(t) < d$. The fuzzy model exactly represents the nonlinear system in the ''local'' region, that is $-d < x(t) < d$(Kazuo and Wang 2001).

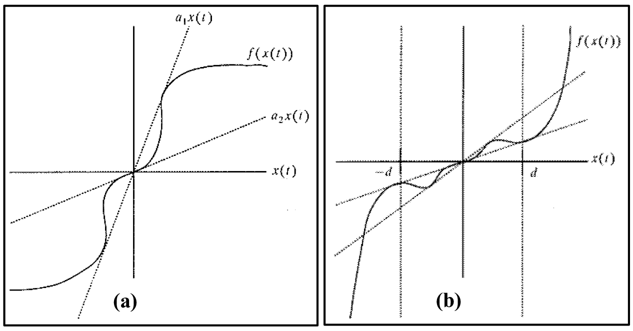


Figure 12:(a) Global sector nonlinearity(b)Local sectornonlinearity

CHAOTIC SYSTEMS

Over the last three decades, chaos has been extensively studied within the scientific, engineering and mathematical communities. Chaos behavior can occur everywhere, even in very simple and low-dimensional nonlinear systems(Ihsan et al. 2010).Chaotic systems have the property of sensitive dependence on initial condition so called butterfly effect.

Most chaotic systems are quite complex in practice and have strong nonlinearities so that it is difficult, if not impossible, to build rigorous mathematical models. Fortunately, a certain class of chaotic dynamical systems can be express in some forms of either a linear mathematical model locally or an aggregation of a set of linear mathematical models (Zhong et al. 2006).

In the regard of representing chaotic systems using TS fuzzy models, Lorenzoscillator, which is a nonlinear, three-dimensional and deterministic systemfrom a technical standpoint, will be representedby the TS modeling framework, then a new method will be presented to exhibit chaotic behavior for RLD circuit.

Lorenz Oscillator

The equations describe Lorenz Oscillator are as follows (Frank 2000):

$$\dot{x}_1(t) = -ax_1(\mathrm{t}) + ax_2(\mathrm{t})$$
$$\dot{x}_2(t) = cx_1(t) - x_2(t) - x_1(t)x_3(t) \qquad (7)$$

$\dot{x}_3(t) = x_1(t)x_2(t) - bx_3(t)$

Where *a*, *b*, and *c* are parameters.

Conventional method can be used to represent this system as (Kevin and Oppenheim 1993), as well,fuzzy structure presents another technique to implement the same system. The following procedure demonstrates that:

Assume that$x_1(t) \in [-d,\ d]$ and $d > 0$. The goal is to derive a TS fuzzy model from the above given nonlinear system equations by the sector nonlinearity approach as the response of the TS fuzzy model in the specified domain exactly match with the response of the original system. Then, the following fuzzy model should exactly represent the nonlinear equation under $x_1(t) \in [-d,\ d]$:

Here $x_1(t)$ is a nonlinear term in the equations, in the last two equations, so we make it as our fuzzy variable. Generally, it denoted as *z* and known as premise variable. For equation (7), it can written as

$$\dot{x}\,(t) = \begin{bmatrix} -a & a & 0 \\ c & -1 & x_1 \\ 0 & x_1 & -b \end{bmatrix} x(t)$$

Where $x(t) = [x_1(t)x_2(t)x_3(t)]^{\mathrm{T}}$. The first step for any kind of fuzzy modeling is to determine the fuzzy variables and membership functions (Kamyar 2008).

To acquire membership functions, the minimum and maximum value of $z(t)$ which under $x_1(t) \in [-d,\ d]$ are max $z(t) = d$; min $z(t) = -d$. Therefore,x_1 can be represent by two membership functions M_1 and M_2as shown in figure (13).

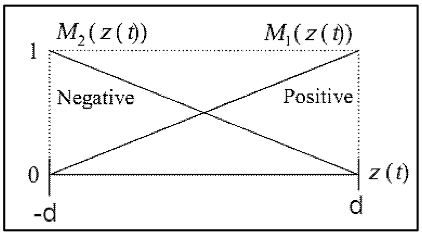


Figure 13: Membership functions for Lorenz oscillator

$$z(t) = x_1(t) = \mathrm{M}_1\big(z(t)\big) \cdot d + \mathrm{M}_2\big(z(t)\big) \cdot (-d)$$

But $\mathrm{M}_2\big(z(t)\big)$ is the complement of $\mathrm{M}_1\big(z(t)\big)$, then

$$\mathrm{M}_1\big(\mathrm{z(t)}\big) + \mathrm{M}_2\big(\mathrm{z(t)}\big) = 1$$

so:

$$\mathrm{M}_1\big(x_1(t)\big) = \frac{1}{2}\left(1 + \frac{z(t)}{d}\right)$$

$$\mathrm{M}_2\big(x_1(t)\big) = \frac{1}{2}\left(1 - \frac{z(t)}{d}\right)$$

The nonlinear system (7) can be model by the following fuzzy rules:

Rule 1: IF $z(t)$ is M_1, THEN $\dot{x}(t) = \mathrm{A}_1 x(t)$

Rule 2: IF $z(t)$ is M_2, THEN $\dot{x}(t) = \mathrm{A}_2 x(t)$
Where the subsystems are determined as:

$$\mathrm{A}_1 = \begin{bmatrix} -a & a & 0 \\ c & -1 & -d \\ 0 & d & -b \end{bmatrix}, \qquad \mathrm{A}_2 = \begin{bmatrix} -a & a & 0 \\ \mathrm{c} & -1 & d \\ 0 & -d & -b \end{bmatrix}$$

Now $\dot{x}(t)$ can be deriving out of defuzzification process as:

$$\dot{x}(t) = \mathrm{M}_1\big(z(t)\big)\mathrm{A}_1 x(t) + \mathrm{M}_2\big(z(t)\big)\mathrm{A}_2 x(t). \qquad (8)$$

This TS fuzzy model can exactly represent the nonlinear system in the region of $x_1(t) \in [-d,\ d]$.

For chaos to be emerge, the nominal values of the parameters are $a = 10$, $b = 8/3$, and $c = 28$ (Frank 2000; Zhong 2006a). The boundary of $x_1(t)$ was [-30, 30], which found through a numerical simulation of equations (7) (Zhong et al. 2006). The initial conditions ($x_1(0)$, $x_2(0)$, and $x_3(0)$) are (10, -10, and -10).

Simulink implementation of sector nonlinearity form for Lorenz oscillator described in equation (8)is presented here as shown in figure (14).

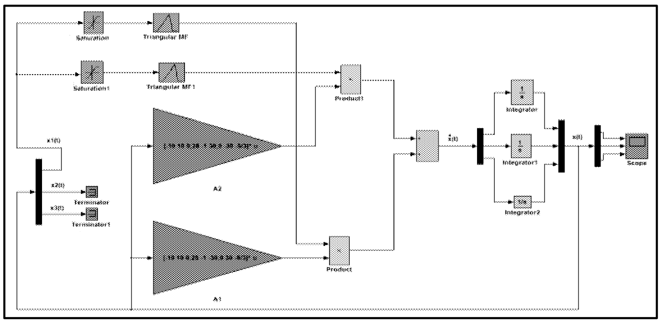


Figure 14: Matlab-Simulink for Lorenz oscillator

The time series resulted shown in figure (15) whichresponded shown that the system behave as chaotic oscillator, neither predictable nor random.

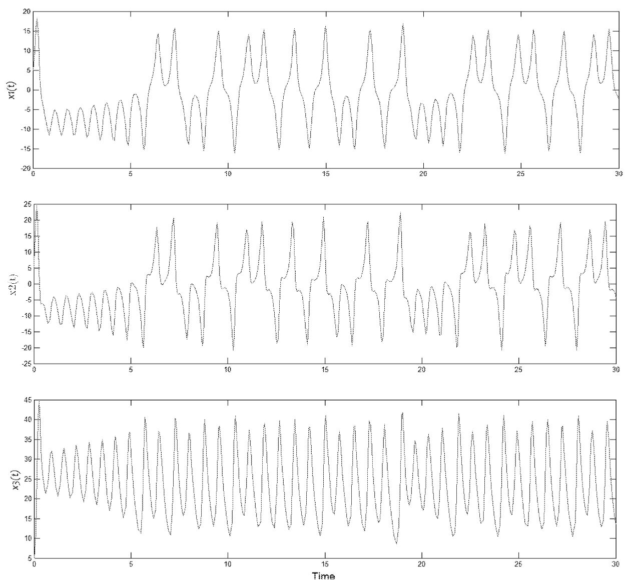


Figure 15: Chaotic behavior of Lorenz oscillator that are implemented on the fuzzy system With a=10, b=8/3, c=28 and $x_1(t)$ bounded by $\pm$ 30

The solution executes a trajectory, plotted in three dimensions, which wind around and around. The trajectory of TS fuzzy model for the Lorenz system is as shown in figure (16), exhibits that the system has a double-scroll chaotic attractor.

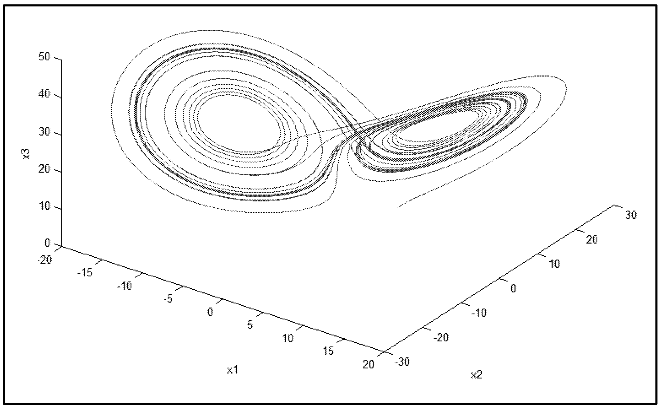


Figure 16: Trajectory of the chaotic Lorenz oscillator

The spectrum of $x_1(t)$, $x_2(t)$, and $x_3(t)$, (calculated by IFFT) are as shown in figure (17). It can be seen that the spectra of $x_1(t)$ and $x_2(t)$ are locally wide-band in the low frequency range and gradually approach to zero as the frequency increases, which shows the existence of chaos in $x_1(t)$ and $x_2(t)$. In comparison, the spectrum of $x_3(t)$ is much simpler: it has a significant spectral peak at the frequency f $\approx$ 1.3 Hz, which implies that $x_3(t)$, is nearly periodic, i.e., weakly chaotic.

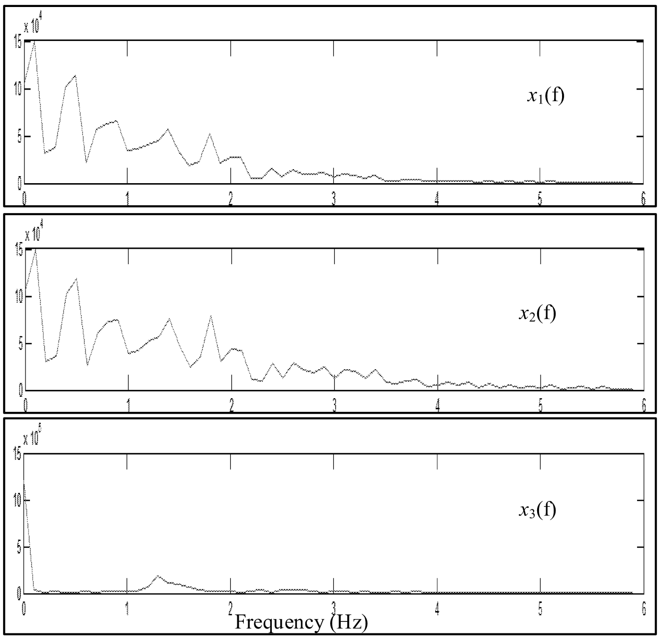


Figure 17: Spectrum of$x_1(t)$, $x_2(t)$, and $x_3(t)$

RLD Circuit

The power of fuzzy modeling appears in the simulation of nonlinear dynamic system .A chaotic behavior in RLD circuit, which is a simple nonlinear electronic circuit, is considered. The circuit requirement to create chaos is a sinusoidal signal generator driving a series circuit with a resonant frequency, a large inductor, a small resistor, and a semiconductor diode. The diode used in this circuit is not an ideal rectifier (Jianxin 2001), and then can be modeled as an ideal diode and a voltage dependent capacitance as shown in figure (18). The signal supplied to drive the circuit is a sinusoid with amplitude 10 Volt peak-to-peak, and frequency 20 KHz(Micah and Whitmer 2003).

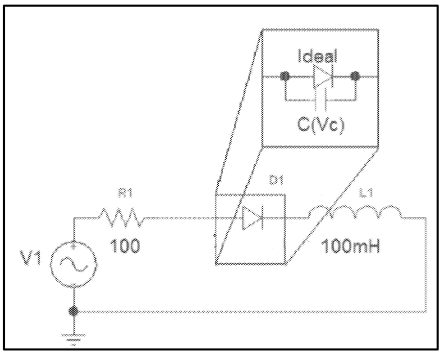


Figure 18: Equivalent circuit for simple RLD circuit

The circuit will behave in two different modes: the first is an RL circuit when the diode is in the forward biasing and the second is RLC circuit when the diode is in reverse biasing. The fuzzy model that represents this circuit also consists of two fuzzy models, (RL and RLC fuzzy models). As shown in figure (19), the fuzzy system represents the RL circuit, yellow color in the figure, consists of three input/one output, while the fuzzy system that represents the RLC, green color in the figure, consists of five input/single output. These two fuzzy systems are working together in the system over all to mimic the RLD circuit showed in figure (18).

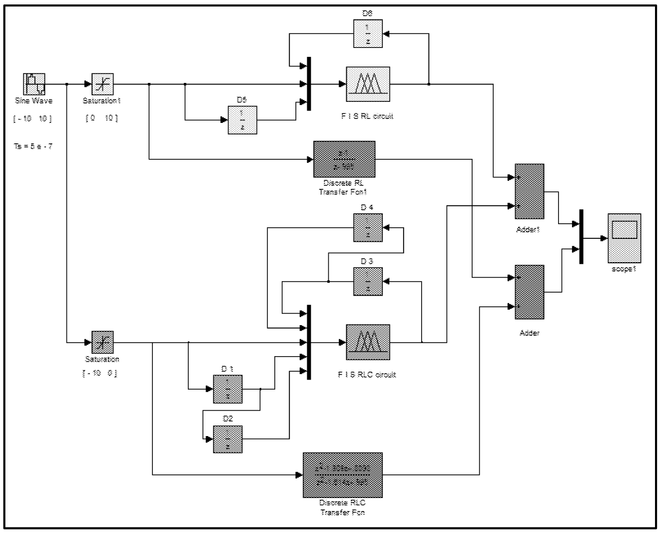


Figure 19: Matlab Fuzzy model of RLD with the transfer function H(z) model shown in violet

The time responses for the fuzzy model as well as the real circuit shown in figure (20) shows that the fuzzy model behave exactly as the behavior gained by the response of the transfer function.

The spectrum of $v_L(t)$, (calculated by IFFT) is shown in figure (21). The spectrum shows that as well as the resonant frequency, there are other frequencies generated and their amplitudes gradually approach to zero as the frequency increases. This mean that the RLD circuit leads to period doubling, which shows the existence of chaos in $v_L(t)$. But at frequency 200KHz, high amplitude oscillations are generated, therefore we can make benefit so that we can generate 10 times the input frequency.

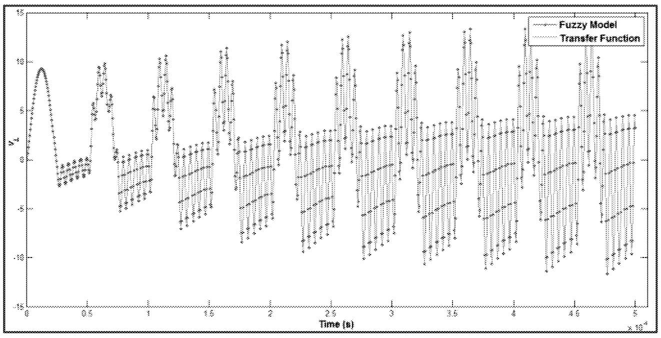


Figure 20: Time response: for the Fuzzy model and the transfer function

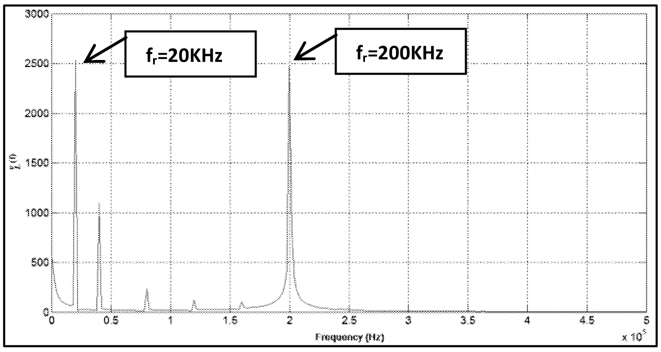


Figure 21: The spectrum of $v_L(t)$

CONCLUSIONS

In this paper, a novel view was proposed to construct models of dynamic electronic devices from the mathematical expressions of these devices. The work concerns the use of Takagi-Sugeno fuzzy structures in the modeling procedures. The results show that the fuzzy circuits, RC and RLC circuit, which are considered as examples of linear dynamic systems, behave as the original circuits. For nonlinear dynamic systems, two examples were taken into account to construct fuzzy models for them. These systems are Lorenz oscillator, a nonlinear dynamic chaotic system, and a chaotic behavior for RLD circuit, a simple nonlinear electronic circuit.Simulink implementation of sector nonlinearity form for Lorenz oscillator is proposed here. The results show that the fuzzy models can exactly represent the original systems.

FUTURE WORK

Design real time fuzzy models for higher complexity of electronic devices such as different types for transistors and diodes.

Design and implementation general real time fuzzy model on FPGA chip that is capable to perform many electronic circuits.

REFERENCES

Frank C. Hoppensteadt 2000. Analysis and Simulation of Chaotic Systems. Springer-Verlag New York.

H. Ying. 2000. Fuzzy Control and Modeling: Analytical Foundations and Applications, IEEE Press series on biomedical engineering

Ihsan Pehlivan, Yilmaz Uyaroglu, and Mesut Yogun 2010. "Chaotic oscillator design and realizations of the Rucklidge attractor and its synchronization and masking simulations." Scientific Research and Essays Vol. 5(16), pp. 2210-2219, Academic Journals.

Jacek Kluska. 2009. Analytical Methods in Fuzzy Modeling and Control. Springer-Verlag, Berlin Heidelberg.

Jianxin Zhang 2001. "Investigation of Chaos and Nonlinear Dynamical Behaviour in Two Different Self-Driven Oscillators." Doctor of Philosophy thesis, Department of Electronic Engineering Queen Mary, University of London.

Kazuo Tanaka and Hua O. Wang. 2001. Fuzzy Control Systems Design and Analysis: A Linear Matrix Inequality Approach.John Wiley & Sons.

Kevin M. Cuomo and Alan V. Oppenheim. 1993."Circuit Implementation of Synchronized Chaos with Applications to Communications."Phys. Rev. Lett., Volume 71, Number 1, pp.65-68.

Leandro dos Santos Coelho and Ernesto Araujo 2007."Fuzzy Model Optimization Based on Nelder-Mead Simplex Method Applied to Identification of a Chaotic System." Proceedings of COBEM, Copyright by ABCM.

Micah Richert and Diane Whitmer. 2003. "Chaotic Dynamics of RLD Oscillator." Unpublished lab report for biophysical measurements lab, UCSD Physics Department.

R. Jang and N. Gulley. 2001. Fuzzy Logic Toolbox for Use with Matlab. The MathWorksInc., Natick, MA.

Zhong Li 2006. "Fuzzy Chaotic Systems Modeling, Control, and Applications", ©Springer-Verlag Berlin Heidelberg.

Zhong Li; Wolfgang A.; Halang; and Guanrong Chen. 2006. Integration of Fuzzy Logic and Chaos Theory. Springer-Verlag Berlin Heidelberg.

BIOGRAPHIES

Basil Shukr Mahmood was born in 1953, finished his B.Sc. in electrical engineering in 1976 , then completed his M.Sc. and Ph.D. in 1979 and 1996 respectively in processors and processing. Now he is a professor of microprocessors and computer architecture in the university of Mosul/College of electronics. His interest is in microprocessors and signal processing, computer architecture, data processing, real time systems and real time networking.

Omar Abd Al Razzak was born in Mosul-Iraq and went to the University of Mosul, where he studied Electronic Engineering and obtained his degree in 1993. After doing his military service, he worked in Bseiso Eng. Corporation co. Amman-Jordan, then as a trainer in Naje-Fonas center in

Benghazi-Libya. In 2007, he worked in Ninawa Planning Unit-Ministry of Planning. In 2009, he began his study for MSc. in Electronic Engineering in Mosul University.

Simulation of Model Reference Neural Network for Network Routing Protocols

Anas Ali Hussien

College of Information Engineering

Nahrain University

Baghdad / Iraq

E-mail: anasali78@yahoo.com

Keywords- Search methods, Dijkstra's Algorithms, Protocols, Routing

ABSTRACT

This paper aims to develop a model reference neural network algorithm to solve a network routing protocol problem. The algorithm has to find the shortest path between the source and destination nodes.

In literature, the routing problem is solved using search graph techniques to find the shortest path. Dijkstra's algorithm is one of the popular techniques to solve this problem. The developed Model reference neural network algorithm is compared with Dijkstra's algorithm to solve a routing problem. Simulation results are carried out for both algorithms using MATLAB. The results confirmed very accurate results of the proposed algorithm. The obtained performance is similar to Dijkstra's algorithm.

INTRODUCTION

Dijkstra's algorithm, conceived by Dutch computer scientist Edsger Dijkstra in 1959 [1] is a graph search algorithm that solves the single-source shortest path problem for a graph with nonnegative edge path costs, producing a shortest path tree [2]. This algorithm is often used in routing.

An equivalent algorithm is developed by Edward F. Moore in 1957 [3]-[4]. For a given source vertex (node) in the graph, the algorithm finds the path with the lowest (i.e. the shortest path) between that vertex and every other vertex. It can also be used for finding costs of shortest paths from a single vertex to a single destination vertex by stopping the algorithm once the shortest path to the destination vertex has been determined. For example, if the vertices of the graph represent cities and edge path costs represent driving distances between pairs of cities connected by a direct road, Dijkstra's algorithm can be used to find the shortest route between one city and all other cities. The shortest path first is widely used in network routing protocols, most notably OSPF (Open Shortest Path First). OSPF is a dynamic routing protocol. It is a link state routing protocol and is part of the interior gateway protocols group. OSPF keeps track of the complete network topology and all the nodes and connections within that network. The basic workings of the OSPF routing protocol are as follows:

A. Startup

When a router is turned on it sends Hello packets to all neighboring devices, and the router receives Hello packets in response. From here routing connections are synchronized with adjacent routers that agree to synchronize.

B. Update

Each router will send an update message called its “link state” to describe its database to all other devices. So that all the routers connected together have an up to date description of each topology that is connected to each router.

C. Shortest path tree

Each router will calculate a mathematical data structure called “shortest path tree” that describes the shortest path to the destination address, this is where OSPF gets its name from. It will try to open the shortest path first. OSPF routing protocol is a very important protocol to consider when setting up routing instructions on the network. As OSPF gives the routers the ability to learn the most optimal (shortest) paths it can definitely speed up data transmission from source to destination. In literature, Dijkstra's algorithm is often described as a greedy algorithm. The Encyclopedia of Operations Research and Management Science describes it as a "node labeling greedy algorithm" and a greedy algorithm is described as "a heuristic algorithm that at every step selects the best choice available at the step without regard to future consequence" [4].

Routing is a process of transferring packets from source node to destination node with minimum cost (external metrics associated with each routing interface). Cost factors may be the distance of a router (Round-trip-delay), network throughput of a link or link availability and reliability expressed as simple unit less numbers. Hence the routing algorithm has to acquire, organize and distribute information about network states. It should generate feasible routes between nodes and send traffic along the selected path and also achieve high performance [5]. Routing process uses a data structure called routing table at each node to store all the nodes which are at a one hop distance from it (neighbor node). It also stores the other nodes (hop count more than one) along with the number of hops to reach that node, followed by the neighbor node through which it can be reached. The router decides which neighbor to choose from the routing table reach a specific destination. In literature, different approaches are applied to solve this problem as: Dijkstra's algorithm [6], dynamic programming technique [4], and genetic algorithm [5], [7].

DIJKSTRA’S ALGORITHM

The Dijkstra’s algorithm calculates the shortest path between two points on a network using a graph made up of nodes and edges. It assigns to every node a cost value. Set it to zero four source node and infinity for all other nodes. The algorithm divides the nodes into two sets: tentative and permanent. It chooses nodes, makes them tentative, examines them, and if they pass the criteria, makes them permanent. The algorithm can be defined by the following steps [6]:

1. Start with the source node: the root of the tree.
2. Assign a cost of 0 to this node and make it the first permanent node.
3. Examine each neighbor node of the node that was the last permanent node.
4. Assign a cumulative cost to each node and make it tentative.
5. Among the list of tentative nodes
a. Find the node with the smallest cumulative cost and mark it as permanent. A permanent node will not be checked ever again; its cost recorded now is final.
b. If a node can be reached from more than one direction, select the direction with the shortest cumulative cost.
6. Repeat steps 3 to 5 until every node becomes permanent. If the algorithm is applied to the network in Figure 1 to calculate the shortest path between the source node a(1) and the destination node b(5); the shortest path will be 1-3-6-5 with cost 20.

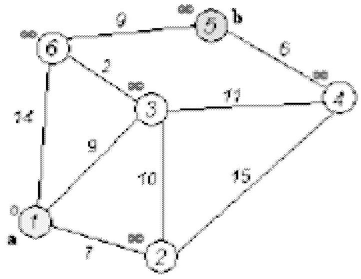


Figure 1 Network topology

PROPOSED ALGORITHMS

The proposed algorithm is based on a model reference neural network (see Figure 2) that are synchronized with the reference clock of the predictable dynamics process in a closed-loop system. The model reference neural network contains a neural network model, which models the real process with predictable dynamic variations.

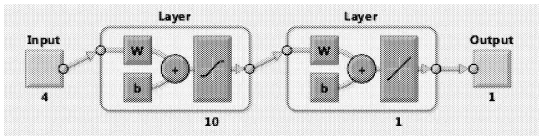


Figure 2 Neural network

The use of neural networks for nonlinear process modeling and identification is justified by their capability to approximate unknown non-linear systems. A nonlinear model that includes a large class of non-linear processes is the following NARMAX model

$$y(k) = f[y(k-1),\ldots,y(k-n),u(k-d-1),\ldots,u(k-d-m)] \quad (1)$$

Where *f (.)* is some nonlinear function, *d* is the dead time in *T/sample*, *n* and *m* are the orders of the nonlinear system model, *u* and *y* being the input and the output of the process with order n and m respectively. A neural network based model, NNARMAX, corresponding to the NARMAX model, may be obtained by adjusting the weights of multi-layer perceptron architecture with adequately delayed inputs [8].

$$y(k) = f^N[u(k-d-1), y(k-1)] \quad (2)$$

Where f^N denotes the input-output transfer function of the neural network, which replaces the non-linear function f in (1) and, $u(k\text{-}d\text{-}1)$ and $y(k\text{-}1)$ have the following structure

$$u(k-d-1) = [u(k-d-1)u(k-d-2)...u(k-d-m)]^T$$
$$y(k-1) = [y(k-1)y(k-2)........y(k-n)]^T \tag{3}$$

For a two layer network, the following expression is obtained from eq. (2)

$$y(k) = \sum_{j=1}^{p} W_j \sigma_j (W_j^u u(k-d-1) + W_j^y (k-1) + b_j) + b \tag{4}$$

Where p is the number of neurons in the hidden layer, σ_j is the activation function for the j-th neuron from the hidden layer, u_j^w the weight vector for the j-th neuron with respect to the inputs stored in $u(k\text{-}d\text{-}1)$, y_j^w the weight vector for the j-th neuron with respect to the outputs stored in $y(k\text{-}1)$, b_j the bias for the j-th neuron from the hidden layer, w_j the weight for the output layer corresponding to the j-th neuron from the hidden layer and b the bias for the output layer. Such structures with a single hidden layer are considered satisfactory for most of the cases.
Since all the industrial processes are working in closed-loop, a closed-loop identification method has been used to obtain the neural model of the process. In order to capture all the nonlinear dynamics of the process, the training data had to be attained around several different operating points such that the entire variation range of the process output to be covered. For this reason, a stepwise reference was chosen and then summed with a pseudo random binary signal generated with a shifting register [9].

In the proposed algorithm, any path from the source node to the destination node is a feasible solution. The optimal solution is the shortest one. Consider the network topology in Figure 3 with 6 nodes. This network will be simulated in the next section. Each node has a number in Figure 3 and the nodes are used to encode a path as a string expressed by the order on numbers. For example, 1-5-4-6 is a feasible path with source node 1 and destination node 6; it might be optimal or not.

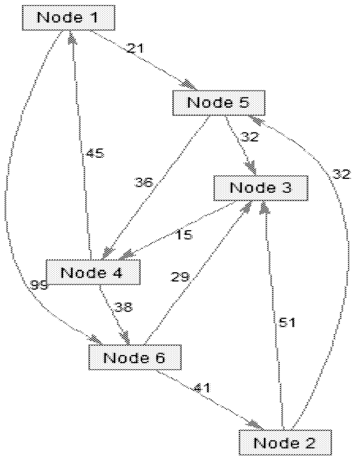


Figure 3 Network topology for simulation

IMPLEMENTATION OF DIJKSTRA'S ALORITHM

The network is created using the following MATLAB command:

net = sparse([6 1 2 2 3 4 4 5 5 6 1],[2 6 3 5 4 1 6 3 4 3 5], 41 99 51 32 15 45 38 32 36 29 21])

The first three row vectors in the sparse command represent source nodes, destination nodes, and equivalent costs respectively. The MATLAB command graphshortestpath is executed to find the shortest path. This command applies Dijkstra's algorithm as the default one to find the optimal solution.

[cost, path, pred] = graphshortestpath(net,1,6)

The obtained solution is:

cost = 95
path = 1 5 4 6
pred = 0 6 5 5 1 4

That means; the shortest path from node 1 to node 6 has to pass via nodes 1, 5, 4, 6 with minimum cost 95. The network is viewed using the following command in MATLAB
[dist,path,pred]= graphshortestpath(net,1,6)

NEURAL MODEL REFERENCE FOR SHORTEST PATH ALGORITHM

The network under consideration is represented as a connected graph with N nodes. The metric of optimization is the cost of path between the nodes. The total cost is the sum of cost of individual hops. The goal is to find the path with minimum total cost between source node and destination node. This paper presents a Model reference neural network algorithm to find the shortest path. The shortest path / coast state matrix is given by

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 21 & 99 \\ 0 & 0 & 51 & 0 & 32 & 0 \\ 0 & 0 & 0 & 15 & 0 & 0 \\ 45 & 0 & 0 & 0 & 0 & 38 \\ 0 & 0 & 32 & 36 & 0 & 0 \\ 0 & 41 & 29 & 0 & 0 & 0 \end{bmatrix}$$

Applying model reference neural network, with the following parameters (Number of epoch =15, μ_u=0.01, hidden layer= 12with performance of $6.61*10^{-6}$, training sample 6000, validation check 6) will obtain the following results.

Figure 4 illustrates the identification between input and output signals using a model reference neural network. After 6000 training samples the results show that they are approximately identical.

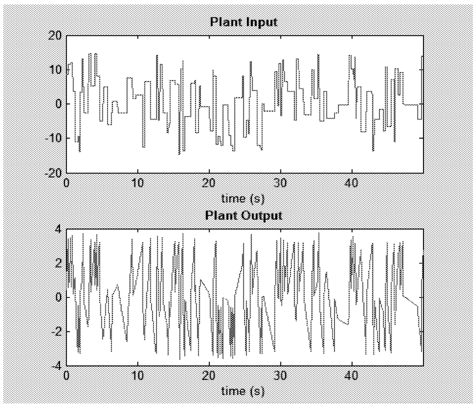


Figure 4 Identification of the network topology

Figure 5 demonstrate the identification of the state transition matrix of shortest path using model reference neural network after 10000 training samples.

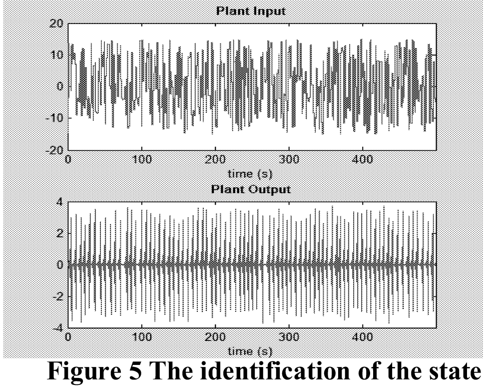


Figure 5 The identification of the state transition matrix of shortest path

The training data for the neural network is illustrated in figure 6.

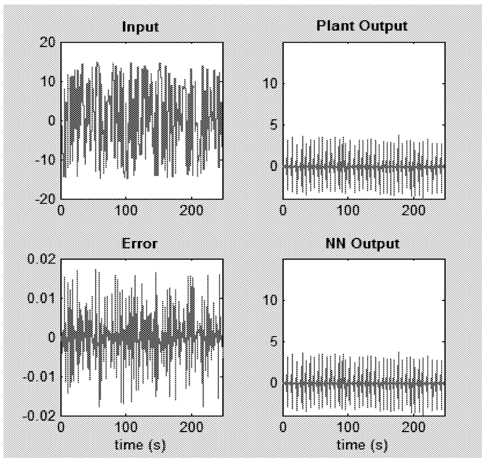


Figure 6 Training data for neural network

From figure 6; the error can be estimated as a very small value, which can be seen a validation and testing error as shows in figures 7 and 8 respectively.

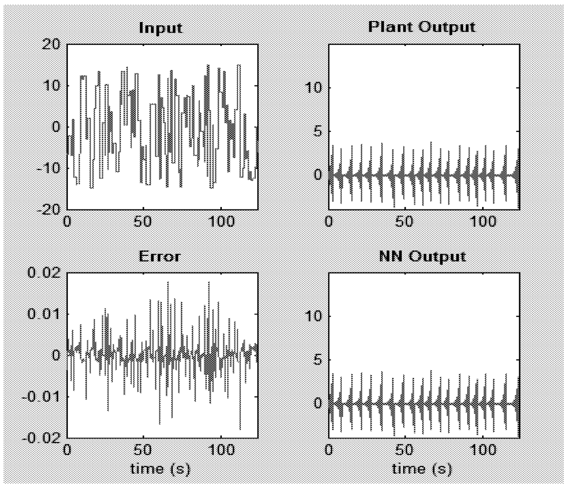


Figure 7 Validation error

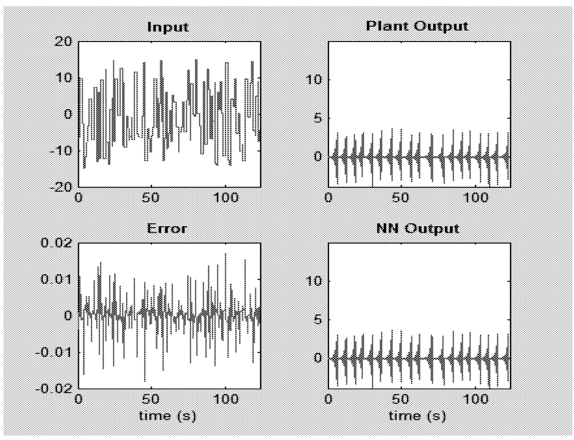


Figure 8 Testing data for NN model reference

Figure 9 illustrates the neural network training performance. From this figure it can be seen that the best validation performance occurred at epoch 9 with 5.6435 *10-6.

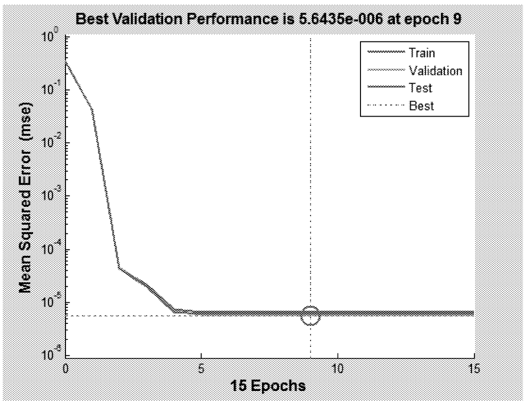


Figure 9 NN training performance

Finally, figure 10 demonstrate the gradient, µu, and validation checks with a number of epochs.

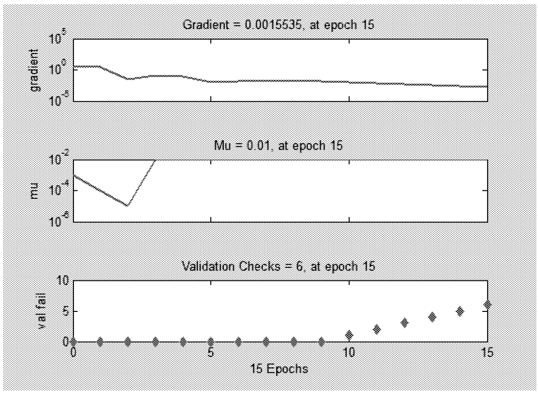


Figure 10 NN training state

In order to estimate the parameters of the neural model, a training sequence was built such that the process output explores its whole operating range.

CONCLUSION

From the above results, it can be concluded that the shortest path can be obtained using model reference neural network with very high accurate assumption, with respect to Dijkstra's algorithm.

REFERENCES

[1] E. W. Dijkstra, "A Note on Two Problems in Connexion with Graphs", Numerische Mathematik l, 269 – 271, l 959.

[2] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, Section 24.3: ‘Dijkstra's algorithm. Introduction to Algorithms (Second ed.)”, MIT Press and McGraw-Hill, pp. 595–601, 2001.

[3] E. F. Moore, "The shortest path through a maze", Proceedings of an International Symposium on the Theory of Switching (Cambridge, Massachusetts, 2–5 April 1957). Cambridge: Harvard University Press, pp. 285–292, 1959.

[4] M. Sniedovich, "Dijkstra’s algorithm revisited: the dynamic programming connexion". Journal of Control and Cybernetics 35 (3): 599–620, 2006.

[5] N. K. Cauvery and K. V. Viswanatha, "Routing in Dynamic Network using Ants and Genetic Algorithm", Int. J. of Computer Science and Network Security, Vol. 9 No.3, pp.194-200, March 2009.

[6] B. A. Forouzan, Data Communications and Networking, Fourth Edition McGraw Hill, 2007.

[7] N. Selvanathan and W. J. Tee, "A Genetic Algorithm Solution to Solve The Shortest Path Problem OSPF and MPLS", Malaysian Journal of Computer Science, Vol. 16 No. 1, pp. 58-67, 2003.

[8] Chen S, Billings S.A. and P.M. Grant, Non- linear system identification using neural networks, Int. J. Control, vol 49, no.3 ,1013-1032,1990. Landau I. D. System Identification and Control Design, Prentice Hall, 1990.

[9] Norgaard, M, Neural Network Based System Identification Toolbox, Technical Report 00-E-891, Department of Automation, Technical University of Denmark, 2000.

SIMULATION METHODS FOR DELAYS AND PACKETS DROPOUT AVOIDANCE IN SOLAR-POWER WIRELESS NETWORK

Waleed Al-Azzawi
Faculty of Technology
De Montfort University
Leicester, UK
Email: walazzawi@dmu.ac.uk

Marwan Al-Akaidi
Arab Open University
P.O.Box: 3322 Safat
13033 Kuwait
Email:m.al-akaidi@arabou.edu.kw

John Gow
Faculty of Technology
De Montfort University
Leicester, UK
Email: jgow@dmu.ac.uk

KEYWORDS

Wireless Network, Robust Control, Neural Predictor, Wi-Fi, WiMAX.

ABSTRACT

Time delay and packet dropout plays a critical role in the performance of real-time Wireless Network Control System (WNCS). Communication networks inevitably commence delays and packet dropouts in WNCS. Hence, it is necessary remove or reduces the effect of delays and packet dropout. In this paper, we compare two methods of simulating delays and packet dropout occurs in wireless network: Robust Model Predictive control (RMPC), and Neural Network predictive Control (NNPC). A numerical example is given to validate the proposed control algorithm and to illustrate the effectiveness of our results.

INTRODUCTION

SWNCS is a form of closed-loop distributed control system where sensors, actuators, and controllers are interlinked by communication wireless networks which are powered by solar energy [1]. SWNCS provides light control, wireless communications network such as Wi-Fi & WiMAX, CCTV surveillance, wireless sensor e.g. weather measurement sensor, speed sensor, etc., and all are powered by solar energy. Due to the many benefits which can be offered by SWNCS, such as simple installation and maintenance, low cost, no wiring, flexibility, clean environment, and high reliability, the SWNCS is appropriate to many applications, for examples, remote process control, altitude measurement for airplane and manufacturing automation [2]. However, the integration of the wireless network in closed-loop control systems presents challenging problems such as network delay, packet dropouts, jitter and limited bandwidth, all of which can degrade the performance of control system and even destabilized the system [3]. One of the majority basic challenges in the SWNCS is the network delay and packet dropouts. In reference [4], a neural network-based predictive controller has been presented for a class of discrete-time multivariable systems. Under the projected adequate conditions, the closed–loop stability of the overall control system has been well established through the Lyapunov stability theory. However references [5-6] have reported a variable sampling time in order to compensate for time delay. A feed forward multi layer Neural Network (NN) is first correctly developed to estimate the time delay for every sample time period. After that, this predicted time delay is taken as the sample time between the present and the future sampling cycles. So as to defeat the influence of network time delay, a novel strategy that uses a new Smith predictor aggregated with single neural adaptive control for the WNCS has been presented in [7]. There are many research deals with the RMPC as in references [8-9] a new linear discrete-time system is proposed which converts the uncertainty of time delay into the uncertainty of parameter matrix. Considering the size of time delay, the NCS with packet losses is modeled as an asynchronous dynamical system with rate constraints. In this paper we have investigated two methods of simulating delays and packet dropout occurs in wireless network. In RMPC the time delays of the SWNCS are considered as stochastic variables controlled by a Markov chain. A discrete-time Markovian jump system with norm unbounded time delay is presented to model the SWNCSs. Based on the SWNCS model, the RMPC based on full state feedback controller can be solved under the framework of LMIs. Other techniques presents methodology that employs NNPC to remove or reduce the effect of time delay in SWNCS, and to employ the value modified by BPNN error predictive model to manage the SWNCS time delays, The NNPC model is used to recursively calculate the output predictions and control signal predictions.

WNCS STOCHASTIC TIME DELAY AND PACKET DROPOUT

A WNCS with process, sensor, actuator, controller and wireless network is shown in Figure 1. The centralized controller communicates with the actuator and sensor over the wireless network**.** The data takes some time to transmit between the controller and actuator and between sensor and controller. The transmission time taken by the data depends upon the topology, routing method and the scheduling strategy used in the system [10]. There are three classes of delays in SWNCs:

a. Time delay between sensor-to-controller, τ_k^{sc}.
b. Computational delay in the controller, τ_k^{c}.
c. Time delay between controller-to-actuator, τ_k^{ca}.

Where, k is possible time dependence of delays Both τ_k^{sc} and τ_k^{ca} are modeled as two homogenous Markov chains that proceeds values in $\mathcal{M} = \{0, 1, \ldots, \tau_k^{sc}\}$ **and** $\mathbb{N} = \{0, 1, \ldots\ldots, \tau_k^{ca}\}$. The transition probability matrices are $\Lambda = [\lambda_{ij}]$ **and** $\pi = [\pi_{rs}]$. That means τ_k^{sc} jumps from mode *i* to j and τ_k^{ca} jump from mode r to s with probabilities λ_{ij} and π_{rs}. Packet dropout occurs due to noise, interference, node failures and packet collisions in the network. It is noted that the Markov chain models has integrated the time delay and packet dropout procedures in wireless networks at the same time, as analyzed and proof in [11].

SIMULATION METHOD FOR TIME DELAY AVOIDANCE IN SWNCS

Suppose the wireless Network control System (WNCS) as shown in Figure 5, which has a physical plant is given by the following system.

$$x(k+1)=Ax(k)+Bu\left(k-\tau_k^{ca}\right)\ldots(1a)$$

$$y(k)=Cx(k)\ldots(1b)$$

Where $x(k)$ is state vector, $u(k)$ is the control input, $y(k)$ is the output A, B, C are system matrices with appropriate dimension and τ_k^{ca} is random delay from controller-to-actuator.

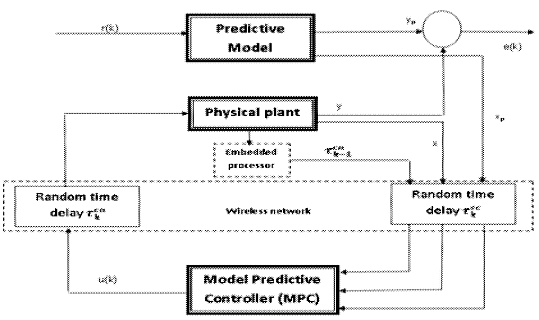


Figure 1: WNCS Model

RMPC Method

In any WNCS, the delay information is important for controller design. The two mode-dependent full state feedback controllers are to be designed by MPC.

$$u(k)=k_1\left(\tau_k^{sc},\tau_{k-\tau_k^{sc}-1}^{ca}\right)x_p\left(k-\tau_k^{sc}\right)+k_2\left(\tau_k^{sc},\tau_{k-\tau_k^{sc}-1}^{ca}\right)x\left(k-\tau_k^{sc}\right)\ldots(2)$$

Referring to the stochastic time delay τ_k^{ca}, achieved the following model:

$$x(k+1)=\tilde{A}X(k)+\tilde{B}u\left(k-\tau_k^{ca}\right)+\tilde{J}r(k)\ldots(3a)$$

$$e(k)=\tilde{C}X(k)\ldots(3b)$$

Where, $X(k)=\left[x_r(k)^T\ x(k)^T\right]$, $\tilde{A}=\begin{bmatrix}A_r & 0\\ 0 & A\end{bmatrix}$, $\tilde{B}=\begin{bmatrix}0\\ B\end{bmatrix}$, $\tilde{J}=\begin{bmatrix}B_r\\ 0\end{bmatrix}$

$\tilde{C}=\left[-C_r\ C\right]$

The SWNCS is achieved:

$$\xi(k+1)=A\left(\tau_k^{sc},\tau_{k-\tau_k^{sc}-1}^{ca},\tau_k^{ca}\right)\xi(k)+J\,r(k)\ldots(4a)$$

$$e(K)=C\,\xi(K)\ldots(4b)$$

Where, $K=\left[k_1\left(\tau_k^{sc},\tau_{k-\tau_k^{sc}-1}^{ca}\right)\ldots k_2\left(\tau_k^{sc},\tau_{k-\tau_k^{sc}-1}^{ca}\right)\right]$

$J=\left[\tilde{J}\ 0\ldots 0\,0\,0\ldots 0\right]^T$ $C=\left[\tilde{C}\ 0\ldots 0\,0\,0\ldots 0\right]^T$

$$A=\begin{bmatrix}A & 0 & \tilde{B}k & 0 & 0 & 0 & \ldots & 0\\ I & 0 & \ldots & 0 & 0 & 0 & \ldots & 0\\ 0 & I & \ddots & 0 & 0 & 0 & \ddots & 0\\ 0 & 0 & \ddots & 0 & 0 & 0 & \ddots & 0\\ 0 & 0 & k & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & \ldots & 0 & I & 0 & \ldots & 0\\ 0 & 0 & \ddots & 0 & 0 & I & \ddots & 0\\ 0 & 0 & \ldots & 0 & 0 & 0 & \ddots & 0\end{bmatrix}$$

At each sampling time k, the controller design scheme can be expressed as in [11]. Implement the first control law $U(k\|k)$ that is

$$U(k+m|k)=U(k|k)=k_1\left(\tau_{k+m}^{sc},\tau_{k+m-\tau_{k+m}^{sc}-1}^{ca}\right)$$

$$x_p\left(k-\tau_{k+m}^{sc}\right)+k_2\left(\tau_{k+m}^{sc},\tau_{k+m-\tau_{k+m}^{sc}-1}^{ca}\right)x\left(k-\tau_{k+m}^{sc}\right)\ldots(5)$$

By substituting the control signal given in system (3), the resulting SWNCS in (5) will become.

$$\begin{cases}\xi(k+1)=A\left(\tau_k^{sc},\tau_{k-\tau_k^{sc}-1}^{ca},\tau_k^{ca}\right)\xi(k)+J\,r(k)\\ e(k)=C\,\xi(k)\\ u(k)_2\le u_{max}\end{cases}$$

In the following, has been presented the MPC for full state feedback controller based on Lyapunov method. We consider the following min-max infinite horizon predictive cost function relating the system (4)

$$\min_{U(k+i|k)_{i=0,1,\ldots,}}\ \max_{\tau_k^{sc}[A,B,A_p]\in\Omega} J_\infty(k)\ldots(6)$$

$$J_\infty(k)=\sum_{m=0}^{\infty}E\left\{\begin{matrix}X^T(k+m|k)QX(k+m|k)+\\ U^T(k+m|k)RU(k+m|k)\end{matrix}|K_k\right\}$$

Where, E[.]-indicates the expectation

$X(k+m\|k)$-indicates the predictive value of state at time k,

$U(k+m\|k)$-indicates the predictive of the control signal at time k, $Q>0$ & $R>0$-are the weighting symmetric matrices, $K_k=\sigma\{x_o,r_o,\ldots\ldots\ldots\ldots,x_k,r_k\}$ is σ algebra created by $\{(x_l,r_l)\ \ 0\le l\le k\}$ and $\Omega=co\{[A_1,A_{1p},B_1]\ldots\ldots\ldots\ldots[A_l,A_{lp},B_l]\}$ is polytypic

Create the subsequent Lyapunov function

$$V(x(k),r(k))=x^T(k)p_{rk}(k)+\sum_{j=1}^{\tau_k^{sc}}x^T(k-j)sx(k-j)+$$

$$\sum_{j=1}^{\tau_k^{sc}}x^T(k-j)K_{rk}^T Tx(k-j)\ldots(7)$$

Where, $s>0$ &$T>0$ are the constant matrices and $p_{rk}>0$ are dependents on the mode $r_k(r_k\in s)$

$$E\{V(x((k+m+1)|k, r_k))(-V(x(k+m|k), r_k)|K_k)\} \le -E$$
$$\begin{Bmatrix} x^T(k+m|k)Qx(k+m|k)+u^T \\ (k+m|k)Ru(k+m|k)|K_k \end{Bmatrix} \ldots(8)$$

$$U(k+m|k) = k_1\left(\tau^{sc}_{k+m}, \tau^{ca}_{k+m-\tau^{sc}_{k+m}-1}|k\right) x_p\left(k-\tau^{sc}_{k+m}\right) +$$
$$k_2\left(+k_2\left(\tau^{sc}_{k+m}, \tau^{ca}_{k+m-\tau^{sc}_{k+m}-1}\right)x\left(k-\tau^{sc}_{k+m}\right)|k\right)x\left(k-\tau^{sc}_{k+m}\right)$$

For the performance index $J_\infty(k)$ to limited, we must have $x(\infty|k)=0$ therefore, we obtain $E\{V(x(\infty|k), r_k)\}=0$

Addition to both sides of equation (6) from m=0 to m=∞ we get:

$$\max_{[A, A_p, B]} J_\infty(k) \le V(x(k), r(k)) \ldots(9)$$

Subsequently the optimization problem of equation (6) above can be converted to the optimization of $V(x(k), r_k)$. Assuming that there exists a positive integer γ and satisfies:

$$\min_{u(k+i|k) i=0,1,\ldots,n-1} \max_{[A, A_p, B]\in\Omega} J_\infty(k) \le \min V(x(k), r_k) \le \gamma_{min} \ldots(10)$$

$$X^T(k)p_i ZX(k) + \sum_{j=1}^{\tau^{ca}_k} X^T(k-j)SX(k-j)$$
$$+\sum_{j=1}^{\tau^{sc}_k} X^T(k-j)K_j^T TK_j X(k-j) \le \gamma \ldots(11)$$

Theorem

For the MJLS uncertain system (5) with stochastic time delay, let the states of the system be measured. Then there exists state feedback controller as in equation (6).Such that the subsequent optimization problem is:

$$\min_{\gamma, Q, W, X, e(k)_i} \gamma \ldots(12)$$

$$\begin{bmatrix} U^2_{max} & e(k)_i \\ e(k)_i^T & Q_i \end{bmatrix} > 0 \ldots(13)$$

From (9), we obtain that the Lyapunov function $V(x(k), r_k)$ is decreasing, it follows that:

$$E\{V(x((k+m+1), r_k)|k) < E\{V(x((k), r_k)|k) \ldots(14)$$

$$E\{V(x((k+1), r_k)|k) < E\{V(x((k), r_k)|k) < \gamma \ldots(15)$$

Due to $x(k|k) = x(k)$, associated with system (4), we obtain:

$$x(k+1|k) = x(k+1|k+1) \ldots(16)$$

At sampling time k+1, we donate $X^{k+1}, P^{k+1}_{rk+1}, Q^{k+1}_i, W^{k+1}, e(k)^{k+1}_i$ as the optimal solution (13) and $X^k, P^k_{rk}, W^k, Q^k_i$ are the feasible solution of (12).

And therefore, obtain

$$E\{V(x((k+1)|k+1, r_k+1)|k) < E\{V(x((k+1|k), r_k)|k) < \gamma \ldots(17)$$

The analyses and proof of the theorem in [11].

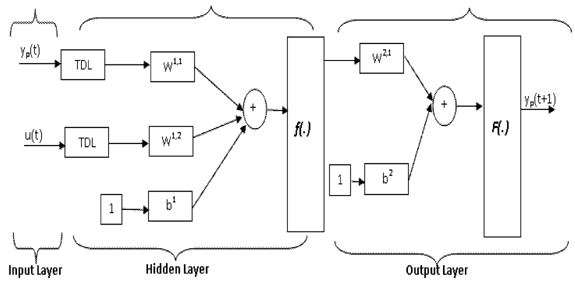


Figure 2: Neural Network Structure for SWNCS

NNPC Method

The NNPC controller that is implemented using the Neural Network Toolbox software employs a neural network model of a physical plant to predict future plant performance. The controller after that computes the control signal that will optimize plant performance over a specified future time horizon. The SWNCS model is employed by the controller to predict future performance. In the following section we will illustrate the system identification procedure. This is tracked by a description of the optimization procedure. Figure 2 presents the structure of the neural network SWNC model which uses previous inputs and previous plant outputs to predict future values of the plant output [12]. Choosing $y_p(t)$ as input vector, $W^{1,1}$ as hidden layer weighting matrix, b^1 as hidden layer bias matrix, $W^{2,1}$ as output layer weighting matrix, b^2 as output layer bias matrix, a^1 as hidden layer output vector and $y_p(t+1)$as output vector of the neural network , the following relations can be written :

$$a^1 = f\left(W^{1,1}y_p(t)+b^1\right) \ldots(18a)$$

$$y_p(t+1) = f\left(W^{2,1}a^1+b^2\right) \ldots(18b)$$

Where, $f(.)$ is tangent hyperbolic function and is defined as:

$$f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \ldots(19a)$$

$$\dot{f} = 1 - f^2(x) \ldots(19b)$$

The term back-propagation refers to the mode in which the gradient is calculated for nonlinear multilayer networks. There are a number of variations on the basic algorithm that are based on other optimization methods, such as conjugate gradient and Newton methods [13]. The BP mathematical approach is summarized as:

$$w^l_{ji}(k+1) = w^l_{ji}(k) - \lambda \frac{\delta F(k)}{\delta w^l_{ji}(k)} \ldots(20a)$$

$$b^l_j(k+1) = b^l_j(k) - \rho \frac{\delta F(k)}{\delta b^l_j(k)} \ldots(20b)$$

Where, l-layer number, $i = 1, 2, \ldots\ldots, \varepsilon_{l-1}$, and $j = 1, 2, \ldots\ldots, \varepsilon_l, \lambda$ and $\rho - learning$ $rates$

$$F(k) = \frac{1}{\varepsilon_2}\sum_{j=1}^{\varepsilon_2} e_j^2(k), \quad e_j(k) = t_j(k) - y^2_{pj}(k)$$

t-target vector and y_p-output vector and ε_2- control horizon

$$\frac{\delta F(k)}{\delta x}=\begin{cases}-2e_j(k)a_i^1, & x=w_{ji}^2\\ -2e_j(k), & x=b_j^2\end{cases}\quad\ldots(21)$$

$$\frac{\delta F(k)}{\delta x}=\begin{cases}-2\left\{\left[\sum_{m=1}^{\varepsilon_2}\left(e_m(k)w_{mj}^2(k)\right)\left(1-a_j^1\right)^2\right] y_i(k)\right\}, & x=w_{ji}^1\\ -2\left\{\sum_{m=1}^{\varepsilon_2}\left(e_m(k)w_{mj}^2(k)\right)\left(1-a_j^1\right)^2\right\}, & x=b_j^1\end{cases}\quad\ldots(22)$$

$$J=\frac{1}{\varepsilon_2 N}\left\{\begin{array}{l}\sum_{j=1}^{\varepsilon_2}\left[t(k+j)-y_M(k+j)\right]^2+\\ \partial\sum_{j=1}^{N}\left[\bar{u}(k+j-1)-u(k+j-2)\right]^2\end{array}\right\}\ldots(23)$$

Where, N- is the horizon over which the tracking error and the control increases are calculated, $\bar{u}$- is the variable provisional control signal, y_M- NN SWNCS model response and ∂- is the value which establishes the part that the sum of the squares of control increases has on the performance index. NN predictive controllers improve tracking performance by reducing the effect of time delays on SWNCS. The basic working principle of the NN predictive controller is to generate a sequence of control signals at each sample interval that optimize the control effort in order to track exactly the desired signal. The neural-network-based predictive Control signal is obtained from minimizing the predictive performance index equation (22) over the specified horizon as follow [14]:

$$u(k)=\left[u(k)\ u(k+1)\ldots\ldots u(k+N)\right]^T\ \ldots(24)$$

The control signal in (23) is found from optimization of the performance index (22) based upon the gradient descent method,

$$\begin{cases}u(k)=u(k-1)+\Delta u(k)\\ u(k)=u(k-1)+\lambda\dfrac{\delta y_p^T(k)}{\delta u(k)}\\ u(k)=u(k-1)+\lambda C^T(k)e(k)\end{cases}\quad\ldots(25)$$

Then the expression of the NNPC controller can be written in the following form

$$u(k)=u(k-1)+\lambda C^T e(k)\ldots(26)$$

where, $C(k)=\left[\dfrac{\delta y_p(k)}{\delta u(k)}\ \dfrac{\delta y_p(k+1)}{\delta u(k)}\ldots\ldots\dfrac{\delta y_p(k+N)}{\delta u(k)}\right]^T$

The control signal sequence is started in the sensor and the real time of the controller is unknown without clock synchronization. Further, is also not required in the proposed approach. Hence, at recent time k, the previous information of the physical plant before $k-\tau_k^{sc}$ is presented on the controller area as follow [15]:

$$\begin{bmatrix}y_p\left(k-\tau_k^{sc}\right)\\ y_p\left(k-\tau_k^{sc}-1\right)\\ \vdots\\ y_p\left(k-\tau_k^{sc}-N\right)\end{bmatrix},\begin{bmatrix}u\left(k-\tau_k^{sc}-1\right)\\ u\left(k-\tau_k^{sc}-2\right)\\ \vdots\\ u\left(k-\tau_k^{sc}-N\right)\end{bmatrix}\ldots(27)$$

Based on (24) and (27), the response predictions can be determined by the subsequent:

$$u\left(k-\tau_k^{sc}|k-\tau_k^{sc}\right)=-\left[\sum_{i=1}^{N}C^T u\left(k-\tau_k^{sc}-i\right)+\sum_{i=0}^{N}y_p\left(k-\tau_k^{sc}-1-i|k-\tau_k^{sc}\right)\right]\ldots(28)$$

$$=-\left[\sum_{i=1}^{N}C^T u\left(k-\tau_k^{sc}+j-i|k-\tau_k^{sc}\right)+\sum_{i=0}^{N}y_p\left(k-\tau_k^{sc}+j-i-1|k-\tau_k^{sc}\right)\right]\ldots(29)$$

The sampling time at sensor node $k-\tau_k^{sc}$ is set into an information packet with the control signal prediction sequence and transmits to the actuator. Obviously, from equation (29) we can generate control signal predictions as follow:

$$U\left(k-\tau_k^{sc}|k-\tau_k^{sc}\right)=\begin{bmatrix}u\left(k-\tau_k^{sc}|k-\tau_k^{sc}\right)\\ u\left(k-\tau_k^{sc}+1|k-\tau_k^{sc}\right)\\ \vdots\\ u\left(k-\tau_k^{sc}+N|k-\tau_k^{sc}\right)\end{bmatrix}\ldots(30)$$

Because the actuator is time-driven, the NNPC executes once at every sampling period despite if the actuator collects new information packets or not. If no information packet is received, the content and sampling time of the actuator schedule remain invariant. When receiving new control signal prediction sequences, the actuator compares the sampling time of the new packets with that of the schedule, and then updates the schedule with the following rule:

$$\begin{cases}U(k)=\bar{\bar{U}}(k) & \textit{if new information packet is recived}\\ U(k)=U(k-1) & \textit{otherwise}\end{cases}$$

Hence, the content of the schedule $U(k)$ is the newest control signal prediction sequence existing in the actuator node at time k, which can be expressed by:

$$U(k)=\begin{bmatrix}u\left(k-\tau_k^{sc}-\tau_k^{ca}|k-\tau_k^{sc}-\tau_k^{ca}\right)\\ u\left(k-\tau_k^{sc}-\tau_k^{ca}+1|k-\tau_k^{sc}-\tau_k^{ca}\right)\\ \vdots\\ u\left(k-\tau_k^{sc}-\tau_k^{ca}+N|k-\tau_k^{sc}-\tau_k^{ca}\right)\end{bmatrix}\ldots(31)$$

where, $k-\tau_k^{sc}-\tau_k^{ca}$-is the sampling time of the control signal predication sequence at actuator node.

In order to remove or reduce the effect of the time delay on SWNCS, the NNPC choses the proper control signal from (38) based on the forward and backward channel delays($i.e.\ \tau_k^{sc}\ \&\tau_k^{ca}$). Thus, the control signal input to the physical plant at time k will be

$$u(k)=u\left(k|k-\tau_k^{sc}-\tau_k^{ca}\right)\ldots(32)$$

SIMULATION RESULT

In order to validate the effectiveness of the proposed method, we consider the following numerical example, where the parameters of the discrete-time physical plant with sampling time T_s=0.01 sec is described as follows:

$$x(k+1)=\begin{bmatrix}1.0123 & 0.0502\\ 0.4920 & 1.0123\end{bmatrix}X(k)+\begin{bmatrix}0.0125\\ 0.5020\end{bmatrix}u\left(k-\tau_k^{ca}\right)+\begin{bmatrix}0.1\\ 0.1\end{bmatrix}r(k)$$

$$y(k)=\begin{bmatrix}1 & 0\end{bmatrix}X(k)$$

Figure 3 and 4 illustrates the output response of the SWNCS for both RMPC and NNPC techniques when WI-FI and WiMAX are employed as wireless networks respectively. Table 1 demonstrates the performance parameters of the SWNCS that obtain when employ RMPC and NNPC techniques. The simulation responses in Figures 3 and 4 demonstrate the effectiveness of the proposed controller at the desired control performance, especially for SWNCS with Wi-Fi/ WiMAX wireless networks. The simulation results show that the proposed method is capable of controlling SWNCS with satisfactory tracking performance under stochastic wireless network delay.

CONCLUSION

In this paper we have investigated two methods of simulating delays and packet dropout occurs in wireless network. In RMPC the time delays of the SWNCS are considered as stochastic variables controlled by a Markov chain. A discrete-time Markovian jump system with norm unbounded time delay is presented to model the SWNCSs. Based on the SWNCS model, the RMPC based on full state feedback controller can be solved under the framework of LMIs. Other techniques presents methodology that employs NNPC to remove or reduce the effect of time delay in SWNCS, and to employ the value modified by BPNN error predictive model to manage the SWNCS time delays, The NNPC model is used to recursively calculate the output predictions and control signal predictions. To analyse the performance of the proposed scheme, a numerical simulation has been carried out. The simulation results have successfully illustrated the effectiveness of the proposed approach.

Table 1: Performance Parameters for SWNCS

Performance parameters	RMPC Technique		NNPC Technique	
	Wi-Fi	WiMAX	Wi-Fi	WiMAX
Maximum Over Shoot	0.32	0.45	0.49	0.76
Settling Time (sec.)	0.24	0.33	0.52	0.66
Steady State Error	$4x10^{-5}$	$4x10^{-5}$	$6x10^{-5}$	$6.5x10^{-5}$

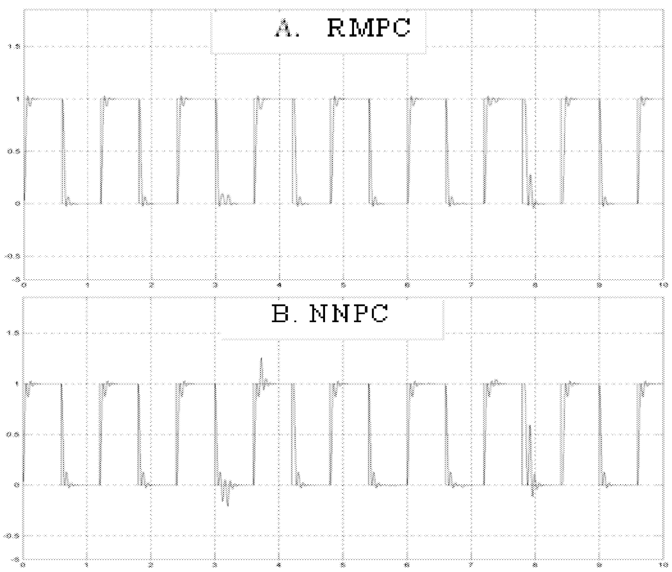


Figure 3: Output Response for SWNCS When Employ Wi-Fi

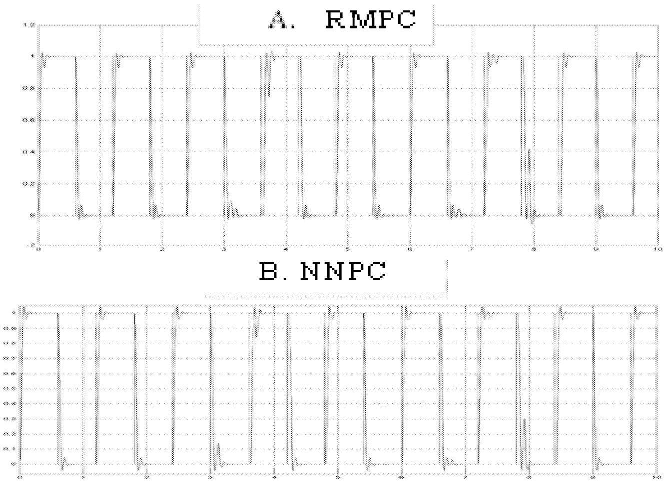


Figure 4: Output Response for SWNCS When Employ WiMAX

REFERENCES

[1] J. P. Hespanha. 2007. "*A Survey of Recent Results in Networked Control Systems*." United States, pp.2352-2357.

[2] W.Zhang, M.S Branickly and S.Phillips. 2001. "*Stability of Networked Control Systems*". IEEE control systems magazine, vol.21, pp.84-99.

[3] P. Kawka and A. Alleyene. 2005. "*Stability and Feedback Control of Wireless Networked systems*". American Control Conference, Portland, OR, USA.

[4] C. Lu, C. Tsai, and C. Liu. 2010. "*Neural-Network-Based Predictive Controller Design: An Application to Temperature Control of A Plastic Injection Modelling Process"*. Asian Journal of Control, Vol. 12, No. 6, pp. 1 12.

[5] N. Sadeghzadeh and A. Afshar. 2008. "*An MLP Neural Network for Time Delay Prediction in Networked Control Systems*". Chinese Control and Decision Conference, IEEE.

[6] A. Antunes and F. Dias. 2008. "*A neural network delay compensator for networked control systems".* IEEE Transcation of auomatic control.

[7] Wencai Du and Feng Du.2009. " *New Smith Predictor and Single Neural Adaptive Control for Wireless Networked Control Systems*." International Conference on Control, Automation and Systems Engineering.

[8] F.Du, W.Du. 2009." *Wireless Networked Control Systems with Generalized Predictive Control and Novel Smith Predictor".* International Conference on Control, Automation and Systems Engineering.

[9] S.Coman and V. Comnac. 2009." *Networked Control Systems for DC Electrical Drives".* Scientific Bulletin of the Petru Maior Universityof Tirgu Mures, Vol.5, pp61-64.

[10] Waleed Al-Azzawi, Marwan Al-Akaidi. 2011. "*Robust Model Predictive Controller for Solar Wireless Network Control System".* submitted to IET control

application and theory Journal, the paper is under review.

[11] Waleed Al-Azzawi, Marwan Al-Akaidi. 2011. "*Solar Wireless Network System Based on Neural Network Predictive Control"*. IEEE Transaction on Automatic Control, the paper is under review.

[12] http://www.doc.ic.ac.uk/~nd/surpise_96/journal/vol4/CS11/report.htm#theback-propagationalgorithm.

[13] Jianqiang Yi and Qian Wang. 2007. "*BP neural network prediction-based variable-period sampling approach for networked control systems"*. Applied Mathematics and Computation, vol185, 976–988.

[14] Menghong Yu, Yang Xu. 2008. *" Design of Predictive Control Compensator for Network Delay in Networked Control"*. IEEE Pacific-Asia Workshop on Computational Intelligence and Industrial Application

INTER-FRAME RETRANSMISSION FOR VIDEO SURVEILLANCE OVER WIMAX

Suherman
Faculty of Technology
De Montfort University
Leicester, UK
Email: suherman@usu.ac.id

Marwan Al-Akaidi
Arab Open University
P.O.Box: 3322 Safat
13033 Kuwait
Email:m.al-akaidi@arabou.edu.kw

Raouf Hamzaoui
Faculty of Technology
De Montfort University
Leicester, UK
Email: rhamzaoui@dmu.ac.uk

KEYWORDS
Transport layer protocol, inter-frame retransmission, video surveillance, WiMAX.

ABSTRACT

Video surveillance is an important application for activity and security monitoring. Surveillance application can take advantage of wireless infrastructure which provides installation flexibility and terminal mobility. However, wireless video transmission is prone to interferences which degrade video quality. This paper proposes an inter-frame retransmission protocol for video surveillance over WiMAX. The protocol reduces packet and frame delay compared to existing protocols.

INTRODUCTION

Worldwide Inter-operability for Microwave Access (WiMAX) technology that offers high bandwidth connectivity and user mobility is a potential network infrastructure for video surveillance application. The surveillance nodes may reach distance as WiMAX is able to cover up to 50 km (Scalabrino et al. 2007). Moreover, the mobility feature of WiMAX enables video surveillance to be attached on moving objects such as public transportation. This paper proposes a transport layer protocol to support the surveillance application over WiMAX.

Transmission Control Protocol (TCP) provides high reliability data transfer, ensuring that each frame is received successfully and sequentially. However, TCP is not suitable for real-time video transmission as wireless interferences and signal disruption may cause significant delay. User Datagram Protocol (UDP) is the most common transport protocol for real-time video transmission over wireless networks (Chughtai et al. 2009). However, UDP does not respond to network conditions which can cause network congestion (Wong et al. 2005).

In order to gain maximum performance for the intended application, the transport protocol should be able to deliver video with sufficient quality as well as maintain low delay connectivity. Many works have been done to improve transport layer protocol performance, whether employing retransmission or congestion control services. The details are reserved in related works section. Since delay is crucial parameter in real time video transmission, video frames should be received in order to avoid delay. The work focuses on how to reduce packet loss by retransmitting dropped packets within one frame before sending the next frame. The NS-2 simulations show that the proposed method is able to reduce packet loss without producing significant delay.

RELATED WORKS

Many works have proposed improvements on the transport layer protocol. RUDP (Reliable UDP), RBUDP (Reliable Blast UDP), UDT (UDP-based Data Transfer) and BVS (Broadband Video Streaming) improve the existing protocol performance by using retransmission. UDP-lite and DCCP (Datagram Congestion Control Protocol) do not retransmit lost packets.

RUDP uses acknowledgement (ACK) as in TCP and provides a congestion control mechanism (Bova and Krivoruchka 1999). However, since RUDP employs almost all features in TCP, RUDP may produce excessive delay as TCP (Tuong et al. 2009). RBUDP (He et al. 2002) and UDT (Gu and Grossman 2007) are datagram protocols that work for high speed bulk data transfer. Both protocols were aimed to solve TCP weakness which underutilize high speed network (Gu and Grossman 2007). RBUDP and UDT employ negative acknowledgement based retransmission. RBUDP waits an additional “DONE” packet before NACK is sent, while UDT uses periodical NACK packets to request retransmission. Such methods work well in high speed networks but not in competing networks. BVS is a semi reliable protocol which applies retransmission when the prioritized packet is lost (Ali et al. 2011). BVS uses a NACK packet to request the sender retransmits lost packets. Frequent prioritized packet loss in BVS results irregular retransmission.

UDP-Lite (Larzon et al. 1999) implements partial checksum for sensitive part of the packets and ignores error in non-sensitive part. However, passing error packets to application layer limits network observation capabilities (Welzl 2005). DCCP improves unreliable connection by providing congestion control mechanisms (Kohler et al. 2006). Two congestion control mechanisms were proposed: TCP-Like and TFRC-Like. DCCP employs acknowledgement service without retransmission which means it does not recover the lost packets.

Many protocols are designed for specific applications such BTP, Bidirectional Transport Protocol (Wirz et al. 2009) and ERT, Embedded Reliable Transport protocol (Wei and Chao 2010). But, only few that are designed for real time video transmission.

THE PROPOSED PROTOCOL

The proposed protocol is intended for video surveillance applications over WiMAX networks. The designed protocol is called as inter-frame retransmission protocol. It employs negative acknowledgements, inter-frame retransmission scheduling and congestion delay. The details are described in the remainder of this section.

Negative Acknowledgement

The proposed protocol uses negative acknowledgement (NACK) to inform the sender that packet loss has occurred and packet should be retransmitted. The NACK packet contains either a list of indices of lost packets or the start and end indices of the lost packets. As soon as the sender receives a NACK packet, it resends the requested packets (Figure 1).

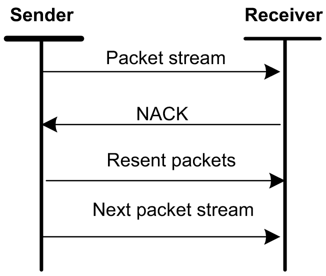


Figure 1: Negative acknowledgement

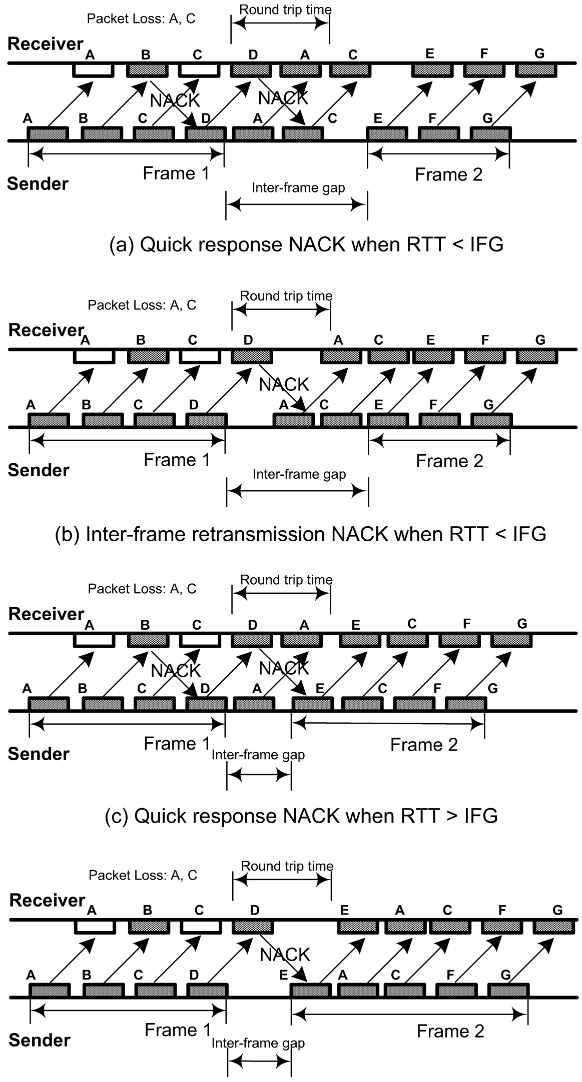
In RBUDP, NACK is used but delivered over a different connection (He et al. 2002). Moreover, RBUDP requires an additional "DONE" packet before sending a NACK. NACK is also implemented in BVS (Ali et al. 2011) which uses quick response scheduling as discussed in the next section.

Retransmission Scheduling

The NACK packet is sent by the receiver when packet loss is detected. The sending time is set according to the scheduling type. There are two NACK scheduling types, Quick Response (QR) and Inter-frame Retransmission (IR). QR requires the receiver to send a NACK packet as soon as packet loss is detected. The packet loss information is determined by two values, the current and previous successfully received packet indices. The sender will check these values to decide which packet to retransmit. For instance, if the current packet index is 7 and the previous one is 4, then packets with indices 5 and 6 should be retransmitted. The advantage of QR is small NACK overhead and responding loss quickly. However, the receiver may generate more than one NACK packet for a frame, which requires more bandwidth and interrupts the sender frequently.

The second scheduling strategy is called inter-frame retransmission (IR). Instead of sending a NACK packet for every detected lost packet, the receiver records indices of packet lost within one frame and sends a NACK packet after receiving the last packet in that frame. If no packet is lost in one video frame, then no NACK packet will be sent. The advantage of IR is that a NACK packet will be sent only once for all lost within one video frame. IR generates fewer NACK packets than QR.

Figure 2 shows the NACK scheduling. One video frame may be sent in several packets. The time distance between the last packet in one frame and the first packet of the next frame is called inter-frame gap (IFG). Figure 2a and 2b assume that the round trip time (RTT) is less than IFG. Packet A and C within frame 1 are lost. In QR, NACK packets will be sent as soon as the receiver receives packets B and D. NACK packets may interrupt the sender frequently and may cause additional delay or another packet loss. On the other hand, IR sends NACK and resends packets during inter-frame gap when the sender is idle.



(a) Quick response NACK when RTT < IFG

(b) Inter-frame retransmission NACK when RTT < IFG

(c) Quick response NACK when RTT > IFG

(d) Inter-frame retransmission NACK when RTT > IFG

Figure 2: NACK scheduling

If RTT is greater than IFG as shown in Figure 2c and 2d, IR interrupts the sender only once. Although IR seems causing the next frame sending time longer, we found that the sender processing time is more sensitive to NACK reception than to packet retransmission. IR has additional requirement that the receiver should be able to detect the last packet in each frame. In case the last packet within a frame is lost, the lost packet will be retransmitted within the next frame.

Prioritized Packets

Unlike TCP which sends an acknowledgement for every received packet, the proposed protocol sends NACK packets only when packet loss occurs. However, if network congestion worsens, NACK packets may be sent more frequently as more loss appears. The frequent packet retransmissions may lead to high delay. To keep delay low, the NACK packet for a particular packet loss will be sent only once. The dropped retransmitted packet will be ignored.

Furthermore, the NACK packet reduction may be applied by sending NACK only for prioritized packets as video coding results non uniform frame significances. An additional packet header is required to flag whether a packet is prioritized or not. Simulation in this paper uses MPEG4 video coding with IPP frame sequence. The prioritized packets are set to be any packets corresponding to I-frames. The conducted simulations show that priority policy significantly reduces delay.

Congestion Delay

Congestion delay (CD) aims at reducing the effect of sender interruption and avoiding another packet loss by postponing the next packet transmission. Congestion delay also makes sure that the current frame arrives before the next frame.

Figure 3a shows retransmission without congestion delay. The sender sends packet E before retransmitting the lost packet C. Packet C which belongs to previous frame may arrive after packet E which belongs to the next frame. This situation results in higher frame delay. In the worst case, packet E can be lost during reception of a NACK packet. By using congestion delay, packet C will be retransmitted before packet E as shown in Figure 3b. This process results in lower delay on packet C and avoids loss of packet E. Although congestion delay introduces more delay for packet E, a small congestion delay value limits this additional delay.

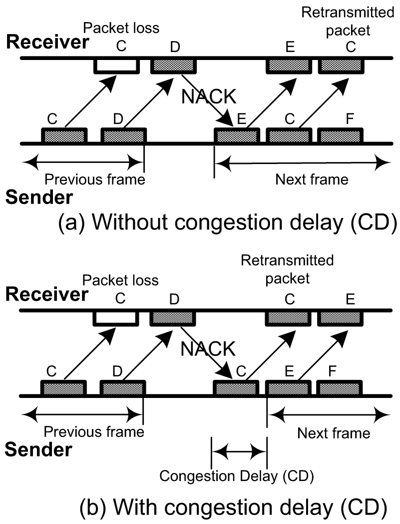


Figure 3: Congestion delay

Congestion delay also acts as instant congestion control by delaying next packet transmission in response to network congestion. Congestion delay produces temporary frame rate reduction as given by Equation 1.

$$FR_{CD} = \frac{1}{1/_{FR_{init}} + CD} \qquad (1)$$

For instance, if the initial frame rate (FR_{init}) is 25fps, and congestion delay (CD) is 0.01s, then the frame rate caused by congestion delay (FR_{CD}) is 20fps. This rate reduction gives the network time to reduce congestion, which potentially reduces packet loss. The congestion delay value should be smaller than inter-frame gap to avoid current frame competing with the next two frames.

SIMULATION ENVIRONMENT

WiMAX System

In order to evaluate the proposed protocol for video surveillance application over WiMAX, we conducted WiMAX simulation using the NS-2 simulator with the NIST WiMAX adds on module (National Institute of Standards and Technology 2007). The WiMAX transmit power and receiver threshold are set to provide 1000 m coverage radius. The modulation is 64 QAM, with two ray ground propagation model. By using this propagation model, measurement is in ideal line of sight path. The cyclic prefix is set to 0, which means no repeating frame preamble to avoid fading. The total cell bit rate is 10 Mbps, 7 Mbps are allocated for video traffics in uplink stream and 3 Mbps downlink are intended for negative acknowledgement services.

The simulated surveillance application has 4 mobile nodes (MN) within one base station. Each node has a different speed to represent some possible surveillance positions. Node 0 is fixed. Node 1 is set to a walking speed, 1.39 m/s. Node 2 and Node 3 are assumed to be attached in vehicles such as a bus or tram. Node 2 moves at 4.44 m/s and Node 3 speed is 6.67 m/s. The hand off process is not presented in this simulation. The single cell network configuration is shown in Figure 4.

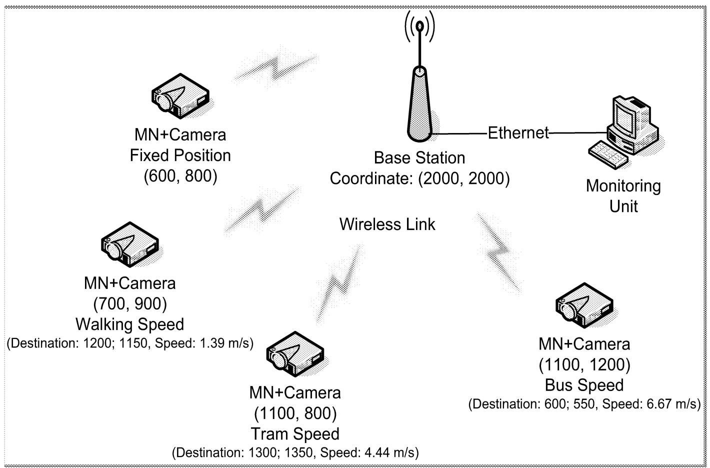


Figure 4: Network configuration

The Observed Protocols

We compared the proposed protocol to UDP, TCP, BVS, DCCP and RBUDP.

Traffic Source

The traffic source uses a video trace which contains a list of packet sequence number, byte length, frame types, and time stamp from real video source, akiyo.yuv with Common Intermediate Format (CIF) resolution 352 x 288. This video trace is used as simulated traffic in simulation, where the received pattern is reconstructed from the received packets based on the original video. The traffic generation and reconstruction in the NS-2 simulator use the Evalvid video evaluation framework (Klaue et al. 2003; Ke et al. 2008). Table 1 shows the simulation parameters.

Table 1: Simulation parameters

Parameter	Value
Sequence	akiyo.yuv
Frame rate	30fps
Frame type	IPP
Video codec	MPEG4
Video bit rate	559.35 Kbps
Packet size	1024 bytes
Group of Pictures	30 frames

Performance Metrics

The performance evaluation was conducted by observing sending and receiving ports of each connection and noting the required values such as packet sequences, sending and receiving times, packet types: acknowledgement packets, and data size. The measurement in NS-2 follows those in (Ke et al. 2008).

The presented performance metrics were obtained as the averages of all nodes. The metrics are:

- Packet delay: one way delay, obtained by subtracting the sending time from the receiving time.
- Frame delay: the latest receiving time of the packets within one frame, subtracted by the frame time stamp.
- Jitter: the absolute value of subsequent delay differences.
- Fluidity: the frame distance, obtained from the difference of current and next frame' receive time.
- Packet loss: number of lost packets divided by the total transmitted packets (in percentage).
- Cumulative throughput: the total number of received bits.
- PSNR: peak signal to noise ratio, obtained by comparing the reconstructed video from the received packets and the original video source.

RESULT AND ANALYSIS

The Protocol Performance

We compared quick response and inter-frame retransmission scheduling. Figure 5 shows the results. Inter-frame scheduling generates lower packet delay and jitter, less packet loss, closer fluidity to the original video, higher cumulative throughput and PSNR than quick response scheduling. Although its frame delay is slightly higher than quick response scheduling, the overall performance of IR scheduling is better than QR scheduling.

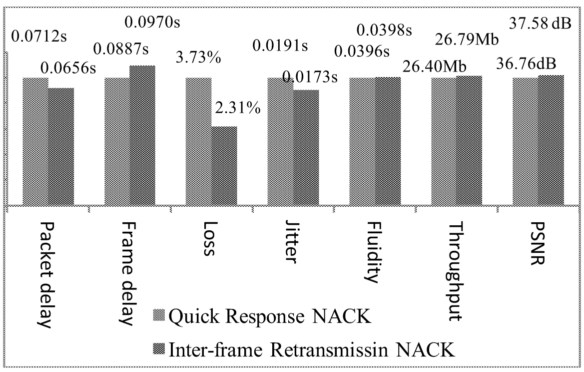


Figure 5: Comparison of QR and IR performance

By applying priority policy to IR scheduling (that is sending NACK packets only if lost packets are parts of the prioritized frames), the protocol is able to reduce packet and frame delays significantly (about 10ms and 32ms in average). Figure 6 shows the comparison of IR scheduling and prioritized IR scheduling. The prioritized one suffers higher packet loss which reduces throughput and video quality. However, in real time video transmission, low packet and frame delays are more important.

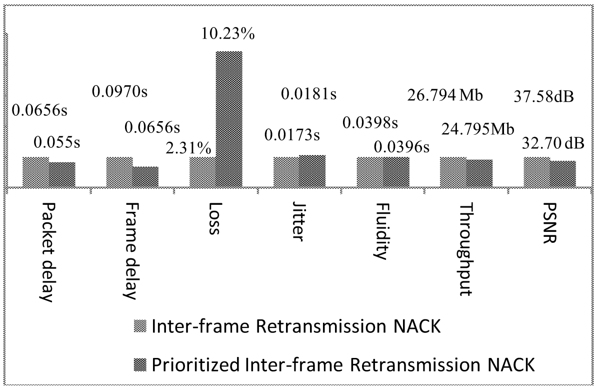


Figure 6: Comparison of IR and prioritized IR performance

The delay parameters gained by prioritized scheduling as shown in Figure 6 should be suppressed further to produce better characteristics for video transmission purpose. As described in Figure 3, congestion delay is expected to achieve the expected performance. Congestion delay should be less than the frame distance which means higher than 0 and lower than 0.004s. The smaller the value, the less the effects to next packet delay. We have tested various CD values as shown in Figure 7. The delay characteristics are relatively constant when CD values are less than 0.001s. However, they change alternately afterwards. In average, congestion delay successfully reduces delay of prioritized IR scheduling.

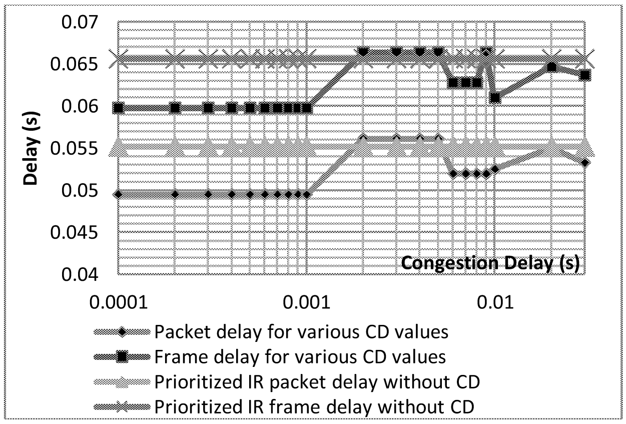


Figure 7: Congestion delay performances

Figure 8 depicts the performance enhancement of the proposed protocol by applying a 0.001s congestion delay. The average packet and frame delays plunge to 0.0495s and 0.0597s respectively. The average jitter is also reduced to 0.0169s. Even if packet loss increased causing a decrease in the cumulative throughput, the congestion delay preserved prioritized frames better. This is shown by the increase of the PSNR, which means that the protocol successfully avoids more loss on prioritized frames and produces better video quality.

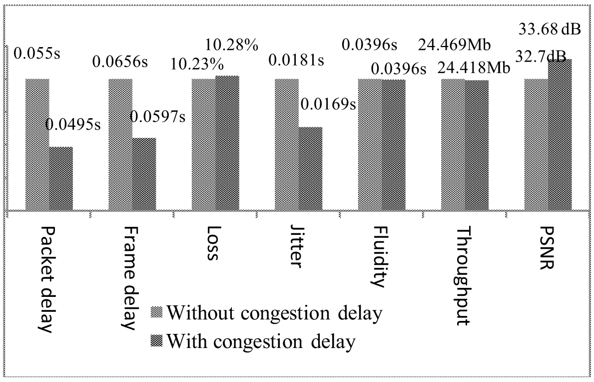
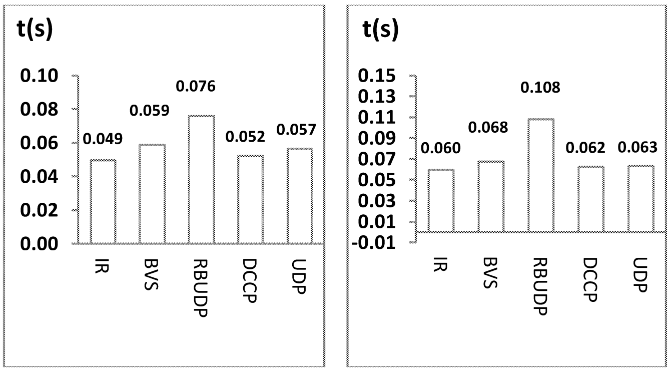


Figure 8: Performance comparison of prioritized IR without and with congestion delay

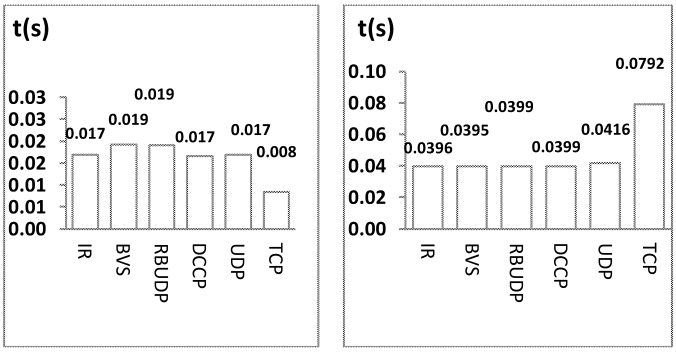
Performance Comparisons

Figure 9 shows the delay characteristics of the examined protocols. Unlike UDP, BVS and RBUDP, the proposed protocol reacts to network congestion by postponing the next packet transmission. This response helps the proposed protocol to reduce network queue and suppress end to end delay. On the other hand, although DCCP and TCP implement congestion control to deal with congestion, these protocols require certain observation periods before reducing or increasing transmission rate. DCCP requires feedback packet containing receiver observation, while TCP implements time out before detecting network congestion. By arranging retransmission time and quickly responding to packet loss, the proposed protocol successfully reduces packet and frame delay. TCP experiences significant packet and frame delays, 0.468 s and 6.4 s respectively (which are not shown in Figure 6).



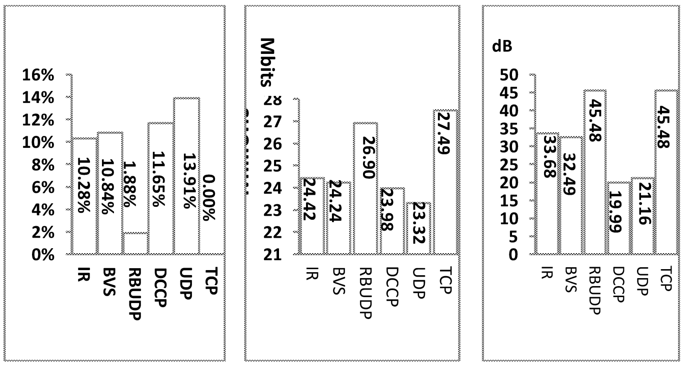
(a) Average packet delay (b) Average Frame delay
Figure 9

Congestion control is proven effective to avoid rough delay caused by congestion. As shown in Figure 10, TCP and DCCP produced low jitter. On the other hand, BVS and RBUDP failed to gain minimum jitter as these protocols inject retransmission traffics without dealing with congestion problem. Although the proposed protocol also streams additional retransmitted packets, congestion delay which deal with congestion is able to hold jitter as low as UDP. TCP yields the worse fluidity as it experiences high packet delay. Other protocols produce almost similar fluidity which shows that the received frames flow is similar.



(a) Average jitter (b) Average fluidity
Figure 10

In comparison to UDP, DCCP and BVS, IR reduces packet loss significantly. The loss is 3.5% lower than UDP, 1.37% lower than DCCP and 0.56% lower than BVS. Therefore the proposed protocol has higher throughput than those protocols. Furthermore, IR is able to preserve priority packets better than BVS which also retransmits priority packets. Consequently, IR produces better video quality as shown in Figure 11c. Although IR has higher packet loss and lower PSNR value than TCP and RBUDP, its low delay characteristics are more desirable for real-time video transmission.



(a) Packet loss (b) Cumulative throughput (c) PSNR
Figure 11

CONCLUSION AND FUTURE WORK

We have proposed an inter-frame retransmission (IR) protocol to reduce packet loss in video surveillance over WiMAX. The prioritized inter-frame scheduling with congestion delay method is able to make the proposed protocol perform better than existing protocols such as BVS, DCCP and UDP. Packet and frame delays as well as packet loss are reduced significantly. The protocol is also able to preserve prioritized frames so that video quality can be maintained.

Since mobile nodes in video surveillance move dynamically, congestion delay should also be dynamically analyzed and updated to enhance performance. Future work may optimize congestion delay values in response to various network conditions. Protocol deployment in other network may lead to different results as the implemented methods within the proposed protocol are optimized only for surveillance scenario over WiMAX.

ACKNOWLEDGEMENTS

This work has been supported by Directorate General of Higher Education (DGHE or DIKTI), Ministry of National Education, Indonesia.

REFERENCES

Ali I.; S. Al-Majeed; M. Fleury; and M. Ghanbari. 2011. "Semi-reliable transport protocol for IPTV over mobile WiMAX", in Proceeding of Computer as a Tool Conference (Lisbon, Portugal, April 27-29), IEEE R8, Portugal, 1-4.

Bova T and T. Krivoruchka. 1999. “Reliable UDP Protocol”, Internet Engineering Task Force (Network Working Group) draft, Cisco Systems, Herdon, VA (Feb).

Chughtai, H. M. O; S. A. Malik; and M. Yousaf. 2009. "Performance evaluation of transport layer protocols for video traffic over WiMAX", in Proceeding of the 3th Multitopic Conference (Islamabad, Pakistan, Dec 14-15), IEEE, Piscataway, NJ, 1-6.

Gu Y. and R. L. Grossman. 2007. "UDT: UDP-based Data Transfer for High-Speed Wide Area Networks", Journal of Computer Networks, Elsevier, vol. 51, issue 7 (May), 1777-1799.

He E.; J. Leigh; O. Yu; and T. A. Defanti. 2002. "Reliable Blast UDP: predictable high performance bulk data transfer", in Proceeding of Cluster Computing Conference (Chicago, Illinois, Sept.), IEEE, Los Alamitos, CA, 317-324.

Ke C.-H.; C.-K. Shieh; W.-S. Hwang; and A. Ziviani. 2008. “An Evaluation Framework for More Realistic Simulations of MPEG Video Transmission”, Journal of Information Science and Engineering, vol. 24, no. 2, (March), 425-440.

Klaue J.; B. Rathke; and A. Wolisz. 2003. "EvalVid, - A Framework for Video Transmission and Quality Evaluation", in Proceeding of the Modeling Techniques and Tools for Computer Performance Evaluation Conference, (Urbana, Illinois, Sept. 2-5), Springer, 255-272.

Kohler E.; M. Handley; S. Floyd; and J. Padhye. 2006. “Datagram congestion control protocol (DCCP),” Request for Comment (RFC) 4340, IETF, (March).

Larzon L; M. Degermark; and S. Pink. 1999. "UDP Lite for Real-Time Multimedia Applications", in Proceeding of the International Conference of Communications (Vancouver, Canada, Jun 6-19), IEEE, Piscataway, NJ, 1999.

National Institute of Standards and Technology. 2007. “The Network Simulator ns-2 add-on - IEEE 802.16 model (MAC+PHY)”, Draft 1.2.1, (Jun).

Scalabrino, N.; F. D. Pellegrini; R. Riggio; A. Maestrini; C. Costa; and I. Chlamtac. 2007. "Measuring the quality of VoIP traffic on a WiMAX testbed", in Proceedings of the 3rd Testbeds and Research Infrastructure for the Development of Networks and Communities Conference (Lake Buena Vista, FL, May 21–23), IEEE, Piscataway, NJ, 1-10.

Tuong L; G. Kuthethoor; C. Hansupichon; P. Sesha; J. Strohm; G. Hadynski; D. Kiwior; and D. Parker. 2009. "Reliable User Datagram Protocol for airborne network," in Proceeding of Military Communications conference (Boston, MA, Oct. 18-21), IEEE, Piscataway, NJ, 1-6.

Wei Z. H. and D W Chao. 2010. "Research of Embedded Reliable Transmission Protocol", in Proceeding of Information Engineering Conference (Heibei, China, Aug 14-15), IEEE, Piscataway, NJ, 272-275.

Welzl M. 2005. “Passing Corrupt Data Across Network Layers: An Overview of Recent Developments and Issues,” in EURASIP Journal on Applied Signal Processing, (Jan), 242–247.

Wirz R.; R. Marin; M. Ferre; J. Barrio; J. M. Claver; and J. Ortego. 2009. "Bidirectional Transport Protocol for Teleoperated Robots", in IEEE Transactions on Industrial Electronics, vol. 56, no. 9, (Sept.), 3772-3781.

Wong C. F.; W. L. Fung; C. F. Tang; and S. H. G. Chan. 2005. "TCP Streaming for Low-Delay Wireless Video", in Proceeding of 2nd Quality of Service in Heterogeneous Wired/Wireless Networks Conference (Orlando, FL, Aug 22-24), IEEE, Los Alamitos, CA, 6.-41.

NUMERICAL MODELING OF LOAD BALANCED MULTICORE SERVERS IN A CLOUD CLUSTER[1]

Vasil Georgiev*, Jouni Karvo**

*Faculty on Mathematics and Informatics, University of Sofia "St. Kliment Ochridski"
1164 Sofia, Bulgaria, v.georgiev@fmi.uni-sofia.bg
**HUS ICT, Helsinki and Uusimaa Hospital District
P. O. Box 750 FIN-00029 HUS, Finland, jouni.karvo@hus.fi

KEYWORDS

Cloud systems, load balancing, Markov chains, numerical iteration, multiprocessing

ABSTRACT

Our paper presents a method for numerical solution of Markov-chain based models of dynamic load balancing schemes for cloud clusters. These schemes were presented in detail in an earlier publication (Georgiev, V. 2010.). Here we are describing one of the threshold-based p2p cases of load balancing along with the fast converging iterative solution of its analytical model. An important feature of described solution is the aggregation of the local load states of cloud servers into a few super-states. In our experience state aggregation approach proves very useful in simplifying the solution itself and its further calculation by simple electronic tables. This approach is illustrated by a case study of the parallelism of the cloud servers. The cluster performance is compared for a different number of server cores varying the rest of system parameters.

LOAD BALANCING MODEL SPACE

The recent advances in service-centric architectures lead to the adoption of the cloud paradigm for distributed server infrastructure. Non-functional subsystems are of great importance for keeping high level QoS in clouds. Their tasks are to improve application performance, ensure application availability, and implement a strategic disaster recovery plan (Triebes K. 2010). Load balancing is helpful in spreading the load equally across the free nodes when a node is loaded above its threshold level.

The whole balancing process is split in three related but independent subprocesses, traditionally called ***control strategies***. These are the ***monitoring***, ***location*** and ***transfer*** strategies which are in a certain relation – the output of each strategy is a precondition or input for the next one in this order (i.e. M $\rightarrow$ L$\rightarrow$ T). Their independence, however, is revealed in the fact that each of the three strategies might be based on different principles (Gong, Z.; P.Ramaswamy; X. Gu; and X. Ma. 2009). One important specific feature of the monitoring strategy is the ***level of precision*** of collected system information. As an example, in load balancing a monitor (and consequently the locator[s] the decisions of which are based on this monitor) may only distinguish two states of each cluster node (or service): underloaded and overloaded – having one threshold parameter as a boundary between these two states (Eager, D., E. Lazovska, J. Zahorian. 1986.) As other examples with less precision would be just the awareness if a given node is in operation or down. A more precise level of information would distinguish three states adding a neutral state between the under- and overloaded ones and having two thresholds. Obviously, the level of precision can go further allowing more precise allocation decisions for the price of incurring bigger system overload.

While the physical connectivity in cloud clusters may allow system-wide broadcasting, the logical connectivity can be limited for performance or security reasons. Thus the locality, the monitoring and decision scope of each node is limited to a degree of neighborhood. We thus consider centralized (cluster-wide) and distributed (p2p) load control schemes where each service has a limited number v of neighboring peers and and each service performs the monitoring and task transfer functions in this limited scope.

This paper presents a fast converging iterative solution for Markov-chain based models for performance evaluation of the whole range of load balancing schemes. For this purpose we consider the system performance measured by the mean service time of the individual tasks by varying only the workload conditions of the cloud cluster for different types of server architecture.

BASIC MODEL FOR LOAD BALANCING

We consider a cloud cluster unifying several (or many) service multicore nodes connected by a broadcasting network. The workload of this cluster consists of a stream of independent tasks which can be executed by any core and node. The ***load balancing task*** is to achieve shorter service time of the cloud for the individual tasks by eliminating or reducing occurrences of temporarily idle nodes/cores while there are tasks waiting in other nodes. In p2p models tasks arrive randomly at any of the nodes (Cybenko G., 1994). In Client-Server (i.e. centralized) models all the tasks arrive in a dedicated balancing server which is performing the load balancing process. Further in this paper we will present the case analysis of a ***distributed load balancing*** scheme and will compare its performance to that of the centralized load balancing, which is common for the cloud clusters (Luis M. Vaquero, Luis Rodero-Merino and Rajkumar Buyya. 2011).

The ***input modeling parameters*** are as follows:

- cluster size: n nodes with c cores each; cloud homogeneity does not exclude the possibility of having nodes that differ in performance because of the different number of cores or because of different external load condition
- task arrival process: global (i.e. cluster-wide) rate Λ and local rate at the nodes λ

[1] This research is part of the work on projects ДДВУ02/22/2010 & 67/2011 of the University of Sofia

- task execution/departure process: local rate of each core ω – i.e. the distribution of the execution time of the consecutive tasks
- global load condition: the ratio between task arrival rate and task and departure rate Φ
- local load condition: the number l of tasks being executed in the node
- locality of the balancing strategies: the number of neighboring peers v and the graph of their logical connectivity (or the adjacency matrix)
- monitoring precision level: the number of distinguishable load states and thresholds values.

For simplicity, we consider three distinguishable load states of the nodes – receiver ($\mathcal{R}$), neutral ($\mathcal{N}$) and sender ($\mathcal{S}$) of tasks – and thus two thresholds: T_1 is the upper boundary of the underloaded (receiver) state and T_2 is the lower boundary of the overloaded (sender) state. In order to define current load state, l is to be compared to $T_{1|2}$. Normally, $c \leq T_1 \leq T_2$.

The ***internal modeling parameters*** (***intermediate results***) are:

- local rate of receiving tasks from overloaded nodes ρ
- local rate of sending tasks to the receiving nodes σ
- mean number of tasks in a node £
- the probabilities of a node to be in the load state of sender, receiver or neutral $P_{\mathcal{S}}$, $P_{\mathcal{R}}$, and $P_{\mathcal{N}}$

The ***output parameters (results)*** of the evaluation of our model are:

- mean service time in cloud T
- mean rate of task transfer messages θ for each node
- mean rate of monitoring messages exchange process ε for a node
- mean idle time of nodes (including time of idle cores)
- optimal system size i.e. the bound of the speedup linearity/scalability

The load balancing modeling and analysis aims at a better performance of the cloud infrastructure. From the user viewpoint this means shorter service time; from the viewpoint of a service provider (which in clouds includes infrastructure-providers) effectiveness and resource utilization are a concern.

ANALYTICAL MODEL OF CLOUD SERVERS ARCHITECTURE

Let us consider the case of a homogeneous cloud cluster consisting of n nodes with c working cores each and connected with a high-speed broadcasting network. Being independent and asynchronous, the task service process is modeled as a stochastic random Poisson process with mean rate of service in a single server core ω; the task arrival process is again a Poisson random process which is balanced between all the cluster nodes in sub-processes with a varying mean rate $\lambda_{\mathcal{R}|\mathcal{N}|\mathcal{S}}$ depending on the local load status. A centralized load balancing process normally keeps only raw or no information about the local load status of all the nodes because of concerns for scalability and system overload (SwiftWater. 2010). That is why we propose to replace it with a local load balancing process which occurs between neighboring nodes and the task of which is the quick elevation of the local load imbalances that occur in the vicinity of each node because of the unpredictability of the individual service time of the tasks (Karagiorgos, G., N. Missirli and F. Tzaferis. 2004). In general each node has v ($v < n$ or even $v << n$) neighboring peer nodes to which it sends the monitoring information about any distinguishable change of the local state (e.g. on transitions between – $\mathcal{R}$, $\mathcal{S}$, etc.) and to which it sends tasks when there are any receivers among the neighbors.

Here we describe a ***Markov-chain based analysis*** of this model. The possible states of each node are presented by the number of tasks in service l that are further reduced to the three distinguishable and monitored states as l passes through the values of T_1 and T_2. The service process is described by the probabilities p_l of the node being in state l. The transitions between the node's states occur with a rate which is derived from the values λ, ω, ρ and σ. ***Figure 1.a)*** shows the state-transition-rate diagram of a random cloud cluster node for the cases. Given the input parameters we have to find the output parameters (passing thorough the intermediate results). Analyzing this case we have the following system of equations (e1) ÷ (e10) for the steady-state probabilities p_l of each node:

$$(\lambda_{\mathcal{R}}+\rho)p_i = (i+1)\omega p_{i+1}, \qquad i \in [0, c] \qquad \text{(e1)}$$

$$(\lambda_{\mathcal{R}}+\rho)p_i = c\omega p_{i+1}, \qquad i \in [c+1, T_1\text{-}1] \qquad \text{(e2)}$$

$$\lambda_{\mathcal{N}} p_i = c\omega p_{i+1}, \qquad i \in [T_1, T_2\text{-}1] \qquad \text{(e3)}$$

$$\lambda_{\mathcal{S}} p_i = (c\omega+\sigma)p_{i+1}, \qquad i \in [T_2, \infty) \qquad \text{(e4)}$$

$$\sum_{i=0}^{\infty} p_i = 1, \qquad \text{(e5)}$$

$$\rho P_{\mathcal{R}} = \sigma P_{\mathcal{S}}, \qquad \text{where} \qquad \text{(e6)}$$

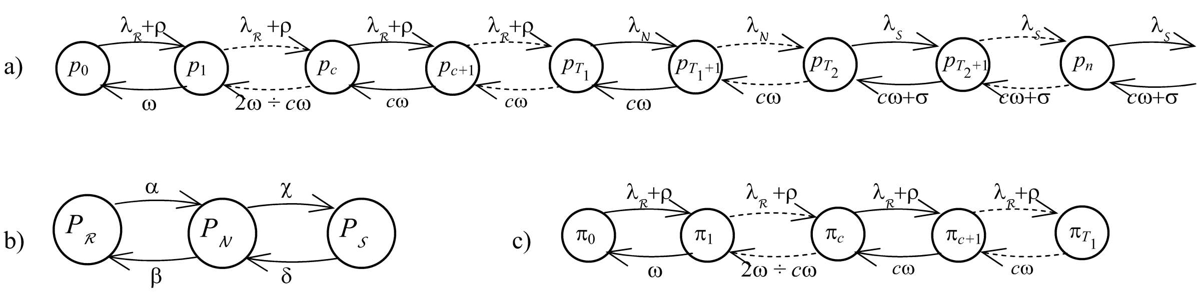


Figure 1. a) State-Transition-Rate Diagram of a Load Balanced Node; b) Same Diagram for the System with Aggregated States; c) Same Diagram for the $P_{\mathcal{R}}$ case

$$P_{\mathcal{R}} = \sum_{i=0}^{T_1-1} p_i , \quad (e7)$$

$$P_{\mathcal{S}} = \sum_{i=T_2+1}^{\infty} p_i , \quad (e8)$$

$\sigma = \lambda_{\mathcal{S}} P_{\Sigma\mathcal{R}} P_{\mathcal{S}}$, where (e9)

$$P_{\Sigma\mathcal{R}} = 1-(1-P_{\mathcal{R}})^v . \quad (e10)$$

This system is complete to p_l from (e1) ÷ (e5) which are the equilibrium conditions of the state-transition-rate diagram in Figure 1.a); (e5) is the law for complete probability. The equations (e6) and (e9) are added in order to define the intermediate rates ρ and σ. The remaining parameters are taken from the input block of the model. Particularly, (e6) with (e7) and (e8) reflect the fact that during the balancing process tasks neither enter nor leave the cloud cluster – they get only redistributed. Thus the overall rate of tasks sent is always equal to that of received tasks. The equation (e9) calculates the rate σ taking into account that the task sending occurs whenever a new task arrives with the rate $\lambda_{\mathcal{S}}$ at an $\mathcal{S}$-node and there is at least one $\mathcal{R}$-node among the neighbors the probability for which is $P_{\Sigma\mathcal{R}}$ in (e10). Here the calculation is based on v Bernoulli trials with failure probability $(1-P_{\mathcal{R}})$ for each.

Usually central balancers perform nothing or little more than random matchmaking between tasks and servers. E.g. they often implement some types of distributed hash tables (DHT) to store and retrieve information in the cluster (Luis M. Vaquero, Luis Rodero-Merino and Rajkumar Buyya. 2011.). For the case of random server assignment:

$$\lambda_{\mathcal{R}|\mathcal{N}|\mathcal{S}} = \lambda = \Lambda/n. \quad (e11)$$

Having solved the system of equations, we have to derive the rest of the intermediate and output results. The mean number of tasks in a node is

$$£ = \sum_{i=T_2+1}^{\infty} i p_i , \quad (e12)$$

The mean service time in cloud T (Little's result) is

$$T = £/\lambda_{\varepsilon}. \quad (e13)$$

The mean rate of receiving monitoring messages' exchange process ε for a node is

$$\varepsilon = (\lambda_{\mathcal{R}}+\rho)p_{T1-1} + \lambda_{\mathcal{S}} p_{T2} + c\omega p_{T1} + c\omega p_{T2}. \quad (e14)$$

The factor of idle state of nodes (incl. the state of idle cores)

$$\Theta = \sum_{i=0}^{c} \left(1-\tfrac{i}{c}\right) p_i . \quad (e15)$$

MODEL ANALYSIS USING STATE AGGREGATION

The obvious way to solve the system (e1) ÷ (e10) is to use numerical solver. Our experiments to do so with Wolfram Mathematica were, however, not successful. This polynomial system of equations was solved within minutes on a two-core desktop only for a limited number of equations e.g. 15-20 which corresponds to limitation of the node states. Increasing the precision (i.e. the number of equations) resulted in unacceptable delay of the solution (hours) and most of the times with insufficient memory error. Note that our performance analysis needs not just a few but tenths or even hundreds of individual solutions of this system of equations. In this sense even a delay of five minutes for each separate solution is hardly acceptable.

The other possibility – which we do recommend – is to solve analytically the dependencies. Here we describe a method to simplify the system solution by dividing the dependencies between system state probabilities in so called aggregated states. Looking at the system (e1) ÷ (e10) the obvious decomposition is into four parts each part corresponding to one equation group from the system (e1) ÷ (e4) respectively. Note that for example (e1) is a M/M/c/c model; (e2) is M/M/1/(T_1-c) model; (e3) – M/M/1/(T_2-T_1) and (e4) is a simple M/M/1 model. As these types of models are well known and studied, so here we only show the solution for the most complex aggregated state of them which is $P_{\mathcal{R}}$. Its definition corresponds to (e1) and (e2) and in fact can be analyzed as two separate substates. However, the analysis is not complicated even if we consider it as one aggregated state.

The idea is to find the probability distribution within each aggregated state as if it is a separate Markov chain. This is the **split phase** of the analysis. E.g. for $P_{\mathcal{R}}$ taken as a isolated aggregated state the state probilities π_i are

$$p_i = \pi_i P_{\mathcal{R}} , \qquad i \le T_1\text{-}1 \quad (e16)$$

$$\pi_i = \begin{cases} \dfrac{1}{i!}\left(\dfrac{\lambda_R+\rho}{\omega}\right)^i \pi_0 , i \le c \\ \dfrac{c^c}{c!}\left(\dfrac{\lambda_R+\rho}{\omega}\right)^i \pi_0 , c < i \le T_1 - 1 \end{cases} \quad (e17)$$

where

$$\pi_0 = \frac{1}{\displaystyle\sum_{i=0}^{c} \frac{1}{i!}\left(\frac{\lambda_R+\rho}{\omega}\right)^i + \frac{c^c}{c!}\left(\frac{\lambda_R+\rho}{\omega}\right)^c \left[\frac{\left(\frac{\lambda_R+\rho}{c\omega}\right)^{T_1-c} - 1}{\frac{\lambda_R+\rho}{c\omega} - 1} - 1\right]} \quad (e18)$$

In (e17) the sum $\sum_{i=0}^{c} \frac{1}{i!}\left(\frac{\lambda+\rho}{\omega}\right)^i$ can be evaluated in closed form using the approach of z-transform which is well explained e.g. in (Geist, R., K. Trivedi. 1983.) Here, we don't need such a derivation as the upper limit of this sum c (the number of processor cores) is a small value typically 1, 2 or 4. For these limits the sum can be coded easily in the electronic table. The distribution of the local steady state priobailities for $P_{\mathcal{N}}$ and $P_{\mathcal{S}}$ is derived in the same manner.

The next step is calculation of the **aggregation phase**. Considering the aggregated states model on **Figure 1b)** one has

$$P_{\mathcal{R}} = \frac{1}{1+\frac{\alpha}{\beta}+\frac{\alpha\chi}{\beta\delta}} , \quad P_{\mathcal{N}} = \frac{\frac{\alpha}{\beta}}{1+\frac{\alpha}{\beta}+\frac{\alpha\chi}{\beta\delta}} \quad (e19, e20)$$

$$P_{\mathcal{S}} = \frac{\frac{\alpha\chi}{\beta\delta}}{1+\frac{\alpha}{\beta}+\frac{\alpha\chi}{\beta\delta}} \quad (e21)$$

where

$\alpha = \pi_{T_1-1}(\lambda_{\mathcal{R}}+\theta)$ (from **Figure 1c**), $\beta = \pi_{T_1} c\omega$, (e22)

$\chi = \pi_{T_2} \lambda_{\mathcal{S}}$, (e23)

$\delta = \pi_{T_2+1} (\lambda_{\mathcal{S}} + \theta)$. (e24)

After this analysis one can proceed further with calculations using the ***iteration procedure*** as defined in the following:

Step 0. Input values: n, c, Λ, $T_{1,2}$, Φ, step $k := 1$;
Derivates: $\lambda_{\mathcal{R}|\mathcal{N}|\mathcal{S}} = \lambda := \Lambda/n$, $\omega := \Lambda/(nc\Phi)$;
Approximations: $\rho^{[0]} := c\omega\Phi - \lambda$, $\sigma^{[0]} := \lambda/\Phi - c\omega$;
go to Step k.

Step k. Solve (e1) ÷ (e5) for $p_0 \div p_{T2}$ using $\sigma^{[k-1]}$ and $\rho^{[k-1]}$;
Calculate $\sigma^{[k]}$ and $\rho^{[k]}$ using (e6) and (e9);
Calculate $\lambda_{\mathcal{R}|\mathcal{N}|\mathcal{S}}^{[k]}$ according global balancing model; for random server assignment $\lambda_{\mathcal{R}|\mathcal{N}|\mathcal{S}}^{[k]} = \lambda, \forall k$);
If ($|\sigma^{[k]} - \sigma^{[k-1]}| < 1\%$ and $|\rho^{[k]} - \rho^{[k-1]}| < 1\%$) then goto End;
else k++ and go to Step k.

End.

The ***iteration procedure implementation*** requires coding of (e17) ÷ (e24) as a row of an electronic table in the order:

1) Calculate π_{T_1-1}, using (e17), (e18) as well as π_{T_1}, π_{T_2} and π_{T_2+1} using the analogous derivations from the submodels for $P_{\mathcal{N}}$ and $P_{\mathcal{S}}$. Note that for the first iteration we need some estimative values $\rho^{[0]}$ and $\sigma^{[0]}$; which we derive from the input model parameters according to the Step 0. above;
2) Calculate α, β, χ and δ using (e21) ÷ (e24);
3) Calculate $P_{\mathcal{R}}$, $P_{\mathcal{N}}$ and $P_{\mathcal{S}}$ using (e19) ÷ (e21);
4) Calculate $\sigma^{[\text{next iteration}]}$ using (e9) and (e10) and $\rho^{[\text{next iteration}]}$ using (e6);
5) Calculate the model output using (e12) ÷ (e16);
6) Repeat 1) ÷ 5) until the end condition is reached.

On ***Figure 2.*** the implementation steps 1) ÷ 5) take each of the numerical table rows (part of the intemediate results are hidden). For step 6) we repeat the iteration (i.e. simply copy that first row) as many times as we need – practically 10 rows are more than enough. By our numerical experiments this simple procedure puts $\sigma^{[k]}$ and $\rho^{[k]}$ within the 1% equilibrium boundaries in less than 10 iterations – usually less than 5 – as can be seen on ***Figure 2.*** in the highlighted column for $\sigma^{[k]}$ (deliberately presented here with a higher precision than other values). The starting approximations for $\rho^{[0]}$ and $\sigma^{[0]}$ were taken from the ergodic conditions for the states $p_c \div p_{T_1}$ and $p_{T_2} \div p_\infty$ respectively. Nevertheless, we found the iterative procedure to be very insensitive to the precise initial values of $\rho^{[0]}$ and $\sigma^{[0]}$. The independent input parameters are highlighted in the second row of the table: system size n, server architecture c, vicinity factor v, the global (i.e. cluster-wide) task arrival process with rate Λ and the global load condition Φ. The rest of the input parameters are derived according to the description in Step 0.

The state aggregation approach presented in this paper is mathematically equivalent to the general solution without aggregation, which we used in (Georgiev, V. 2010.), but state aggregation yields more compact formulae and it is less vulnerable to errors in the derivation process as well as in the coding. Having built the iteration table we can instantly obtain any needed output value of the vast problem space – an advantage which is not available by brute use of solvers, or by simulation and experimental performance evaluation.

CASE STUDY

We have done a few numerical experiments with our model. These experiments evaluate the effect of the system architecture i.e. the system size and the number of processor cores on the cluster performance regarding three parameters presented as a function of the global load factor Φ: the mean system service time T of tasks, the mean rate of receiving monitoring messages' exchange process ε and the fraction (i.e. probability) of idle state of nodes (including the state of idle cores) Θ. We consider a cluster of 800 concurrently working cores spread in three ways: 200 quad-core nodes OR 400 dual-core nodes OR 800 uni-core nodes and servicing an arrival process with mean rate of Λ=8000 tasks per minute. The vicinity factor of the nodes $v = 4$ which corresponds to a logical topology of 4-dimensional hypercube. The output model parameters T, ε and Θ are evaluated for each of these types of servers by two conditions: with and without application of the presented distributed load balancing scheme.

The obtained results are presented in three graphs as follows. The dependency of the mean service time from the load conditions $T(\Phi)$ is presented on ***Figure 3***. We see that the locally balanced clusters work always faster. Furthermore: the p2p load balancing tends to turn the growth of the service time to sub-linear speed while the tendency for no load balancing is clearly super-linear. The best performance is achieved by the uniprocessor cluster, closely followed by the cluster of dual-core servers and leaving the quad-core cluster far behind in performance. The interpretation of such a result is that our model does not take in consideration the

N	c	v	T1	T2	n	Λ	Φ 0 = Λ/ncω		λ	ω	ρ0	σ0	(λ+ρ0)/ω					
	2	**4**	3	5	**200**	**8000**	**0.9**		40	22.2	20	*36*	2.7					
	Σπ0	π0	π1	πc	πT1-1	πT1	πT2	πT2+1	**PR**	**PN**	**PS**	PΣR	ρ	σ	£	*T*	ε	Θ%
0	7.3	0.136	0.368	0.5	0.496	0.526	0.426	0.503	0.356	0.45	0.19	0.83	3.395	*6.325974*	4.71	0.02	34.57	9.6
1	4.9	0.206	0.402	0.39	0.392	0.526	0.426	0.212	0.347	0.25	0.4	0.82	15.09	*13.09763*	5.99	0.04	22.51	46
2	6.6	0.153	0.378	0.47	0.469	0.526	0.426	0.305	0.315	0.35	0.34	0.78	11.31	*10.5317*	5.45	0.03	28.21	28
3	6	0.167	0.386	0.45	0.446	0.526	0.426	0.272	0.323	0.32	0.36	0.79	12.69	*11.38947*	5.6	0.03	26.39	33
4	6.2	0.162	0.384	0.45	0.455	0.526	0.426	0.284	0.32	0.33	0.35	0.79	12.23	*11.08802*	5.55	0.03	27.05	31
5	6.1	0.164	0.385	0.45	0.452	0.526	0.426	0.28	0.321	0.32	0.36	0.79	12.4	*11.19212*	5.57	0.03	26.82	31
6	6.1	0.163	0.384	0.45	0.453	0.526	0.426	0.281	0.32	0.33	0.35	0.79	12.34	*11.15589*	5.56	0.03	26.9	31
7	6.1	0.163	0.384	0.45	0.453	0.526	0.426	0.281	0.321	0.32	0.35	0.79	12.36	*11.16845*	5.56	0.03	26.88	31
8	6.1	0.163	0.384	0.45	0.453	0.526	0.426	0.281	0.32	0.32	0.35	0.79	12.36	*11.16408*	5.56	0.03	26.89	31
9	6.1	0.163	0.384	0.45	0.453	0.526	0.426	0.281	0.321	0.32	0.35	0.79	12.36	*11.1656*	5.56	0.03	26.88	31
10	6.1	0.163	0.384	0.45	0.453	0.526	0.426	0.281	0.321	0.32	0.35	0.79	12.36	*11.16507*	5.56	0.03	26.88	31

Figure 2. Electronic Table of the Iteration Procedure in 10 Steps for a System of 200 dual-core Nodes.

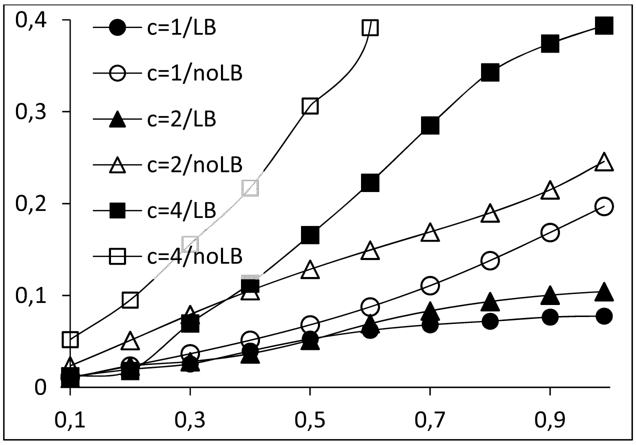


Figure 3. Dependency of the Mean Service Time from the Load Conditions $T(\Phi)$.

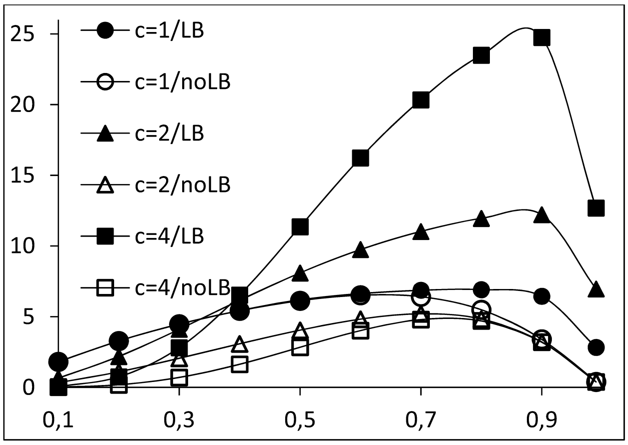


Figure 4. Dependency of the Mean Rate of Monitoring Messages' from the Load Conditions $\varepsilon(\Phi)$.

intraprocessor balancing and control which is performed by the operating system core locally.

The dependency of the mean rate of receiving monitoring messages' $\varepsilon(\Phi)$ is presented on ***Figure 4***. This parameter in fact evaluates the communication overload which the load balancing scheme imposes on the system. Obviously the bigger number of neighbors incurs a higher exchange rate. This dependence is even deeper for moderate to higher (but not highest) level of system load when predictably the node passes between the load states $\mathcal{R}$, $\mathcal{N}$ and $\mathcal{S}$ more frequently.

Finally we have calculated the Θ% the ratio of idle time of the servers including the server cores, shown in ***Figure 5***. Predictably the faster services' case – that of the locally balanced unicore servers yields lowest percentage of idle processor time.

CONCLUSION

We have studied the cloud processing performance by a model of distributed load balancing. This paper presents the major performance parameters as well as an aggregation method for easy and fast numerical solution of the model which we find very practical compared to other methods for solution of the Markov-chain based models. We didn't report

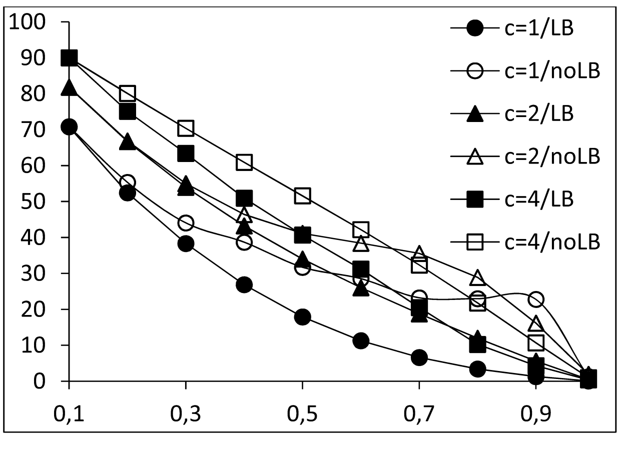


Figure 5. Dependency of the Mean Rate of Monitoring Messages' ε from the Load Conditions $T(\Phi)$.

verification data of the obtained results but the studied parameter space demonstrates logical consistency with the studied phenomena like service delays, system overloads, etc. An important direction of our further studies is the development of these models toward realistic presentation of the load balanced servicing by including of the impact of the communication overload generated by the monitoring strategies in the global (cluster-wide) model.

REFERENCES

Cybenko G., 1994. Dynamic Load Balancing for Distributed Memory Multiprocessors. Journal of Parallel and Distributed Computing, 7, 1989, pp. 279 – 301.

Eager, D., E. Lazovska, J. Zahorian. 1986. A Comparison of Receiver-Initiated and Sender-Initiated Adaptive Load Sharing. Performance Evaluation, Vol. 6., pp. 53-68, 1986.

Geist, R., K. Trivedi. 1983. The Integration of User Perception in the Heterogenous M/M/2 Queue. Performace'83. pp. 203 – 216.

Georgiev, V. 2010. Load Diffusion and Brownian Models for Cloud Balancing: between C-S and p2p. In Proceedings of ESM'2010, Hasselt, Belgium, 25th – 27th October, 2010. pp. 170 – 177.

Karagiorgos, G., N. Missirli and F. Tzaferis. 2004. The generalized diffusion method for the load balancing problem. Advances in Parallel Computing, Volume 13, Elsevier B.V. 2004, Pages 225-232

Luis M. Vaquero, Luis Rodero-Merino and Rajkumar Buyya. 2011. Dynamically Scaling Applications in the Cloud. ACM SIGCOMM Computer Communication Review. 41/1, January 2011.

WEB REFERENCES

Gong, Z.; P.Ramaswamy; X. Gu; and X. Ma. 2009. SigLM: Signature-Driven Load Management for Cloud Computing Infrastructures. 17th International Workshop on Quality of Service, IWQoS 13-15 July 2009. (http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp? punumber=5201378)

SwiftWater. 2010. Cloud computing, load balancing, and extending the data center into a cloud. (http://vburke.wordpress.com/2010/03/26/cloud-computing-load-balancing-and-xtending-the-data-center-into-a-cloud/)

Triebes K. 2010. Cloud Balancing: The Next Generation of Global Server Load Balancing. Virtualization Review (http://virtualizationreview.com/blogs/app-delivery-ondemand/2010/06/cloud-balancing.aspx)

ENGINEERING SIMULATION

NUMERICAL SIMULATION FOR THE FORMABILITY PREDICTION OF THE LASER WELDED BLANKS (TWB)

Hossein Mamusi
Abolfazl Masoumi
Ramezanali Mahdavinejad
School of Mechanical Engineering
College of Engineering
University of Tehran
Tehran, P. O. Box 11155-4653
Iran
E-mail: amasomi@ut.ac.ir

KEYWORDS
TWB, Forming Limit Diagram, Necking criteria, ABAQUS/Explicit

ABSTRACT

Tailor-welded Blanks (TWBs) are tailor made for different complex component designs by welding multiple metal sheets with different thicknesses, shapes, coatings or strengths prior to forming. In this study the Hemispherical Die Stretching (HDS) test (out-of-plane stretching) of TWBs were simulated via ABAQUS/Explicit to obtain the Forming Limit Diagrams (FLDs) of Stainless steel (AISI 304) laser welded blanks with different thicknesses. Two criteria were used to detect the start of necking to determine the FLD for TWBs and parent sheet metals. These two criteria are the second derivatives of the major and thickness strains that are given from the strain history of simulation. In the other word, in these criteria necking starts when the second derivative of thickness or major strain reaches its maximum. With having the time of onset necking, one can measure the major and minor strains at the critical area and determine the forming limit curve.

INTRODUCTION

Tailor-welded Blanks (TWBs) are tailor made for different complex component designs by welding multiple metal sheets with different thicknesses, shapes, coatings or strengths prior to forming (Chan et al. 2007). Then the prepared blanks are formed to the desired shapes by approperiate forming method. For example in automotive industries, the final blank is stamped to the desired shape for the car body panel. In this technique, one can use the stronger or thicker sheets where needed and in this way a local stiffness is obtained which leads to product weight reduction without loss of stiffness and safety. The TWB technique has benefits for the producers, consumers and the environment due to weight reduction that causes less material and fuel consumption. Forming behaviour is the most important factor in applying TWB in the automotive industries, although the cost should be studied, too. The Forming Limit Diagram (FLD) has been accepted for the formability prediction of sheet metals and could be used for TWBs. A forming limit diagram, also known as a forming limit curve, is used in sheet metal forming for predicting the forming behavior of sheet metal (Llewellyn and Hudd 1998; Marciniak et al. 2002). The concept of the FLD which was developed by Keeler (Keeler 1965) and Goodwin (Goodwin 1968) and then was used by industry could be achieved theoretically and experimentally. To find the FLD, the sheet metal is subjected to various combinations of principal stresses (σ_1 and σ_2) to create different combination of principal strains (ε_1 and ε_2). For this purpose, usually the sheet metal specimens are stretched with constant length and variable widths via a hemispherical punch (out-of-plane stretching) or flat punch (in-plane stretching), or stretched with single geometry specimens with different lubricants. As the experimental method is both expensive and time consuming, in this paper a numerical simulation method is introduced by which precise, rapid and less expensive FLDs are produced only with applying the mechanical properties that resulted from uniaxial tensile test.

In recent years, many researches have attempted to develop precise and reliable models to find FLCs of base metals and several necking criteria have been proposed to predict the onset of localized necking (Situ et al. 2006; Petek et al. 2005). For the predicting forming limit of Tailor Welded Blanks, (Narayanan et al. 2010) introduced some necking criteria, namely the effective strain rate based criterion (ESRC – RC1), major strain rate based criterion (MSRC – RC2), thickness strain rate based criterion (TSRC – RC3), and thickness gradient based criterion (TGNC – RC4). For the present work these criteria were evaluated and among them two criteria were prefered to develop an accurate model to find FLCs of TWBs.

METHODOLOGY

The Hemispherical Die Stretching (HDS) test via ABAQUS/Explicit FE code is simulated in 3D space to evaluate and analyze the formability of TWBs. Dry friction state was assumed and the friction coefficient between the blank and the punch was assumed to be μ=0.1. The die modeled as rigid with 100mm diameter of punch and 105.6mm the diameter of matrix.

Holder and the punch were allowed to move in the Z direction along the axis of the punch and the matrix is fixed. The modeled die is shown in the Figure 1 and the blanks were modeled as deformable solid and meshed with the C3D8R elements.

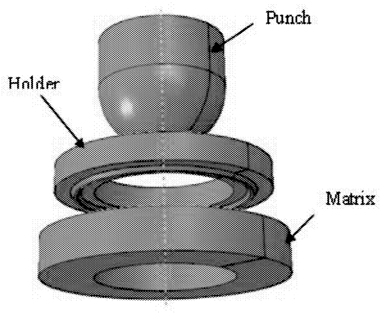


Figure 1: Modeled HDS die

In this simulation the weld line was not modeled because calculations shows that both with and without the incorporation of soft zone and fusion zone (FZ) properties in the FE simulation shows similar LDH results (less than 0.1 pct difference in the LDH). During the FE modeling, therefore, the soft zone and fusion hard zone can be omitted (Panda et al 2009).

The stainless steel (AISI 304) was selected for the formability analysis based on the experimental results of the Chan et al (Chan et al. 2007). The mechanical properties of this material are shown in the Table 1 and Figure 2 shows engineering stress-strain curves, but in the modeling the material true stress-strain data have been used. The density also considered as 7900 kg/m^3 and to define the elastic properties, Young's modulus is taken as 197 GPa and Poisson's ratio is taken as 0.29.

Table 1: Mechanical properties of sheet metals

Thickness(mm)	1	1.2
Yield Strength(Mpa)	355	372
Tensile Strength(Mpa)	1112	1205
Necking Strain	0.37	0.39
Normal Anisotropy(r_m)	1	1
Strain Hardenning Exponent, (n)	0.45	0.45

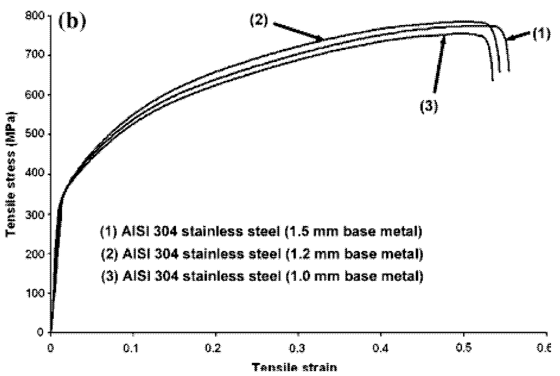


Figure 2: Stress Strain Curves of Stainless Steel Base Metals (Chan et al. 2007)

NECKING CRITERIA

Selecting a suitable necking criterion is the essential problem that must be solved to find the limit strains. As mentioned earlier, two criteria were chosen to detect onset necking. These criteria that are based on the strain history of the simulation are the second derivative or acceleration of major strain and thickness strain. After the simulation, the strain history must be extracted and evaluated. Specimen of width 100 mm and length 200 mm of stainless steel AISI 304 having 1 mm thickness were used as an example to analyze this simulation process and to determine the onset of necking. After the simulation the strain history data were extracted as shown in the Figure 3 and fitted with Gaussian equations which showed good adaptivity with the strain history data. After the curve fitting and plotting first and second derivative of strain history, one can determine onset necking time and fracture time.

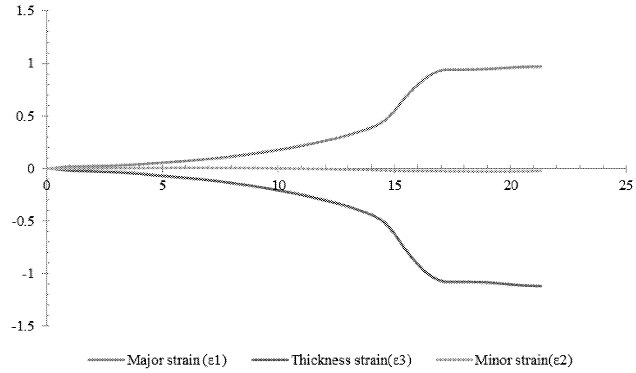


Figure 3: Major, Minor and Thickness Strains

SECOND DERIVATIVE OF MAJOR STRAIN

Figure 4 presents the contours of major strains of the stretched TWB. As it is shown in Figure 4, the major strain gradient near the weld line is more than in other places, but this necking place depends essentially on the friction coefficient, lessening the friction coefficient will result in a closer necking, and necking also occurs in the thinner blank. Here, the second derivative of major strain or major strain acceleration determines the onset of the necking.

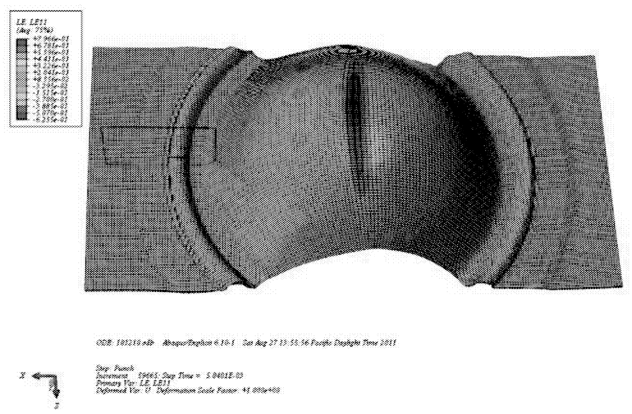


Figure 4: Contours of Major Strain of a TWB (1mm/1.2mm, 100×200mm)

In Figures 5 and 6 the evolutions of major strain rate and major strain acceleration of a necked element is plotted, respectively. It can be seen that major strain acceleration in the 14.72 s, reaches its maximum, that means the onset necking, and major strain rate in the 15.50 s, reaches its maximum that means the fracture of specimen.

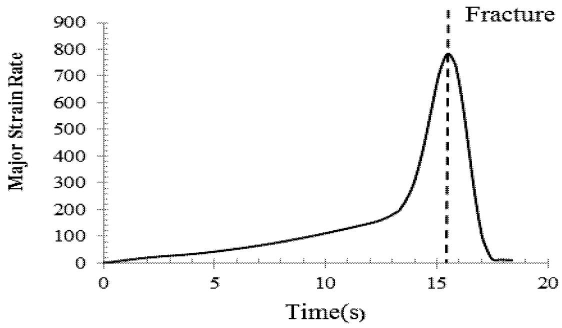


Figure 5: First Derivative of Major Strain

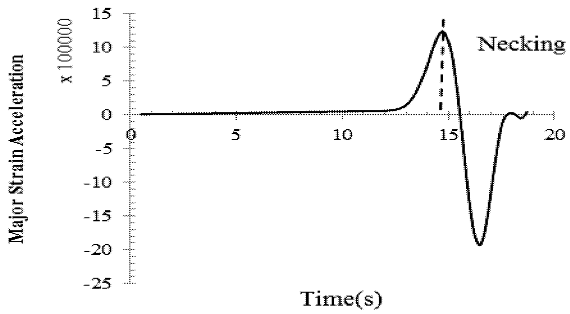


Figure 6: Second Derivative of Major Strain

SECOND DERIVATIVE OF THICKNESS STRAIN

As shown in Figure 7 and in comparison with Figure 4, the thickness strain gradient and the major strain gradient take place in the same place. Comparing these figures with Figure 3, one can find that their changes are proportionate. The method of finding the onset necking time with this criterion is similar to the first criterion. But notwithstanding of these similarities, these results are different together as shown in the Table 2. Figures 8 and 9 show first derivative and second derivative of thickness strain respectively. Figures 8, 9 and Table 2 show that with the second derivative of thickness strain criteria, the necking onsets at 14.62 s and fracture occurs at 15.52 s.

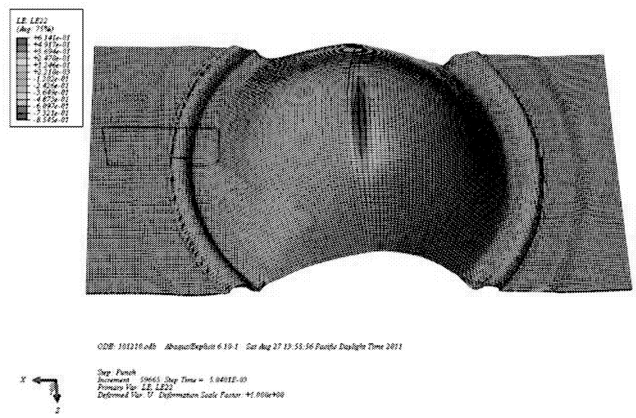


Figure 7: Contours of Thickness Strain of a TWB (1mm/1.2mm, 100×200mm)

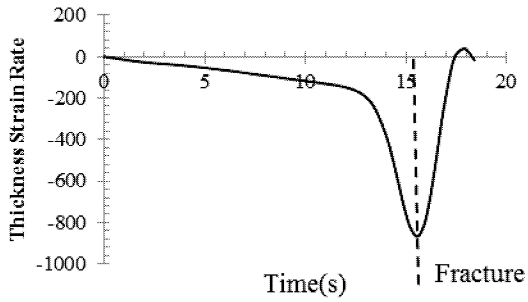


Figure 8: First Derivative of Thickness Strain

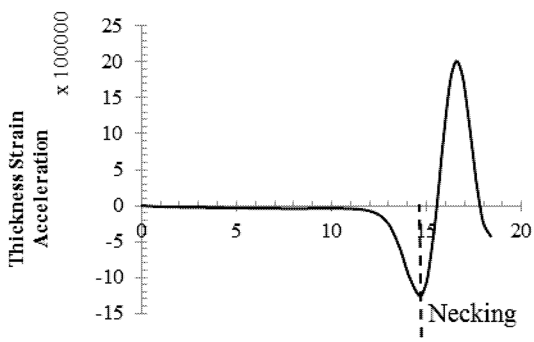


Figure 9: Second Derivative of Thickness Strain

Table 2: A Comparison between the Results of the Analysis via Two Criteria

Necking Critrion	Major Strain Second Derivative($\ddot{\varepsilon}_1$)	Thickness Strain Second Derivative($\ddot{\varepsilon}_3$)
Necking Start Time(s)	14.72	14.62
Fracture Time(s)	15.50	15.52
Major Strain(ε_1)	0.48	0.47
Minor Strain(ε_2)	0.02	0.02

LINEARITY OF THE STRAIN PATH

For research the linearity of the strain path until the forming limits, two points of 1.0/1.5 laser welded blanks have been considered, one from the right hand side of the diagram (tension-tension side), and the other from the left hand side (tension-compression side). Figures 10 and 11 show the principal strain history of both points. The relation between the major strain (ε_1) and minor strain (ε_2) that named the strain path is shown in Figures 12 and 13, too. The results indicate that the strain path of both points first is linear but not until the target point (necking) and then the strain path become non-linear and this may be due to the friction factor.

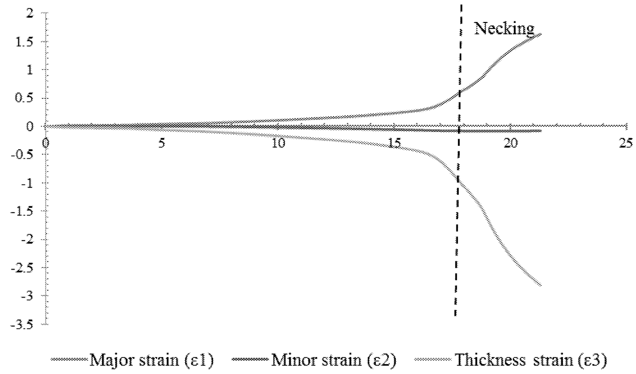


Figure 10: Principal Strains of 80×200 mm specimen of TWB 1.5/1.0 mm

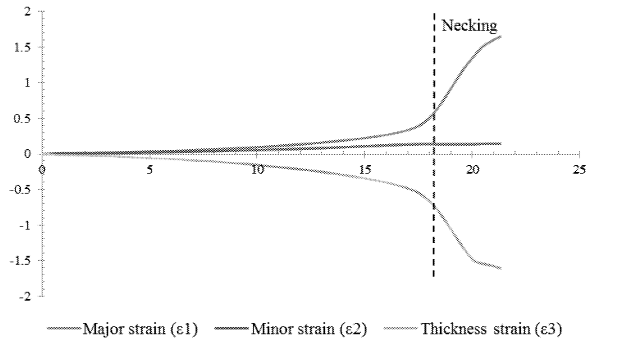


Figure 11: Principal Strains of 140×200 mm Specimen of TWB 1.5/1.0 mm

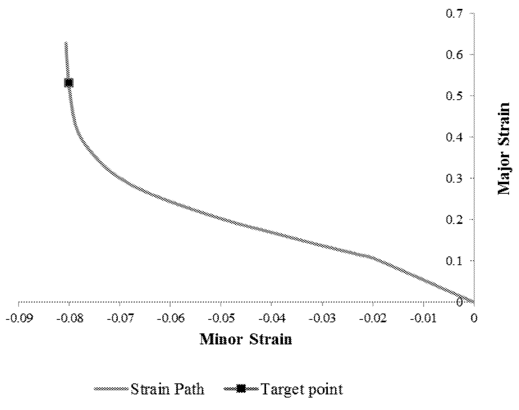


Figure 12: Strain Path of Left Hand Side of the FLD

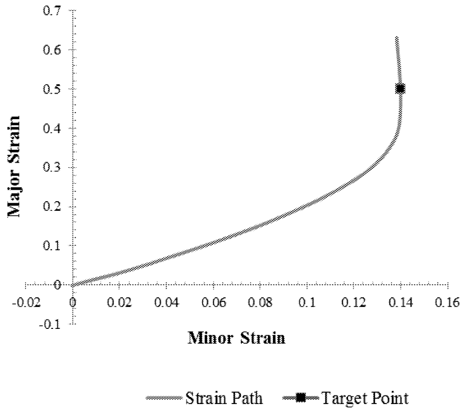


Figure 13: Strain Path of Right Hand Side of the FLD

RESULTS AND DISCUSSION

The forming limit diagrams of TWBs are predicted using two necking criteria, second temporal derivatives of major strain and thickness strain. The Figures 14 to 18 show the analytical calculated FLCs of stainless steel (AISI 304) base metal and TWBs with both first criteria (second derivative of major strain) and second criteria (second derivative of thickness strain), in comparison with experimental results from Chan et al (Chan et al. 2007).

Majority of results show that the results of second criteria are closer to the experimental results. But the comparison between the Figure 4 and Figure 7 show that both criteria detect the same place as the necked zone.

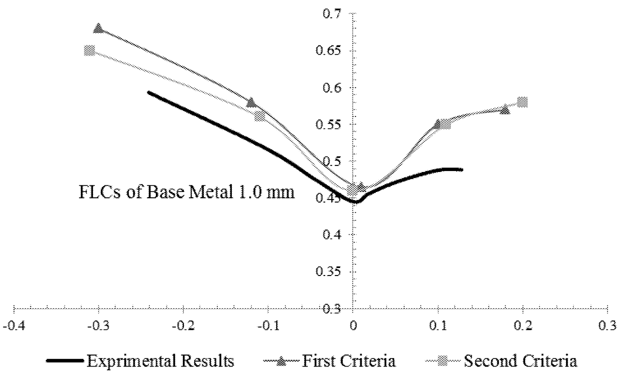


Figure 14: FLCs of Base Metal 1.0 mm

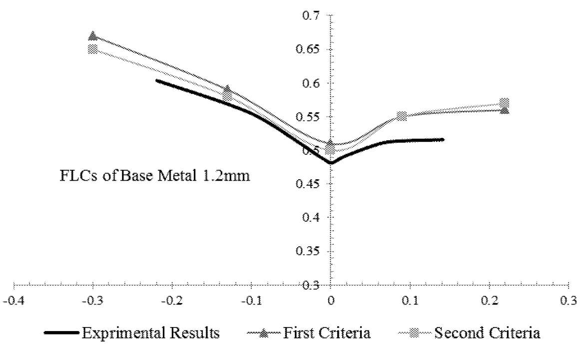


Figure 15: FLCs of Base Metal 1.2 mm

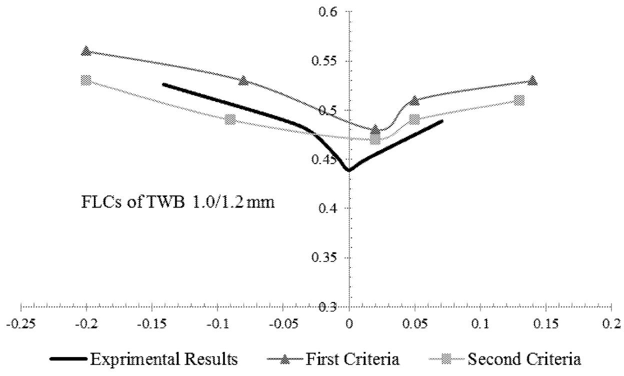


Figure 16: FLCs of TWB 1.0/1.2 mm

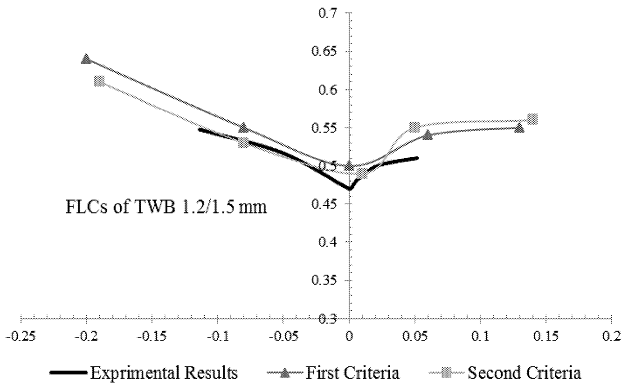


Figure 17: FLCs of TWB 1.2/1.5 mm

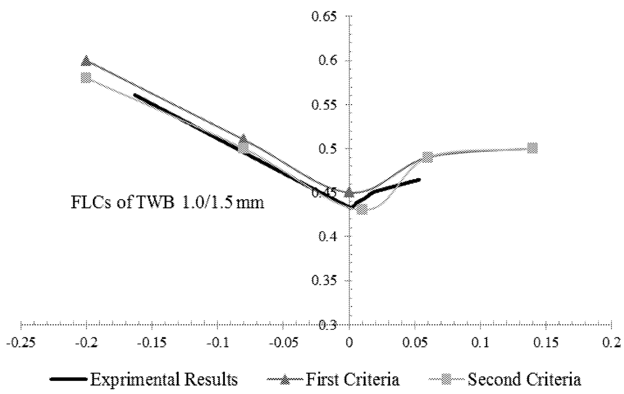


Figure 18: FLCs of TWB 1.0/1.5 mm

CONCLUSIONS

The main aim of the work was the presentation of a FE model for predicting the formability of TWBs with two necking criteria that may be more applicable in the automotive industry that use TWB technology. Following conclusions can be drawn from the study.

- Both criteria show the same results for necking place and fracture time. As shown in Figures 4, 7 and Table. 2 both the second derivative of major and thickness strains with good accuracy are similar in predicting the necking zone and fracture time. Therefore, both of them could be used for predicting the necked zone and fracture time in the industrial applications.
- The second derivative of thickness strain, however present more reliable results. Figures 14 to 18 show the forming limit results of second criteria (second derivative of thickness strain) are closer to the experimental results while the second criteria show better accuracy. Therefore, when it is needed to predict the forming limit more accurately, it is better to use the second derivative of thickness strain as the necking criteria for modeling and simulation of the process.

REFERENCES

Chan, L. C.; C. H. Cheng and C. L. Chow 2007. "Weldment properties evaluation and formability study of tailor-welded blanks of different thickness combinations and welding orientations." *Journal of Materials Science* **42**(15): 5982-5990.

Goodwin, G. M. 1968. "Application of strain analysis to sheet metal forming problems in the press shop." SAE paper 680093,

Keeler, S. P. Determination of forming limits in automotive stampings. *Sheet Metal Ind*., 1965, 42, 683–691.

Llewellyn, D. T. and R. C. Hudd 1998. "Steels : metallurgy and applications." Oxford, Butterworth-Heinemann. pp. 28.

Marciniak, Z.; J. L. Duncan and S. J. Hu 2002. Mechanics of sheet metal forming. Oxford, Butterworth-Heinemann. pp. 75.

Narayanan, R. G.; B. S. Naik and P. J. Ramulu 2010. "Application of a few necking criteria in predicting the forming limit of unwelded and tailor-welded blanks." *Journal of Strain Analysis for Engineering Design* **45**(2): 79-96.

Panda, S. K.; V. H. B. Hernandez; M. L. Kuntz and Y. Zhou 2009. "Formability Analysis of Diode-Laser-Welded Tailored Blanks of Advanced High-Strength Steel Sheets." *Metallurgical and Materials Transactions a-Physical Metallurgy and Materials Science* **40A**(8): 1955-1967.

Petek, A.; T. Pepelnjak and K. Kuzman 2005. "An improved method for determining a forming limit diagram in the digital environment." *Strojniski Vestnik-Journal of Mechanical Engineering* **51**(6): 330-345.

Situ, Q., M. Jain and M. Bruhis 2006. "A suitable criterion for precise determination of incipient necking in sheet materials."*Aluminium Alloys 2006, Pts 1 and 2* **519-521**: 111-116.

BIOGRAPHIES

Hossein Mamusi: MSc student at School of Mechanical Engineering, College of Engineering, University of Tehran

Abolfazl Masoumi: Associate Professor at School of Mechanical Engineering, College of Engineering, University of Tehran

Ramezanali Mahdavinejad: Associate Professor at School of Mechanical Engineering, College of Engineering, University of Tehran

INVESTIGATIVE MODEL REFERENCE CONTROLLER FOR SERVOMOTOR TRACKING SYSTEM BASED ON NEURAL NETWORK TECHNIQUE

Raaed K. Ibrahem
College of Electrical and Electronics Techniques
Foundation of Technical Education
Baghdad/Iraq
Email: raid_khalid2000@yahoo.com

KEYWORDS
Neural Network, Model Reference control, Servomotor

ABSTRACT

Neural network are appropriate for the modeling and control of multifaceted physical system since their capability to manage multifaceted input-output charting without thorough mathematical model of the systems. Demonstrating a non-linear active charting, Model Reference Controller Neural Network (MRCNN) is appropriate to manage active non-linear complications. In this paper, a MRCNN have been employed in a sun-seeker tracking system. First, the MRCNN will be used as identifier to recognize the opposite model of the system to be controlled through supervised, and then the MRCNN is used such as a feed forward controller, to create control voltage to power the sun-seeker to track pre-selected routes of location.

INTRODUCTION

For there are variation of parameters, such as load torque, inertia and mechanical friction, low speed sun seeker tracking system is a typical nonlinear system. It is difficult to get accurate system’s mathematical model due to unstructured uncertainties for the unmodelled dynamics like nonlinear friction. It is impossible to get high accuracy response by means of traditional control method based on system’s mathematical model, such as PID. The neural network area consists of a very promising direction to solve the problem relating to unknown nonlinear dynamic system. Hence, neural networks appear as a powerful tool for learning highly nonlinear dynamic systems. Their massive parallelism, very fast adaptation, and inherent approximation capability, have attracted extensively researchers in the field of system identification and control [1-4]. Many existing neural control laws for mechanical systems suffer from important shortcomings. Such as intensive computation effort and high storage capacity [5-6], so it is impossible to apply neural control algorithm to system such as sampling time is very short. This is a main cause for neural network controllers have not been widely used so far. This paper presented a neural-network based control scheme for a low sun seeker tracking system. The control system of variable pitch has the character of non-linear, time-varying parameter, multi-disturbance and time-lag due to the frequent movement of variable pitch switch, the inertial element of high quantity of servomotor driven mechanism and the load disturbance caused by the external sources. Therefore, the advanced control methods, such as the sliding variable structure control, the neural network control and et al, are applied in the servomotor tracking system to solve these problems. The neural network in the intelligent control field is a high non-linear dynamic system, which can be seen as a tool between the logical reasoning and numerical calculation, and it is independent of mathematical model. The neural network is formed by the interconnection of much neurons presented based on the neurobiology cognitive science studying on the information processing [7-10], so it has the characters of strong adaptability, non-linear mapping ability and robustness. To reduce calculating time, a MRCNN is used to estimate the inverse model of plant to be controlled; parameters of neural network are updated on-line according to the dynamic back propagation algorithm (DBP). To ensure system’s initial robustness and close-loop stability, a fixed gain feedback controller is used.

Problem Formulation

The Model Reference Adaptive Control (MRAC) configuration [11] employs two neural networks: a controller network and a model network. The model network can be trained off-line using historical plant measurements. The controller is adaptively trained to force the plant output to track a reference model output. The model network is used to predict the effect of controller changes on plant output, which allows the updating of controller parameters. Plant NN Plant Model NN Inverse Plant Model Reference Model Adaptation Algorithm NN Controller Command Input Plant Disturbance Sensor Noise Plant Output Noise & Disturbance at Plant Output Tracking Error NN Plant Controller Robustness Filter Command Input Plant Output Predicted Plant Output Control Input NN Plant Model. Figure 2 shows the details of the neural network plant model and the neural network controller as they are implemented in the Neural Network Toolbox software. Each network has two layers, and you can select the number of neurons to use in the hidden layers. There are three sets of controller inputs:

- Delayed reference inputs
- Delayed controller outputs
- Delayed plant outputs

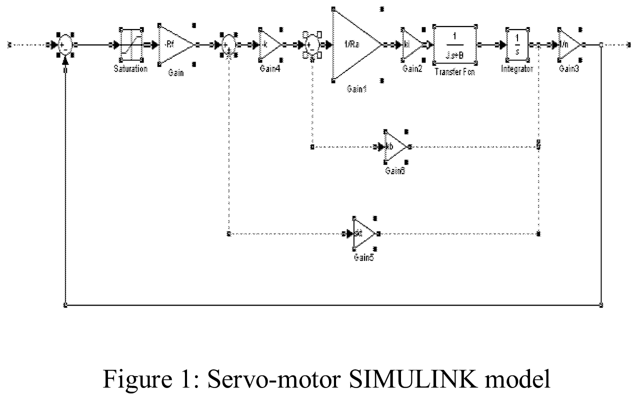


Figure 1: Servo-motor SIMULINK model

For each of these inputs, you can select the number of delayed values to use. Typically, the number of delays increases with the order of the plant. There are two sets of inputs to the neural network plant model:

- Delayed controller outputs
- Delayed plant outputs

As with the controller, you can set the number of delays. The next section demonstrates how you can set the parameters

Error discriminator

The low speed system's composed of an error discriminator, operational amplifier, controller, servo amplifier, DC motor, and gear train. The servo-motor tracking systems demonstrate in Figure 1. When the vehicle is aligned with the sun,

$$oa = \frac{W}{2} + L \tan \alpha(t)$$
$$ob = \frac{W}{2} - L \tan \alpha(t) \quad \ldots (1)$$

Where,

oa- the width of the sun ray that shines on cell A.

ob- is the same on cell B.

α (t)- is the error angle between solar axis and vehicle axis.

Since the current i_a (t) is proportional to oa, and i_b (t) is proportional to ob, we have

$$i_a(t) = I + \frac{2LI}{W} \tan \alpha(t)$$
$$i_b(t) = I - \frac{2LI}{W} \tan \alpha(t) \quad \ldots (2)$$

Operational amplifier

The relation between the output of the operational amplifier and the current $i_a(t)$ and $i_b(t)$ is

$$e_o(t) = -R_f[i_a(t) - i_b(t)] \ \ldots. (3)$$

Servo amplifier

The gain of the servo amplifier –K. the output of the servo amplifier is expressed as

$$e_a(t) = -K[e_o(t) + e_t(t)] = -Ke_s(t) \ \ldots (4)$$

The output voltage of the tachometer e_t is related to the angular velocity of the motor through the tachometer constant K_t.

Tachometers

The output voltage of tachometer e_t is related to the angular velocity of the motor through the tachometer constant K_t.

$$e_t(t) = K_t \omega_m(t) \ldots. (5)$$

The angular position of the output gear is related to the motor position through the gear ratio 1/n.

$$\theta_o = \frac{1}{n}\theta_m \ldots (6)$$

DC Motor Model

The dc motor has been modeled as

$$e_a(t) = R_a i_a(t) + e_b(t)$$
$$e_b(t) = K_b \omega_m(t)$$
$$T_m(t) = K_i i_a(t) \quad \ldots (7)$$
$$T_m(t) = J\frac{d\omega_m(t)}{dt} + \beta\omega_m(t)$$

Where J and β are the inertia and viscous-friction coefficient seen at the motor shaft.

MRCNN CONTROLLER MODEL

As shown in figure 2, the basic control scheme consists of a feed forward MRCNN controller and a fixed gain feedback controller. The MRCNN is first used as an identifier to emulate the inverse dynamics of the dc servo system, and this network is called as MRNNI, it is trained off-line and on-line. When MRNNI is trained, it is used as a feed forward controller called as DRNNC. The system control voltage *U* is composed of the feed forward controller output voltage *Un*

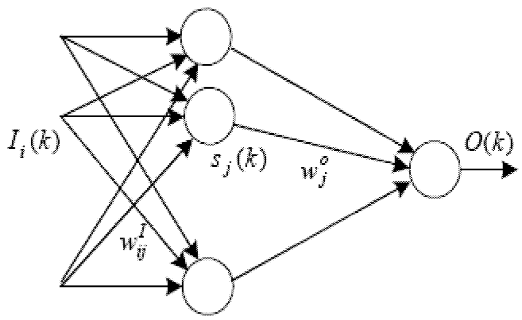


Figure 2: DRNNC three layer Neural Network

and the feedback controller U_P. If the MRNNI has learned the inverse model of the system, the MRCNN alone provides all the necessary voltage for the system to track the desired trajectory and output of the feedback controller will tend to zero.

LEARNING MRNNI

In this paper we used a three-layered neural networks with 13th hidden layer can approximate any nonlinear function to any desired accuracy [1]. MRCNN networks superior to multiplayer feed forward static neural networks to deal with dynamic problems [11]. The structure of three layers MRCNN is shown in figure 3. It consists of an input layer, an output layer and one recursive hidden layer.

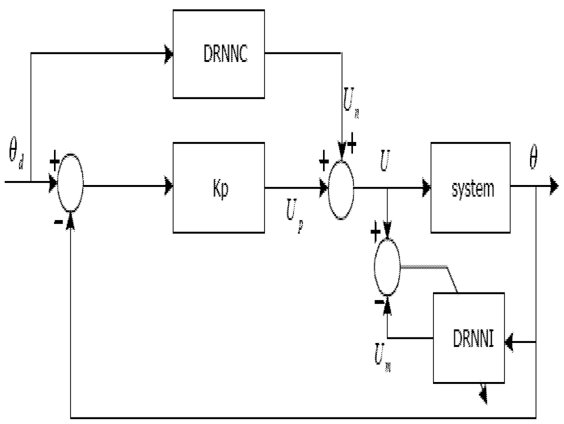


Figure 3: DRNNC control for the servo-motor tracking system

Where *j ijoj I*(*k*),*w* , s_j and *O*(*k*) are the ith input to the MRCNN, the connecting weight between jth recursive neuron and the output of networks, connecting weight between ith input to network and the jth hidden neuron, the output of jth hidden neuron and the output of the MRCNN. The mathematical model of MRCNN is shown below:

Where x_j (*k*) *is* the output of jth recursive neuron, D_j *w* is the recursive weight of jth hidden neuron, *f* (□) is sigmoid function. When DRNN is used as MRNNI, output of networks *O*(*k*)= U_m (*k*) . When MRNN is used as MRCNN, *O*(*k*)= U_n (*k*) .

$$s_j(k) = w_j^D x_j(k-1) + \sum_i w_{ij}^I I_i(k) \;...\; (8)$$

The cost function to train MRNNI is defined as:

$$J = \frac{1}{2}(U - U_m)^2 = \frac{1}{2}e_m^2 \;....\; (9)$$

The objective of the learning process is to adjust the network parameters (weights) so as to minimize the cost function *J* over the entire train set. The back propagation algorithm is given below [12].

$$\Delta w(k) = -\eta \frac{\partial J}{\partial w}$$

$$= \eta e_m(k) \frac{\partial U_m}{\partial w}$$

$$= \eta e_m \frac{\partial O(k)}{\partial w} \;...\; (10)$$

Where *w*(*k*) is any weight of MRNNI, *h* is the learning rate of this weight. Define the output gradients with respect to output, recurrent, and input weight, respectively as below

$$\frac{\partial O(k)}{\partial w_j^o} = x_j(k) \;...\; (11)$$

$$\frac{\partial O(k)}{\partial w_j^D} = w_j^o P_j(k) \;...\; (12)$$

$$\frac{\partial O(k)}{\partial w_{ij}^I} = w_j^o Q u(k) \;...\; (13)$$

$$P_g(k) = \frac{\partial x_j(k)}{\partial w_j^D} = f'(s_j) x_j(k-1) \;...\; (14)$$

$$Q_{ij}(k) = \frac{\partial x_j(k)}{\partial w_{ij}^I} = f'(s_j) I_i(k) \;...\; (15)$$

From above equations, learning algorithm of weight w_{ij} , D_j *w* and O_{jw} can be got. The learning rate can be chosen properly [13].

IDENTIFICATION SCHEME

For the dc system position tracking, the MRNNI is used to identify the unknown system dynamics (dc motor, amplifier, and the mechanical friction) that mapping the control voltage *U* to the motor position. Because the MRNNI is used to identify the inverse model of the DC servo system, the inputs to feed forward controller MRCNN is a desired position trajectory and the output of DRNNC is control voltage for system to tack the desired trajectory. From function (7), the relation between control voltage and the motor position can be written as a difference equation below

$$U(k) = d_1\theta(k-3) + d_2\theta(k-2) + d_3\theta(k-1) \;....\; (16)$$

If the aim is to track the desired speed, similarly can get the difference relationship between control voltage and the speed of dc motor as below

$$U(k) = e_1 w(k) + e_2 w(k-1) + e_3 w(k-2) \;...\; (17)$$

Where d_1 , d_2 , d_3 and e_1 , e_2 , e_3 are system parameters .
Function (16) and (17) can be written in this form

$$U(k) = h'(\theta'(k-1)', \theta'(k-2), \theta(k-3))\theta \;...\; (18)$$

$$U(k) = g(w(k), w(k-1), w(k-2)) \;....\; (19)$$

The MRNNI is trained to emulate the unknown function *h*(t) or *g*(t) . For position tracking, the inputs to the MRNNI are *q* (*k* 1),*q* (*k* 2) and *q* (*k* 3) .For speed Tracking, the inputs to the MRNNI are *w*(*k*),*w*(*k* 1) and *w*(*k* 2) . When the DRNNI is trained, it is used as a feed forward controller MRCNN. For position tracking, the inputs to MRCNN are desired trajectory dq (*k* 1),dq (*k* 2) and dq (*k* 3) . For speed tracking, the inputs to DRNNC are desired speed (*k*), (*k* 1) *d d w w* and (*k* 2) *d w.* Control voltage *U,* is the sum of the MRCNN,U_n , and the feedback controller, U_p .

$$U = U_n + U_p \;...\; (20)$$

SIMULATION RESULTS

With the purpose of confirm the efficiency of the proposed method, we consider the following numerical example, the closed-loop transfer function for the servomotor with controller, motor and tachometer are:

$$G(s) = \frac{100}{2s^2 + 3s + 0.076}$$

The model reference model for sun seeker tracking is shown in Figure 4. The testing data for NNMRC is illustrated in Figure 5, validation data of NNMRC as shown in Figure6. Figure 7 show the training data for NNMRC. Furthermore,

CONCLUSION

This paper presents a real-time control of a low speed sun seeker tracking system. It is shown that, MRCNN is efficient for system identification and control, and this system. Through this proposed method, can tack any selected trajectories with high performance under strong mechanical friction and other nonlinear factors. Also, it can be seen that with the higher system frequency of programming environment, the better control results will be got with this method proposed in this work. This paper has given a brief introduction to the use of neural networks in control systems. In the limited space it was not possible to discuss all possible ways in which neural networks have been applied to control system problems. We have selected one type of network, the multilayer perception. We have demonstrated the capabilities of this network for function approximation, and have described how it can be trained to approximate specific functions. We then presented three different control architectures that use neural network function approximations as basic building blocks. The control architectures were demonstrated on three simple physical systems. There have been many different approaches to using neural networks in control systems, and we have not attempted to provide a complete survey of all approaches in this paper.

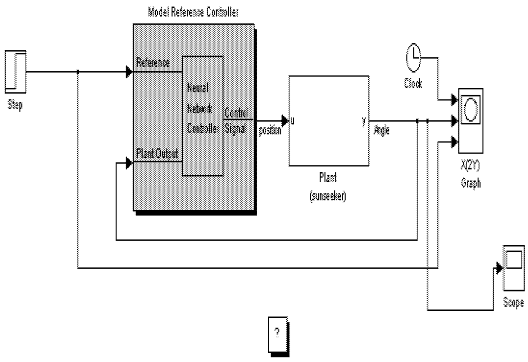


Figure 4: the model reference neural network model for servomotor system

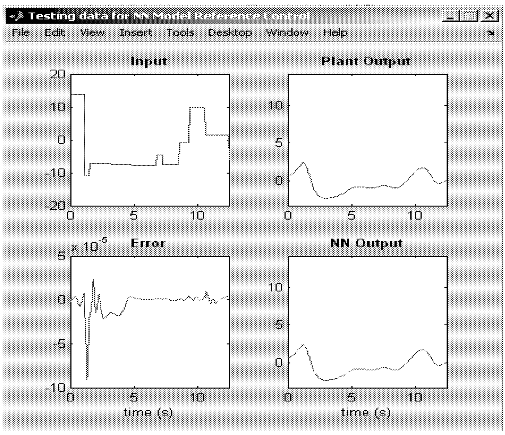


Figure 5: testing data for NNMRC

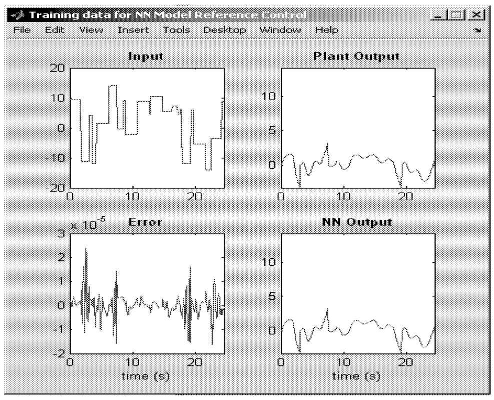


Figure 6: Training data for NNMRC

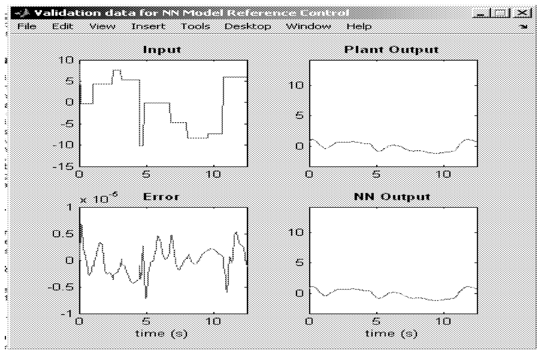


Figure 7: validation data of NNMRC

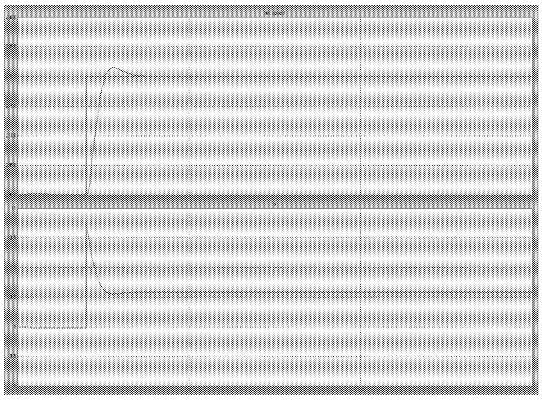


Figure 8: Output Response for servomotor with neural network

REFERENCES

Hornik,K.,Stinchcombe, M.,and White,H. 2009." Multilayer feed forward networks are universal approximations". Neural Networks, 359-366.

Psaltis and Yamamura. 2001. "multilayered neural network controller". IEEE Control Syst. Mag., pp 17-21.

K.S.Narenda and K.Parthasaathy, 1990. "Identification and control of dynamical systems using neural network". IEEE Trascation on Neural Network.

Siri Weerasooriya and M.A.EI-Sharkawi. 2001. "Identification and control of a DC motor using backpropgation neural networks". Proceeding of IEEE/PES Winter meeting. Pp1-7.

Benjamin C. Kuo and Farid Golnaraghi. 2003. "Automatic Control Systems". Eight Edition. Wiley

Jui-Hong Horng. 2005. “Neural adaptive tracking control of a DC motor”. Information Sciences. Vol.118. pp13.

Shi K L, Li H. 2004. “A novel control of a small wind turbine driven generator based on neural networks”. IEEE Power Engineering Society General meeting.2:166-169.

Liu H, Li S, Chai T. 2004. “New design method of fuzzy and PID hybrid controller and its application in power plant control system”. *Power Engineering*, 24(1): 78-82.

Zhang X, Xu D, Lu Y. 2004. “Adaptive fuzzy control for large-scale variable speed wind turbines”. Journal of system Simulation. 16(3): 573-577.

Kong Yigang, Wang Zhixin. 2008. “Design and simulation of fuzzy sliding-mode robust controller for large wind generating unit”. *Proceedings of the CSEE*, 28(14): 136-141.

K.S. Narendra, and K. Parthasarathy. 2000.” Identification and Control of Dynamical Systems Using Neural Networks”. *IEEE Transactions on Neural Networks*, vol. 1, pp. 4-27

L.Jin, P.N.Nikiforruk, and M.M.Gupta. 2007. " Direct adaptive output tracking control using multilayered neural networks". IEE Proceeding-D. Vol.140. pp.393-398.

Chao-Chee Ku and Kwang Y. Lee. 2005.” Diagonal recurrent neural networks for dynamic system control”. IEEE transcation on Neural Network. Vol.6. pp.114-156.

A SELF-EVOLVING CONTROLLER FOR A PHYSICAL ROBOT: A NEW INTRODUCED AVOIDING ALGORITHM

Dan Marius Dobrea
Adriana Sirbu
Monica Claudia Dobrea
Faculty of Electronics, Telecommunications and Information Technologies
"Gheorghe Asachi" Technical University
Bd. Carol I, no. 11, Iaşi, România, 700506
E-mail: mdobrea@etti.tuiasi.ro

KEYWORDS
Genetic algorithms, controller, robot, avoidance algorithm, autonomous behavior

ABSTRACT

One of the major problems when building robots capable to operate in real-world environments is the ability of the robots to deal with a continuous stream of unexpected real-time events. In this paper, we present a real reactive robot able to avoid obstacles. The obstacle avoidance behavior is tested in real environments and it is mainly based on the movement knowledge database extracted previously in a simulation-based learning process. Exactly, a new obstacle-avoiding genetic algorithm implemented in a configurable simulation environment for differential wheeled robots is used to obtain the database of movements for the real robot. The extracted knowledge database comprises the main set of rules that directly maps the sensor information into the engine commands. In real environments, however, for those singular cases in which none of the simulation-based extracted rules applies (i.e. this is the case of the situations never seen before), new rules solving properly the imminent collisions are obtained online and with the same GA algorithm.

INTRODUCTION

In the last 20 years, due to the advances of the chipset industry that mainly boosted the processors market (e.g. higher working frequencies, new architectures, multiples cores etc.) the soft computing techniques received a powerful support. The soft computing techniques that deal mainly with imprecise and uncertain data are very well fitted with real environments (which have the same characteristics).

Fuzzy-logic (FL), neural networks (NNs) and genetic algorithms (GA) are all soft computing techniques largely used in real-environment applications such as real-time robot control applications. Genetic algorithms, for example, are extensively used to develop controllers for robots – especially the hybrid ones, like GA-fuzzy or GA-neural controllers. In these hybrid approaches, the GA adjusts the parameters of the robot controller. The different values of these parameters, obtained during an evolutionary process, generate different behaviors of the robot. In the genetic evolution, the robots with the best performances have more chances to spread their characteristics to the offsprings. At the end, after several iterations, a robot with good performances is obtained. There is a large number of successful reported applications in which the GA tunes the fuzzy logic controllers (Tan et al, 2011), (Martíneza et al, 2009). In the designs of the fuzzy controllers devised for obstacle avoidance, other type of algorithms like the ant colony enhances the used GA techniques (Chiou et al, 2010).

In robotics, one of the main applications of the GA algorithms consists in finding the optimal path that should be followed by a robotic system in order to reach a particular goal (Hosseinzadeh and Izadkhah, 2010), (Repoussis et al. 2009). In order to have a reliable navigation algorithm for an autonomous robot, the latter must be able to: (a) localize its current position, (b) execute (independently from a human operator) a local collision-free motion within its environment and (c) find out a global optimal path to its final goal. Out of all these last three objectives our work, presented in this paper, is dedicated only to develop a robot-improved ability to execute a local collision-free motion trajectory within real environments. For this, we take advantage of a simulation environment for robotic systems in which a genuine GA is used to extract the core movement knowledge-database. Finally, the effectiveness of the obtained controller-evolved strategies is tested on a real robot. The main concept of the new self-evolving controller is similar to that of the FL approach (e.g. the existence of a set of rules). Unlike the FL approach, in our case the rules are extracted using a GA not using the knowledge of some human experts.

The main application of our GA controller is in the field of the remote controlled robots. In such applications, the GA controller adjusts autonomously the remote command whenever the user command put the robot in the imminent danger of colliding.

OBSTACLE AVOIDANCE

The uncertainty of the real environments is the biggest challenge of each autonomous robot. The autonomous robotic controllers must deal with a large number of factors like the robotic system mechanical and electrical characteristics and the environment complexity (e.g. the variation of spatial and temporal parameters, the nonlinearities and uncertainty manifested through chaotic and random dynamics of the environment objects etc.). Consequently, each autonomous robot must be able to explore autonomously its environment, recover from failures, solve new avoidance problems and, most of all – all these tasks should be done in real-time.

Our robot learns, adapts and generalizes the knowledge related to both itself and environment with the help of a genetic algorithm and thus becomes able to execute a collision-free motion. The learning and adaptation processes are similar to those in a human being that reacts and learns through experience.

The robot

The robot has three degrees of freedom (3DOF), being able to execute two basic movements: rotation and translation. In the rear part of the robot there are placed two motor-driven wheels that provide locomotion through differential drive mechanism, see Figure 1. The two direct current engines that drive the wheels are, in turn, controlled by a microcontroller system. An unpowered wheel, placed in the frontal-central part of the robot, ensures stability. The robot has an average top speed of 0.3 m/s.

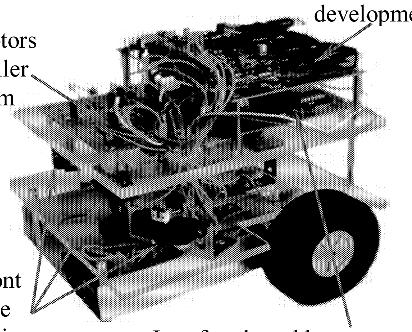


Figure 1: A Picture of the Robot Hardware Implementation

The robotic system has four infrared, IR, proximity sensors – GP2D120XJ00F. The IR sensors have an active distance measuring range from 4 up to 30 cm. Three of the IR sensors are placed in front of the robot, supervising the left, the central and the right part of the frontal environment, and one is placed in the central-rear position of the robot, see Figure 1. The sensors return a voltage value proportional to the distance to the detected obstacles. One of the main IR sensor problems is that the distance characteristic function is a non-linear one. As a result, the sensor characteristic was linearized based on a regression algorithm (Dobrea and Dobrea, 2010a).

The robot sustains the GA auto-organizing controller that is built on a powerful 32-bit MCF5213 microcontroller. The microcontroller belongs to ColdFire^{TM} family and it is based on a RISC architecture comprising a large number of peripheral equipments like eight PWM channels, four 32-bit timers with DMA request capability, eight channels ADC, 3 UARTs, 1 CAN etc.

The robot is designed to learn, in an adaptive manner, based on the information supplied by the sensors. The speed of the wheels is updated at 3.3 Hz, through two PWM channels that command two H-bridges. The program is developed in C language, using the CodeWarrier IDE. More, we use the microcontroller interrupt sub-system in order to reduce the computational burden.

Genetic algorithm

Genetic algorithms (GA) are global search and optimization techniques inspired from the natural selection mechanism existing in the nature.

A GA is based on a population of candidate solutions, called chromosomes. Each chromosome is evaluated and ranked by a fitness evaluation function. The fitness function provides information of how good each chromosome is. The evolution of the GA from a generation to the next one involves three steps: fitness evaluation, chromosome selection and building the next generation (mainly based on GA operators like reproduction and mutation). The next generation is created with the only goal of improving the population fitness.

The avoidance algorithm

The avoidance algorithm has a GA as a main engine. In one chromosome we code left and right engine commands, using for this two signed chars variables; thus, c_j = {$left\_eng_j$, $right\_eng_j$} for the j^{th} chromosome. As a result, each chromosome is represented on 16 bits. Both chromosome sections, of 8 bits each, representing the left and the right engine commands, can take values only within the interval [–128, 127]. The values for the engine commands have the following significance: a value of **+127** denotes full forward power engine, **–128** represents full back power engine and **0** corresponds to a stop command for the engine.

The values, s_i, of the IR sensors are first acquired, linearized (Dobrea and Dobrea, 2010a) and, finally, normalized in [0, 0.9] range, with 0 denoting no obstacle and 0.9 denoting an imminent collision; more, due to the noise, each sensor value smaller than 0.02 is forced to take a 0 value. The analytic form of the fitness function, $f(\cdot)$, is presented in equation (1). The relation (1) complies with the fundamental paradigm that defines the fitness in the genetic algorithm field: the fitness takes the lowest value (zero in our situation) only when a chromosome successfully solves the problem. In other case, the fitness measures the ability of each chromosome to solve more or less the problem –

being in a direct proportional relationship with the vicinity of an obstacle, as it is in our case.

$$f_j(c_j[n]) = \frac{1}{4}\sum_{i=1}^{4} s_i[n] \qquad (1)$$

After 300 ms of robot movement, based on the c_j chromosome containing the engine commands, the fitness is calculated. A collision-free motion, specified by a chromosome c_j, is characterized by values of 0 for all four sensors (this corresponds to the "no obstacle" case, in the vicinity of the robot); in such a situation the fitness function, $f(c_j[n])$, is also zero.

To obtain – based on a continue interaction of the robot with its environment, and without any human intervention – a set of adaptive movements rules, in a **first main stage** the robot has to go straightforward. When the robot comes close enough to an obstacle (i.e. the obstacle lies in the active range of its sensors), the GA starts to find the best solution in order to avoid the obstacle; this last case corresponds to the **second main stage** of the algorithm. In this last stage, the most important goal of the real-coded genetic algorithm is to find the best chromosome(s) that encapsulates those engine commands that minimize the fitness function. As a result, the GA finds the best way to avoid the obstacle.

The GA works with population of chromosomes. In other researches, each chromosome characterizes a single robot (Messom, 2002). In our case, we have only one robot that operates the GA, which can be considered a major practical limitation. To avoid this limitation, in a first step, the robot will move in one direction based on the engine commands encapsulated in the first chromosome from the population. After 300 ms the robot stops and the fitness value associated with the chromosome is determined. Then, the robot comes back into the initial position and the algorithm proceeds, in the same way, for the others chromosomes.

Table 1: Movement Knowledge Database – Real Recordings

Sensors				Engines	
Left	Center	Right	Back	Left	Right
⋮	⋮	⋮	⋮	⋮	⋮
0.54	0.19	0	0	105	-100
0	0	0.57	0	-109	117
0.49	0.37	0	0	103	-109
0	0	0.61	0	47	85
0.34	0	0	0	93	10
0	0	0.41	0	-104	56
0.43	0.62	0.31	0	-128	-72
0	0	0.40	0.32	75	91
⋮	⋮	⋮	⋮	⋮	⋮

After each generation, the fitness value(s) of the best chromosome(s) from the entire population improves. At the end, the GA obtains the best obstacle-avoiding solution(s) for that obstacle-encountered particular situation. Based on the best-obtained chromosome, the robot moves accordingly and it correctly navigates without any kind of collision. After it successfully avoids the obstacle, the robot switches back to the **first main stage** of the algorithm, moving only in a forward direction. However, this will happen only up to the moment when a new obstacle will come into the sight of the sensors range. In this last case, the robot will switch to the **second main stage** of the algorithm, in which the GA will have to solve the new avoidance problem. Several of these complete stages will compose the so-called **learning phase** of the robot (we will discuss this, in detail, later).

Each time when the GA solves an avoiding task, the obtained solution is saved into a table – the so-named **movement knowledge database**. Each row of this table contains the sensors information preceding the GA execution (four values – a value for each sensor) and the information stored into the best chromosome (two values – a value for each engine command), see Table 1.

In the **avoiding phase**, when an obstacle comes across, a Euclidean distance is computed between the actual sensor values of the robot and the sensor values stored in the robot movement knowledge database. If one of the computed distances is lower than a predefined threshold the associated engine commands are executed. Otherwise, the robot enters into the **learning phase** and a new movement rule is extracted and locally stored in the movement knowledge database. In this way, the self-evolving controller becomes endowed with the capacity to adjust itself to any new previously unseen situation encountered in real environments. More, from the cases presented above one can deduce that the avoiding phase can take place with or without the learning process, this fact being elected by the meeting of the predefined threshold criteria. The **learning phase** is composed of several GA instances. The number of the GA instances is equal with the number of the rules from the **movement knowledge database.**

RESULTS AND DISCUSSIONS

The first implementation of the self-evolving controller

In a first attempt, all the steps previously presented were implemented and the self-evolving controller of the real robot (see Figure 1) was tested. The method we chose for the chromosome selection was the stochastic universal sampling, which exhibits both no bias and a minimal spread (Baker, 1987). The crossover method was two points. Table 2 shows the list of the main parameters of the GA algorithm as they were implemented on the robot.

Table 2: The Main Parameters of the Genetic Algorithm

Parameter	Value
Population Size	10
Generations	20
Generation gap	0.9
Probability of Mutation	0.0437
Crossover Points	2
Crossover Probability	0.7

As a result, we obtained finally a functional prototype. One big disadvantage of the self-evolving GA implemented controller was the significant computation time spent by the robot in the learning phase. In its turn, this large learning time generated other additional problems like the wear and tear of the mechanical parts of the robot as well as the necessity of a large number of rechargeable cycles for the batteries. To reduce this time we made use of some practical observations. For illustration, not every time the GA needed 20 generations to converge. In many situations after fewer generations (in several cases even in the first generation) the fitness function took a zero value. Based on this observation we imposed the GA to stop whenever the fitness on a single chromosome took a zero value. In this mode, the learning time was considerable diminished.

However, even after all the software enhancements the time elapsed in the **learning phase** remained high for our real implemented robotic system. Moreover, a robotic system tested in real environments usually raises some difficulties mainly related to the possibility to make debug on the software modules. These difficulties cumulated with the long time required for each practical testing made our first solution to be quite impractical. Therefore, an alternative approach had to be considered.

In order to overcome all the above-presented disadvantages and thus, to further be able to confirm the main concepts of the new-introduced self-evolving GA controller, we chose the following complementary approach that speeds up the development process of the robotic system: (1) **the (**main part of the) **learning phase** is done in a simulation environment that finally provides the extracted movement knowledge database, and (2) **the avoiding phase** (we mean by this the testing of the controller) is done in a real environment, using for this the previous extracted movement knowledge database.

The simulator

One problem with the new approach used to extract the movement knowledge database resides in the possibility of the self-evolving controller to learn an avoiding behavior that is adjusted to the particular features of the simulated robotic system – features that does not necessary has an identical correspondent in the real world.

To avoid this problem we had to model both the robot (its mechanical, electrical and electronically characteristics) and the environment, as realistic as possible. To achieve this goal we chose the MobotSim simulator, version 1.0.03. The MobotSim is a configurable 2D simulator of differential drive mobile robots. The core language for MobotSim is the Sax Basic language – a Visual Basic for Applications^{TM} compatible language. In the MobotSim environment we configured for the robot the following parameters: the platform diameter, the distance between wheels, the wheels diameter, the number of the sensors, the angle between the sensors, the sensor ring radius, the radiation cone, the sensors range and the percentage of misreading. All of these parameters were set accordingly with the real robot mechanical and electronic features.

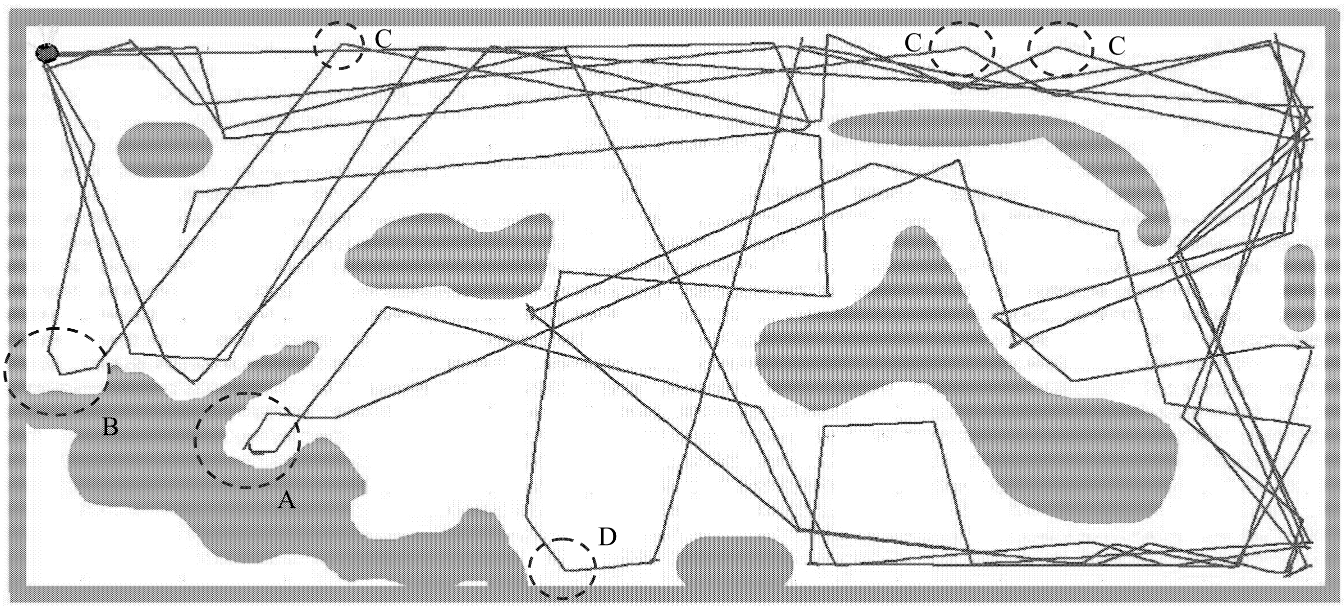


Figure 2: The **Learning Phase** of the Robot

The second implementation of the self-evolving controller

In the second implementation of the robot we split, as previously presented, the learning phase and the avoiding phase. For the learning phase, we used a simulator environment while the avoiding phase was tested in real environments.

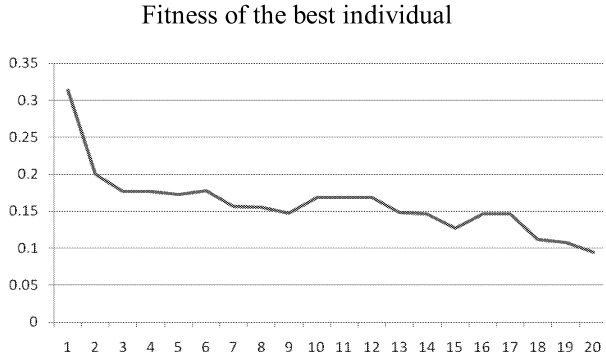


Figure 3: The Evolution of the Fitness for the Best Chromosome in Each Generation – Case D in Figure 2

Figure 2 presents the results obtained in the **learning phase**, on more than 200 GA instances. The time needed to obtain this figure on an Intel(R) Core(TM) i7 CPU at 2.8 GHz, with 8 GB of RAM was 12 hours and 24 minute. Figure 3 presents the evolutions of the GA (i.e. the evolution of the fitness of the best chromosome on each generation) that correspond to the case D displayed in Figure 2. Out of this figure, one can notice the ability of the GA algorithm to improve the avoidance performances and to solve each of the particular indicated obstacle-avoidance tasks.
One of the main parameters of the GA (with an important impact on both the convergence time to a solution and the quality of that solution) is the population size. This is even more critical in our case because when a new, previously unseen situation is encountered by the robot a local GA have to solve the new avoiding task; in this last case and, especially when speaking about real applications – a fast convergence time and a good avoiding solution provided by the GA algorithm proves to be of great value in critical situations. After a large number of tests, a population of 10 chromosomes seemed to comply with the both previously presented constrains.

In Figure 2 one can observe that the GA self-evolving controller is able to solve all the collision situations, like the ones when the robot is placed in corners or even in the hard situations similar with the ones marked with A and B. The ability of the real robot to solve real situations is given by the diversity of the situations encapsulated in the **movement knowledge database**. From this reason, the **learning phase** has to be stopped not when a time value is elapsed or when a number of GA instances is reached. The learning phase will be considered completed only when the simulated robot had been already solved a large number of different types of obstacle avoidance cases. Otherwise, the **movement knowledge database** will comprise only solutions for the most common encountered avoiding situations – as the one marked with C in Figure 2.

The learning environment for the robot consisted in a delimited zone (of around 2 m x 2 m) having one obstacle randomly placed within. This learning zone was build from a number of boxes, of different dimensions, placed as to obtain a close perimeter. From the entire movement knowledge database obtained in the learning process presented in Figure 2, in the real robot only the first 50 instances of the GA were used. The first 50 rules includes A and B avoidance cases, presented in Figure 2.

Based on these rules the robot was able to navigate correctly in the real environment, without any kind of collision. A very interesting characteristic consists in the fact that the robot (with only 50 movement knowledge rules) never triggered the GA in the real environment. More, comparing to other existing approaches, one of the most important features of this robotic self-evolving controller remains its ability to navigate in real environment without any kind of collision which still exists in the learning phase of other reported solutions based on ANNs (Dobrea and Dobrea, 2010a), (Dobrea and Dobrea, 2010b), (Tan et al. 2008).

CONCLUSIONS

This paper reports the development of a new intelligent self-evolving controller for a robotic system able to auto-evolve to a behavior that allows to sense, reason and act – all these in order to avoid all obstacles from a real world environment.

As we mentioned previously, a single evaluation of the entire population, even for a small population of 10 chromosomes, requires significant computation time. From this reason, the main set of movement knowledge database was extracted inside a **simulation environment**. This set of rules supports, on the first stage, the robot dynamics. In the second stage, if an unseen situation occurs, a genetic algorithm, similar with the one implemented in the simulation environment, will solve the problem and the new extracted rule will be used in order to avoid in the future similar situations.

In classical approaches, transferring controllers evolved in simulation environments to physical robots is a very difficult task (Messom, 2002). In our approach, to transfer the **movement knowledge database** obtained inside a simulation environment to the real robot was a success and it represents a new way to solve this problem. The approach presented in this paper has another big advantage: it is very easy to implement it in real robotic systems as it has already been proven.

ACKNOWLEDGEMENT

This work was supported by the CNCSIS – UEFISCSU project number PN II – IDEI 1552/2008.

REFERENCES

Tan, S.; Yang, S.X.; and A.M. Zhu. 2011. "A Novel GA-Based Fuzzy Controller for Mobile Robots on Dynamic Environments with Moving Obstacles.", International Journal of Robotics & Automation, Vol. 26, Issue 2, 212-228

Chiou, J.S.; Wang, C.J.; Wang K.Y.; Hu, Y.C.; Cheng, S.W.; and C.H. Chen. 2010. "Hybrid Algorithm of FLC Design for Robot Soccer.", International Journal of Nonlinear Sciences and Numerical Simulation, Vol. 11, 119-122

Martíneza, R.; Castilloa, O.; and L.T. Aguilarb. 2009. "Optimization of interval type-2 fuzzy logic controllers for a perturbed autonomous wheeled mobile robot using genetic algorithms.", Information Sciences, Vol. 179, Issue 13, 2158-2174

Hosseinzadeh, A. and H. Izadkhah. 2010. "Evolutionary Approach for Mobile Robot Path Planning in Complex Environment.", International Journal of Computer Science Issues, Vol. 7, Issue 4, No 8, 1-9.

Repoussis, P.P.; Tarantilis, C.D.; and G. Ioannou. 2009. "Arc-Guided Evolutionary Algorithm for the Vehicle Routing Problem With Time Windows.", IEEE Transactions on Evolutionary Computation, Vol. 13, Issue 3 (June), 624 - 647

Dobrea D.M. and M.C. Dobrea. 2010a. "An auto-organization bio-inspired robotic system.", *In Proceedings of the International Conference on Future Information Technology* (Changsha, Ch, Dec. 14-15). IEEE, Picataway, N.J., 354-358.

Baker J.E., 1987. "Reducing bias and inefficiency in the selection algorithm.", *In Proceedings of the Second International Conference on Genetic Algorithms* (Cambridge, USA). L. Erlbaum Associates Inc. Hillsdale, NJ, pp. 14-21, July 1987

Tan A.H.; Lu N.; and D. Xiao. 2008. "Integrating Temporal Difference Methods and Self-Organizing Neural Networks for Reinforcement Learning With Delayed Evalutive Feedback.", IEEE Transaction on Neural Network, Vol. 19, 230-244

Messom C. 2002. "Genetic Algorithms for Auto-tuning Mobile Robot Motion Control.", Research Letters in the Information and Mathematical Sciences, Vol. 3, 129-134

Dobrea D.M. and M.C. Dobrea. 2010b. "An Autonomous Robotic System", *In Proceedings of the 9th International Symposium on Electronics and Telecommunications* (Timiș oara, Ro, Nov. 11-12) . IEEE, Picataway, N.J., 107-110

BIOLOGICAL SIMULATION

MODELLING OF THE BIODYNAMIC RESPONSES OF THE FEET TO VERTICAL VIBRATION

Naser Nawayseh
Mechanical and Mechatronics Engineering Department,
College of Engineering, Dhofar University,
PO Box 2509, Postal Code 211
Salalah, Oman
Email: nnawayseh@yahoo.co.uk

INTRODUCTION

The biodynamic responses of the human body to vibration can be broadly presented by two different frequency response functions. The first function is called the driving point function and is obtained from measurement of two different parameters, such as acceleration and force, at the same point, usually the point of contact between the body and the vibrating surface. The apparent mass (the complex ratio of force to acceleration) and mechanical impedance (the complex ratio of force to velocity) have been used as driving point biodynamic response functions during vertical vibration (e.g. Holmlund et al. 2000; Nawayseh and Griffin 2003) and horizontal vibration (e.g. Nawayseh and Griffin 2005a).

The second function is called the transmissibility and is obtained from measurement of the same parameter at two different locations. For example, the transmission of vibration through the human body can be obtained by measuring the acceleration at the seat surface and the vertical, horizontal and/or rotational acceleration on the head (e.g. Paddan and Griffin 1988a; Paddan and Griffin 1988b).

Absorbed power is another form of biodynamic responses that has received attention recently (e.g. Mansfield et al. 2001; Nawayseh and Griffin 2010) although some studies on absorbed power were published in the sixties (e.g. Pradko et al. 1966).

The majority of the studies on biodynamic responses to vibration focused on measuring the biodynamic responses at the seat surface (e.g. Boileau et al. 1998) while a few studies measured the apparent mass at other interfaces between the human body and the vibrating surface such as at the backrest (e.g. Nawayseh and Griffin 2005b) and the footrest (e.g. Nawayseh and Griffin 2003). Contact with a backrest and footrest showed significant effect on the apparent mass and the cross-axis apparent mass measured on the seat. In some cases (especially with off-road vehicles) high relative motion exists between the seat (the body) and the feet and in order to understand the response of the whole coupled system to vibration and its relation to comfort, all parts of the system interfacing with a vibration surface may have to be included in the model. Despite that, the human body models in response to vibration are mostly for apparent mass or mechanical impedance measured on the seat surface (e.g. Broman et al. 1996; Kitazaki and Griffin 1997; Matsumoto and Griffin, 2001) while hardly any modelling can be found for the apparent mass measured at the backrest or footrest.

In this paper, a mathematical model that can predict the response of the feet to vertical vibration is presented. The effect of the vibration magnitude and the effect of the sitting posture on the parameters of the model will be studied.

THE EXPERIMENTAL DATA

The model presented in this paper is used to predict the apparent mass measured experimentally with three different sitting postures and three different vibration magnitudes (Nawayseh and Griffin 2003) in order to study the effect of posture and vibration magnitude on model parameters. The sitting postures were the maximum thigh contact posture (Max) where the subjects sat on a seat with their heels just in touch with a footrest, the average thigh contact posture (Ave) where the subjects sat on the seat with their upper legs horizontal and lower legs vertical, and the minimum thigh contact posture (Min) where the height of the footrest increased by 160 mm from the average thigh contact posture position. Three vibration magnitudes were studied with the average thigh contact posture only; 0.25, 0.625, and 1.25 ms^{-2} r.m.s. More details of the experimental setup can be found in (Nawayseh and Griffin 2003).

THE PROPOSED MODEL

The model used in this study is a quantitative two-degree-of-freedom model with only vertical degrees of freedom (Figure 1). Although some individual experimental data showed more than two peaks, which implies the need of more than two-degree of freedom model, two of the three peaks were more pronounced than the third. Hence, only two-degree of freedom model will be used to summarise the previously measured data. Since this model is not a mechanistic model, none of the parameters of the model is intended to represent part of the lower body segments. The models parameters are described as follows:

m_1 and m_2 are the masses of mass 1 and mass 2, respectively.
k_1 and c_1 are the vertical stiffness and damping beneath mass 1.
k_2 and c_2 are the vertical stiffness and damping beneath mass 2.
z_1 represent the vertical motion of mass 1
z_2 represents the vertical motion of mass 2.
z_b represent the vertical motion of the footrest where the force and acceleration were measured.

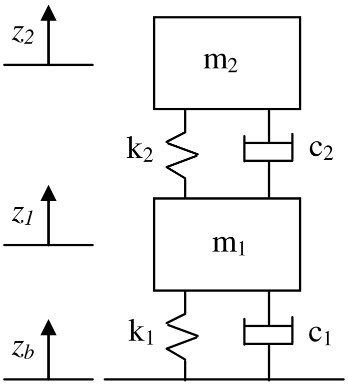


Figure 1 Two-degree of freedom model.

Mathematical equations of the model

The equations of motion of the model were derived using Newton's second law of motion. The following two equations describe the motion of the model:

$$m_1 \frac{\mathrm{d}^2 z_1}{\mathrm{d}t^2} + c_1\left(\frac{\mathrm{d}z_1}{\mathrm{d}t} - \frac{\mathrm{d}z_b}{\mathrm{d}t}\right) + c_2\left(\frac{\mathrm{d}z_1}{\mathrm{d}t} - \frac{\mathrm{d}z_2}{\mathrm{d}t}\right) + k_1(z_1 - z_b) + k_2(z_1 - z_2) = 0$$

$$m_2 \frac{\mathrm{d}^2 z_2}{\mathrm{d}t^2} + c_2\left(\frac{\mathrm{d}z_2}{\mathrm{d}t} - \frac{\mathrm{d}z_1}{\mathrm{d}t}\right) + k_2(z_2 - z_1) = 0$$

The total vertical force of the model at the footrest level can be found from:

$$f_z(t) = m_1 \frac{\mathrm{d}^2 z_1}{\mathrm{d}t^2} + m_2 \frac{\mathrm{d}^2 z_2}{\mathrm{d}t^2}$$

The apparent mass can be calculated by finding the Laplace transform of the above three equations assuming zero initial conditions and solve for the complex ratio between $F_z(s)$ and $Z_b(s)$ where $F_z(s)$ is the Laplace transform of the total vertical force, $f_z(t)$, $Z_b(s)$ is the Laplace transform of the vertical displacement of the base of the model, $z_b(t)$, and s is $j\omega$ where j is $\sqrt{-1}$ and ω is the angular frequency. The apparent mass (AM) is:

$$AM = m_1 \frac{BD}{DA - C^2} + m_2 \frac{BC}{DA - C^2}$$ Where,

$A = m_1 s^2 + c_1 s + c_2 s + k_1 + k_2$
$B = c_1 s + k_1$
$C = c_2 s + k_2$
$D = m_2 s^2 + c_2 s + k_2$

The apparent mass modulus and phase can be calculated from the real and imaginary parts of the apparent mass complex function as:

$$AM_{\mathrm{mod}}(\omega) = \sqrt{(ReAM(\omega))^2 + (ImAM(\omega))^2}$$

$$AM_{\mathrm{ph}}(\omega) = \tan^{-1} \frac{ImAM(\omega)}{ReAM(\omega)}$$

Optimisation of the parameters of the model

All model parameters (i.e. masses, stiffness, damping coefficients) were optimised using the Nelder Meade simplex search method within MATLAB. The parameters of the model were optimised by comparing the experimentally measured moduli and phases of the vertical apparent mass (Nawayseh and Griffin 2003) with the calculated moduli and phases from the model as shown in the following error equation:

$$\mathrm{err} = w_1 \sum_{i=1}^{N} (AM_{\mathrm{mod\_c}} - AM_{\mathrm{mod\_m}})_i^2 + w_2 \sum_{i=1}^{N} (AM_{\mathrm{ph\_c}} - AM_{\mathrm{ph\_m}})_i^2$$

where,

$AM_{\text{mod-c}}$ and $AM_{\text{mod-m}}$ are the moduli of the calculated and measured vertical apparent masses, respectively.
$AM_{\text{ph-c}}$ and $AM_{\text{ph-m}}$ are the phases of the calculated and measured vertical apparent masses, respectively.
w_1 and w_2 are arbitrary weighting factors employed to improve the prediction.

RESULTS AND DISCUSSION

Prediction of individual data

The model was employed to predict the moduli and phases of the vertical apparent masses of 10 subjects seated on a rigid flat vibrating surface and adopting the average thigh contact posture (Ave). The calculated (i.e. from the model) and measured (i.e. from the experiment) apparent masses were in good agreement for both the moduli (Figure 2) and phases (Figure 3). The model replicates the two peaks in the data. Some subjects showed 3 peaks (e.g. Subject 5 and Subject 10 but the model can produce only two of them as the model has just two degrees of freedom. The high variability in each parameter of the model (Table 1) is not surprising given the high inter-subject variability shown in Figure 2 and Figure 3 produced by using subjects with different mass, build, stature, and age.

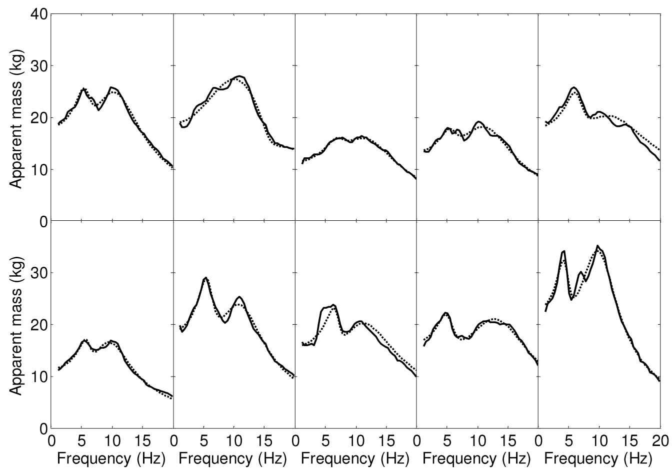


Figure 2 Measured and calculated apparent mass magnitude of the feet of 10 subjects with the average thigh contact posture at 1.25 ms^{-2} r.m.s. ——, measured; ·······, calculated.

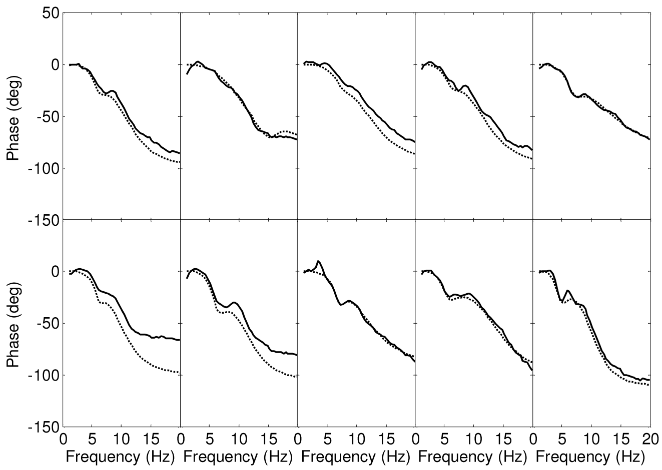


Figure 3 Measured and calculated apparent mass phase of the feet of 10 subjects with the average thigh contact posture at 1.25 ms^{-2} r.m.s. ——, measured; ·······, calculated.

Table 1 Model parameters for 10 subjects adopting the average thigh contact posture.

Subject	m_1	m_2	k_1	k_2	c_1	c_2
	kg	kg	N/m	N/m	Ns/m	Ns/m
1	16.3	1.9	78069	2751	856	35
2	17.5	1.2	102840	11352	1544	28
3	10.7	0.7	64197	1631	693	18
4	12.0	1.4	67204	2081	701	32
5	16.0	2.2	101890	4117	1377	47
6	10.6	0.8	41839	1227	489	11
7	15.6	3.1	77009	4799	711	60
8	14.5	1.2	81223	2467	991	17
9	13.6	3.2	105010	3770	875	66
10	19.4	3.8	79331	3236	712	56
Median	15.1	1.7	78700	2993	784	33

The optimised parameters of the model for 10 subjects are shown in Table 1. The mass of Mass 1 of the model was much greater than the mass of Mass 2. On average, the mass of Mass 1 accounted for 89% of the total mass of the model while the mass of Mass 2 accounted for only 11% of the total mass of the model. The stiffness and damping coefficient beneath Mass 1 were much greater than those beneath Mass 2; on average the stiffness and damping coefficient beneath Mass 2 were only 4.4% of the stiffness and damping coefficient beneath Mass 1 (Table 1).

The stiffness and damping coefficient beneath Mass 1 (if thought of as representing those beneath the foot) are approximately one magnitude order smaller than those found by modelling the dynamic response of standing subjects (Subashi et al. 2008). The greater parameters obtained when standing could be attributed to the more weight on the feet of the subjects in that position.

Effect of sitting posture on model parameters

The data for subject 10 was used to investigate the effect of the sitting posture on the parameters of the model. It is a usual practice to use the median or the mean response to do such investigation. However, with a multi-peak response such as that of the average thigh contact posture and the minimum thigh contact posture, where the peaks happening at different frequencies for different subjects, the characteristics of the response might be concealed due to the averaging process. The choice of subject 10 was arbitrary but one should expect that the behaviour to be the same with any other subject as the studied model is a linear model.

Figure 4 shows good agreements between the measured and predicted apparent masses (moduli and phases) of the feet with three different sitting postures. Table 2 presents the optimised parameters for each posture. The total mass of the model represents the mass of the feet, the mass of the lower legs, and the mass of part of the upper legs carried by the footrest. The change in the total mass of the model is consistent with the change in the total mass carried by the footrest as a result of change in the sitting posture; with the Max sitting posture (where the heel are just in touch with the footrest), the mass on the footrest will be least. With increasing the height of the footrest more mass will be supported on the footrest.

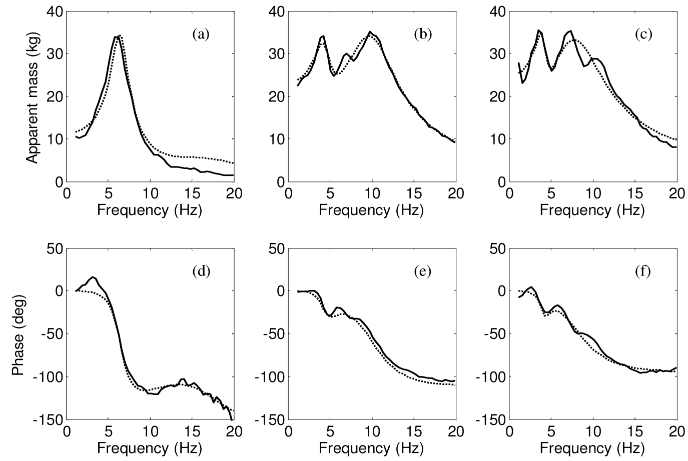


Figure 4 Measured and calculated apparent mass moduli and phases with three different sitting postures. (a) and (d), maximum thigh contact posture; (b) and (e), average thigh contact posture; (c) and (f), minimum thigh contact posture. ——, measured; ·······, calculated. (Data for subject 10 at 1.25 ms^{-2} r.m.s).

Table 2 Effect of the sitting posture on the parameters of the model. Parameters are for subject 10 at 1.25 ms^{-2} r.m.s. "Max" refers to maximum thigh contact; "Ave" refers to average thigh contact; "Min" refers to minimum thigh contact.

Posture	m_1	m_2	***Total mass*** (m_1+m_2)	k_1	k_2	c_1	c_2
	kg	kg	kg	N/m	N/m	Ns/m	Ns/m
Max	4.6	6.7	11.3	40606	16026	39	197
Ave	19.4	3.8	23.2	79331	3236	712	56
Min	22.3	2.3	24.6	62188	1665	973	20

Although the model is not a mechanistic model, one might think of the stiffness and damping coefficient beneath Mass 1 as being those beneath the feet. The increase in k_1 and c_1 with the change of posture from Max to Ave is consistent with the increase in stiffness with more applied load and increase in damping with the increase in support of the legs. The decrease in stiffness when changing to the Min posture might be attributed to the significant effect of legs angle as reported previously (Kitazaki 1997).

Effect of vibration magnitude on model parameters

The apparent mass of the feet of subject 10 was also used to investigate the effect of vibration magnitude on the parameters of the model (Figure 5). The model predicts well two of the three apparent mass peaks found experimentally for the feet response. It has been shown experimentally that the human body behaves non-linearly when exposed to vibration; the frequency of the peaks decreases with the increase in vibration magnitude (Nawayseh and Griffin 2003). As the natural frequency (and resonance frequency) of any system is related to its mass and stiffness (for simplicity think of a single degree of freedom system), one would expect that for the model to replicate the effect of vibration magnitude on the frequency of the peak (i.e. increase in vibration magnitude decrease the peak frequency), the mass of the system should increase and/or the stiffness of the system should decrease with the increase in vibration magnitude. However, the mass of the feet, lower legs and the parts of the upper legs supported on the footrest is constant and does not depend on the vibration magnitude. The increase in the mass of the system with the increase in vibration magnitude (Table 3) is insignificant as a 400% increase in vibration magnitude induced an increase of less than 9% in the total mass of the system.

The effect of vibration magnitude on the stiffness of the model is shown in Table 3. The increase in vibration magnitude caused the stiffness of the system to decrease which in turn reduced the peak frequency following the same trend found experimentally. The increase in vibration magnitude from 0.25 to 1.25 ms^{-2} r.m.s. reduced k_1 by about 28% and reduced k_2 by about 56%. The increase in vibration magnitude also increased c_1 by 35% which helped in reducing the frequency of the peak while there was no trend for the effect of vibration magnitude on c_2.

Table 3 Effect of the vibration magnitude on the parameters of the model. Parameters are for subject 10 with the average thigh contact posture.

Vibration Magnitude ms^{-2} r.m.s.	m_1 kg	m_2 kg	***Total mass*** (m_1+m_2) kg	k_1 N/m	k_2 N/m	c_1 Ns/m	c_2 Ns/m
0.25	16.9	4.4	21.3	109420	7282	530	79
0.625	18.0	4.9	22.9	96630	5927	578	90
1.25	19.4	3.8	23.2	79331	3236	712	56

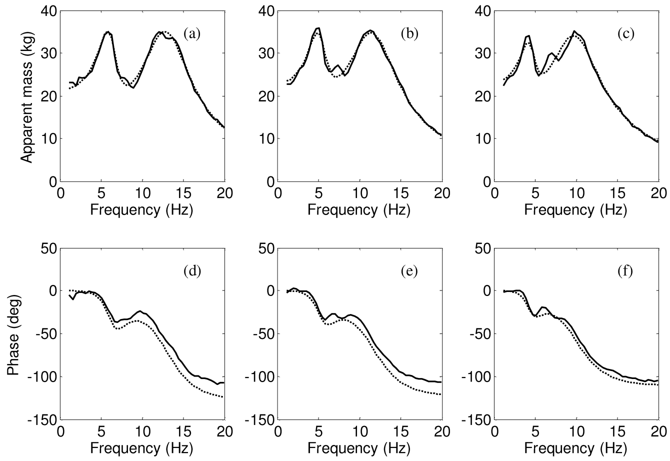


Figure 5 Measured and calculated apparent mass moduli and phases with three different vibration magnitude. (a) and (d), 0.25 ms^{-2} r.m.s.; (b) and (e), 0.625 ms^{-2} r.m.s; (c) and (f), 1.25 ms^{-2} r.m.s. ——, measured;········ , calculated. (Data for subject 10 with the average thigh contact posture).

The parameters of the model presented in this paper are applicable to the foot positions and the vibration magnitudes used in this study. More work should be carried out to identify the parameters of the model in foot positions such as that adopted by car drivers. The mathematical models of the apparent mass of the feet can be used to design vehicle floors that attenuate or prevent the vibration from reaching the human body through the car floor-feet interface.

CONCLUSIONS

A two-degree of freedom model was found to predict well the moduli and phases of the apparent mass of the feet. The effect of the sitting posture and the vibration magnitude on the model parameters was studied. The effects were found to be consistent with the experimental observations reported in previous studies.

REFERENCES

Boileau, P.E. and S. Rakheja. 1998. “Whole-body vertical biodynamic response characteristics of the seated vehicle driver-Measurement and model development.” *International Journal of Industrial Ergonomics* 22, 449-472.

Broman, H., M. Pope. and T. Hansson. 1996. “A mathematical model of the impact response of the seated subject.” *Medical Engineering and Physics* 18, 410-419.

Holmlund P, R. Lundström and L. Lindberg. 2000. "Mechanical impedance of the human body in vertical direction". *Applied Ergonomics* 31, 415-422.

Kitazaki S. 1997. "The apparent mass of the foot and prediction of floor carpet transfer function." In proceeding of 1997 *United Kingdom Group Meeting on Human Response to Vibration*, 17-19 Sept, University of Southampton, Southampton, United Kingdom,.

Kitazaki, S. and M.J. Griffin. 1997. “A modal analysis of whole-body vertical vibration using a finite element model of the human body.” *Journal of Sound and Vibration* 200, 83-103.

Mansfield, N.J., P. Holmlund, and R. Lundström. 2001. "Apparent mass and absorbed power during exposure to whole-body vibration and repeated shocks." *Journal of Sound and Vibration* 248, 427-440.

Matsumoto, Y. and M.J. Griffin. 2001. “Modelling the dynamic mechanisms associated with the principal resonance of the seated human body.” *Clinical Biomechanics* 16, 31-44.

Nawayseh, N. and M.J. Griffin. 2003. “Non-linear dual axis biodynamic response to vertical whole-body vibration.” *Journal of Sound and Vibration* 268, 503-523.

Nawayseh, N and M.J. Griffin. 2005a. “Non-linear dual axis biodynamic response to fore-and-aft whole-body vibration.” *Journal of Sound and Vibration* 282, 831-862.

Nawayseh. N and M.J. Griffin, 2005b. "Tri-axial forces at the seat and backrest during whole-body fore-and-aft vibration". *Journal of Sound and Vibration* 281, 921-942.

Paddan, G.S. and M.J. Griffin. 1988a. "The transmission of translational seat vibration to the head.I. Vertical seat vibration." *Journal of Biomechanics* 21, 191-197.

Paddan, G.S. and M.J. Griffin. 1988b. "The transmission of translational seat vibration to the head. II. Horizontal seat vibration." *Journal of Biomechanics* 21, 199-206.

Pradko, F, R. Lee and V. Kaluza. 1966. "Theory of human vibration response". ASME. Paper No. 66-WA/BHF-15.

Subashi, G.H.M.J., Y. Matsumoto and M.J. Griffin. 2008. "Modelling resonances of the standing body exposed to vertical whole-body vibration: Effect of posture." Journal of Sound and Vibration 317, 400-418.

AUTHOR BIOGRAPHY

NASER NAWAYSEH was born in Jordan and studied Mechanical Engineering at the University of Jordan and graduated in 1993. He obtained his MSc by research in Solar Water Desalination in 1996 and published several journal and conference papers in this field. He worked as a site engineer, as a design engineer and as a lecturer between 1993 and 2000. In 2000, he started his PhD on biodynamic responses to vibration at the Institute of Sound and Vibration Research (ISVR) of the University of Southampton in The United kingdom and graduated in 2004. He was, immediately appointed as a Research Fellow at the Institute of Sound and Vibration Research where he worked for 3 years on different small and big European and International projects investigating biodynamic responses, seating dynamics, postural stability, motion sickness and performance of anti-vibration gloves. He published several journal papers, conference papers, and technical reports in those areas. In September 2007, he joined Dhofar University in Oman as an Assistant Professor and he has been the Chairperson of the Mechanical and Mechatronics Engineering Department at Dhofar University since July 2010.

TERRITORIAL INTELLIGENCE ENGINEERING

European Structural Funds and regional interactions, which convergences for the European regions?

Bernard Elissalde
Patrice Langlois
Dominique Goyat
University of Rouen (France),
UMR CNRS IDEES (6228)
E-mail: bernard.elissalde@univ-rouen.fr

KEYWORDS
cellular automata, regional policy, regions, regional inequalities

ABSTRACT

The aim of this paper is about the trends of regional disparities in the European Union who can be considered as a complex system.. For modelling the uncertain efficiency of the regional policy we attempted to use a cellular automata (Geocells) developed by P.Langlois). Methodologically, this cellular automata Geocells is based on interrelated processes between variables (like time periods, growth rates in the GDP per head, flows of public investments) and three geographical levels (european level, national level and regional level). This three levels are used to lay the emphasis on the fact that the EU structural expenditure, and its spatial impact, work as a rules-based system. Simulations were made, in order to evaluate, on the one hand, the specific role of each level and each variable, and on the other hand how some change in one part affects the whole. In this perspective we could underline the role of increasing or decreasing budgets, the weigth of national policies and national economic trends and the contiguity effect according to the geographical location of each eligible region.

1. INTRODUCTION

The aim of this paper is to present the functioning of GeoCells : the geographic simulation platform, as well as a first application enabling to model the European regions' behaviour according to the variation in aid granted by the European Union and to neighbourhood effects. The platform's overall diagram relies on the use of cellular automata techniques.

The taking into account of regional disparities by the European Commission's structural policies, the analysis of European regions relative positions from the angle of macroeconomic and budgetary indicators, will allow an evaluation of the overall effectiveness of the aforementioned policies, and to measure the influence in the modification of granting rules. The different settings offered by the simulation platform ensure to simulate both the impact of aid on the overall evolution of the Europe of the Fifteen regions (measured by indicators such as beta or sigma convergence), as well as on each region's specific future.

The introduction of simulation and forecasting methods in the debates on the European Union's regional policy does not aim to enable to find the one and only response to the problem of European regions' unequal development, but to suggest a range of credible options as a decision support tool for territorial solidarity and economic and social cohesion, in a European space in perpetual evolution. Even though they belong to an interdependent group such as the European Union, these spatial units each own their specific trajectories, in which the reaction delays, the transformation rhythms strongly vary from one region to another. All of these differentiated laps of time will build a European regional mosaic, making it unlikely to happen a mechanical adjustment between the impulsions of the European Union's regional policy (Structural Funds, Cohesion Funds, etc.) and the regional readjustment initially planned.

2. COMMUNITY SOLIDARITY POLICIES AND THE REGIONS' FUTURE

The issue of the solidarity effort between State members and the regions, as well as their adequacy to the cohesion principles displayed in the European texts and treaties is at the centre of the debates on European regional policy. The main questioning is about the European public policies' ability to adjust disparities produced by the single market. How can we improve redistribution and territorial equity in an Union of low economic growth ? In such an economical context, should we limit the solidarity effort of rich countries or on the contrary emphasize it in order to accelerate the catching-up of backward regions ?

2.1 The issue of allocated Funds' effectiveness

In order to evaluated the effectiveness of passed programmes and to discuss the necessity of possible reorientations in the granting of structural Funds within the European Union, it is essential to recall budgetary stakes and a few functioning principles of European intervention as far as regional policy is concerned. Since the creation of the ERDF in 1975, redistribution rules and amounts granted, within the framework of what we put behind regional policy have changed during the last twenty years and have not affected the same regions during the same periods of time. The general conditions of eligibility have changed over the reforms, and individually for each region, because economic results made them eligible or not from one period to another. Various reforms in 1984, then in 1988, carried along by the imperative of economic and social cohesion, have changed the amount of granted budgets. A Cohesion Fund concerning four countries (Spain, Portugal, Greece, and Ireland), was born subsequently from 1994 onwards, in order to offset these countries' budgetary effort, with the aim of a convergence towards criteria of participation to the Economic and Monetary Union. These so called cohesion countries have received ever since aid of a substantial weight in their GDP.

For many years, surveys have tried to evaluate the consequences of the European economic integration consolidation over regions, as well as the effectiveness of the various European Structural Funds' reforms1. They all agree to admit, at country scale, the alternation of catching-up stages, notably for the Cohesion countries (Spain, Portugal, Greece, and Ireland), more particularly for Ireland which has reached the European average from 1997, with more uncertain stages. Regional convergence would have known a slowing down in the first half of the 1980's, then an improvement at the end of this same 1980's, followed by an overall divergence period in the 1990's. This changing diagnosis, at European scale, is, in most cases, tempered by a maintaining, ore even an increase in disparities between regions of a same country. Since regional growth dynamics are not exactly the same as the ones at a national level, we have been wondering after the works of Martin (2000), Maurel (1999), and Riou (2002) about the spatial consequences of the European economic integration and about the reasons of an unequally distributed and polarizing growth. The conditions for the reduction of this gap between the scopes of interregional solidarity and territorial cohesion and reality, outlined in many official texts such as the European Spatial Development Perspective or the Second report on economic and social cohesion (Eurostat 2000), as well as in the results of our previous works (Elissalde 2005) were the purpose of different simulations achieved in this paper.

In addition to the issue of European Structural Funds' effectiveness, this shift between scale levels makes it wonder in the future about the solutions to mobilize in order to bring down the development gaps significantly increased by the progressive transition from fifteen to twenty seven members in the European Union. The use of a simulation platform through cellular automaton aims at answering the question of knowing on which conditions (of settings in terms of budgetary redistribution), according to which duration of financial aid programs, and according to which objective levels of reduction, convergence, or adjustment, European solidarity policies could be effective.

Two hypothesis can be considered: either the results of financial aid programs at European level are due to other factors such as the differentiated effects of each country's particular national situations (lets not forget that regional aid work according to the principle of additionality), either the selectivity of growth dynamics would be due to the region's relative position inside a wider spatial unit, affected itself by a situation of growth or stagnation. With regard to the analysis presented above, we have therefore added the hypothesis that neighbourhood effects between regions played a role in the spreading of growth or in the reduction of disparities.

2. 2. The issue of regions' convergence

In order to evaluate the consequences of different parameters settings noted above on simulation results, we have used the concept of convergence and two indicators in order to refine diagnosis on the reality of disparities reducing between European regions. Following other economic works, Beine and Docquier (2000) suggest three declensions of this idea of convergence :

- absolute convergence which supposes that per capita income would converge towards each other independently from initial conditions and led policies
- conditional convergence which puts forward the hypothesis that identical territories in terms of demographic growth, public policies, but with different initial conditions, are supposed to converge towards the same stationary state the ones according to the others.
- « club » convergence which puts forward the hypothesis that convergence is not achieved in a global manner but by group of regions. Per capita incomes would not converge at an identical pace for all regions but through groups of regions, which would distance themselves by an original growth dynamic.

As for evaluation, we adopted the distinction of X.Sala-i-Martin (1996) which opposes:

- sigma convergence C_σ measuring the evolution of the GDP per capita over n years, calculated as the annual variation gap of the coefficient of variation between two periods t0 et tn, where mi is the GDP mean of the year i and σ_i its standard deviation. Thus we have:

$$C_\sigma = \frac{\frac{\sigma_n}{m_n} - \frac{\sigma_0}{m_0}}{n}$$

- beta convergence C_β which measures the relation between the GDP per capita (logarithmical) variation over a given period compared to the initial level.

If we call:

$$x_k = \log PIB_{k0}$$

the GDP neperian log of the region k for the initial year t0 and

$$y_k = \log(PIB_{ki}) - \log(PIB_{k0}) = \log\left(\frac{PIB_{ki}}{PIB_{k0}}\right)$$

the GDP logarithmical variation of k between t0 and ti, beta convergence between two periods t0 et tn, is then defined by the slope a of the straight line Y = aX + b of the variable $Y = \{y_k\}_{k=1,..,K}$ in relation to the variable $X = \{x_k\}_{k=1,..K}$. Therefore, we can write:

$$C_\beta = \frac{Cov(X,Y)}{Var(X)}$$

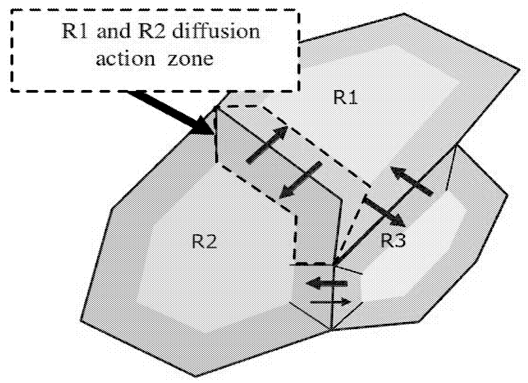
Lets note that the more these indicators are negative, the more they indicate a better convergence.

The diffusion of wealth between regions (if we suppose it exists) tends to make the regions' standards of living converge. This mechanism is opposed to the regions' internal growth, which increases exponentially from these wide range of wealth levels between regions. This growth then tends to make them diverge. The combination of these two effects is generally to the benefit of growth. Even though the growth of the wealthier improves the wealthiness of the neighbours, the gap deepens nevertheless. The financial aid policy towards the most underprivileged regions seeks therefore to counterbalance the predominant effects of this divergence. Lets note also that the way Europe measures the "convergence" (through beta-convergence) allows in theory to mask the exponential widening of absolute gaps between regions using relative gap logarithms… However, simulations show that the behaviour of these indicators are not so simple to interpret.

These indicators were already used in the research works of Charleux (2003) and Le Gallo (2006) and initially in France by F.Maurel's team (1999) from the General Commissariat for National Planning in order to test the hypothesis of an « absolute » catching-up of the most underprivileged regions, then by introducing complementary explanatory variables (ex: facilities endowment) in order to evaluate their weight in growth regional gaps. These calculations carried out by comparing separately convergence tests at States' level and at regions' level lead simultaneously to the empirical observation of a convergence between countries and of a regional divergence internal to countries.

3. MODEL ANS SIMULATIONS

3.1. Description of the growth-diffusion model between European regions



We will now clarify the diffusion model a bit unusual that we have used. Let's note Xi the GDP of the region i, Pi its population and Yi = Xi/Pi its GDP per capita at a moment t.

We put forward the following hypothesis. Each cell has the **aim** to homogenize, through time, its standard of living Y in relation to its neighbours. But the standard of living is not comparable to a physical magnitude capable of diffusing like a flow. It is through the variation of wealth (X) symbolized by the GDP (by internal growth and by diffusion) or through the variation of population (P) (also by internal growth or by migrations) that each region can work in order to achieve its goal. In this model, at first approximation, we have considered that population was constant throughout time. It is therefore here only on the variation of X that relies the diffusion mechanism in order to reach the goal.

Another hypothesis is to consider that only a small fringe close to the borderline (area in dotted line on the figure) takes part in the diffusion of wealth, by the levelling-out of standards of living of the two neighbouring border fringes. Since we do not have any information on the spatial distribution of the populations inside a region, we must put forward the hypothesis of an uniform distribution. Consequently, instead of launching forth into geometric calculations of insane zoning, we use a simple proportionality parameter, accessible in the user interface by a cursor, **called diffusion rate**, which sets the rate k (of surface area, population, and wealth all at once, since we consider them as uniformly distributed over the region's surface area), which takes part in the diffusion between regions. This rate defines therefore the part of the region's surface area matching the border fringe, in which the standard of living is going to attempt to level up through time, with the neighbouring regions' homologous fringes.

In order to model the diffusion between two regions i and j, we then introduce the coefficient kij which is the surface area's proportion i matching the intersection between the border fringe defined by k and the proportion pij of its land borderline shared by the region j , defined by $p_{ij}=\frac{l_{ij}}{\sum_{k\in N(i)} l_{ik}}$, where lij is the borderline's length between i and j.

We then have : $k_{ij}=k\,.\,p_{ij}$

If the wealth on the two sides of the border fringe between i and j was evenly distributed like connected vessels, we would obtain a levelled-out standard of living (which is not the average of the two previous standards), defined by:

$$Y_{ij}=\frac{k_{ij}.X_i+k_{ji}.X_j}{k_{ij}.P_i+k_{ji}.P_j}$$

We can then define the variation dXij (positive if it emits or negative if it receives) of the diffusion from the region i towards the region j during a short lapse of time dt as being proportional to the concerned population (kijPi) and proportional to the difference between the current standard of living (Yi) and the (local) aim of levelling-out (Yij) of standards of living i and j. This can be translated into the following equation :

$$\frac{dX_{ij}}{dt}=K\,.\,k_{ij}.P_i\left(Y_i-Y_{ij}\right)$$

The value of K is set internally (since we can already play on k).

By adding the border fringes of the region i, we note down:

$$dX_i=\sum_{j\in N(i)} dX_{ij}$$

One should notice that this diffusion is, by construction, preservative of the mean $\sum_{i=1}^{n} X_i$. (because one can verify easily that for any couple (i, j) we have : dXij + dXji = 0)

Moreover, the variable Xi is subjected to an a priori exponential internal growth, $\frac{dX_i}{dt}=C_i X_i$

Internal growth is adjustable, either individually region by region through the attribute table, either on the whole as being the same for all regions with the help of a cursor present in the user interface.

The final growth-diffusion equation is thus given by :

$$X_i(t+dt)=X_i(t)+(C_i.X_i(t)+K.k_{ij}P_i(Y_i-Y_{ij}))dt$$

The lapse of time for the discretization of growth and diffusion processes are small compared to redistributing flows, because they correspond to continuous processes. We have selected the month as lapse of time, that also matches the time unit that we chose, so dt =1. (Ci is then the twelfth of the annual growth rate).

The equation with this lapse of time is then written :

$$X_i(t+dt)=(1+C_i)X_i(t)+K.k_{ij}P_i(Y_i-Y_{ij})$$

The model introduced here attempts to give an account of the crossed recursion of the processes' effects contributing to regional dynamics : the region's peculiar growth, the redistribution mechanism linked to its membership to a wider territorial group (State, Union), and finally neighbourhood effects. Many economic models try to isolate the various sectoral variables in the final growth. Here, we suppose that the three components presented above create system effects introducing a large part of uncertainty in terms of growth.

We will not describe in detail the part of the model concerning aid and contributions, insofar as they show through clearly enough in the settings of the user interface.

3.2 Simulator's settings

Given the data available for the group of regions NUTS2 of the fifteen member States European Union, the model retained, as the main indicator, the variation of the GDP per capita of each European region. The variation of this magnitude linked only to the variation of the GDP (we have made the choice of a constant population), is subjected within the platform to several influences adjustable for each simulation:

- The GDP variation rate is, either specific to the region, either identical to the group of regions of a same country, either, by simple hypothesis, identical for the whole group of regions.
- The terms of public intervention include the mechanisms relating to contributions (Countries and EU), to the aid linked to regional policy, such as eligibility thresholds (75%) for Structural Funds.
- The European budget weight was taken into account, stabilized around a threshold of 1% of the European total GDP since fifteen years (threshold reached since 1984). From this average budget, simulations were able to make the Community budget weight vary from 0,5% to 3% of the EU total GDP.
- The principle of additionality between the States and the European Union in the Structural Funds financing was also taken into account, as well as the variability of the relative importance of regional policy in the Community expenditures.
- Finally, the rule of 4% maximum weight of European aid in the GDP of a region or of a State was applied.
- To these principles officially ratified by the European Commission, we have added to our model a spatial dynamic parameter: the hypothesis of the role of spatial interactions and of contiguity effects in the regions' trajectories.

The diffusion by contact with neighbouring regions, made possible by the functioning of the cellular automaton, is carried out therefore naturally in one way or another. Many regional growth models analyse the region as a stand-alone unit and ignore spatial interaction phenomena linked to proximity, neighbourhood, or contiguity effects. What is happening in the neighbouring regions is ignored, while sensitivity to exchange and migration distance is very large. Now, many works have shown that economic interaction and territorial interaction acted in a multiplicative way (Heylen.C et al, 2001), and some empirical assessments evaluating the spatial auto-correlation degree between European regions, as far as GDP per capita is concerned, confirm the pertinence of the reasoning process (Elissalde, 2005). The existence of territorial cores matching either regional areas or national spaces having similar development characteristics and trajectories corroborates this idea. While a situation of spatial competition between activities and between territorial units exists, the taking into account of contagion, of mimicry phenomena, of power struggles linked to neighbourhood effects proves to be necessary.

3.3 Test of the simulator

In order to test the validity of the model, one of the first priorities was to attempt to "calibrate" the simulator's results in relation to the regions' real variations. Therefore, we have compared, given the most recent data available, the 2004 GDP per capita actual results with a simulation fot the period 1996-2004 (see diagram figure 5). It emerges from the comparison between simulated and observed variations that the platform shows an important degree of credibility, both in its overall results (values of the coefficient of determination) and for regions taken individually. For certain settings of the simulator, the differences between the simulated and the observed variations, reveals a trend towards significant results. They underestimate the big metropolitan regions' final values (Brussels, London, Hamburg, Paris area, etc.), and on the contrary overestimate the less-developed regions' ones (Alentejo, Epire, Calabria, Estremadura). These results reproduced on the diagram match settings including a 10% GDP per neighbourhood diffusion rate. By repeating simulations over the same period with a diffusion rate reduced to 5%, but with a doubling of the Community budget (2% of the GDP), added to a substantial increase of regional aid in the aforementioned budget, an overall result is reached, which, this time, underestimates, with one or two exceptions, the entire values of the regions' GDP per capita. The more the GDP diffusion rate is decreased from one region to another, the more the underestimation is important. Several interpretations can henceforth be suggested. It seems that the simulator gets closer to reality, when it includes an important degree of permeability from the neighbouring regions GDP (through the diffusion rate), giving indirectly an account of exchanges and interdependencies between them. Moreover, it seems necessary to wonder about the overall impact (and not only at a particular region's level) of structural policies in relation to the co-variation general dynamics of the Europe of the fifteen entire group of regions. From these results, it emerges that the reducing of disparities are potentially plausible for regions that are eligible for European Funds, and that the options of European regional policies comes within a choice between egalitarianism by readjustment, equity without hierarchical upheaval, and lack of solidarity.

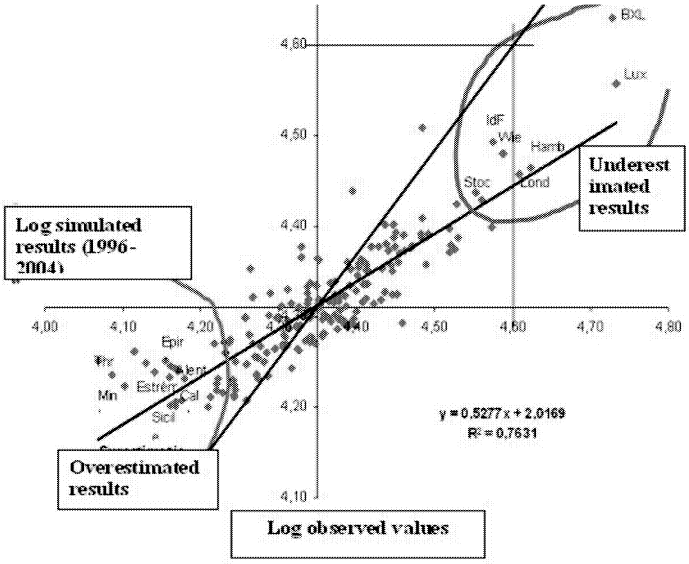


Figure 2: Comparison between simulated (1996-2004) and observed (2004) variations of the GDP per capita

4. TOWARDS AN OPTIMIZATION OF THE TWO CONVERGENCE INDICATORS VIA SIMULATIONS?

If multiple regression calculations achieved above brought out the most discriminating variables, the next aim of the work consists in selecting specific settings for each one of the four variables that lead to the most effective results for the two aspects of the convergence idea. Knowing that the two chosen indicators give an account of various convergence process, reduction of dispersion and disparities for the sigma indicator and improvement of development levels through time for the beta indicator, variations in settings lead to more or less close or more or less far, not to say completely opposite results.

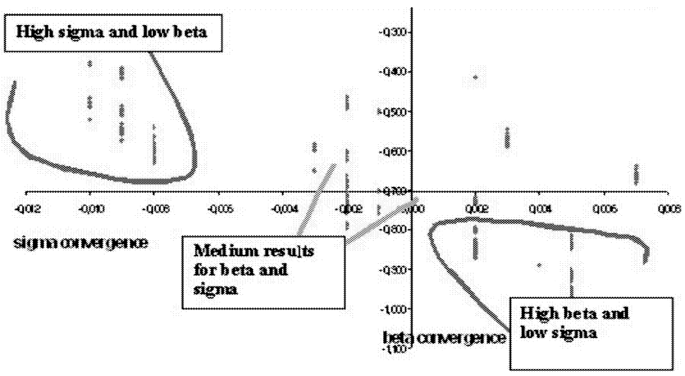


Figure 3: Comparison of the beta and sigma convergence indicators (over 440 simulations)

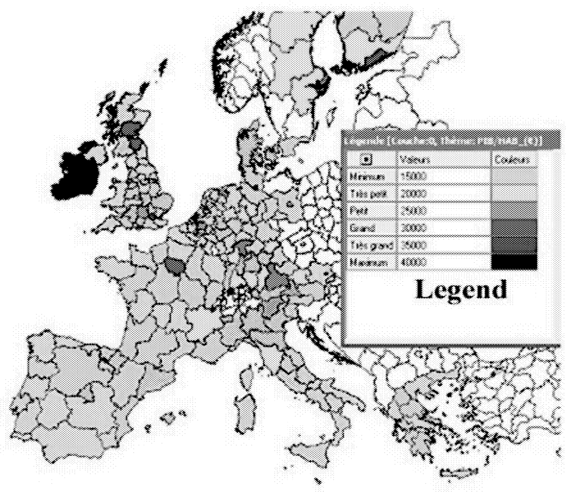
By reproducing on the above diagram the variation in values taken by beta and sigma convergence over a series of more than 400 simulations, it comes out that there are no simulation for which settings lead to high values **both** for sigma and beta convergence. On the contrary, opposite values, divergence for an indicator, convergence for the other, are the most frequent. While high values for both would correspond, on the contrary, to an ideal situation of regional growth and of convergence for the entire group of regions.

This apparent paradox results from each indicator's and each variable's functioning modes. When diffusion mechanism by contiguity are activated between two regions, the weakest receives an additional contribution, but on its side, the strong region carries on to develop itself. In other words, there is not automatically a reducing of disparities (sigma) even if there is progression of every regions compared to the previous situation (beta around -0,9). Conversely, when by assumption, neighbourhood effects are almost nonexistent, weak regions are simulated externally only by the aid of Community Funds, the overall gap between regions decreases but with a beta convergence that does not improve much (values around -0,4). We find here again the strongly discriminating impact of neighbourhood effect (already pointed out with multiple regressions) compared to the Community aid factor.

When a certain proximity exists between two indicators, it is significant to notice that it does not match necessarily the maximum settings offered by the four variables. The tendencies towards an (ideal) complete convergence does not result, according to our simulations, from the one and only increase of the Community budget and from the allocated Funds to the regional policy, but seems more likely due to particular mixing that take into account interregional regulations. Hence the sometimes deceiving or "unexpected" impact of structural policies which does not bring the best effectiveness in terms of reducing of regional disparities within the EU. These efficient mixing are the indication of a multicriteria complexity which relies on specific proportions between Funds levels and the average durations of Community aid (unlike short and standardized durations of current programs), and above all on the taking into account of the interdependency between neighbouring regions. For the reason of the co-evolution of the European regions system's various magnitudes, seldom are situations favourable to complete convergence processes of wide regional groups.

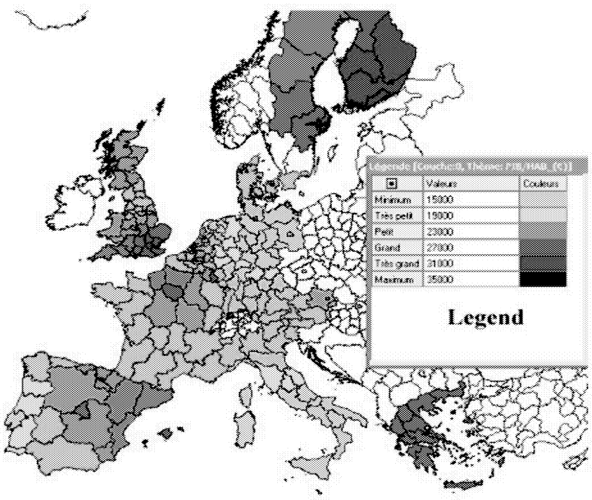
5. WHICH CONFIGURATION FOR TERRITORIAL COHESION?

The cartography of the various types of simulation gives concrete expression to the impact of territorial cohesion modes' geographic distribution chosen by each option :



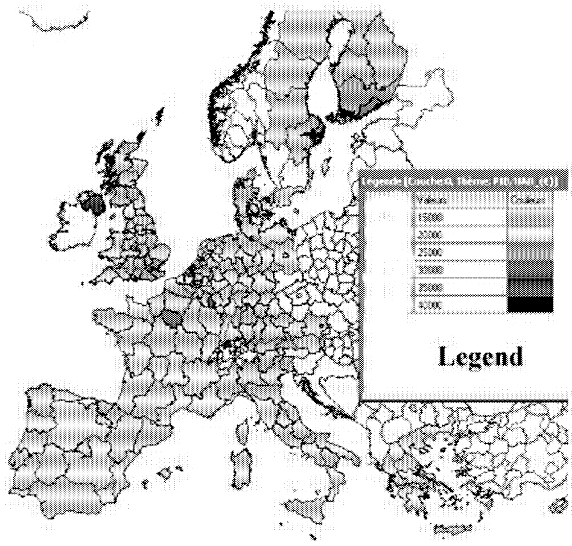
Map 1 - GDP per capita value after a simulation over 10 years without diffusion

This first map reproduces the result of a situation where values of beta convergence and sigma convergence evolve in a completely opposite way. The simulation concerns a Community budget brought up to 2% of the GDP during a 10-year period with, by hypothesis, a total lack of GDP diffusion by contiguity between neighbouring regions. This hypothesis of regional growth's watertightness gives a good sigma with a low dispersion of incomes between regions, since backward regions saw their GDP per capita rise up, but, on the other hand, a bad beta convergence, since the GDP per capita of the entire group of regions, except for a few metropolitan regions, increased only slightly.



Map 2 - GDP per capita value after a simulation over 25 years with a diffusion by neighbourhood of 40%

This second type of setting concerns a 25-year period with a budget representing 2% of the Fifteen's GDP, and a diffusion rate by neighbourhood of 40%, distributed among every neighbouring regions. The results of this type of setting show a good beta convergence and a bad sigma convergence. Representative of an inegalitarian growth, this setting reveals a certain effectiveness at national level, but establishes itself as less homogeneous at European level. The high GDP diffusion rate between contiguous regions reflects the importance of European spaces' growing integration. It produces prosperous regions aggregation phenomena by expansion very often from metropolitan regions : Parisian Basin, South of England, North East of Spain (hence the high values of beta convergence). But on the contrary, neighbouring effects also work between peripheral and poor region areas that do not pull out of their backwardness (hence a bad sigma convergence). Growth develops itself by clusters of regions, but the development gaps are not on the whole being shortened.



Map 3 - GDP per capita value after a simulation over 15 years with a diffusion by neighbourhood of 30%

This third map represents the result of settings combining values of beta and sigma convergence close to each other ensuring an average progression of underprivileged regions and a reducing of the overall disparities. It is obtained with a 15-year duration of policies application, relying on a doubling of the European budget, by devoting to it half of the regional policy. While diffusion by integration of neighbouring regions concerns 30% of the GDP. Here, it is about an attempt of a compromise trying to reconcile the beneficial effects of each of the two types of convergence. From this setting, emerges a configuration of the European space which is relatively homogeneous, but dominated by a few very big metropolis (London, Paris, Brussels, Luxemburg, Stockholm), which development level stands out clearly from the rest of the regions.

From these two indicators, in theory complementary, and often found in literature about regions' convergence, the introduction of a growth propagation variable by contiguity, changed the expected scenarios worked out by the instigators of the Community policies. This introduction of spatial interaction by neighbourhood transforms the Community policies determinist projections in a system of regional units reacting according to a multi-level and a multi-localized complexity. To the multiplicity of settings offered by the simulation platform answers a few seldom co-occurrence probabilities of the two forms of convergence. The taking into account of the neighbourhood effects reveals the existence of an auto-organized process, which is only seldom holder of a global convergence at the level of a group of regions as well as of the entire European space.

6. CONCLUSION

The use of the GeoCells cellular automaton is an attempt in order to simulate the combined effects of the role of neighbourhood interactions and of the variations in the European regional policies on the reducing of disparities and the European group global cohesion. GeoCell's multi-layer organization is adapted to the specificity of this policies' functioning involving a transfer of resources between State members via the European Union's budget and the neighbourhood effects linked to the integration in the European space. Starting from the double assessment that European Union's spatial dynamic revealed an inegalitarian growth and strongly polarized mode in which a third of European regions produce two thirds of the Community GDP, and that structural aid did not have, in this context, the expected consequences at regional level, we attempted to optimize the combination of duration and of public expense levels in order to make probable the reducing of disparities. The GeoCells model contributed to highlight the incidence of spatial interactions and the complexity induced by diagnosis brought up by two type of indicators of economical convergence. The fact that all of the European regional system's components vary together by interactions between scale levels and between neighbouring regions, make it difficult to achieve the coincidence between improvement of the growth of the whole and the reducing of disparities between regions. It is however at these conditions that the cohesion policy, which makes up one of the pilar of the European building will be likely to answer the challenge of territorial integration in an enlarged Union. On the contrary, there is a high probability of a perennization of an inegalitarian growth mode, associating consolidation of the European integration and regional divergence process partially compensated by redistribution funds.

REFERENCES

Batty.M , Goodchild.M , Maguire.D, 2005, GIS, Spatial Analysis and Modelling, ESRI Press, Redlands, California

Beine.M, Docquier.F (eds), 2000, Croissance et convergence économique des régions, De Boek, université

Charleux . L., La politique régionale de l'Union européenne: des régions à l'espace?, Thèse de doctorat, université J.Fourier , Grenoble, 2003.

Elissalde.B , « Géographie, temps et changement spatial », L'espace géographique, n°3, 2000.

Elissalde.B, « Temporalités et modélisation des dynamiques régionales : le cas de l'Union Européenne » in Modélisations géographie, Déterminismes et complexités, Guermond Y.(dir), Hermès-Science, 2005.

Elissalde.B et al, Géopolitique de l'Europe, ed. Nathan, 2006.

Engelen.G, Neis.B, Whalen.J, White.R , 2004, An individual based model of the Lobster Fishery in St John Bay Newfoundland Canada, , (Research Institute for Knowledge Systems, RIKS)

Eurostat ; Deuxième rapport sur la Cohésion économique et sociale, Commission des Communautés européennes, 2000.

Gross C., Opérations topologiques et géométriques sur les multicartes combinatoires ; application à la cartographie thématique, thèse Université de Strasboug, 1989.

Heylen.C et al, Transports de marchandises par la route au niveau régional, Eurostat, Commission européenne, 2001.

Kiel,L.D et Elliott. E, Chaos theory in the social sciences, University of Michigan Press, 1997.

Langlois P., « Formalisation des concepts topologiques en géomatique », revue internationale de géomatique, Vol 4, n°2, 1994.

Longley.P , Batty.M, 1996, Spatial Analysis: Modelling in a GIS Environment, Wiley, New York

Martin. P, A quoi servent les politiques régionales européennes?, Economie internationale, n°81, 2000

Maurel. F, Scénario pour une nouvelle géographie économique de l'Europe, Commissariat général du Plan, ed. Economica, 1999.

Riou,.S, « Géographie, croissance et politique de cohésion en Europe », Revue française d'économie, n°3, vol XVII, 2002.

Salah-i-Martin,X, « Regional Cohesion: Evidence and Theories of Regional Growth and Convergence », European Economic Review, n°40, 1996.

Young.A.T, Sigma convergence versus Beta convergence: evidence from U.S. County-Level Data, Emory Economics, univ. Atlanta, Sept 2003

From the event to the ephemeral city

Françoise Lucchini
Bernard Elissalde
Sylviano Freire-Diaz

Department of Geography, University of Rouen
7 Rue Thomas Becket 76821 Mont-Saint-Aignan, France
Email : Francoise.lucchini@univ-rouen.fr
Bernard.elissalde@univ-rouen.fr
Sylviano.freire-diaz@univ-rouen.fr

KEYWORDS: mobile phone network, events, city, chronotope

ABSTRACT

This article aims to promote the notion of an ephemeral city, and an urbanity based on relationships. Thus the city can be apprehended by way of occasional festive or social interactions generated by human gatherings in a limited piece of territory. This draws us away from the approach to the city that gives precedence to the perennial nature of its functions and activities. The original data provided by mobile phone use make it possible to approach temporary urban situations in time and in space. Temporary variations in urban life can be caught and visualised in a refined spatio-temporal sequencing that gives the "pulse" of the city through the varying locations of activities and individuals across the urban environment. The study of a cultural event such as the Armada in Rouen in 2008[1] by way of mobile phone data evidences the spatial and temporal arrhythmia generated by a well-known cultural event in the daily life of the city, and also gives an idea of the heterotopic dimension of place.

SEEING THE CITY DIFFERENTLY

This article centres on a different approach to the urban object that we have chosen to call the ephemeral city. In this case, the entity known as the "city" is not apprehended by approaches that explore the various architectural forms, residential change, the interpenetration of economic functions, or changes in activity. The focus here is on a category of spatial and temporal data that is liable to provide a reading of the city through its prime ability to be a central place, and to create opportunities for encounter, unexpected mobility patterns, unforeseen happenings, conviviality, and social intermingling. The data collected from the behaviours of mobile phone users[2] are particularly well-suited to this scientific question. We hypothesised that not only are there mobility configurations and forms of spatio-temporal groupings that are the product of general processes inherent in the daily functioning of the city and the city system, there are also forms of gathering that draw away from the usual daily functioning patterns, in particular at the time of major festive or cultural events or well-known gatherings. Thus the city produces changing and ephemeral forms. Seen by way of the presence of mobile phone users, the observation of these gatherings of individuals at certain moments makes it possible to capture these ephemeral forms of the city and the displacement of its centralities in fairly great detail. It is precisely this spatio-temporal variability of places in the city, and the heterotopic dimension that accompanies it, that we set out to apprehend.

This scientific viewpoint is intended to define the workings of the “ephemeral” city. For this purpose, three types of exploration are required. The first concerns the relationship that becomes established between a city and a cultural or heritage-related event. The second broaches the issue of the suitability of the type of data used to the study of a particular occasion or event in urban environment. Finally, reflection is required on the line of research thus offered to urban geography research to enable the city to be apprehended in its spatio-temporal variations.

The first of these dimension looks at the way in which, in urban geography, it is possible to apprehend the relationships that form between a city and a cultural "event" in a wide sense. This dimension relates to forms of heritage-building[3] that are grounded in natural

[1] This is a gathering of tall ships that has been organised periodically since 1989, drawing large numbers of people.

[2] This anonymous data is made available by the provider Orange, and gives indications on incoming and outgoing calls and text messages

[3] In French *patrimonialisation*, from *patrimoine*, or heritage. This describes a process of collective appropriation whereby an "object" is endowed with a value; thus the "object" becomes a heritage by way of an "inverse affiliation" (Pouillon J., 1998), since the relationship that is (re)constructed is with the past.

and cultural elements of a heritage present in the geography and the history of a city. But this dimension relates likewise to a form of urban creativity, an ability to give life to a territory, and a potential for touristic and economic attractiveness, which have been widely developed by European cities since the 1990s in an attempt to distinguish themselves one from the other. This event-culture, very frequent in cities, pinpoints one aspect of the cultural domain, more readily conceded today, and that is the recognition of the considerable role of the cultural dimension in economic development and in the positioning of the city in terms of territorial image. In our contemporary societies, pervaded by environmental and economic issues and in search of solutions for sustainable development, culture and heritage have become obvious choices and assets for cities, in particular in cases where their economic development has undergone setbacks and decline. The need to find new activities, or again the desire to reshape urban spaces and develop new centralities, can be grounded in culture and heritage. This creativity deployed by cities in their use of the assets provided by their heritage and culture, and the forms taken on by cultural events and gatherings such as those relating to the European Capitals of Culture, are clearly different from the "creative city" described by Richard Florida (Florida R., 2002). The "ephemeral" city generates an urban creativity that cannot be reduced to residences of creative individuals providing "technology, talent and tolerance". According to Florida's approach, the city is seen as being creative when there is a balance or match between the presence of a cultural "milieu", made up in part of professionals of abstract production functions (cognitive capitalism), and the economic successfulness of the territory in which it is located. Yet it must be conceded that this static manner of apprehending urban creativity completely ignores the three quarters of the urban population making up the city and taking part in its functioning. The creativity of a city, its renewal and its dynamism are indeed subject to economic, technological, political and cultural conditions, but they cannot be completely dissociated from the overall population present in the city, or from its behaviours. Urban renewal and rehabilitation can take on different forms, lasting or temporary, but they all contribute to shaping a new geography of the city, and to the construction of a new collective imagination of the city: abandoned areas are stimulated by new activities, the old city centres are renovated, city quarters are developed to favour greater citizen flexibility and mobility, the image of the city is changed for the duration of an event, new forms of sociability are deployed on the occasion of a festival or a large gathering. Among the "lasting" or perennial forms of urban renewal, numerous cities have made the choice of major architectural projects, and the use of internationally renowned architects, where the media coverage also bolsters the reputation of the city and its touristic attractiveness. Among the non-permanent, ephemeral forms aiming to renew the city, there are the large festivals, cultural gatherings and heritage-based events.

The "ephemeral" city first of all is a form of urban renewal, by manufacturing an imagined city that cannot be reduced to its physical form. Thus events provide different figures of the city, which contribute to the urban project, and can give an image of the city that differs from the usual perception, made up of monuments, museums or focusing on economic or political visibility. The dialogue established for the duration of an event between an urban site and that event can be seen as the search for a point of balance between a particular geographical site, a local history, and a universal link liable to gather a large number of people. The cultural event provides moments for sharing a lost collective memory: in the instance of the Tall Ships Armada in Rouen, every five years there is a re-invention of a shared memory, and a reactivation of the transmission of the link that existed in the past between the city, its inhabitants, its river, and its fluvial and maritime identity (Lucchini F., Hucy W., 2008; Lucchini F., 2010). In this respect, large events play a part in structuring the social sphere, punctuated by memorable events in the form of festive or heritage-related gatherings, which bring with it an intermingling of populations and conviviality, while at the same time integrating these moments into the longer time-span (that of the city memory), and the short time-span (that of city rivalries). Secondly, the "ephemeral" city is also a city with changing forms, as a result of conspicuous differences that can be observed in the frequentation of places in the city, in the shifts of centralities in time and space, and in the distribution of people and activities across the urban fabric, variable from one hour to the next. These mobility patterns, and the variable concentrations of people in time and space turn the city into an entity that takes on ephemeral, non-permanent forms. The urban "pulse" is the product of the daily functioning of the city, which can undergo change and arrhythmia on the occasion of a major event organised in the city.

THE TALL SHIPS GATHERINGS AND URBAN GOVERNANCE

In the case of the Tall Ships event, known as the *Armada*, the signals sent out by the docking of tall sailing ships from all over the world (Mexico, Japan, Russia, Chile etc) recall the earlier port activity along the Seine quaysides, which have been deserted for decades, and have recently undergone some renovation. Here there is no monumental heritage, but rather, for the duration of a popular event, the reconstruction of a shared memory and imagination in the city of Rouen, originally an up-river port. When cities cannot claim any clear economic development, inter-city competition and the quest for stronger arguments for singularity lead cities to construct and valorise a "symbolic capital" (Gravari-Barbas M., Jacquot S., 2007). The 2008 Armada in Rouen belongs to the "maritime" category of events initiated twenty-five years ago by "Douarnenez 86" and "Douarnenez 88'. These were at the time gatherings of old sailing vessels, assembled thanks to the perseverance of a few enthusiasts, mainly but not exclusively belonging to the "tall ship" category[4]. Since these pioneering events, flotillas have enlarged and diversified, and the gatherings have been repeated and have spread to other French ports (gulf of Morbihan, Paimpol, Brest, Toulon) and abroad (Portsmouth in 2007, Tacoma in the Puget Sound in 2008, Nova Scotia in 2009 and San Diego in 2010), demonstrating the interest in the maritime heritage on the part of a wide public. By way of these objects recalling the maritime epics of the past, a strong message is sent out, serving the reputation of the organising city, and adressing the local population. The docking of these tall ships, an event with a major symbolic content, contributes to reinventing or reinforcing the port and maritime identity of the city concerned. In the course of the 20th century, the diversification of the economic bases of port cities has in certain instance led to the de-connection of the port from the economic and social evolutions of the city. The quest for a means to generate identity and to make use of the local heritage has led to this renewal of lies between cities and their port areas, often on the occasion of temporary events, and more rarely in lasting manner by way of a maritime museum (for instance Liverpool, Cherbourg or Rouen). Unlike races out at sea, whether or not they line up the old sailing ships (the Cutty Sark race or the Tall Ships race) or modern ships, the docking of tall ships at the quayside for some ten days, as in the case of the Armada, sometimes accompanied by processing downriver to the estuary or in the roads, provides visitors with proximity to the ships and their crews which gives these gatherings considerable success. In addition to the spectacle of the ancient riggings, the Armadas increasingly also offer concerts and other events not directly related to past navigation.

There is widespread agreement among researchers today on the fact that one of the major alterations in the last decade in public policies for the city has been the shift from a "technician" management of the city to a global apprehension of the urban object. In this perspective, the management of large events is a fundamental element of urban governance. According to P.Chaudoir and S. Ostrowetsky, the organisation of large-scale celebrations and events in an urban environment has four main purposes: a) giving meaning to places, b) bringing history into the city, c) developing partnerships among actors in the city and among citizens, and d) developing a founding myth and a specific identity for the city (Chaudoir P., Ostrowetsky S., 1995).

To the importance of the meaning of place, of the interplay of actors, and of relationships between local elected representatives and citizens, A. Bourdin adds the omnipresence of "events" in our contemporary societies (Bourdin A., 2005). He shows how varied forms of value enhancement, are sought, by way of original interactions between events, their temporal patterns, their particular locations, and their consequences on the urban space. "*Large exhibitions, fairs, exceptional trading events, sports championships, centenary or bicentenary celebrations, festivals, and so forth, as well as everything that can associate the contemporary artistic logics of installation and performance with effects on the practice and use of the city and urban amenities, all form the ideal field for the event [...] The event "produces" the city in three ways: it drives urban actors and leads them to cooperate and organise; it produces cycles, movements, and occasions in city life; it gives value to places and develops a whole array of logistics, which contributes to the flexibility of the city and gives it life*" (Bourdin A., 2005,p43). The consumer citizen, long since formatted to consume material goods, also consumes events, and enters into a form of addiction to the event. The attraction of these particular occasions leads citizens to frequent places that, outside the event, are not part of their living space, or may even carry negative connotations in the collective urban imagination. The large gatherings (whether cultural, sporting or other) amount to a broadening of the urban offer, but they also reduce mental and cultural distances within the city, and alter (momentarily?) the tropisms inside the urban space.

FROM SOCIO-RESIDENTIAL LOCALOZATION TO TIME-RELATED GEOLOCALIZATION

Social indicators relating to income or age show cities as subdivided into parts, appearing as mirrors of the improbable cohesion of the society overall. This approach to the city nevertheless leads to a view of the functioning of the city as being the reflection of social determinism: the city is an effect, rather than a place that contributes to social mingling, or even to increasing opportunities for interaction for each individual. These ways of viewing the city do not enable the different temporal patterns of the city to be apprehended. There

[4] Sail Training International classifies its A-class of tall ships as "all square-rigged vessels and all other vessels over 40 metres (131 ft) length overall (LOA)" (i.e. length excluding bowsprit and aft spar)

is indeed an implicit time scale in approaches that focus on urban structures. This corresponds to the apparent diachroneity between changes in the urban space and changes in the social trajectories of the populations, and the adjustments that are, or are not, made to the living environment[5].Yet this focus of analyses on projecting social parameters onto time in itself justifies adding other angles of vision to apprehend the synchrony of urban social interaction. According to place and according to time of day, the city changes, and is experienced and lived-in according to rhythms and time scales that no longer be read solely according to inter-census periods. Instead of a city that appears static, a focus on urban spaces in their variability according to time of day, or between night-time and day-time activities (Cauquelin A.,1990), generates the image of a shifting, moving city. In this perspective, mobile phone data can enable the pulse of the city to be captured in different manner. The rhythms of the city have hitherto been captured either by durable indicators, or by sample-based surveys. There is now data available to attempt to capture thus urban pulse in a different way.

The analysis of the flows of mobile phone calls and their geo-localization is a new and original source of information enabling capture hour by hour of the variable intensity of urban life and of the uses of the city. The anonymous data made available to us by the provider Orange give information on incoming and/or outgoing calls and text messages. Variations in users concentration are recorded for the geo-localization of the surface area covered by the relay aerial detecting the call. From these aerials, the presence of each subscriber in the urban space is only registered in case of an "occurrence" (i.e. call or text message) is registered on the subscriber's phone. This is therefore an attempt to approach temporary urban situations in time and space. The aim is to capture the patterns of variation in urban life from data that can provide a detailed spatio-temporal sequencing giving the position of individuals in the urban environment.

In addition to its technological dimension, the mobile phone is a fundamental social phenomenon of present-day lifestyles in the same way as Internet. The acceleration of human time scales has changed our relationship with the world and with others. The culture of immediacy, by way of real-time communication, whether for purposes of information, distraction, emotion or rationality, is gaining precedence over memorisation and *post-hoc* narrative. These nomadic technologies enable any user to remain in contact with his or her "tribe" (family, friends, professional environment), and they are breaking down the barriers between the private and the public spheres. The interactive nature of these means of communication seems perfectly suited to the individualisation of choice, and the apparent freedom of decision that subscribers demonstrate in the use of their mobile phones. Since the individual can be, virtually, anywhere and everywhere, and in the instant, with his or her own, we hypothesise that the spectacle provided by exceptional events is likely to stimulate the communication fervour, and thus make it possible to measure variations in the volumes of visitors in the urban space during the 2008 Armada in Rouen.

There has been some earlier work on data recorded by mobile phone providers. Outside France the first experiments conducted by a M.I.T team headed by Carlo Ratti centred on how to capture the intensity of flows on mobile phone networks, for instance from increased frequentation generated by an art exhibition in the city of Graz in Austria (Ratti C. et al, 2005). This laboratory also developed cartographic representations of the intensity of mobile phone communications in the city of Rome on the occasion of another cultural event (the Madonna concert), and a sporting event (the triumphant procession of the Italian football team after winning the world championship in 2006). This work also endeavours to capture the city "in real time", so as to plan better functioning by urban services, using geo-localization applications. In the pioneering work in France, the orientations differ from this perspective. They are more concerned with the study of social networks, categories of mobile phone users, and the issues of time-management for users of mobile agendas (organizers) and "real-time" technologies. The SENSe laboratory in the Orange-Labs network thus observed, at the time of the *fête de la musique*[6] in June 2008, the gathering of crowds in different quarters in Paris[7] and the forms of sociability produced through the categories of social networks observed. These lines of research aim to look more closely at the potential occurrence of the unforeseen and the unexpected in individual routines by way of the technologies of nomadic communication (Aguiton C. et al, 2009; Smoreda Z. et al, 2010).

All these options for mobility on individual scale have an effect on the potential use of urban space, which thereby has an impact on the spatio-temporal structures of the city. In our approach to the Armada in Rouen by way of anonymous data provided by Orange, certain choices are in line with the analysis generally set out by Time Geography (Hägerstrand T., 1967): we look at the localization of individuals in space, at the extension in the surface areas involved in individuals' displacements, and the duration of their presence and

[5] See M.Roncayolo: "the city is constantly manufactured, destroyed and recomposed" in Duby G., 1985, *Histoire de la France Urbaine*, vol.5

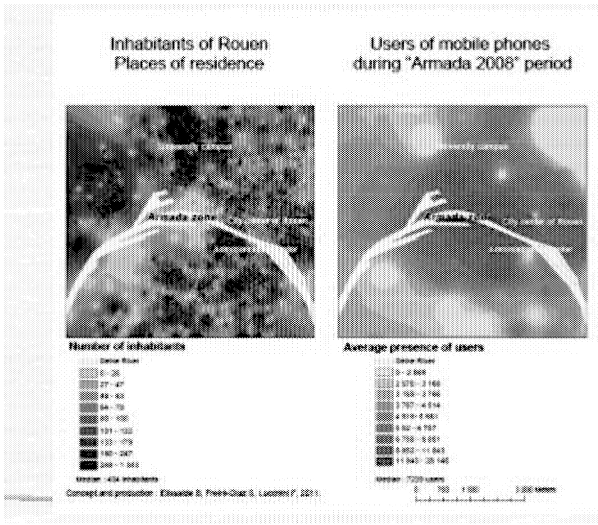
[6] A day of mainly informal music-making organised across France

[7] See the website: www.urbanmobs.fr

their activities. Nevertheless, the work on the Armada differs from work grounded in Time Geography in that the aim is more to observe unusually large and unexpected concentrations of mobile phone users, rather than the more well-known and more regular concentrations linked to the day-to-day functioning of the city. Cultural events do indeed appear to possess the ability to increase intra-urban mobility, and to generate spatial and temporal itineraries that are different from the usual paths taken by individuals in a city. What we propose here mobilises the categories of space and time on the scale of a city. We have however chosen to base the analysis on aggregates of telephone users or telephone calls, rather than presenting detailed behaviours at individual level. These aggregates of mobile phone use are enumerated in the areas covered by the 579 relay aerials situated across the Rouen urban area. The observation of the variations in the volumes thus obtained across the urban space according to the time of day shows a change in the pace of life in the different city areas.

THE EMERGENCE OF THE EPHEMERAL CITY

To cast light on the information obtained about the entity "the city" from anonymous mobile phone data in comparison with data from classic enumeration methods, we deliberately superimposed two categories of cartographic images of the city over the investigated space in Rouen. The first cartographic representation corresponds to the population of Rouen according to place of residence, on the scale of the urban block. The other cartographic representations show the distribution of French users of mobile phones during and outside the Armada period in 2008. The urban image projected by the first cartographic representation suggests a city made up of empty and filled spaces, and above all the very numerous segmentations and groupings of people. The second type of cartographic representation, derived from information on communication exchanges and behaviours of mobile phone users in the city of Rouen, generates an image of the city not in its ability to produce segmentation, but in its scope for being a central place where populations intermingle. The image of the city of Rouen generated by mobile phone data, even in static form as it is here[8], is clearly very different from population distribution according to place of residence, and relates more to an interpretation of the city that is centred on its ability to maximise exchanges and reduce partitioning. While in residential terms concentrations can reach some thousand inhabitants, the mean density of users of mobile phones over a large part of the city is greater than ten or even twenty thousand. The extent of zones of concentration of



mobile phone users is likewise far greater than that of areas of high residential density, and curiously takes on the form of a circular wave centred on the heart of the city of Rouen. The distribution differential of mobile phone users and their behaviours during and outside the period of an event should also be underlined. During the 2008 Tall Ships Armada, the quayside area in Rouen is one of the spaces where mobile phone users concentrated, while users completely desert this area outside the duration of the event, and appear more in the historic and economic centre of the city.

The effect of an event like the Armada is clearly different from the routine of city life. The disruption generated in the usual pulsation of the city by the free access to some fifty tall ships over 10 days on the Seine quaysides in the heart of Rouen can be clearly seen in the behaviours of mobile phone users. This questions the usual attraction potential of the various city quarters. The frequentation of the city by mobile phone users during the Armada differs from the itineraries usually used. The flows observed show very different geographical configurations at the time of the event and outside the time of the event. The 3D representation of mean volumes of users during the 10 days of the 2008 Armada, and the mean volumes of users over an equivalent period outside the event is eloquent: while the concentrations of users are on a similar level for the city centre whatever the period envisaged, the most marked levels of concentration during the Armada are very specifically located on the quaysides of the Seine, generally very little frequented. These high concentrations of users are particularly visible on the left bank of the Seine during the event, while this zone is thus far only very partially affected by renovation operations undertaken by the municipality and the *Rouen-Elbeuf-Austreberthe Communauté d'Agglomération* ("*Seine-Ouest*" project). Among mobile phone users, French subscribers make up the bulk of

[8] Because of its spatial and temporal detail, mobile phone data enables the development of animated maps representing the course of a week, or contrasting night and day over 24 hours (Elissalde B., Lucchini F., 2011)

the visitors to the Armada. In addition to the sailors and the crews, foreign mobile phone users visiting the 2008 gathering account for a fairly constant proportion of 5% of the 5 million visitors over the ten days' duration. What is more remarkable is the international range of foreign visitors to the Rouen urban area during the Armada. The spectrum of nationalities covered by the origins of foreign mobile phone users is wide, reaching over one hundred nationalities (116), while at the same time the European group is the most represented. The nationalities of mobile phone users are a good indicator of the attractiveness of a festive or cultural event. In fairly logical manner, the most numerous foreign nationals were British, Italian and Belgian, amounting to several thousand individuals. The foreign mobile phone users are also fairly regular in their spatial behaviours throughout the event, giving precedence to the historical city centre and the Seine quaysides. These two tropisms, one towards the river and the other towards the urban heritage, outline the boundaries of tourist deployment generated by an urban event. And this is particularly true when the tropism towards the river suddenly ceases as soon as the boats leave on the 14th of July.

THE DISPLACEMENT OF URBAN CENTRALITIES

Beyond this first change in general pattern generated by the event in the usual functioning of the city, other urban phenomena emerge. The peak of mean human presence at the time of the 2008 Armada conceals other variabilities which will now be presented. The study of the mobile phone relay aerials capturing telephone exchanges and the presence of alias can provide considerable detail from both the spatial and the temporal point of view. Normal urban functioning observed via the behaviours of mobile phone users shows the existence of zones of spatial concentration, and likewise there are also moments of intense telephone exchanges. Thus weekdays see larger numbers of mobile phone exchanges than weekends and holidays, and in a given day the communication peaks are generally situated around 11 a.m. and 6 p.m. If these spatial and temporal data are crossed, the existence of chronotopes[9] in the city can be evidenced, according to the principle that "something is happening at a given moment in a given place" (Crang M., 2001; Lefèbvre H., 1992; Elissalde B., Lucchini F., 2011).

By implementing a reading of several place in the city of Rouen in the form of a cross-section from the north-east outer boulevards to the quays on the left bank of the Seine, a decentring of the attractiveness of different quarters is observed, and the crowd of visitors can be followed according to the time of day. The retail activity of Fridays and Saturdays leads to peaks of frequentation in the historic centre, while the activities and festivities around the Armada (snacks and meals, children's games, free concerts from 6 p.m., fireworks from 10 p.m. etc) lead to a reversal in the daytime and night-time frequentation. It can be noted that there is a contrast between the decrease in intensity of telephone exchanges after 8 p.m. in the historic centres, while the quaysides are even more active between 8 and 10 p.m., reaching a peak in the evenings of the two Saturdays, and above all the evening preceding July 14th (Bastille day).

The chronotopes that can be seen in the city are thus sequences of time that individualize parts of the city as being under- or over-animated, introducing a differentiation in relation to the usual urban activity patterns. Thus these chronotopes can be apprehended by way of observation of unusual features in user frequentation, such as the peak of 2000 extra phone calls observed in the heart of the city on the afternoon of Saturday July 12th in comparison with the previous Saturday: this peak seems to have been generated by the traditional procession of the ships' crews through the main streets of the right bank in Rouen. The chronotopes also show up other forms of variability for a given place, it is not unusual to note marked oscillations in frequentation, upwards and downwards alike. This is illustrated by marked deviations from the mean numbers of calls, measured according to an observation of the aerials at a given time, for instance every day at 6 p.m.

The festive activities on the quaysides culminate on the two Saturdays, July 5th and 12th 2008, and in particular at the time of the traditional "*bal*" (popular open-air dance) on the evening of July 13th. On that same evening, towards midnight, aerial 1211 situated on the left bank quayside registered the maximum number of calls for the whole 10 days of the Armada. The intensity of the night life was still producing around 4000 calls and texts towards 2 a.m. The dually festive nature (the Armada plus Bastille day) of an occasion of this sort has the ability to shift the traditional celebrations of the evening before July 14th towards the Seine quayside, not normally very attractive. However the ephemeral character of any mass event entails a sudden falling-off of spectators in the festive venues. For all the aerials near the quays, the departure of the tall ships on the morning of July 14th signals a return to the usual status of the quaysides as a place for a stroll or to go jogging.

URBAN TROPISMS: WHAT PROPORTION IS A RANDOM PATTERN?

Since everything appears to change, and the spatio-temporal concentrations of mobile phone users show

[9] The chronotope is a concept developed by M.Crang, and used here in its geographical application.

very marked variability across the city, it can be wondered how far they are driven by random elements, or by phenomena whose explicative dimension could be more clearly apprehended. Using the aerials and the collective movements observed though them, an indicator of the "urban pulse" can be developed from the concentration recorded by an aerial on a given day of a proportion of users in relation to the 10 days of the event as a whole. This indicator, calculated from the contingency table for frequentation of users for each aerial and for each day, gives the following calculation:

$$\text{Index P} = \frac{X_{ij}}{X_{..}} * 100 \quad \text{where}$$

X_{ij} is the number of mobile phone users counted on aerial i on day j

$X_{..}$ is the total number of mobile phone users over the 10 days

This daily urban "pulse indicator" shows the existence of concentrations of mobile phone users for a given day and a given aerial. To determine whether these concentrations could result from chance requires a test of independence for the 579 aerials registering telephone communications across the Rouen urban area, these being physically installed in 307 places across the city area (a place corresponds to a mast on which several aerials with different orientations can be installed). This test is equivalent to comparing the actual numbers of users registered on these aerials with the theoretical numbers that should be found if the distribution was random, in other words if it is postulated that the position of the aerials does not have a role in the human concentrations observed. In some cases it can be observed that the concentrations exceed the numbers of users expected for a given aerial on a given day by more than 9000. In other instances, there are 8000 fewer users than expected if the distribution were random. It appears that the distribution of users observed in reality is widely different from the distribution that would result from a random distribution of frequentation (the Chi^2 observed is 84 time greater than the theoretical Chi^2). Mobile phone users thus did tend to gather and disperse around certain aerials. User densities in the city that are clearly different from behaviours that would result from chance are thus clearly identified. These densities signal a geographical zone defined by aerials situated on the Seine quaysides, that is to say precisely where the cultural event was taking place, acting as a temporary attractor.

EPHEMERAL CENTRALITIES

Beyond the variability of urban occupation from one day to another, we were also interested in the question of the displacement of the intensity of urban activity according to the time of day. An attempt was made to evaluate this intensity by way of an indicator for relative over-frequentation by mobile phone users in certain quarters for each hour of the day. Partly derived from the indicators developed by Florence or Isard, this indicator uses the following formula:

$$IndiceF = \frac{X_{ij} / X_{.j}}{X_{i.} / X_{..}}$$

where

X_{ij} is the total number of users observed around the aerial i for the hour j

$X_{.j}$ is the total number of users observed around all the aerials for the hour j

$X_{i.}$ is the total number of users observed around the aerial i for all hours

$X_{..}$ is the overall total for all the users and all the aerials over the 24 hour periods

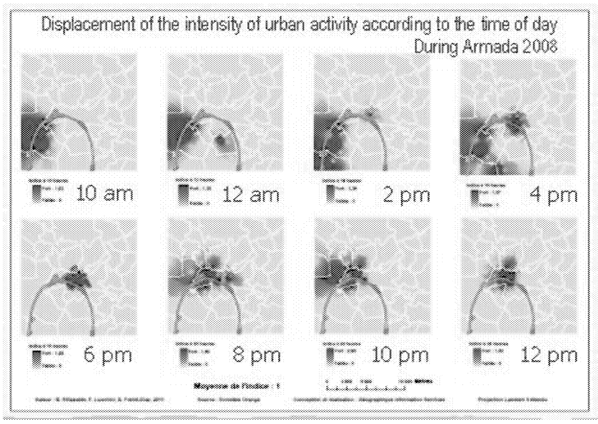
With the F index, we can measure the relative magnitude of the mean number of users around an aerial in a given hour in relation to all the aerials in the same hour. This concentration is then weighted by the mean weight of that aerial, whatever the hour, in relation to all the aerials in the urban area. This indicator make it possible to determine whether an aerial is under- or over-used at a given time in relation to its mean weight among the aerials overall whatever the hour. Values greater than 1 indicate a relative over-representation relating to an hour in the day, that is to say a moment when the number of mobile phone users around the aerial registers a value that is greater than the mean weight of the aerial in question. The successive maps, for each hour and for the 20 highest indicator values, show the displacement of centralities according to the hour in the day, differentiating night and day, and according to their functional orientation.

If observations are started in the morning, it can be seen that the aerials in the industrial and port zone (Petit-Quevilly, Petit-Couronne) and in Sotteville-lès-Rouen, are prominent for the hours from 10 a.m. to 4 p.m. (maximum values ranging from 1.34 to 1.97). As these are aerials covering important activity zones rather than residential or retail trade zones, it is not surprising that these zones should stand out during working hours, and the fall away after 6 p.m. The contrast with the rest of the urban area shows up particularly in the morning at a time when overall calls (private and professional) and the aerials associated with mixed zones, have not reached the maximum, which is situated in the afternoon. The Seine quaysides, in contrast, register high values for the night life accompanying the Armada (maximum values ranging from 1.4 to 2.9). They stand out from the rest of the urban area (residence and work) where the phone activity decreases after 7 p.m. Nor are these hourly shifts of centralities unimodal, since the roundabout of mobile phone calls moves from one block to another, at several moments also involving the city centre on the right bank.

This cartography(see map on the next page) of the extreme values for hour by hour concentrations of users, rather than global volumes that would merely evidence the "Armada effect", casts light on the hourly variations in the displacement of activities across the city. These temporal sequences, during which one quarter or one block sees its pulse accelerating, shape the urban space, which is no longer differentiated merely by way of contiguous surface areas, assumed to be stable over time. They show a city that is shaped by chronotopes of varying duration and intensity.

CONCLUSION

By way of the example of the Tall Ships Armada in Rouen, we considered that it could be interesting to try to capture the upheavals generated by a major event in the urban routine. If urban life is the concrete manifestation of daily patterns, urban events introduce variations in the city's pulse, sometimes leading to a form of arrhythmia within the urban space. While traditionally the city is understood by way of functional approaches, we have shown that, paradoxically, it is from these ephemeral situations that the images of the city alter, and that new representations are formed of the different quarters and of the city as a whole. The spatio-temporal precision of mobile phone data widens our understanding of the city by way of its changes in pace and rhythm, and enables this entity to be apprehended differently. The hours-by-hour and calendar variability of spatial concentrations shows how urban centralities are displaced, and shows the city as a place of change. The ephemeral city thus orders itself around a combination of numerous chronotopes. These chronotopes that individualise certain parts of the city by their under- or over-frequentation are generated by ephemeral tropisms. The portions of the city concerned, receiving new life for the time of an event, form contemporary heterotopias, or localised utopias where several spaces and several temporalities appear to cohabit in one and the same place. The meaning of these places evolves as the event proceeds. It can be integrated (superimposition), or appear distinct from the daily functioning of the city (the "Armada effect"). Thus from the examination of a large event, it is possible to focus on the appearance of a ephemeralB city, where the emphasis is on a form of urbanity that is based on the life of relationships, and on changes in urban sociabilities, in which individuals are less guided by the constraints of their belonging, than by personal opportunities.



REFERENCES

Aguiton C., Cardon D., Smoreda Z., 2009, « Living Maps ». *Proceedings of the First international Forum of application and Management of Personal Electronic Information*, MIT, Cambridge, MA, US.

Bourdin A., 2005, *La Métropole des individus*, Editions de l'Aube.

Cauquelin A., 1990, *La ville la nuit*, PUF.

Chaudoir P., Ostrowetsky S., 1995, « L'espace festif et son Public, L'intervention Culturelle En Espace Public ». *Annales De La Recherche Urbaine*, déc. 1995.

Crang M ., 2001, « Rhythms of the city ». May J., Thrift N., 2001, *Timespace, Geographies of Temporality*, London, Routledge.

Elissalde B., Lucchini F., 2011 in press, « Pulsations urbaines. Comment un évènement transforme-t-il la ville ? ». *in Données Urbaines,* 2011 éditions Anthropos.

Gravari-Barbas M., Jacquot S., 2007, « L'évènement, outil de légitimation de projets urbains ». *Géocarrefour*, vol 82/3, 2007.

Hägerstrand T., 1967 [1953], *Innovation diffusion as a spatial process, translation* by Allan Pred, University of Chicago Press.

Lefebvre H., 1992, *Eléments de rythmanalyse*, Syllepse Ed.

Lucchini F., Hucy W., 2008, « L'image de la ville ». Guermond Y. coord., 2008, *Rouen, la métropole oubliée ?,* l'Harmattan, 73-107.

Lucchini F., 2010, *La fabrique des lieux culturels*, Mémoire d'Habilitation à Diriger des Recherches, Université de Rouen.

Pouillon J., 1998, *Le cru et le su*, Col. La librairie du XX[e] siècle, Paris, Seuil.

Ratti C., Sevtsuk, Huang, Pailer, 2005, « Mobile Landscapes: Graz in Real Time ». *Proceedings of the 3rd Symposium on LBS & TeleCartography*.

Smoreda Z., Aguiton C., Fourestié B., Morlot F., 2010, *Taking the Urban Pulse with Mobile Networks*, ParisTech Review, juin 2010.

MODELING URBAN CULTURAL SYSTEMS WITH THE INTEGRATIVE SIMULATION PLATFORM *ROUANTS*

Rawan Ghnemat
Princess Sumaya University for Technology
Amman, Jordan
email: rawan.ghnemat@gmail.com

Françoise Lucchini
UMR IDEES/MTG
University of Rouen, France
email: francoise.lucchini@univ-rouen.fr

Cyrille Bertelle
LITIS
University of Le Havre, France
email: cyrille.bertelle@gmail.com

KEYWORDS

Urban Dynamics, Cultural Sites, Swarm Intelligence, Geographical Information System, Urban Spatial Environment.

ABSTRACT

This paper emphasizes the interest of understanding the functioning of dynamics of collective cultural forms in urban area, as social reality for everyone. Geographers have focused on understanding the interrelationships that people forged with their territory. They work on experiments on social systems belonging to specific territories, such as cities. City is generally understood as a complex system that can be described with several organizational levels: micro, meso and macro levels. At the micro level, the system is the set of interrelationships linking intra-urban elements such as individuals, households and institutions. At the meso level, the city is considered as a systemic entity with functional responses generally included in the mechanisms of competition and/or complementarity with other neighboring entities. Finally, at the macro level, urban systems are composed of interdependent cities belonging to national and international territories. We position this research at the intra-urban level. For that purpose, Rouants simulation platform concerns spatial analysis of urban dynamics described here by services development and their practice by users. A study case allow to analyse cultural sites development applied to the French urban area of Rouen.

INTRODUCTION

The study presented in this paper del with the description of cultural behavior of citizen according to cultural sites on a real urban environment. We focus on describing how the spatial configuration of a city may compel individual behavior within social systems (Lucchini 2010; 2006; 2002). In this paper, we explore, for a specific intra-urban area (Rouen, French urban area), the complex spatial mechanisms linked to an urban activity development. We specifically study the development of cultural sites in time and on a specific city space, as well as the practice of social individuals to these cultural sites. We are also interested in understanding and modeling the adaptive mechanisms of these cultural sites according to their user practices (Ghnemat 2009).

COMPLEXITY OF SPATIO-TEMPORAL SYSTEM DESCRIBING URBAN CENTER DEVELOPMENT

Territory, functional system of interrelations between men and their space

This works results from dialogue between geographers and computer scientists, in order to understand territory functionnalities. Territory is usually grasped by geographers as a system of interrelations linking itself with its inhabitants. "How territory works as system?" is the core of our approach. First, we need to understand what are the relevent components of this system, then to highlight the associated mechanism of actions on them to finally being able to design operational objectives of management. Temporal and spatial dynamics act jointly on all territories. For instance, urban change is a permanently renewed process that geographers have integrated since long time ago within their analysis of urban territories (Benenson and Torrens 2004, Batty 2005). Dynamics of such changes can be expressed in following terms: how urban territory evolves when a new activity takes place on a specific urban location? For instance, what will be the impact of new cultural structures like a cinema multiplex or a new music scene in requalifying urban zone produced by industrial and harbour brownfields? It is precisely what's happend on the urban area of the French city of Rouen. The process of requalifying old sites into new uses began in Rouen since 2009,

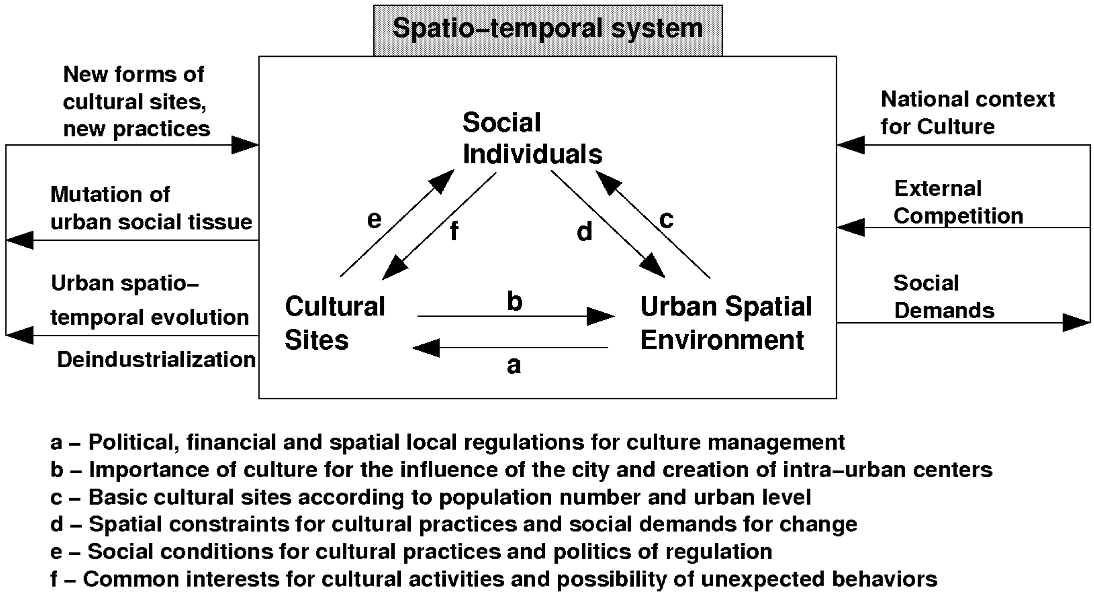


Figure 1: Spatio-temporal system to model the actors involved in urban cultural site development

particularly in an important part of its central area, previously devoted to intensive industrial and harbour activities on each side of the Seine river. "Seine Ouest" project deals with the reality of urban rehabilitation by the development of cultural centers and by their attractive potentiel they are able to generate in terms of visitor number on this urban zone.

Spatio-temporal model of relations between producing actors and users within cultural centers

To reflect the complex reality of the creation and practice of these cultural sites, we have built a theoretical model of relationships between actors who produce and use the cultural sites. For this purpose, the model proposes to experiment a spatio-temporal system consisting of three groups of actors; the social individuals, the cultural sites and the urban spatial environment, as described on Figure 1. The complexity results from the interleaving of different types of interactions:

- The complexity coming from interactions between social individuals and producing self-organization (Bertelle et al. 2009);
- The inherent complexity of social individuals characterized by multiple criteria which are themselves in interaction (age, gender, educational level, social level, etc.);
- The game of competition and/or complementarity in the attractiveness of cultural sites varying on several levels of attractivity depending on their nature and their location (for example, the attractive power of a cinema in comparison with the attractive power of an opera, or, as another example, the attractive power of a site at central urban location in comparison with another site at peripheral urban location);
- The effect of the general spatial configuration of a city being more or less constraining and so acting on the formation of self-organized dynamics (as examples, the barrier effect from a river or from an industrial area, the attractor effect coming from a municipal center or from a high density residential area, the accessibility produced by road transport planning and by public transport planning).

Spatial dynamics and temporal dynamics are mixed. It is usual to observe that, after some periods, bifurcations appear in urban planning: for instance, it is the case of the reorganization of brownfields into cultural centers, within spaces previously dedicated to productive activities and so repulsive to cultural sites.

ROUANTS SIMULATION PLATFORM TO MODEL USAGE DYNAMICS OF CULTURAL SERVICES

Multi-center multi-criteria dynamical model specification

The goal is to understand and to analyse the cultural practices in city. *Rouants* is a dynamical model for studying attractiveness of cultural sites in city. It is built on the achievements of the sociology on culture in order to identify the behavior of people depending on

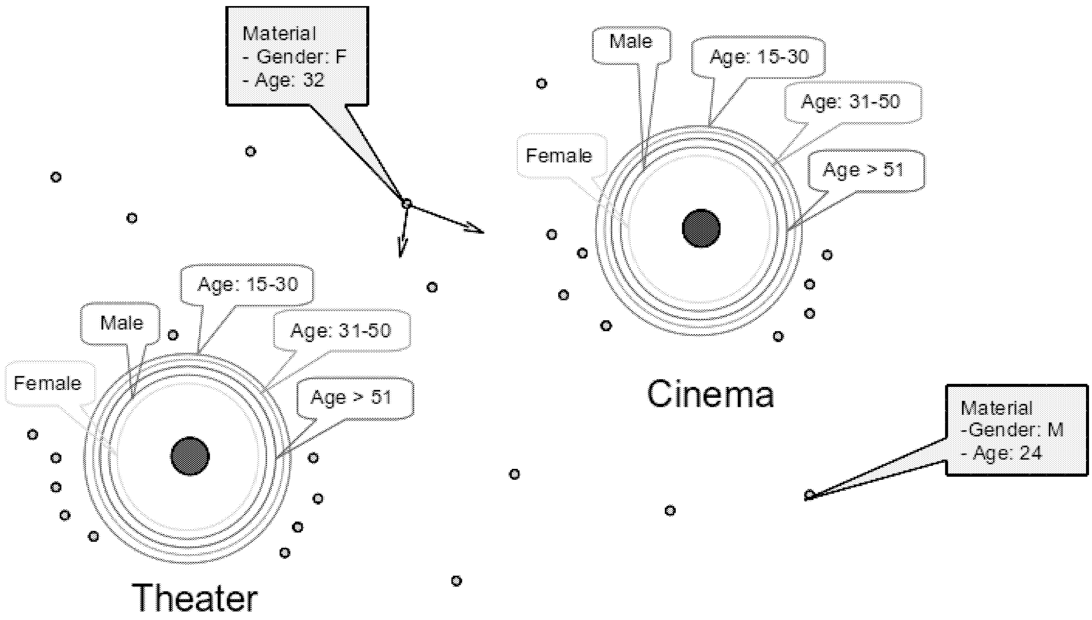


Figure 2: Cultural equipment dynamics modeling

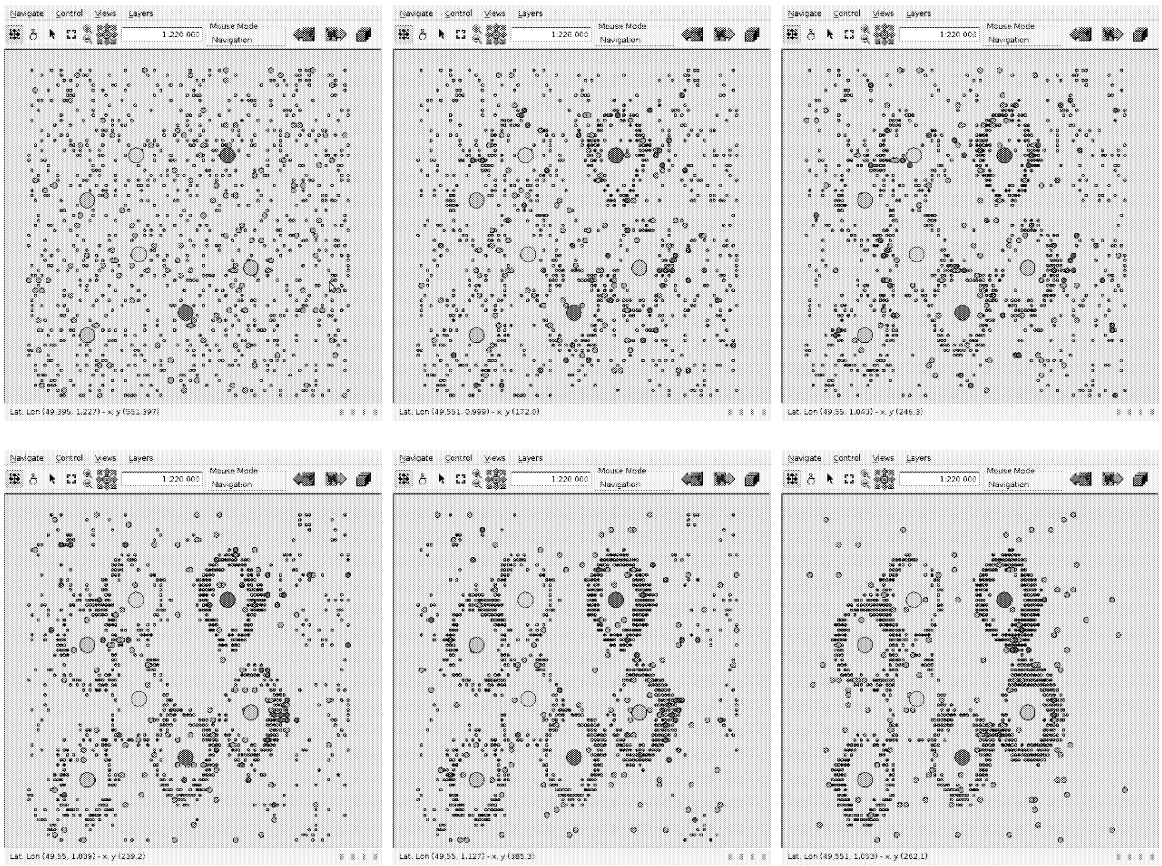


Figure 3: Virtual urban area prototype simulation computation on Repast, at successive steps from top to bottom and left to right: iteration 0, iteration 152, iteration 250, iteration 370, iteration 601, iteration 1601.

cultural sites. The model was applied to the design of cultural offer for the city of Rouen. We use swarm intelligence methods and stigmergy concept to model the dynamics of cultural center development. Specifically, we rebuild, by dynamical simulation, a multi-centre and multi-criteria system. The people dynamics, movement and decision, are inspired by self-organized processes existing in collective movements designed by swarm intelligence and stigmergy (Bonabeau et al. 1999, Camazine et al. 2001). These processes correspond to the dialogue between individuals and environment for a collective benefit. The model includes a mixture of deterministic and probabilistic behaviors, corresponding to the part of unpredictability in human behavior.

Even if ants are used in this bio-inspired model, we don't consider that ant behavior is able to represent human behavior. Ant behavior is used to implement self-organization processes; human behavior is introduced in complex mechanism linking many criteria and specific knowledge expressed by statistical values of social practice described by the pheromone attraction function.

Figure 2 is a schematic representation of an easy-to-understand simulation composed of two cultural centers, a cinema and a theater. We represent on this figure, the pheromone/attraction functions as circles around the centers and we label them with the associated characteristics/colors. Materials represent here users which will be carried and moved by artificial ants, major components of our decentralized computation: their move are controled by a complex spatial system of attraction functions mixed with individual decision making. Mathematical formula of this computation can be found in (Ghnemat 2009)

Virtual urban area prototype

Figure 3 is describing Virtual urban area prototype as an experimental configuration with seven centers and initially random places for the peoples/materials and the ants. Each center is emitting heigh attractive pheromone functions associated to a color labeled from 0 to 7. On this figure, we show the result of one simulation where ants progressively aggregate the materials around the centers, following pheromone trails and clustering algorithm. On the left top sub-figure, we see the initial distribution of materials and ants. In the other sub-figures, we see successive steps of simulation. We can observe the formation of material affectation to each center in order to respect the attraction process, according to the material characteristics. In (Ghnemat 2009), the analysis of the attraction phenomena is detailed, allowing to better understand how the computation generates self-organization and distributes the material between the centers.

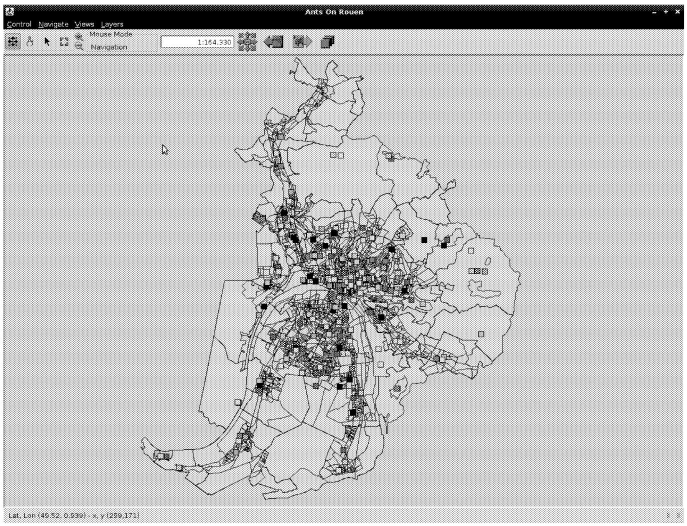
Rouants simulation platform

Rouants simulation platform allows to dynamically experiment the complex interrelationships mixing the three types of actors previously described: social individuals, cultural sites, urban spatial environment. This framework implements decentralized approaches and so is devoted to built self-organized mechanisms of collective cultural behavior. The results of the simulation allow to reveal the emergence of spatio-temporal forms from the interleaving of multiple types of interactions and from the spatial constraints coming from urban system(Ghnemat et al. 2007, Ghnemat 2009, Crooks 2007).

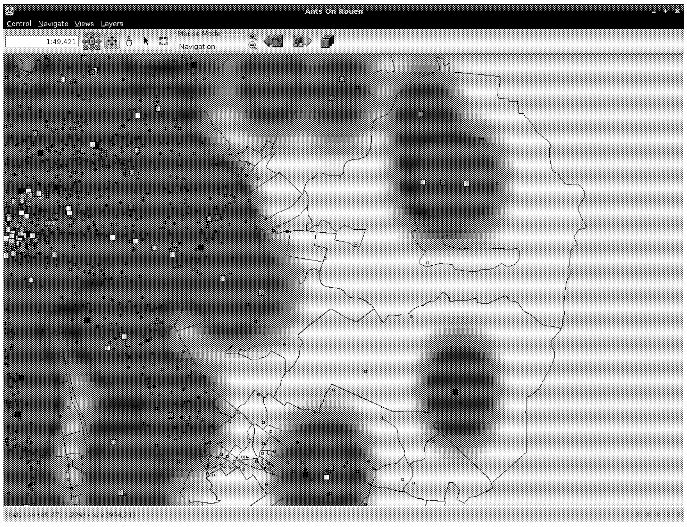
Rouants implements the previously described algorithms of collective intelligence and integrates them inside a specific GIS, representing the cultural sites and the block housing over the urban area of Rouen in High-Normandy. Figure 4 shows different visualizations produced by this platform. On the sub-figure (a), we see the whole distribution of cultural sites and centers over the urban area; the sub-figure (b) is a sumple vue allowing to observe how the pheromones are spreading over the space; the sub-figure (c) described an analysis of some specific places in this agglomeration, tracing on the graphes on the left side of this sub-figure, the temporal series of number of materials being attracted in two zones, according to the dominant color/criteria of the materials.

Conclusion and Perspectives

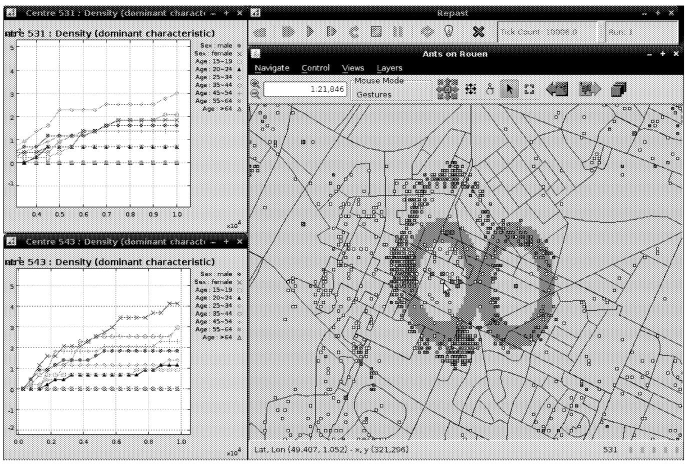
We illustrate in this paper the possibility of experimenting dynamics of multiple services-users in city. *Rouants* simulation platform is an effective tool for approaching cultural practices in urban situations. Decentralized mechanisms reinforce the suitability of the type of data and methods used to the study of socio-spatial complex systems like cities. We observe emergence of multiple cultural aggregation phenomena in relation to the sociology of culture and to the specific urban and social environment of the city of Rouen. The simulation highlights territorial situations of rivalry and complementarity in cultural practices. The modeling concept of the platform allows also to experiment adaptive capacity and feedback process like politics of regulation of cultural centers, according to the user practices. The social and territorial impacts of these mechanisms are one of our further perspectives and analysis. There is also a general evidence: the use of cultural heritage is a vector for both cultural policy and cultural development.



(a) Rouen Cultural equipments GIS



(b) pheromone density visualization



(c) result output on GIS and graphics

Figure 4: Visualization of *ROUANTS* simulation platform

Acknowledgements

This work is part of the multidisciplinary project RISC (French acronym for Complex Networks and Complex Systems), supported by the French Ministery of High-Education, the Region High-Normandy and its Research Network on Transportation, Logistics and information Processing (GRR TLTI) and by European Union with FEDER grants. With this support, a part of the final implementation of the framework *Rouants* has been made by Florent Marchand de Kerchove. Docteur Rawan Ghnemat was granted by a French Government Scholarship.

REFERENCES

Batty M., 2005. *Cities and Complexity.* The MIT Press.

Benenson I. and Torrens P.M., 2004. *Geosimulation.* Wiley.

Bertelle C.; Duchamp G.H.; and Kadri-Dahmani H. (Eds.), 2009. *Complex Systems and Self-Organization Modelling.* Understanding Complex Systems. Springer.

Bonabeau E.; Dorigo M.; and Theraulaz G., 1999. *Swarm intelligence - from natural to artificial systems.* Santa Fe Institute Studies in the Sciences of Complexity. Oxford University Press.

Camazine S.; Deneubourg J.L.; Franks N.R.; Sneyd J.; Theraulaz G.; and Bonabeau E., 2001. *Self-Organization in Biological Systems.* Princeton University Press.

Crooks A.T., 2007. *The Repast Simulation/Modelling System for Geospatial Simulation.* Tech. rep., UCL Working Papers Series. Paper 123.

Ghnemat R., 2009. *Adaptive modeling for spatial emergence within complex systems.* Ph.d. thesis, University of Le Havre.

Ghnemat R.; Bertelle C.; and Duchamp G.H., 2007. *Adaptive Automata Community Detection and Clustering.* In *ICCIIS 2007 within World Congress on Engineering (WCE 2007).* Newswood Limited, International Association of Engineers, London, U.K., vol. Volume 1, 25–30.

Lucchini F., 2002. *La culture au service des villes.* Villes. Anthropos, Economica.

Lucchini F., 2006. *Un SIG pour la culture : un instrument de réflexion sur la localisation des équipements culturels dans une agglomération.* *Géographes Associés*, 30, 237–243.

Lucchini F., 2010. *La fabrique des lieux culturels.* Hdr, University of Rouen.

SIMTRAP
A MULTI-AGENT MODEL FOR DECISION SUPPORT IN SPATIAL PLANNING OF THE PASSENGERS' SPACES IN PUBLIC TRANSPORT SYSTEMS

Jérémy Fiegel
Cyrille Bertelle
LITIS - Université du Havre
25 rue Philippe Lebon - BP 540
76620 Le Havre - France
jeremy.fiegel@hotmail.fr
cyrille.bertelle@univ.le-havre.fr

Arnaud Banos
UMR Géographie-Cités - CNRS
13 rue du Four
75006 Paris – France
arnaud.banos@parisgeo.cnrs.fr

KEYWORDS
Multi-agent model, spatial planning, pedestrian simulation

ABSTRACT

In order to provide decision support in spatial planning of the passengers' spaces in public transport systems and through a partnership between the University of Le Havre, the CNRS and the RATP, a new multi-agent model has been developed. It allows the modeling and the simulation of corridors, platforms, interiors of trains, etc. and the behavior of travelers who walk on them.
The decisional process of these simulated pedestrians is described: it depends on their local environment, but also their primary and secondary objectives and their state of satisfaction. These rules will also result in emergent behaviors such as the formation of lines of people along corridors or "funnels" near train doors.
Finally, while still being validated, an example of results is presented: the comparison of two versions of the interior of a Paris subway train.

INTRODUCTION

Large companies of public transportation, such as RATP (public transport system in Paris), continually look to optimize the organization of the spaces used by their customers (spatial arrangement of trading rooms, corridors, platforms and interiors of the vehicles) and to answer questions like: How many seats should be put in a train car? What shape and orientation should they have? Is it possible to temporarily close an exit of a platform without overcrowding the others? How large a construction zone on a platform can be without creating a risk for travelers, even in rush hours?
This research is complex: for each new idea, it is necessary to ensure that the new spatial arrangement is worthwhile in terms of efficiency, comfort and safety, for all types and numbers of travelers. Furthermore, experimentation is often impossible or very expensive: it is rare to be allowed to close a station or to build a new train only for testing this type of ideas. Finally, the pedestrians' reactions may be contrary to what the engineers had expected.
To assist them in their decisions, these companies turned to modeling and simulation of their spaces and of the flows of passengers who use them.

But the models they still use nowadays to deal with this problematic often work on a macroscopic level (crowd simulation) and the resulting loss of precision and details is increasingly felt. Indeed, some common situations can be dramatically changed in seconds by the actions of a single passenger using a small element of the environment. For example, the departure of a train may be delayed because a single person, sitting on a folding seat near a door, refused to stand up quickly and therefore, caused a slowdown of the exchanges near his position.

BANOS and CHARPENTIER (2010) describes the various constraints to be included in a model to manage to get finer simulations at a microscopic level, focusing on individuals. Such a model must include: highly detailed and scaled environments, a spatial and temporal precision fine enough for the studied situation, the ability to simulate a realistic number of pedestrians with some physiological and behavioral heterogeneity and to integrate both local interactions (between individuals and between an individual and the environment) and global interactions (collective behavior).

The model described here, called SimTRAP (for Simulation of Exchanges between TRains and Platforms), was created through a partnership between the University of Le Havre (LITIS), CNRS (Geographie-Cites and ISC-PIF) and RATP. Its primary goal is to simulate the exchanges of passengers between platforms and subway or suburban trains as accurately as possible, within the constraints identified above. These exchanges are one of the most complex and most critical steps in the journey of the passengers, especially during rush hours. SimTRAP is a multi-agent model. Such models take better account of the diversity and complexity of individual behavior (different strategies, motivations, etc.) and this choice follows the example many other works on this theme (HAKLAY et al. (2001); KERRIDGE et al. (2001); DAAMEN (2002); BATTY (2003); LACROIX et al. (2006); PELECHANO et al. (2007); BANOS and CHARPENTIER (2010)). It is based on the multi-agent system NetLogo (WILENSKY (1999)).

SIMULATION ENVIRONMENT

A large number of objects (or obstacles, such as walls, seats, folding seats, doors, handrails, etc.) and spaces (such as corridors, entrances, exits, platform edges, etc.) found in areas of transport can be represented in SimTRAP. Each element is represented by a polygon. Their shape and position are generated from a file from a Geographic Information System (GIS) and they have several additional attributes like size, density, bonuses or penalties to the speed of their occupants, etc. Manipulable obstacles such as doors and folding seats can also be opened or occupied on certain conditions. Figure 1 shows an example of an environment used by SimTRAP.

These elements are arranged on a grid of square "patches" which are small enough to maintain a high level of detail. They usually have the size of the smallest item: the diameter of handrails (5 or 10 cm). Each of these cells computes their distance to the different attractors (using gradients and propagation) that are entrances, exits, seats, handrails, etc. As these areas or obstacles may not be attractive throughout an entire simulation and for all passengers, several values are computed. These affect travel choices of individuals.

PASSENGERS' BEHAVIORS

During the simulations, a number of passengers are created in these environment and try to fulfill their goal. The model can generate only a few of them (off-peak situations), up to hundreds (rush hours) or thousands of travelers (incidents). These individuals (or agents) have several characteristics and follow a few simple rules allowing them to choose their actions and then, to accomplish them. Finally, thanks to this individual strategy, emergent behaviors can be observed, such as the formation of "lines" when passengers walk towards the same goal and "funnels" when some of them wait to board and there are others wishing to go down.

After their creation, each passenger receives a mission (or primary objective), a size (shoulder width) and a maximum speed. The mission may by to reach an exit of the environment and thus to leave the simulation, or another area and wait there until the end of the simulation, while trying to maximize comfort, especially in approaching an attractor.
It is also possible to specify some attractors to be the preferred ones for some choices and thus create "types" of travelers. For example, regular passengers can receive a higher maximal speed: in the real world, these people rush about in the corridors or platforms. If their journey is short, they will also choose to stay near doors, rather than sitting down far away.

In addition to the above characteristics, at each step of simulation, each traveler is updating its secondary goal, state of satisfaction and, after evaluating his available actions, his position if he decides to move.
Each mission is broken down into several secondary objectives (or steps) that the traveler must necessarily fulfill to achieve his goal. For example, if a person on a platform want to be in a train, he has to get on board, after choosing a door, reaching it and opening it. At each step, the attractors may differ and so, some agents may have the same mission, but take different paths. In addition, the fulfillment of some objectives can be delayed or prevented by an element of the environment. In the previous example, the success of the boarding step depends on the ability of travelers to open their choosen door at the right time. Finally, there are secondary objectives in which only one attractor must be selected from a huge set: train doors, train seats...

Then, the secondary objectives allow travelers to compute their satisfaction. This attribute is used to represent a kind of "comfort": people dissatisfied with their position or with their local environment will try to act in order to be satisfied again.
Dissatisfaction can have two origins: because of the current position, which does not allow the fulfillment of the current secondary objective, or, during waiting steps on a platform or in a train, because of local discomfort, such as the too close presence of other agents.
Once dissatisfied, the traveler enters in a process adapted from the eco-resolution system of FERBER (1995), in which he will evaluate the different new positions that he may choose to go (up to a certain radius) and decides to stay where he is or to move to the best destination he found. The best new position is determined through a comparison based on the distance and the angle between the possible destinations and the optimal one (without obstacle or inconvenience on the way).

There are some specific actions added to these general rules, such as the decision to stand and fold up the occupied seat when the local density exceeds a certain threshold.

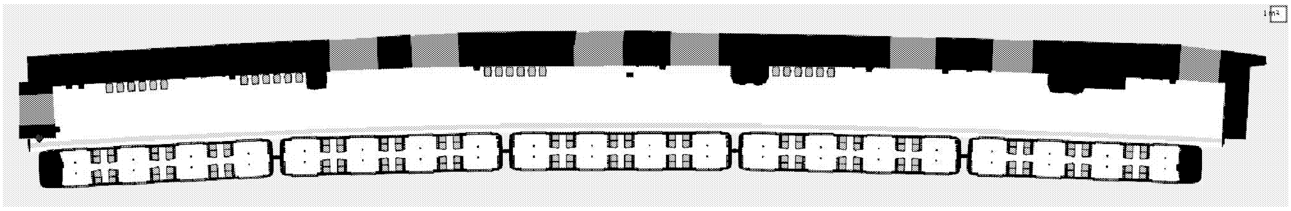


Figure 1: Example of environment used by SimTRAP representing the platform towards Nation of the subway station Charles de Gaulle – Etoile on line 2 in Paris and a parked MF67 train (w: 82 m, h: 12 m).

It is important to note that the SimTRAP model is stochastic. Indeed, between the initialization of passengers which is almost never the same due to the variability of their characteristics, and their choices depending on their previous actions and on those of other people (objectives and positions), two simulations have very little chance to be identical.

Finally, even if simulated people seem to care only for their own state, some collective behavior may emerge, especially in wide environments with not many obstacles (corridors or platforms).
First, in the same way as in other studies, as in the social force model of HELBING and MOLNAR (1995), it can be observed the formation of "lines" of people wanting to reach the same area or to go through the same places (example in figure 2).

Similarly, in simulations of exchanges of passengers between a platform and a train, the agents wishing to get on board, in the presence of other agents wishing to get down, create some kind of queue called "funnel" on both sides of the doors, but only below a certain density that remains to be computed. As in reality, at some point, travelers end up directly in front of the doors and then, they greatly slow down the exchange.

EXAMPLE OF RESULTS AND ANALYSIS

The model, with the multi-agent system NetLogo, can save a lot of different data: local and global densities, speeds, number of travelers who have reached their objective, etc. These results can be displayed in the form of means, extrema or lists to see their changes over time. This is interesting in that a change in the environment may result in changes in values outside of those originally involved. For example, increasing the capacity of a train by changing the configuration of the seats (gain) may result in a slight decrease of the boarding speed (loss), which is a drawback not always expected.

A concrete example of its use is the comparison in terms of maximum density and boarding speed of two differents arrangements of the same area, such as the different interiors of a MF67 train car of the Paris subway. Theses cars, equipped with four openings on the platform side, are available in two versions: the first contains six sets of four regular seats facing pairs by pairs with four folding seats attached to their back (configuration shown in figure 1); the second contains only three of the previous sets, the others have been replaced by three sets of four regular seats fixed to the cabin and with no folding seat. In this second version, fewer people can sit, but, theoretically, there should be more that can stay in standing areas.
After some simulations, during which a large number of passengers tried to get on board, it has been observed that there are not more people who can fit in the second version of the train, but less (see Figure 3). In fact, the gained surface is mainly between sets of seats or near the doors. These areas are less attractive than seats and they do not compensate the loss of some of them. The results are identical to those noticed by the RATP, after counting some exchanges in real life.

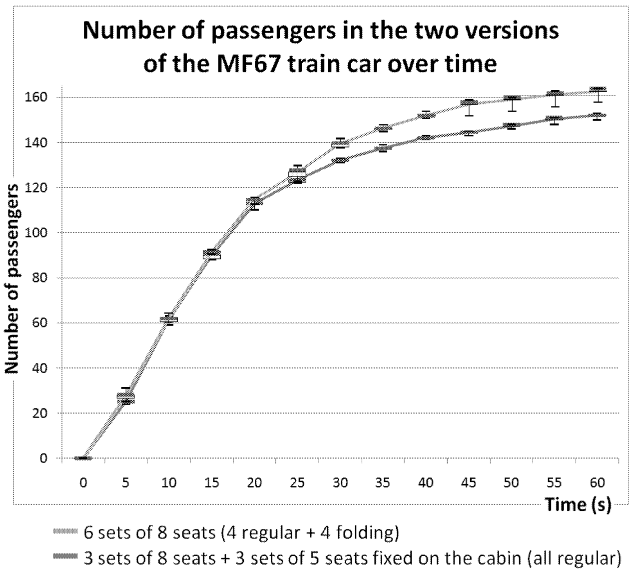


Figure 3: Simulation results of the filling of two different arrangements of the same MF67 train car.

But this example is one of the simplest tests achievable

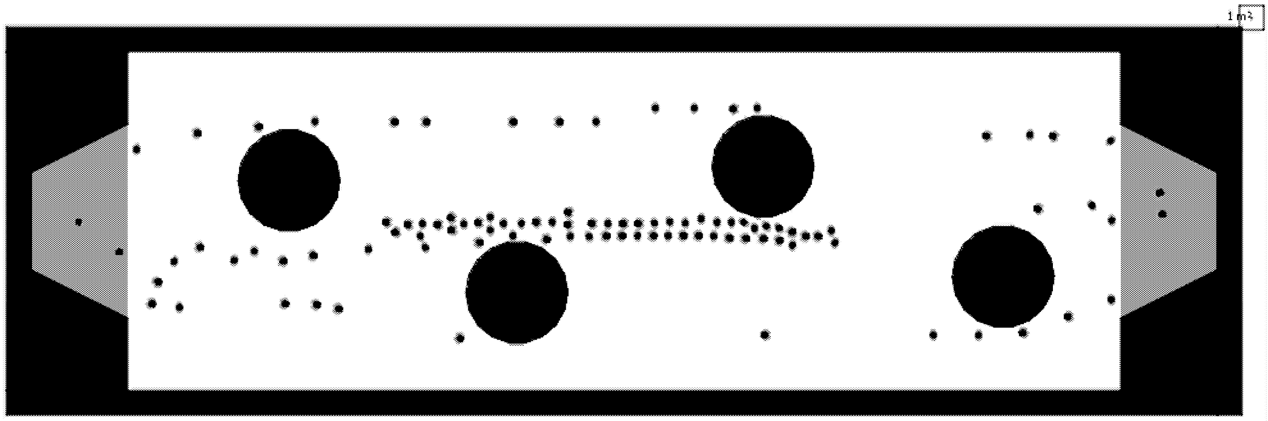


Figure 2: Example of the formation of two lines of travelers with the same objectives in a simulation (the upper line wants to go to the red area on the right, the lower line wants to go to the one on the left).

by the SimTRAP model. In many other situations and because the model is stochastic, it is necessary to compute a large number of simulations to obtain reliable data, according to all possible configurations of population. For example, to study the case of an exchanges of passengers between a platform and a train thoroughly, it is necessary to vary at least the number of passengers wishing to get on board and the number of those wishing to get down.
This is an aspect of the SimTRAP model that it remains to be tested. A new collaboration with the Institute for Complex System of Paris-Ile de France (ISC-PIF) and the use of their model, called OpenMole REUILLON et al. (2010), which allows the easy use of grid computing, will allow to do large numbers of simulations with SimTRAP.

REFERENCES

BANOS A. and CHARPENTIER A., 2010. *Simulating pedestrian movement in dynamic environments. Cybergeo.* URL `http://cybergeo.revues.org/index23131.html`.

BATTY M., 2003. *Agent-Based Pedestrian Modelling*, UCL Centre for Advanced Spatial Analysis, chap. 5. 26. URL `http://www.casa.ucl.ac.uk/working_papers/paper61.pdf`.

DAAMEN W., 2002. *SimPed: a pedestrian simulation tool for large pedestrian areas.* In *European Simulation Interoperability Workshop 2002.* URL `http://www.tudelft.nl/live/ServeBinary?id=ce418bda-1404-47ac-bcb3-ddbd7250efb2&binary=/doc/EuroSIW02.pdf`.

FERBER J., 1995. *Les systmes multi-agents : Vers une intelligence collective.* InterEditions. URL `http://www.lirmm.fr/~ferber/publications/LesSMA_Ferber.pdf`.

HAKLAY M.; O'SULLIVAN D.; THURSTAIN-GOODWIN M.; and SCHELHORN T., 2001. *"So go downtown": simulating pedestrian movement in town centres. Environment and Planning B: Planning and Design*, 28, 343–359. doi:10.1068/b2758t. URL `http://eprints.ucl.ac.uk/10974/`.

HELBING D. and MOLNAR P., 1995. *Social force model for pedestrian dynamics. Physical Review E*, 51, 4282–4286. doi:10.1103/PhysRevE.51.4282. URL `http://link.aps.org/doi/10.1103/PhysRevE.51.4282`.

KERRIDGE J.; HINE J.; and WIGAN M., 2001. *Agent-based modelling of pedestrian movements: the questions that need to be asked and answered. Environment and Planning B: Planning and Design*, 28, no. 3, 327–341. doi:10.1068/b2696. URL `http://EconPapers.repec.org/RePEc:pio:envirb:v:28:y:2001:i:3:p:327-341`.

LACROIX B.; MATHIEU P.; and PICAULT S., 2006. *Time and Space Management in Crowd Simulations.* In *Proceedings of the European Simulation and Modelling Conference.* 315–320. URL `http://lacroixbenoit.free.fr/papers/crowd_simulation_esm06_lacroix.pdf`.

PELECHANO N.; ALLBECK J.; and BADLER N., 2007. *Controlling individual agents in high-density crowd simulation.* In *Proceedings of the 2007 ACM SIGGRAPH/Eurographics symposium on Computer animation.* Eurographics Association, Aire-la-Ville, Switzerland, SCA '07, 99–108. URL `http://portal.acm.org/citation.cfm?id=1272690.1272705`.

REUILLON R.; CHUFFART F.; LECLAIRE M.; FAURE T.; DUMOULIN N.; and HILL D., 2010. *Declarative task delegation in OpenMOLE.* In *HPCS.* 55–62.

WILENSKY U., 1999. *NetLogo.* Tech. rep., Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL.

Multi-agent constraint optimization approach to modeling of demand responsive transport systems

Habib Abdulrab
LITIS, INSA-Rouen

Avenue de l'Université - BP 8,
76801,Saint-Étienne-du-Rouvray,
France
E-Mail: abdulrab@insa-rouen.fr

Eduard Babkin and Eugene Ivlev
Department of Information Systems
and Technologies
National Research Univeristy „Higher
School of Econoimics"
B.Pechorskaya 25/12,
603155, Nizhny Novgorod,
Russia
E-Mail: eababkin@hse.ru

KEYWORDS
Multi-Agent systems, Demand Responsive Transport, Distributed Constraint Optimization Problem, Dynamic Programming Optimization Protocol.

ABSTRACT

The paper discusses new Multi-Agent approach to intellectual management of distributed Demand Responsive Transport systems based on Distributed Constraint Optimization. It shows how we can formulate problem in terms of constraints and solve it with Dynamic Programming Optimization Protocol. Also the paper describes the software design of the suggested model prototype and simulation results. The purpose of this work is the developing theoretical and algorithmic foundations for a decentralized Multi-Agent system which achieves a global optimal solution in Demand Responsive Transport problem.

INTRODUCTION

Today in big cities a problem of passenger transportation is quite difficult. Contemporary transportation services provided by private companies can't solve their tasks because of several issues and don't satisfy citizen requirements. The reasons for this are traffic jams, weak connection of central areas and outskirts by regular transport and high prices. For citizens will be actual such transportation that is cheaper than the taxi, but one is able to process individual requests. At the same time the competition is constantly increasing among the companies which are providing public transportation services. Therefore costs decreasing by using smart information systems becomes very actual task.
To solve above pointed problems researchers have suggested different models referred to as «Intelligent Transportation Systems» (ITS) (Chevrier et al. 2006). One of these is «Transport on demand» (or «Demand Responsive Transport», DRT) (Abdulrab et al. 2008). Usually in DRT model customers are able to place an order of their transportation from one point to other point within the boundary of some geographical area (e.g. in a city). The transportation company processes the placed order (request) and calculates the optimal solution taking into account all requests arrived for the current time. The optimal solution consists of assigning each request to any vehicle with the optimal route of travelling to all vehicles so that the company is decreasing its total operational costs.
In practice DRT model is implemented as a complex information system. In order to make such the system practically relevant and beneficial researchers need to be working into two main directions. First of all, the system should provide passengers-customers with convenient and clear end-user tools implementing "anywhere and anytime" paradigm: the passengers should be able place the orders, analyze and select the results using multiple emergent communication technologies like mobile phone call, sms-services, or some software on their computers. On the other hand, the system should provide the transportation company with effective decision support tools for costs optimization and dynamic change of the number of vehicles according to the variable demands.
In our research we pursue the goal of developing appropriate decision support algorithms for the aforementioned systems, and seek affordable solutions of two main problems of decision support in the DRT domain: (1) routing and (2) distribution of requests.
Until now different approaches were proposed for those problems. Some approaches consider DRT as a centralized system (Jin 2009; Cubillos et al. 2006), others consider it as a decentralized or distributed system (Abdulrab et al. 2010; Cubillos et al. 2006). Recent works explored such different methods for modeling routing and distribution as combinatorial auctions (Abdulrab et al. 2010) and refinement planning (Kambhampati 1997). Among different generic approaches our attention was attracted to DCOP method (Petcu and Faltings 2004) and DPOP optimization algorithm (Petcu and Faltings 2005).
In this work we offer new multi-agent distributed DRT model designed using DCOP method, description of the main algorithms incorporating DPOP algorithm and design solutions for the corresponding information system. Our proposed algorithms resolve both mentioned DRT problems (the routing problem and the problem of optimal requests distribution among vehicles) in such a manner, that transportation companies will obtain maximum profit and

citizens will be fully satisfied by the service. User satisfaction coheres with decreasing cost of the trip and its duration. The transportation company gains more due to using of a smart planning and allocation algorithms. We consider that task as dynamic and distributed so that we can use Multi-Agent paradigm for the modeling. Following that paradigm we introduce two types of agents: vehicles and customer requests. Also we add two types of constraints between these agents which will be described in details in the next sections.

This paper consists of five sections. In Section 2 we outline our foundations – DCOP method and DPOP optimization algorithm for solution of the distribution problem. Section 3 formally defines proposed DRT model in terms of the paradigm of Multi-Agent Systems (MAS). Section 4 provides for offered software design of multi-agent information system which corresponds to the proposed DRT model, and some simulation results. Section 5 concludes this work and outlines major achievements and directions of future work.

FOUNDATIONS: DCOP METHOD AND DPOP ALGORITHM

Adoption of Multi-Agent paradigm for solution of DRT problems implies using nontrivial discrete optimization algorithms. For solution of the most complex task in our model, namely, the distribution problem, we decided to apply generic DCOP method (Distributed Constraint Optimization Problem), originally proposed by A. Petcu in (Petcu 2004). DCOP is an extension of well-known Distributed Constraint Satisfaction Problem (DisCSP) (Yokoo et all. 1992) for the problems which include quality, or cost as some of the problem dimensions, becoming a kind of an optimization problem. To formalize a particular problem DCOP method uses notions of variables and constraints which bound their values. Constraints are divided in two types – soft constraints and hard constraints. Hard ones have to be observed, soft ones should be optimized (minimized or maximized in different problems).

In order to solve the problems formalized in accordance with DCOP method researches have developed a number of algorithms like Asynchronous Distributed Optimization (ADOPT) (Modi et al. 2005), Optimal Asynchronous Partial Overlay (OPTAPO) (Mailler and Lesser 2005), Distributed Optimization Procedure for Tree-structured networks (DTREE) (Petcu and Faltings 2004), Dynamic Programming Optimization Protocol (DPOP) (Petcu and Faltings 2005) and others.

Whereas other algorithms require an exponential number of messages to be sent between agents, DPOP requires a linear number of messages. This is important in distributed settings because sending a large number of small messages (like search algorithms usually do) typically entails large communication overheads.

DTREE is correct only for tree-structured networks. In general, for networks with arbitrary topology the most appropriate solution is an algorithm on the base of DPOP. It has been proven to be correct and complete, what means that it will find the best solution of a problem if a solution to this problem exists.

As described in (Petcu and Faltings 2004) DPOP consists of three phase:

- DFS arrangement;
- UTIL propagation;
- VALUE propagation.

In phase 1, a DFS traversal of the graph is done using the algorithm of pseudo-tree construction (Faltings et al. 2008). The obtained DFS plays a role of communication structure for the remaining two phases of the algorithm: UTIL propagation (UTIL messages travel bottom-up on the tree), and VALUE propagation (VALUE messages travel top-down on the tree).

A pseudo-tree arrangement of the graph G is a rooted tree with the same vertices as G and the property that adjacent vertices from the original graph fall in the same branch of the tree (Petcu and Faltings 2004). DFS algorithm needs election of the root vertex. Since we deal with independent intelligent agents we shouldn't assign that agent which will be a root. For example, we give them an opportunity to elect root vertex with Leader Election algorithm described in (Faltings et al. 2008).

Phase 2 - UTIL propagation: this is a bottom-up process, which starts from the leaves and propagates upwards only through tree edges. In this process, the agents send UTIL messages to their parents. These messages summarize the influence of the sending agent and its whole sub-tree on the rest of the problem. The semantics of such a message is similar to an n-ary relation having as scope the variables in the context of this message (its dimensions). The size of such a message is the product of the domain sizes of the variables from the context.

Phase 3 - VALUE propagation top-down, initiated by the root, when phase 2 has finished. Each agent determines its optimal value based on the computation from phase 2 and the VALUE message it has received from its parent. Then, it sends this value to its children through VALUE messages. After the VALUE propagation is ended each agent has the single value assignment from the domain of its variable.

FOUNDATIONS: DCOP METHOD AND DPOP ALGORITHM

DRT system is usually determined as an intermediate type of transport between classic regular transport and taxi system. A customer notifies desire to be transported from one point to another assuming that the request will be performed within a defined time window. The transportation company receives requests from all customers and allocates them between vehicles taking into account some constraints. Routes, number of vehicles and other parameters are dynamically defined according to the current state of the transportation network.

Using general paradigm of multi-agent systems we are going to describe such problem in a formal way. In our DRT model the transport network is defined as an oriented graph whose nodes are bus stations, arcs are roads associated with distance between the stations. M – is set of customers, N – is set of vehicles. We define two types of the agents as constituents of the typical model. First type of agents is *vehicle-agent*, second type of agents is *request-agent*. In our model each passenger request is defined by four parameters:

$$(StartPoint_Z,\ EndPoint_Z,\ price_Z,\ wt_Z) \qquad \mathbf{(1)},$$

where *StartPoint, EndPoint* – start and end locations of the requested travelling, *price* is an amount of money paid by customer for the trip and *wt* is waiting time of vehicle arrival at the starting point, $z \in$ M. Each vehicle also has some parameters as follows:

$$(Point_i,\ Speed_i,\ Capacity_i) \qquad \mathbf{(2)},$$

where $Point_i$ – current location of the vehicle, $Speed_i$ – average speed and *Capacity* is the maximum allowable number of passengers for the vehicle.

There is also one control agent in our model – the *agent manager*. It is responsible for receiving requests and supporting collaboration between all other agents.

To facilitate coordination and making collective decisions among the agents we suggest to apply DCOP (Petcu and Faltings 2004) – a method of formalization that allows to use DPOP algorithm (Petcu and Faltings 2005) for achieving the global optimal outcome in the distributed Multi-Agent system.

As the first approximation in the course of application of DCOP method to our DRT model, let us imagine that each request can be processed by each vehicle. Vehicle-agents and request-agents have one variable per agent. Vehicle-agent's variable domain includes all possible request assignments to this vehicle. For example, if the number of requests equals to 2, each vehicle's variable domain consists of values: 00, 10, 01, 11. Value 00 means that neither request won't be assigned to this vehicle, on the contrary 11 means that both requests will be served by this vehicle. The defined variable is associated with a unary constraint, which maps requests assignment to correspondent profit. Vehicle's profit is defined by the following equation:

$$P_i = \sum P_{KM} D_i - P_F D_A \qquad \mathbf{(3)},$$

where P_{KM} – the price per kilometer (defined by certain pricing policy of the company), D_i is the distance between starting and end points of the customer *i* in a route which is built by the vehicle-agent for this assignment. P_F is the fuel cost per kilometer and D_A is an overall distance of the route.

For the request-agent's variable, in this case, its domain will cover all vehicles. So if we have 2 vehicles with UIDs V_1 and V_2, then each request-agent's variable has the following possible values: $\{V_1, V_2\}$.

In addition to the constraint considered above we have to add two kinds of hard constraints. We use the binary constraints to ensure that a request is taken by at most one vehicle. When two vehicles can process the same request, a binary constraint is added to prevent them to choose assignments that are conflicting. However, because of the exploited optimization DPOP algorithm (it does not use hard constraints), we need to reformulate those hard constraints as correspondent soft constraints, that is giving a penalty to invalid combinations of values. Hard constraints can be naturally simulated by utility functions which assign 0 to feasible tuples of values and $-\infty$ to infeasible ones.

Therefore we may reformulate original hard constraints as follows:

Rule 1. Whenever the chosen assignment of the vehicle-agent does not contain a request which it should contain according to the corresponding request-agent, a penalty is given.

Rule 2. Whenever the chosen assignment of a vehicle-agent contains a requests which it should not contain according to the corresponding request-agent, again a penalty is given.

Application of the proposed constraints is explained in the Example 1.

Example 1

There are two vehicle-agents with UIDs V_1,V_2. Two request-agents exist with UIDs R_1, R_2 accordingly. Both request-agents have domain: $\{V_1,V_2\}$. Both vehicle-agents have the same domain: {00, 10, 01, 11}. The constraints are shown on Figure 1.

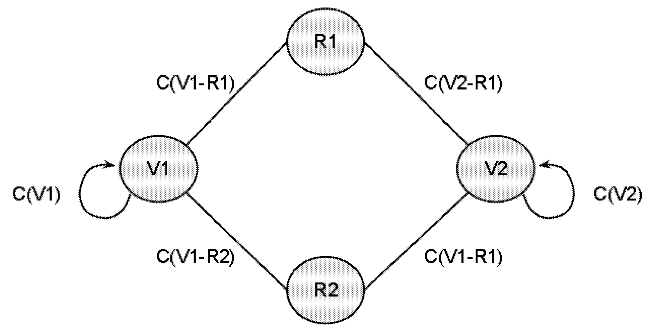


Figure 1: The Problem Graph

Soft constraints corresponding to the edges on Figure 1 are defined by Rules 1 and Rule 2. Their values, by the example $C(V_1\text{-}R_1)$, are shown in Table 1.

Table 1: The Constraint $C(V_1\text{-}R_1)$

R_{V1}^{R1}	R_1		
	$\{R_1, R_2\}$	V_1	V_2
V_1	{1,1}	0	$-\infty$
	{0,1}	$-\infty$	0
	{1,0}	0	$-\infty$
	{0,0}	$-\infty$	0

Let's consider Table 1 with four columns and six rows. Values of $-\infty$ in column 3 and column 4 of the table are assigned by Rule 1 and Rule 2 accordingly.

Given such problem definition in DCOP-style we may expect that DPOP algorithm will produce a solution which summarize overall system utility and achieves the highest profit available for the transportation company. Customers get reward at the expense of payments for the trip by the shortest way.

In accordance with the generic structure of DPOP algorithm, outlined earlier in Section 2, we suggest following sequential steps for making the global solution in our DRT model:

- Agent manager (AM) receives several demands of trip from from request-agents, which are started on any end-user device. The main task of AM agent is establishing of connection between vehicle-agents with free seats and request-agents at this time moment. Thus we have the problem graph such as shown in Figure 1.
- Each vehicle-agent plans its optimal route relying on the list of received requests.
- Each vehicle-agent generates unary and binary constraints taking into account its optimal routes for first ones and Rule 1 and Rule 2 for second ones.
- Agents elect a root using distributed leader election algorithm.
- As the root is elected it starts DFS traversal phase.
- Agents network performs UTIL and then VALUE propagation phases of DPOP optimization algorithm, generating the global optimal solution.

Consider these steps in action.

Example 2

We have some map shown in Figure 2. Nodes are bus stations, edges are the roads between them associated with distance. In our example two vehicles are located at node 1 and at node 8 correspondingly.

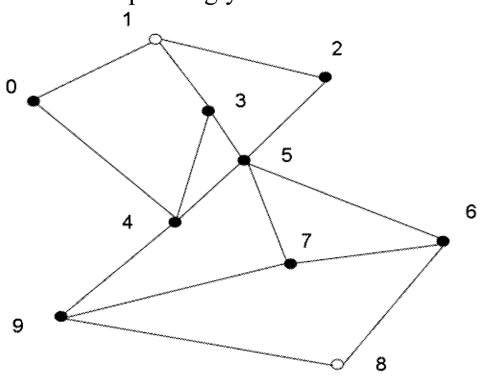


Figure 2: Example of the map

Also there are two requests (Table 2).

Table 2: The Received Requests

#	Origin	Destination
1	5	9
2	4	1

Each vehicle builds optimal routes for each assignment in their domains (Table 3).

Table 3: The Routes and Utilities

Assignment	V_1		V_2	
	Utility	Route	Utility	Route
{0,0}	0	-	0	-
{0,1}	14	13431	20	865431
{1,0}	13	13549	18	86549
{1,1}	25	13549431	30	86549431

All agents constraints shown in Figure 1. The constraint $C(V_1\text{-}R_1)$ shown in Table 1. Other constrains are built by a similar manner. Upon construction of the model agents start leader election procedure and DFS arrangement which results in the agents network designed in Figure 3b.

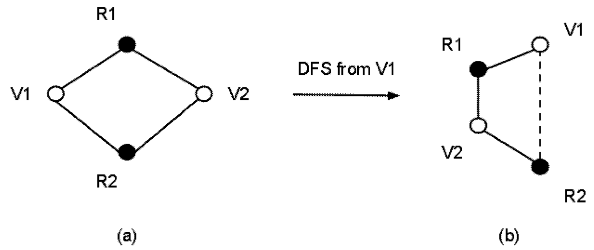


Figure 3: DFS Arrangement

In that pseudo tree structure root vertex is V_1, the sole leaf vertex is R_2. In accordance with DPOP algorithm node R_2 must form the following JOIN structure:

$$JOIN_i^{P_i} = JOIN_i^{P_i} \oplus R_i^{P_i} \oplus (\oplus_{X_j \in PP_i} R_i^j) \quad \textbf{(4)}$$

Therefore it combines constraints with parents and pseudo parents:

$$JOIN_{R_2}^{V_2} = null \oplus R_{R_2}^{V_1} \oplus R_{R_2}^{V_2} \quad \textbf{(5)}$$

Thereafter it makes projection along R_2 dimension:

$$UTIL_{R_2}^{V_2} = JOIN_{R_2}^{V_2} \perp R_2 \quad \textbf{(6)}$$

Then it sends the constructed UTIL message to its parent node, V_2. Corresponding UTIL message shown on Table 4.

Table 4: The Structure of $UTIL_{R2}^{V1}$

$V_1 \backslash V_2$	{1,1}	{0,1}	{1,0}	{0,0}
{1,1}	-∞	-∞	0	0
{0,1}	-∞	-∞	0	0
{1,0}	0	0	-∞	-∞
{0,0)	0	0	-∞	-∞

Then agents repeat application of combine and projection operations with its relations and send UTIL messages to their parents. At the end of UTIL propagation agent V_1 receives UTIL message from R_1, joins it with unary constraints (see Table 5). After that V_1 must choose value for itself. It chooses {1,0} because of that value corresponds to maximum utility 33. Agent V_1 sends chosen value to its child R_1, so the next agent could assign single value to its variable.

Table 5: The Structure of $JOIN_{V1}$

V_1	{1,1}	{0,1}	{1,0}	{0,0}
	25	32	33	30

After the VALUE propagation phase is finished we have the following assignment:

• V_1 = {1,0} • $R_1 = V_1$

• $V_2 = \{0,1\}$ • $R_2 = V_2$

It is not difficult to check that the assignments are consistent and the result ensures the user gets an optimal profit in respect to a value of the utility function.

SOFTWARE IMPLEMENTATION AND SIMULATION

In order to prove the theoretical assumptions, we have developed a program prototype which demonstrates practical implementation of our proposals for DRT model and for the multi-agent optimization algorithm. Software is written in Java programming language. We also use JADE framework (Belligemine and Rimassa 1999) as a Multi-Agent platform and Google Maps technology for visualization purposes. All agents control is performed with developed web-interface.

Computational experiments with the developed prototype confirmed expected results. Table 6 shows results of our simulation procedure with various number of agents. As mentioned in (Petcu and Faltings 2005), DPOP is the most appropriate algorithm for DCOP in the case of loosely coupled networks. Analyzing results of our simulations we can say that in such problem formulation our approach can be applied for systems with the small total number of agents. In order to increase the performance of the system we could introduce some heuristics. For example, we can add a planning domain to each vehicle-agent as it is described in (Abdulrab et al. 2010). So vehicle-agent will be responsible to serve of restricted geographic area. It permits significantly decrease the graph coupling.

Table 6: The Experimental Results

Number of vehicles	Number of requests	Response time, sec.
2	4	6
2	8	106
3	4	20
3	5	more 5 min
4	3	7
4	4	more 5 min
5	3	35
9	1	8
18	1	10

CONCLUSION

In this article we proposed a new solution of DRT problem, providing original multi-agent DRT model designed in accordance DCOP method, and specific multi-agent optimization algorithm which specializes generic DPOP optimization algorithm. The proposed DRT model and algorithms were implemented inside the prototype of multi-agent information system. The performed simulations demonstrate that the model is viable in the case of loosely coupling transportation networks. With several improvements the results can be applied to more complex problems because of the model and the algorithms inherit scalability properties that DCOP ensures.

The model of DRT system in DCOP formulation can be easily extended to satisfy more number of constraints, such as the traffic jams. In the future works we plans to add the support of that constraints.

The research was supported by Scientific Fund of NRU HSE (grant #10-04-0009).

REFERENCES

Abdulrab, H., Babkin, E. and Satunin, S. 2010. “A hybrid multi-layered approach to demand responsive transport systems modeling”. *In Proceedings of System of Systems Engineering , 2010 5th International Conference*.

Abdulrab, H., Itmi, M., and Xu Jin. 2008.”A multi-agent based model for urban demand-responsive passenger transport services”. *Neural Networks, 2008. IJCNN 2008*. (IEEE World Congress on Computational Intelligence).

Belligemine F. and Rimassa, G. 1999. “JADE - A FIPA-compliant Agent Framework.” In *Proceedings of PAAM ’99*. London, pp. 97-108.

Chevrier, R., Canalda, P., Chatonnay, P. and Josselin, D. 2006. “Comparison of three algorithms for solving the Convergent Demand Responsive Transportation Problem.” In *Proceedings of IEEE ITSC 2006 IEEE Intelligent Transportation Systems Conference*, Toronto, Canada, pp. 17-20, September.

Cubillos, C. and Guidi-Polanco, F. 2006. ”An Agent Solution to Flexible Planning and Scheduling of Passenger Trips.” *Artificial Intelligence in Theory and Practice*, IFIP International Federation for Information Processing. Volume 217/2006. pp. 355-364.

Faltings, B., L´eaut´e, T. and Petcu, A. 2008. ”Privacy guarantees through distributed constraint satisfaction.” In *Proceedings of IAT’08*. December 9–12.

Jin, X. and Zhe, H.. 2009. “An intelligent model for urban demand-responsive transport system control.” *Journal of Software*. Vol. 4, no. 7, pp.766-776.

Kambhampati, S. 1997. ”Refinement planning as a unifying framework for plan synthesis.” *AI Magazine*, vol. 18(2), pp. 67-97.

Mailler, R. and Lesser, V. 2005. “Asynchronous partial overlay: A new algorithm for solving distributed constraint satisfaction problems.” *Journal of Artificial Intelligence Research (JAIR)*.

Modi, P.J., Shena, W., Tambea, M. and Yokoo, M. 2005. “Adopt: asynchronous distributed constraint optimization with quality guarantees.” *Artificial Intelligence*, Vol. 161, pp. 149–180.

Petcu, A. and Faltings, B. 2005. “DPOP: A scalable method for multiagent constraint optimization.” In *Proceedings of the 19th International Joint Conference on Artificial Intelligence, IJCAI-05*, pp. 266–271, Edinburgh, Scotland.

Petcu, A., Faltings, B. 2004. ”A distributed, complete method for multi-agent constraint optimization.” In *Proceedings of CP 2004 - Fifth International Workshop on Distributed Constraint Reasoning (DCR2004)*, Toronto, Canada, September.

Yokoo, M., Durfee, E.H., Ishida, T. and Kuwabara, K.1992. ”Distributed constraint satisfaction for formalizing distributed problem solving.” In *Proceedings of International Conference on Distributed Computing Systems*, pp. 614–621.

Multi-scale Analysis of the Vulnerability of Road Networks

Cyrille Bertelle, Antoine Dutot, Michel Nabaa, Damien Olivier
LITIS
25 Rue Philippe Lebon
76600 Le Havre, France
email: firstname.lastname@univ-lehavre.fr

KEYWORDS
Complex system, network, risk, hazard, self-organization, GIS

ABSTRACT

In this work, we deal with risk prevention and crisis management requiring evacuation. Road network structure and mobility process are studied by using swarm intelligence methods. Risk prevention is considered by analyzing the vulnerability of the network represented by a multigraph, at different scale levels thanks to betweenness centrality and hierarchical clustering methods. Crisis management during an evacuation is also considered. A microscopic traffic model is developed and a micro-evacuation simulator is proposed to organize an evacuation, beginning from crisis occurrence until vehicles evacuation. Our model is adaptive, it takes into consideration multigraph structure and its elements vulnerability at different scale levels, and unexpected events during an evacuation. Finally, the developed tool helps decision makers managing risks.

INTRODUCTION

In recent decades, the population exposure to natural or anthropic risks has increased, the recent events like Fukushima or the Xintia storm show it. The main reason is the demographic pressure which generates an urbanization in flood or industrial areas. The damage are directly caused on people and their environments. In the same time, the concept of safety and its perception is evolving, thus during a long time it was based on beliefs, but knowledge has progressed drawing on the latest advances in information technology and communications. The result is the definition by States of a global safety policy with local implementation.

In France, territorial local authorities have more and more responsibilities and to assume them they adopt tools and skills to identify hazards and to make preventive information to populations and to define contingency plans. In this context, the CODAH[1] has created in 2001 the DIRM[2].

The study developed, in this paper, puts forward a contribution taking into account the road networks with a systemic approach considering territory and risk in their complexity. The purpose is to protect the population by handling the flow generated by mobility of persons in the environment constituted by the road network. These flows modify the state of network which engenders a feedback on the flows themselves. We search to highlight structures, forms and self-organizations.

The main goal is to provide, through analysis of the dynamics of the considered situation, elements of decision support for planning control with relevant informations.

VULNERABILITY AND ROAD NETWORKS A SYSTEMIC APPROACH

We must consider vulnerability in relation with the territory. Indeed, vulnerability must be studied with its spatial dimension and bearing in mind the interactions between the multiple components using the space. Territory is a complex system for which a systemic approach is possible searching in particular the emergence of self-organizations.

Vulnerability and Morphodynamic of Road Networks

To study the vulnerability of the road network we must analyze its material resistance but also examine it structural resistance to hazards that surround it. Potential damage of the structure can generated dysfunctions that lead to lose some functions of the network. The analysis of the network structure will therefore allow us to identify road sections that are more weak than others *ie.* the most used roads and crossroads. The authorities must have a great attention on these sections especially to deal with the phenomena of saturation (tailback, traffic jam ...) during an evacuation. Except material damage and structural, indirect damages due to loss of one or more weak links make parts of the network unavailable jeopardizing its functioning by decreasing its effectiveness to link the different parts of the territory.

[1]COmmunauté De l'Agglomération Havraise: it groups 17 communes including Le Havre. The goal is to to develop many common projects in territory management, health and hygiene, public transport, risk management...

[2]Major risk management team

An urban mesh constituted by a grid structure like Manhattan allows to find easily alternative paths when a part of the network malfunctions, without adding an important displacement overhead. By contrast, a similar malfunction that affects a weak link of a network with the same size, but with a structure less regular may require the vehicle to make an important roundabout way to reach the destination (see figure 1).
To highlight the vulnerability of the road network two approaches are considered: the first one is based on random walks in the graph which modelizes the network and the second uses informations coming from mobility survey, with this two methods we try to identify the paths which are more taken than the others by the vehicles. The authorities must ensure that traffic on these roads is flowing freely, if possible. Both methods are complementary. On the one hand the structure analysis is pessimistic, some paths in the network are considered as weak even if they are rarely used and on the other hand this vision is corrected by the knowledge of the population mobility which is introduced in the simulation.

Multi-scale Structural Analysis of the Vulnerability of the Networks

The structure of the road network is modeled by a oriented multigraph, which is a graph in which multiple edges between nodes are either permitted. The structure influences the flows both in normal conditions and in crisis. When an hazard occurs on a part of the road network and damages it, this leads to change the network structure and by the way the functioning influencing the entities using the road network. So three vulnerabilities must be considered : material, structural and functional (Gleyze (2005), Reghezza (2006)). Material damage depends on hazards and level of exposure. This is material vulnerability. The road network, by its structure, is exposed and therefore vulnerable. Thus the network structure plays a positive or negative role in reducing vulnerability, for example. As mentioned previously, a network which has a mesh structure is less vulnerable than others because we can often find alternative paths to avoid damaged parts of the network. In a different scale, structural damage affects the function of the road network and the mobility. This reveal the functional vulnerability which can also affect the structure of network. Three scale levels of the road network are considered: micro, meso and macro.
At the micro level, we study how an accident can affect the efficiency of the road. To do that, we use two method, one based on clustering coefficient, the other computes the cost of the detour when an arc has been lost. The local clustering coefficient of a vertex in a graph quantifies the probability that the neighbors of this vertex are also connected to each other. These measures allow to determine the importance of an edge. Thus, if we consider an edge with an high clustering coefficient, it offers more alternative paths if it is damaged, than an edge with a low coefficient.
At the mesoscopic level, we compute the betweenness centrality of each edges. For a given edge, this corresponds to the number of shortest paths from all vertices to all others that pass through that edge. This calculation, adapted from the method of calculation of centrality in a graph (Freeman (1979)), measures the importance to the mesoscopic scale of each arc in the graph.
At the macro level, we use a community detection algorithm adapted to multigraph. It highlights the edges which link two communities. These edges are weak, for example the removed edge on figure 1 is a link between two communities. They are essential for the flows and they are important singularities in the global map of vulnerability. We have also developed a method based on a biased random walk on the multigraph, allowing to identify the area of the network that may be congested. The walk is trapped by the structure of the multigraph.

ENVIRONMENTAL MODEL, DYNAMIC GRAPHS AND GIS

Figure 2 describes the environmental model processing methodology used. It extracts from GIS layers a graph inside which dynamic processes will allow to describe simulations. This task is handled using the dynamic graph programming library GraphStream, which is developed inside the LITIS laboratory (Dutot et al 2007).

Simulations and Evacuation Handling

Figure 3 presents one of the outputs of the proposed framework, showing the clustering computation on the city of Le Havre. It consists here in a macroscopic analysis of the urban morphology that automatically detects the network's more connected areas, named clusters or communities, therefore allowing to isolate the inter-communities edges which are typically the more vulnerable.
Figure 4 presents another output of the framework showing a snapshot at a given time of a dynamic map which computes random walks: these dynamic processes randomly travel through the graph and reveal spatio-temporal structures in the network which emerge from their potential use.
On figures 5 and 6, we illustrate an example produced by a micro-simulation scenario inside which a malfunction of the network is generated, and whose effect is to make a part of its edges unusable. We then measure the impact using a morphodynamical analysis of the graph that computes the evolution of the betweenness centrality during the malfunction. We typically observe on the well known "La Brecque" crossroads, at the Le Havre city entry point and shown on figures 5 and 6, that these

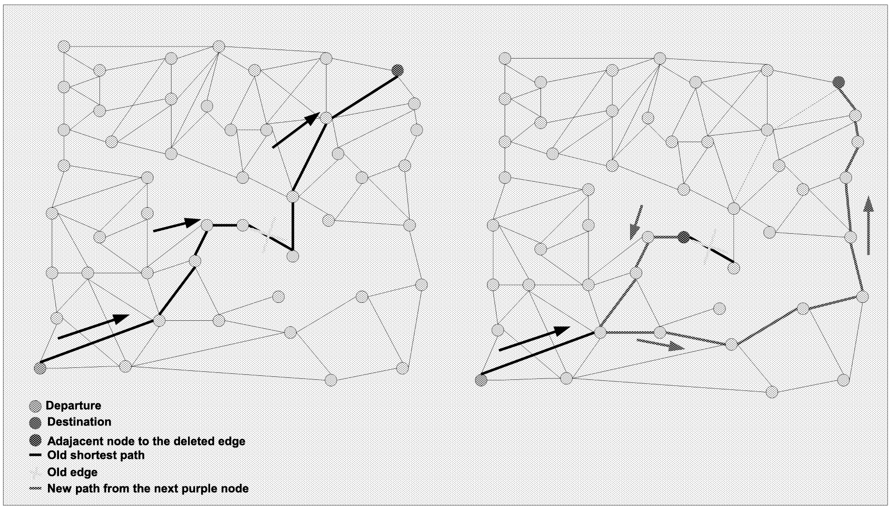


Figure 1: Influence of the loss of an edge on an non regular graph.

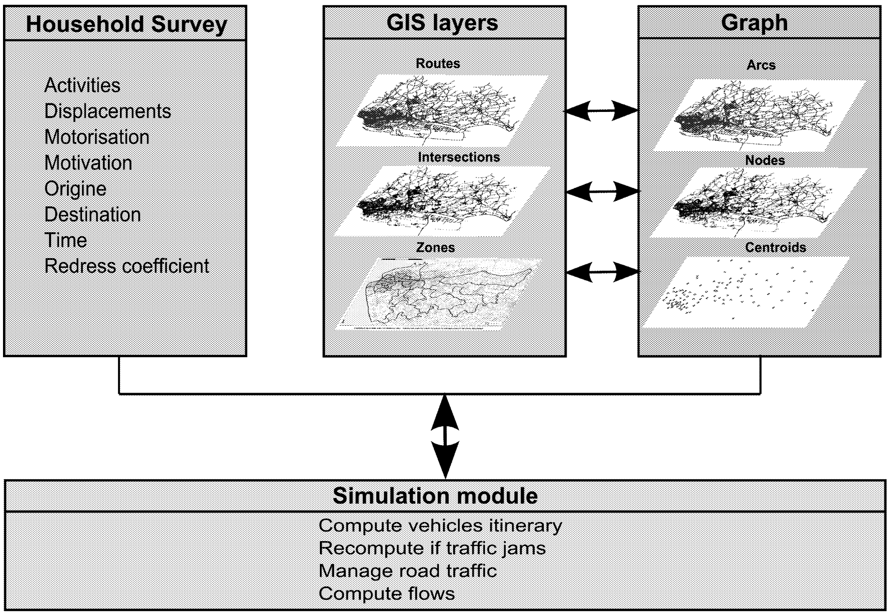


Figure 2: Graph extraction from a GIS and coupling with a vehicle displacement micro-simulation

centralities are heavily modified by the malfunctioning. This allows us to quantify the studied area vulnerability.

CONCLUSION

The work presented in this paper, provides modeling tools of road networks and a simulation of the flows which are likely to dysfunction. An analysis of morpho-dynamics of the network is realized over several levels of scale, providing elements to characterize the vulnerability of the territory. The result is a "smart" dynamic map of the territory highly adaptable to the network changes that may be a natural or anthropic accident. This tool is also used to specify bypasses when a local accident arises. The simulation then offers dynamically recalculations of the flows, useful decision-making at crisis management in case of major risks.

Thanks

This work has been funded by the Haute-Normandie region and the metropolitan area of Le Havre.

REFERENCES

Banos A., 2005. *A la conquête de la fourmilière urbaine : quand le géographe se fait explorateur de mondes artificiels.* Tech. rep. URL `[Online].Available:http://fig-st-die.education.fr/actes/actes2005/index.htm`.

Bertelle C.; Dutot A.; Nabaa M.; and Olivier D., 2008a. *Detection of organizations in large graphs : Application on risk management. Agent Based Spatial Simulation.*

Bertelle C.; Dutot A.; Nabaa M.; Olivier D.; and Mallet P., 2008b. *Modeling of the vulnerability related to the dynamic road traffic.* In *ESM'08*. Le Havre.

Bourcier J. and Mallet P., 2006. *Allocation spatio-temporelle de la population exposée aux risques majeurs, contribution à l'expologie sur le bassin de risques majeurs de l'estuaire de la Seine : modèle pretresse. Revue internationale de Géomatique*, 16, no. 10, 457–478.

Church R. and Cova T., 2000. *Mapping evacuation risk on transposion networks using a spatial optimization model. Transportation Research Part C: Emerging Technologies*, 8, no. 1-6, 321–336.

Cova T. and Church R., 1997. *Modeling community evacuation vulnerability using GIS. International Journal of Geographical Information Science*, 11, no. 8, 763–784.

Cutter S.; Mitchell J.; and Scott M., 2000. *Revealing the vulnerability of people and places: a case study of georgetown county, south carolina. Annals of the Association of American Geographers*, 90, no. 4, 713–737.

Daudé E., 2005. *Systèmes multi-agents pour la simulation en géographie : vers une Géographie Artificielle*, chap. 13. 355–382.

Dutot A.; Guinand F.; Olivier D.; and Pigné Y., 2007. *Graphstream: A tool for bridging the gap between complex systems and dynamic graphs.* In *EPNACS: Emergent Properties in Natural and Artificial Complex Systems.*

Fèbvre L., 1956. *Pour l'histoire d'un sentiment : le besoin de sécurité. Économies, sociétés, civilisations*, 11.

Freeman L., 1979. *Centrality in social networks : conceptual classification. Social networks*, 1, no. 3, 215–239.

Gleyze J., 2005. *La vulnérabilité structurelle des réseaux de transports dans un contexte de risque.* Ph.D. thesis, Université Paris VII, Laboratoire COGIT-IGN.

Nabaa M., 2011. *Morphodynamique de réseaux viaires. Application au risque.* Ph.D. thesis, Université du Havre.

Provitolo D., 2007. *A proposition for a classification of the catastrophe systems based on complexity criteria.* In *European Conference Complex Systems - EPNACS'07, Emergent Properties in Natural and Artificial Complex Systems.* Dresden, 93–106.

R. N. and North M., 2004. *Repast for GIS.* In *Agent 2004: Social Dynamics: Interaction, Reflexivity and Emergence.* University of Chicago and Argonne National Laboratory, IL, USA.

Reghezza M., 2006. *Réflexions autour de la vulnérabilité métropolitaine : la métropole parisienne face au risque de crue centennale.* Ph.D. thesis, Université Paris x – Nanterre.

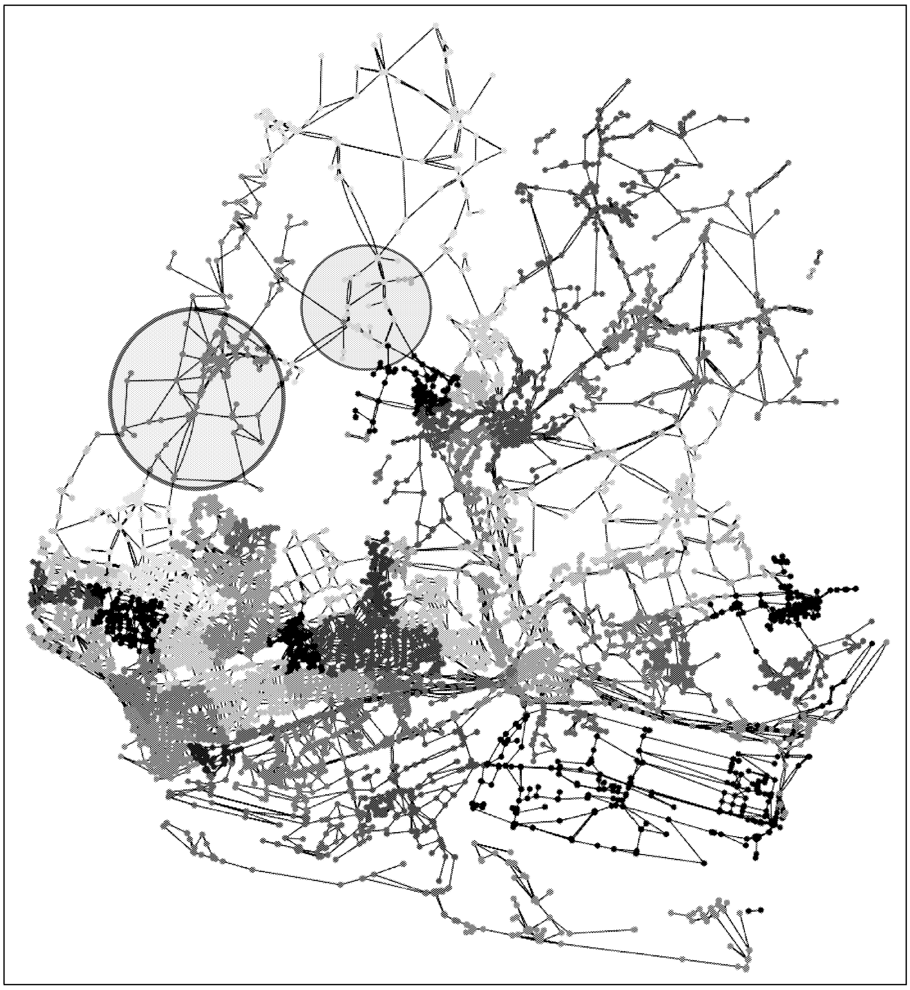


Figure 3: Hierarchical clustering on the Le Havre city agglomeration graph

Figure 4: Random walk computation on a road network extracted from a Le Havre city GIS. The edge color shows the traffic importance (dark blue for low usage, to red for high usage).

Figure 5: Zoom on the large crossroad named "La Brecque" at the entry of Le Havre city

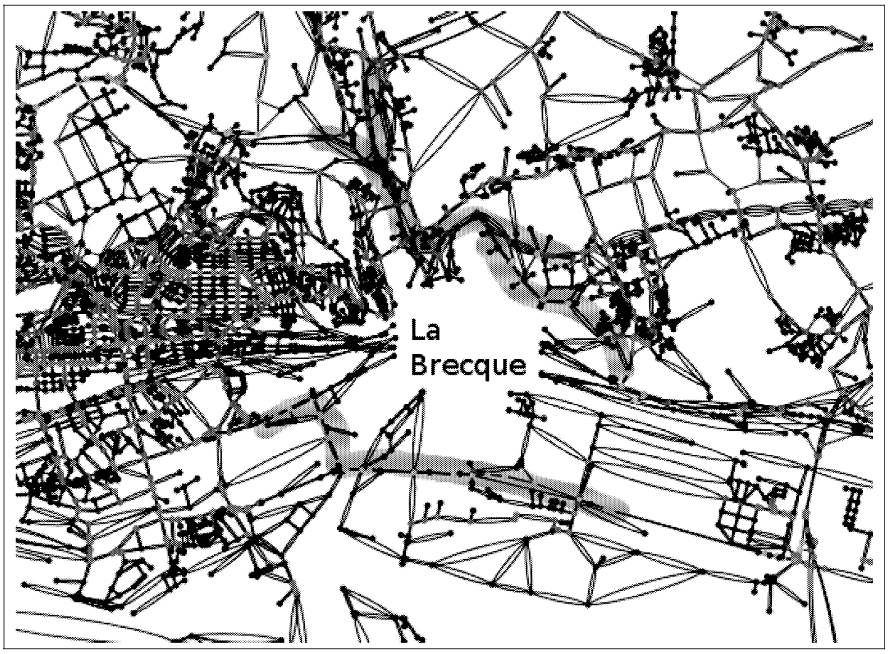


Figure 6: Traffic flow reorganization around the malfunctioning crossroad. The simulation framework shows a flow reorganization and by-path areas as grey overlays.

E-LEARNING

A Framework Model for an Accessible e-Learning Platform

Jehad Al-Sadi
Arab Open University
E-mail: *j_alsadi@aou.edu.jo*

KEYWORDS
e-Learning, Learning Management System, Web Accessibility, Online Educational Environment.

ABSTRACT

e-Learning platform is the advance developed tool which will generate new era of education under the concept of anyone can study in anywhere at anytime. On the other hand, e-learning well enable disabled students to access more materials and work more independently than they could use traditional method. However, the information and activities posted were not appropriately accessible with adaptive technology used by general disabled students. Therefore, all eLearning platforms should focus on the issue of web accessibility to facilitate disabled students an effective access.
This paper proposes an investigation of existing concepts and tools of web accessibility for eLearning. The paper also presents a web accessibility model to support disabled students to access e-learning. e-Learning system composed of three main flavors: a content management system (CMS), learning management system (LMS), and a set of activity tools. We also present some existing models including the EU4ALL Content Personalization System, and also introduce a LMS framework model to provide accessibility for disabled students to access e-learning platforms.

INTRODUCTION

Information plays an important role in all aspects of current society. Information and Communication Technologies open up opportunities for spreading knowledge among the society. It is also more convenient for governments, institutions, public and private enterprises to use the Internet to provide information services. Moreover, ICT offers great opportunities of social inclusion. Technological development can enable people with disabilities to improve their quality of life [12-16]. They can accomplish tasks that would be impossible to do without the computer, such as: writing a letter, communicating, drawing a picture, etc. Moreover, the digitalization of many public services such as education (school, university), shopping, banking, library, or even sending a letter allows people with disabilities to live in much the same way as those who are not disabled. They can acquire an "independent life" and achieve social integration.

Nowadays, the number of colleges and universities which provide courses and degree programs via distance education has been growing dramatically. The advantages of the online learning have been widely described in the literature. However, little literature addressed the issue of accessibility for people with disabilities in using online distance education environment [17].

According to the Individuals with Disabilities Education Act [19], assistive technology must be considered when teams plan individualized education programs for students with disabilities [18]. This requires expanding the narrow conception of assistive technology to include uses in improving knowledge and skills in content area classes by providing better access to the general education curriculum.

Most online educational environments are still not accessible to students with disabilities. Furthermore, the information and activities posted were not appropriately accessible with adaptive technology used by disabled students [20]. As a result, all eLearning websites should be focused on the issue of web accessibility to facilitate disabled students to effectively gain access. To create web accessibility in any website for people with disabilities, there are several components required to create web accessibility to facilitate this group of users.
Therefore, this paper would be proposed to investigate existing concepts and tools of web accessibility for eLearning, and to create and develop the web accessibility model support disabled students in eLearning. Section 2 of this paper presents the importance of web accessibility, section 3 focuses on Assistive technology and authoring tools, section 4 introduces the skills and knowledge gained from using e-learning platforms, finally section 5 concludes the paper.

THE IMPORTANCE OF THE WEB ACCESSIBILITY INITIATIVE

The Internet for students is the main and the largest information source. One major advantage of the communication tools over the internet is that they can play an important role in social relationship. For a person with a disability, the internet communication tools can be a viable opportunity to interact with others [3]. Furthermore, disabled persons are able to interact with others with similar interests and concerns.

According to the Royal National Institute for the Blind [8] “the internet is one of the most significant communication developments since the invention of Braille. For the first time ever, many blind and partially sighted people have access to the same wealth of information as sighted people and on the same terms”. Web publishing of many public services such as education (school, university), government, may allow people with disabilities to live in much the same way as those who are not disabled.

On the other hand, if the information and services provided by governments, institutions and public enterprises are not fully accessible, then there is an information gap which creates a digital divide between those who can benefit from opportunities provided by ICT and those who cannot. The accessibility issue of the Internet has grown significantly; users who need to access the net without GUI interface are in some ways like the blind users [9].

WAI (the Web Accessibility Initiative) is a positive part of a larger universal trend to standardize certain aspects of web content and interface design. WAI is international in scope and its impact on interface design and consequently on the design and development of online educational content itself cannot be underestimated. WAI is a set of recommendations intended to standardize certain key aspects of interface design so that physically and cognitively challenged users can be on an equal footing with regular users when they access information and interact with others across the Internet [2].

There are standards in American and also Canada similar to WAI initiative [4]. Canada has also developed its accessibility standard and thereby contributing to the trend to standardization. The main target for these standards is guide educators who develop online content ethically, and pedagogically bound to take as many cognitive and physical disabilities into account as possible when designing and mounting content online, or any course for that matter. HCI (Human Computer Interaction) specialists also considered standards for disabilities while designing which has many other measurable benefits as well.

Online course content generally defined as the textual; visual or aural content within a Content/Learning management system (CMS/LMS), and also might be included other things such as text; images; sounds; videos; and animations. The Web Content Accessibly Guide (WCAG) is a document issued by W3C in 2006 that explains how to make Web content accessible to people with disabilities. It contained eight guidelines effectively supported to disabled people in general and to visually-impaired people in specific directly to access any website if the web developers rely on to create their web content, these eight guidelines are [1, 5]:

- Providing text alternatives for all non-text content.
- Providing synchronized media equivalents for time dependent presentations.
- Ensuring that information/functionality/structure are separable from presentation.
- Distinguishing clearly between foreground words and images from the background in visual presentation.
- Making all functionality operable via a keyboard or other input interface.
- Facilitating the ability of users to orient themselves and move within the content.
- Ensuring that the meaning of content can be determined.
- Organizing content consistently from page to page.

In addition to WCAG, User Agent Accessibility Guidelines or UAAG was part of a series of accessibility guidelines published by W3C [5]. The UAAG guideline documents has explained how to make user agents accessible to people with disabilities including visually-impaired users and particularly to increase accessibility to Web content, due to user agents include of web browser together with media players and assistive technology which were software that some people with disabilities use in interacting with computers.

The guidelines UAAG are as following:

- to follow applicable specifications and conventions.
- to facilitate access by assistive technologies.
- to ensure that the user interface is perceivable.
- to ensure that the user interface is operable.
- to ensure that the user interface is understandable.

ASSISTIVE TECHNOLOGY AND AUTHORING TOOLS

Cook and Hussey [6] define Assistive technology as a generic term that includes assistive, adaptive, and rehabilitative devices and the process used in selecting, locating, and using them to support independence for people with disabilities by enabling them to perform tasks that they were formerly unable to accomplish or had difficulty accomplishing by providing enhancements to or changed methods of interacting with the technology needed to accomplish such tasks.

Authoring tool can be defined as any software or collection of software components that authors could use to create or modify Web content for use by other people. When recommended design strategies for the website were implemented which would be supported websites especially eLearning website (LMS) to be fully accessed by every people including visually-impaired people, one of the greatest barriers to access was the lack of authoring tools that support web developers which in line with existing accessibility guidelines. As a result, W3C had developed the particular guideline to guide the developers to develop and use accessible software supporting to create accessible web content so called Authoring tools accessibility guidelines ATAG [7].

In general, the ATAG has provided guidance for developers of software which created content for the web or in a web-based markup language and the purpose of the Guidelines were to assist developers in designing authoring tools that generate accessible web content and to assist developers in creating an accessible authoring interface [1].The guidelines includes:

authoring tool must facilitate access by assistive technologies,

authoring tool user interface must be perceivable,

authoring tool user interface must be operable,

authoring tool user interface must be understandable,

production of accessible content must be enabled,

authors must be supported in the production of accessible content,

and accessibility solutions must be promoted and integrated.

Based on all above web accessibility tools mentioned, Pisit [1] generated a conceptual model of web accessibility to support visually impaired students on eLearning which is shown in Fig.1

The EU4ALL Content Personalization system

Cooper and Heath presented in [10] the EU4ALL (European Unified Approach for Accessible Lifelong Learning) project. In this approach, the EU4ALL CP Module is the decision engine of the system as shown in Fig.2. It decides what content is served to the student through the VLE following a given request, depending

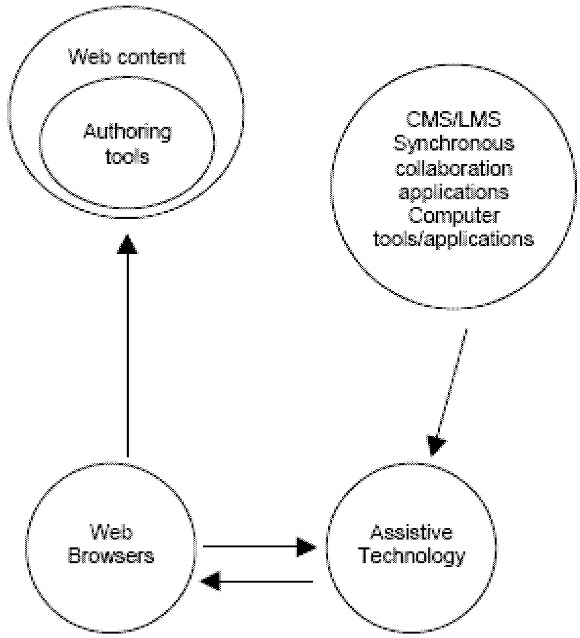


Fig1. Web Accessibility Model E-learning System

on their expressed needs. The VLE has embedded or associated a content repository of some form where the resources including any alternatives are stored. The User Model (UM) stores the preferences and needs of all registered students' Personal Needs and Preferences (PNP) and Device Model (DM) standardized descriptions of device properties. Because the EU4ALL approach is to develop a flexible framework and not a dedicated system a Metadata Repository (MR) is included in the system to facilitate the storage of suitable indexed metadata sets of the content metadata needed for personalization Digital Recourse Description (DRD). It is envisaged long-term that such content metadata will reside in the VLE or repository, with the content. At run-time the CP module creates a series of relevant facts from the instances of the metadata. The CP module determines if there are alternative resources available on the system for the one being requested that better meets the user's profile and the device they are using. The CP module gets the identifier for the best resource, and if this is valid for the user and their device sends its identifier to the VLE. If no valid resource is identified a message is sent to the end-user and the systems administrator so remedial action can be taken as required [10].

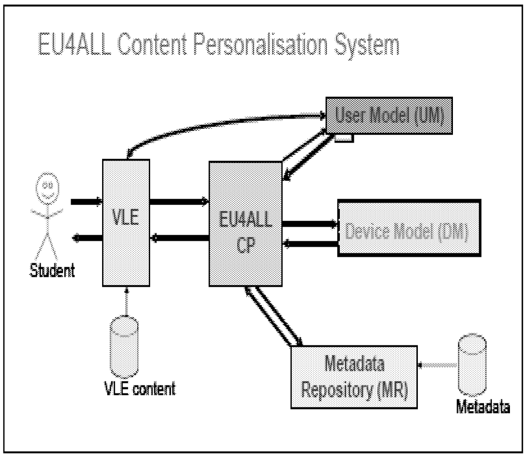


Fig 2. Schematic of EU4ALL Content Personalization System

The Proposed Module

Figure 3 shows our proposed module where the LMS/CMS has its common components which contain content of different types, tools and activities such as forums, chats, quizzes, etc. In addition to the common components, there are specialized assistive tools such as tools related to authoring content for disabled persons, also special tools for disabled persons. The assistive technology interface can be hardware, software, or both. The main functionality of this interface is to make the content and common activities of the LMS/CMS accessible by disabled users. Finally it is important to mention here that the web browser is complied with W3C WAI initiative.

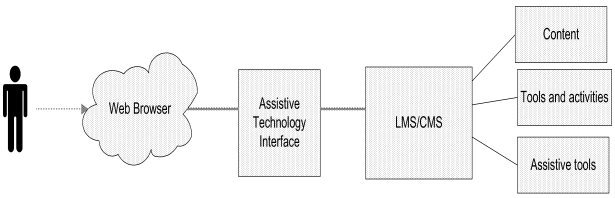


Fig 3: A Proposed Module LMS/CMS with assistive technology interface

In an educational context designing the content interaction is extremely important in order to reach a learning goal. Moreover, in online learning the methodology is crucial. For example, a tool may meet technical accessibility requirements, but it may be unusable for a disabled student because it is designed with a regular interface in mind. Likewise, the design of a lesson could be perfect if it is delivered using a multimedia system, but may be poor if it uses adaptive technologies like a speech synthesizer.

Therefore, it is very important to redesign traditional pedagogical approaches by integrating information and communication technologies into courses to make sure they comply with the accessibility guidelines.

Furthermore, the LMS e-content and also related activities have its own pedagogical aspect to author such content. Such pedagogy is different that the pedagogical aspects on traditional authoring content. Now since the content is going to be used by disabled users then there will be a new era of research for structuring a new pedagogy of authoring content complied with assistive technology.

CAN E-LEARNING HELP STUDENTS WITH DISABILITIES

This section explores how e-learning can help disabled students to gain learning and knowledge. According to many previous studies, there are many usefulness of making e-learning platforms accessible for disabled students, e.g. Barrett in [11] presented the following list as an overview of the skills, abilities, and knowledge gained by students with the use of e-Learning and online course contents.
Computer skills in terms of learning and operating software and hardware
Ability to manipulate data
Ability to operate various types of computer programs
Ability to create, manage, send, and analyze data
Ability to surf the Web and discover even more about learning and the world around them
Ability to search, analyze, distinguish and use information gathered on the Internet
Students are more knowledgeable of current technological offerings and changes in the workplace then ever before.
Students with disabilities are now able to compete more in the learning environment with the implementation of E-Learning. E-Learning has helped to create a more level field of learning for these students.
Students can effectively use course knowledge and apply it to real work applications.
Case in point – recent supervision of a master's mini-thesis completed by a visually impaired student

CONCLUSION

This paper presented the importance and the need of applying assistive technology to access e-Learning platform which will generate new era of education under the concept of anyone can study in anywhere at anytime. e-Learning and assistive technology tools well enable disabled students to access more materials and work more independently than they could use traditional method. Therefore, this paper pointed up that all eLearning platforms should focus on the issue of web accessibility to facilitate disabled students an effective access.

This paper presented an investigation of existing concepts and tools of web accessibility for eLearning. Furthermore, we introduced a LMS framework model to provide accessibility for disabled students to access e-learning platforms.

REFERENCES

[1] Pisit Prougestaporn, Web accessibility Model for eLearning Website, 5th international Conference on eLearning for Knowledge-Based Society, Dec 2008, Thailand, pp 43.1-43.5.

[2] Paolucci, P. Should Online Course Design Meet Accessibility Standards? Educational Technology & Society, 7 (1), 6-11, 2004.

[3] M.Arrigo, E-Learning Accessibility for blind students, International Conference on Multimedia and ICT in Education, Lipson, April 2009.

[4] Even Grounds, Accessibility Consulting, What is Section 508, http://www.evengrounds.com

[5] W3C Ubiquitous Web Applications Group, see: http://www.w3.org/2007/uwa/

[6] Cook, A.M. and Hussey, S. (2001). Assistive Technologies: Principles and Practice, 2nd Edition. Mosby.

[7] W3C. (2006). "Authoring tools accessibility guideline." http://www.w3.org

[8] Royal National Institute for the Blind, Communicating with blind and partially sighted people. Peterborough (2004)

[9] T. Wattenberg, Beyond legal compliance: Communities of advocacy that support accessible online learning, Internet and Higher Education, 7 123–139 (2004).

[10] Cooper and Heath, 2009 Cooper, M., & Heath, A. (2009) Access for all to eLearning, Multimedia and ICT in Education Conference (m-ICTE), 2009.

[11] Bob Barrett, Using E-Learning as a Strategic Tool for Students with Disabilities, Sixth International Conference on Higher Education and Disability, 2007.

[12] K. Williamson, S. Wright, & A. Bow, The Internet for the blind and visually impaired, Journal of Computer-Mediated Communication, 7 (1) (2001).

[13] H.S. Kaye, Computer and Internet Use Among People with Disabilities, Disability Statistics Report, Washington DC: U.S. Department of Education, National Institute on Disability and Rehabilitation Research (2000).

[14] L. Johnson, E. Moxon, In Whose Service? Technology, Care and Disabled People: The Case for a Disability Politics Perspective, Disability & Society, 13(2), 241-258 (1998).

[15] Berry, J. Apart or a part?: Access to the Internet by visually impaired and blind people, with particular emphasis on assistive enabling technology and user perceptions. Information Technology and Disabilities, 6 (3), 1-16.

[16] S. Burgstahler, Distance learning: Universal design, universal access, *AACE Journal,* 10(1), 32-61. (2002).

[17] S. Kinash, S. Crichton, & W.S Kim-Rupnow, A review of 2000-2003 literature at the intersection of online learning and disability: Electronic curb-cuts., American Journal of Distance Education, 18(1) (2004).

[18] Okolo, C. M., 2005. "Interactive Technologies and Social Studies Instruction for Students with Mild Disabilities", Handbook of Special Education Technology Research and Practice, pp. 623 - 641.

[19] Individuals with Disabilities Education Act of 2004.

[20] Arrigo, M. "E-Learning Accessibility for blind students." Italian National Research Council, Institute for Educational Technology, 2005.

BIOGRAPHY

Jehad Al-Sadi received his B.Sc. degree in Computer Science from Tennessee State University, USA, in 1989, and M.Sc degree in Computer Science from Jackson State University, USA, in 1993. From graduation till 1996, he was a lecturer in the Department of Computing Science at the Al-Isra University, Jordan. He gained his Ph.D. in Computer Science from the University of Glasgow, U.K. 2002. He joined the Computer Science Department at. The Zarka Private University, Jordan from 2002 until 2004. He is currently an associate professor and the Chairman of IT and computing department at Arab Open University, Jordan. His research interests are: interconnection networks, optical networks, fault-tolerant routing, e-learning, and learning management systems.

Adopting e-Learning Strategy at The Arab Open University

Bayan Abu Shawar
Information Technology and Computing Department
AOU-Amman
Jordan
b_shawar@aou.edu.jo

Keywords:
e-learning, LMS, Moodle, SIS

ABSTRACT

Arab Open University is one of the first organizations that adopt an e-learning methodology in the Arabic region. AOU has partnerships with the United Kingdom Open University (UKOU) and other national educational institutes. In this paper, our experience of using two e-learning platforms FirstClass, and Moodle is clarified. Further more, we discuss the developments and enhancements that have been done in-house over the learning management system; specifically Moodle; to employ our regulations and rules over this platform and to facilitate its usage for our students. A description of integrating the learning management system with other electronic systems such as student information system and human resource system is discussed. In addition to that the quality assurance strategy of the AOU is presented.

INTRODUCTION

The growth of Internet-based technology have brought new opportunities and methodologies to education and teaching represent in e-learning, online learning, distance learning, and open learning. These approaches are typically use in place of traditional methods and mean that students deliver their knowledge though the web rather than face-to-face tutoring.

E-learning is a new trend of education system, where students deliver their materials through the web. E-learning is the "use of internet technology for the creation, management, making available, security, selection and use of educational content to store information about those who learn and to monitor those who learn, and to make communication and cooperation possible." [1].

Kevin kruse [17], addressed the benefits of e-learning for both parties: organization and learners. Advantages of organizers are reducing the cost in terms of money and time. The money cost is reduced by saving the instructor salaries, and meeting room rentals. The reduction of time spent away from the job by employees may be most positive shot. Learning time reduced as well, the retention is increased, and the contents are delivered consistently. On another hand, learners are able to find the materials online regardless of the time and the place; it reduces the stress for slow or quick learners and increases users' satisfaction; increases learners' confidence; and more encourages students' participations.

Recent advancement in mobile technology has improved many areas such as commerce [2]-[5], services [6] [7] and recently in Education. The use of mobile technology in education is also known as “mobile education” or “m-education” [8]. Mobile technologies have provided unique opportunities for educators to deliver educational materials efficiently, and to support the cognitive and social process of student learning. Educational materials can be delivered to students through mobile devices. Mobile technology can also be integrated into learning management system to improve interactivity in the classroom and also in distance learning.

In this paper the following topics will be presented: e-learning platform of the AOU is described; the integration process between learning management system and student information system; the requirements and benefits of using the mobile technology in teaching; how to enhance the mobile technology in AOU.

THE E-LEARNING PLATFORM OF THE AOU

Arab Open University was established in 2002 in the Arabic region, and adopted the open learning approach. An open learning system is defined as "a program offering access to individuals without the traditional constraints related to location, timetabling, entry qualifications."[12].

The aim of AOU is to attract large number of students who can not attend traditional universities because of work, age, financial reasons and other circumstances. The "open" terminology in this context means the freedom from many restrictions or constraints imposed by regular higher education institutions which include the time, space and content delivery methods.

Freed et al. [9] claimed that the "interaction between instructors and students and students to students remained as the biggest barrier to the success of educational media". The amount of interaction plays a great role in course effectiveness [10]. For this purpose and to reduce the gap between distance learning and regular learning, the AOU requires student to attend weekly tutorials. Some may argue that it is not open in this sense; however the amount of attendance is relatively low in comparison with regular institutions. For example, 3 hours modules which require 48 hours attendance in regular universities, is reduced to 12 hours attendance in the AOU.

In order to give a better service to students and tutor, to facilitate accessing the required material from anywhere, and to facilitate the communication between them, an e-learning plat form is needed. A learning platform "is software or a combination of software that sits on or is accessible from a network, which supports teaching and learning for practitioners and learners."[1] A learning platform is considered as a common interface to store and access the prepared materials; to build and deliver learning activities such quizzes and home-works; support distance learning and provide a set of communication possibilities such as timetables, videos, etc.

AOU has partnerships with the United Kingdom Open University (UKOU) and according to that at the beginning the AOU used the FirstClass system as a computer mediated communication (CMC) tool to achieve a good quality of interaction. The FirstClass tool provides emails, chat, newsgroups and conferences as possible mediums of communication between tutors, tutors and their students, and finally between students themselves. The most important reason behind using FirstClass was the tutor marked assignment (TMA) handling services it provided. However, the main servers are located in the UKOU which influences the control process, causes delays, and totally depends on the support in UKOU for batch feeds to the FirstClass system [11].

To overcome these problems, AOU use Moodle nowadays as an electronic platform. Moodle is an open-source course management system (CMS) used by educational institutes, business, and even individual instructors to add web technology to their courses. A course management system is "often internet-based, software allowing instructors to manage materials distribution, assignments, communications and other aspects of instructions for their courses." [13] CMS's, which are also known as learning management systems (LMS) or virtual learning environments (VLE), are web applications, meaning they run on a server and are accessed by using a web browser. Both students and tutors can access the system from anywhere with an Internet connection. The Moodle community has been critical in the success of the system. With so many global users, ther is always someone who can answer a question or give advice. At the same time, the Moodle developers and users work together to ensure quality, add new ,modules and features, and suggest new ideas for development [14].[15]. Moodle also stacks up well against the feature sets of the major commercial systems, e.g., Blackboard and WebCT [16]. Moodle provides many learning tools and activities such as forums, chats, quizzes, surveys, gather and review assignments, and recording grades.

Moodle was used in AOU mainly to design a well formed learning management system which facilitates the interaction among all parties in the teaching process, students and tutors, and more over to integrate the LMS with the student information system (SIS).

In addition that Moodle is easy to learn and use, and that it is popular with large user community and development bodies. Moodle is flexible in terms of:

- Multi-language interface,
- Customization (site, profiles),

---

[1] http://www.elearningproviders.org/HTML/pages/link.asp?id=15

- Separate group features, and pedagogy.

The unified image of the e-learning platform of the AOU from the starting web page shown in figure1, the users will be able to:

- Connect to the SIS, where they could do online registration, seeing their grades and averages as presented in figure2.
- Perform learning activities through the LMS, such as submitting assignments, do online quizzes, etc.
- Retrieve resources through AOU digital library subscriptions.

Figure 1. The Unified Image Of The AOU E-Learning Systems

IN-HOUSE DEVELOPMENT AND ENHANCEMENTS

To fit the AOU requirements and specification, a number of modifications and customizations were made (see Figure3), including:

- Log records. Logs are replicated into other isolated tables, to increase performance, and to keep track records for long period, while removing these log records from original tables.
- Activate questionnaire for students. Redirect students to fill the questionnaire of each course they study.
- Export questionnaire results to Excel files in special format.
- Students attendance and absences sheets are provided.
- Grades customizations (fractions) excel sheets are available.
- Randomly captured assignments for quality assurances purposes.
- WAP (wireless application protocol) services (grades, schedule, financial issues, news) are presented.

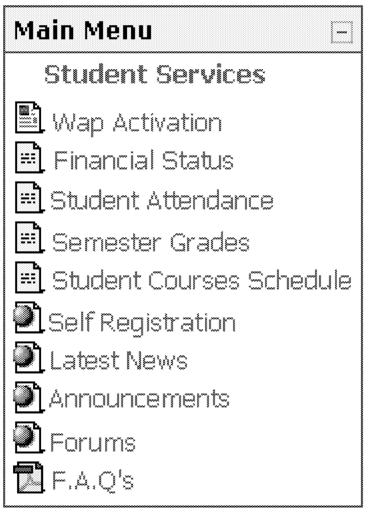


Figure 2. The SIS Of The AOU

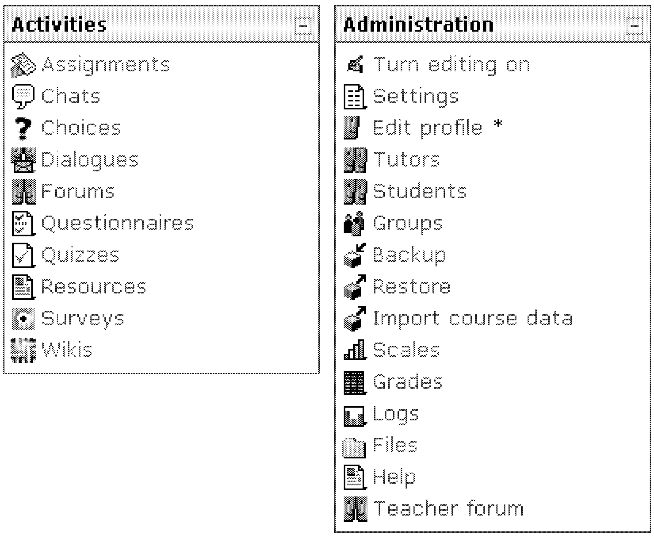


Figure 3. LMS Course Activities and Administration

INTEGRATING LMS WITH SIS AND HUMAN RESOURCE

The learning management system (LMS) is software that automates the administration of training events. The term LMS is now used to describe a wide range of applications that track student training and may include functions to:

- Manage users logs, course catalogs, and activity reports
- Provide basic communication tools (email, chat, whiteboard, video conferencing)
- Manage competency (e-Tests, e-Assignments)
- Allow personalization (user profiles, custom news, recent activity, RSS)
- Enable monitoring activities (QA, accreditation, external assessment).

The usefulness of the LMS could be summarized as follows:

- Simplicity, easy creation and maintenance of courses.
- Reuse, support of existing content reuse.
- CMC, TMA, Tests, Progress, learner involvement.
- Security, secure authentication/authorization
- Administration, intuitive management features
- Technical support, active support groups
- Language, true multi-lingual
- Affordability, maintenance and annual charges.

The student information system (SIS) is an Oracle based program which provides the necessary information such as students' information, courses registered, faculties, grades, etc. LMS integration with SIS (or LMS-SIS) is a system used inside the university to reducing accessing time, automatically generating accounts, minimizing faults, mistakes and errors to null, obtaining availability of requirements and simplifying registering, entering and filling process as shown in figure 4.

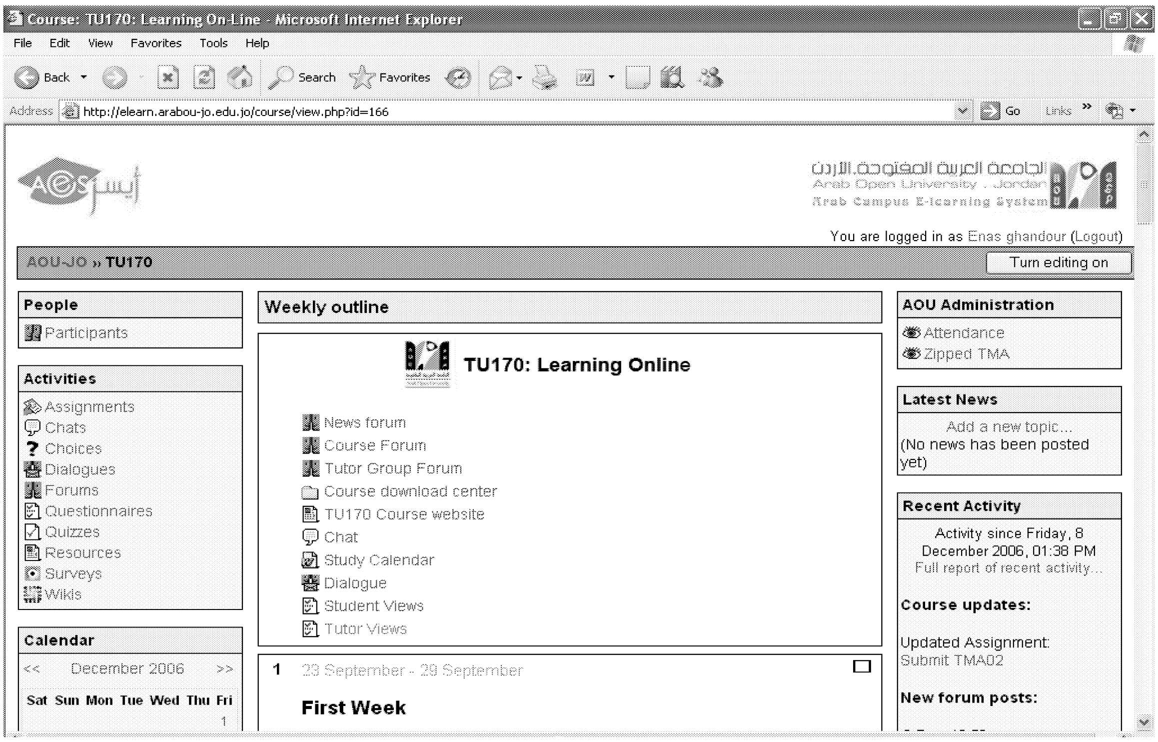


Figure 4. The LMS of AOU

The integration process added a lot of facilities which reduces time and cost in the following ways:

- Automatic structure enrollment: each student is provided with a username and password which enable students to register automatically.
- Automatic course enrollment: students are automatically enrolled into LMS courses they have been registered.
- Automatic group enrollment: students are automatically enrolled into LMS courses group, as they registered this group in the university.
- Automatically withdraw students from courses where students want to drop or have some financial problems.
- Student semester grades: students are enabled to see their grades through the LMS rather than bringing it from registrar.
- Students registered courses: where students could see the registered courses information such as their groups, time, course names and short names.
- Student's financial issues: where students could see their financial status and payment schedule.

The system is intended to satisfy the special needs and methodology adopted at the AOU. The SIS is flexible enough to adapt to the specific needs of branches while maintaining a unified standard that facilitates the interoperability of the system amongst branches and the headquarters. The SIS performs all aspects of students' information functions from filing an application to admission up to graduation, within the AOU methods. The SIS deals with all the entities involved and facilitate an easy and reliable way of the entities to perform their functions.

QUALITY ASSURANCE STRATEGY AT AOU

Arab Open University with the collaboration with UKOU performs a number of procedures to guarantee the quality of learning process. The description of these procedures is summarized in the following points:

- TMA marking template: Tutor marked assignment template is a form filled by the tutor of a course for each submitted TMA by students. It contains the deserved grade for every part of the TMA along with the feedback comments to the students.
- TMA monitoring: A form filled by the course coordinator and the program coordinator designed for monitoring tutors marking and filling the TMA templates.
- TMA samples: Three TMA samples should be collected for each section. One with a good grade, one is average, and one with a low grade.
- Quiz samples: Three samples should be collected for each quiz.
- Final exam samples: Three samples should be collected for each section of every course.
- Student questionnaire: A questionnaire filled by the students of every section to monitor the tutor, the course, and the tutoring environment.
- Tutor view questionnaire: A questionnaire filled by tutors to monitor the course content and the tutoring environment.
- Face-to-face preview: A form filled by the program coordinator to monitor tutor performance after attending a tutoring session of a specific tutor.
- Final grade statistics and distributions: grades reports and distributing of grades generated by SIS system after submitting student final grades.

At the end of each semester, each course coordinator has to prepare a complete folder that contains the following documents:

- Three samples of a marked TMA for each tutor in the course, each sample should be associated with its marking template and its monitoring form approved by the program coordinator. Notice that the three samples should be selected randomly; one is good, one is average and one is weak, and this is done automatically nowadays.
- Three samples from each quiz during running the course. One sample from

each of the good, average and weak categories.

- Three samples from the marked final exam of the course. One sample from each of the good, average and weak categories.
- 4-The face-to-face monitoring form for each tutor.
- The tutor monitoring forms
- Results of student questioners on the course level and for each tutor.
- Students' grades
- Grade distributions and statistics.

One of the duties of the program coordinator is to supervise the preparation of the above documents for all courses in the program and send them to the headquarter of the university to be reviewed from the external examiners whom usually come from UKOU.
Notice that preparing and performing such documents consume the time and efforts of many administrational and educational members of the university including tutors, course coordinators, program coordinators, and secretaries.

CONCLUSION

The open learning and distance learning become widely used as a way of teaching in the education community. The need for learning management systems to deliver the courses online becomes a significant issue. We discussed the efficient features of Moodle as a learning management system used in the Arab Open University. In this paper, a complete description of the improvements that have been conducted over the learning management system at AOU is introduced. The university strict regulations on the learning process to assure the quality of delivering all learning activities in an optimal way. Accordingly, there is a need to improve the existing learning management system to guarantee the implementation of such quality assurance regulations electronically to save effort and to perform all required procedures. We presented the integration process between learning management system and student information system that has been applied in the Arab Open University. Having consistent data is one of the main cores of the integration process in addition to the saving of efforts of time and cost. Moreover, this integration facilitates a lot of services which were done manually, basically the automatic enrolment process.

Future Work

Our new trend at AOU is to integrate VLE system with mobile technology, where students could receive the important notifications and messages through their mobile devices. We are investigating the needs and requirements to manage to do that.

REFERENCES

[1] Mikic, F., & Anido, L. 2006. Towards a standard for mobile technology. Proceedings of the International Conference on Systems and International Conference on Mobile communications and Learning Technologies (ICNICONSMCL'06) - Volume 00. Pp. 217-222.

[2] Siau, K., Lim, E, Shen, Z. 2001. "Mobile Commerce – Promises, Challenges, and Research Agenda,". Journal of Database Management, vol. 12, no. 3, pp. 4-13.

[3] Krogstie, J., Lyytinen, K., Opdahl, A., Pernici, B., Siau K., Smolander k. 2003. "Mobile Information Systems - Research Challenges on the Conceptual and Logical Level," Lecture Notes in Computer Science – Advanced Conceptual Modeling Techniques, vol. 2784, pp. 124-135.

[4] Siau, K., Shen, Z. 2003. "Building Customer Trust in Mobile Commerce," Communications of the ACM, vol. 46, no. 4, pp. 91-94.

[5] Galanxhi-Janaqi, H., Nah, F. 2004. "U-Commerce: Emerging Trends and Research Issues," Industrial Management and Data Systems, vol. 104, no. 9, pp. 744-755.

[6] Siau, K., Shen, Z. 2002. "Mobile Commerce Applications in Supply Chain Management," Journal of Internet Commerce, vol. 1, no. 3, pp. 3-14.

[7] Siau K., Shen, Z. 2003. "Mobile Communications and Mobile Services," International Journal of Mobile Communications, vol. 1, nos. 1/2, pp. 3-14.

[8] Siau, K., Nah, F. 2006. Mobile Technology in Education, IEEE Transactions on Education, Special Issue on Mobile Technology in Education, Vol 49, No.2.

[10] Rovai, A.P., & Barnum, K.T. 2003. On-line course effectiveness: an analysis of student interactions and perceptions of learning. Journal of Distance Education, 18(1), 57-73, 2003.

[11] Hammad, S., Al-Ayyoub, A.E., & Sarie, T. 2005. Combining existing e-learning components towards an IVLE. EBEL 2005 conference.

[14] Giannini-Gachago D., Lee M., & Thurab-Nkhosi D., 2005. Towards Development of Best Practice Guidelines for E-Learning Courses at the University of Botswana, Proceeding Of Computers and Advanced Technology In Education, Oranjestad, Aruba.

[15] Louca, S., Constantinides, C., & Ioannou, A. 2004. Quality Assurance and Control Model for E-Learning, Proceeding (428) Computers and Advanced Technology in Education

[16] Cole J., Using Moodle, O'Reilly, First edition, July 2005

[17] Kruse, K. The benefits of e-learning. 2003. [Online] : http://www.executivewomen.org/pdf/benefits_elearning.pdf.

WEB REFERENCES

[9] Freed, K. 2004. A History of Distance Learning, Retrieved June 25, 2004 from http://www.media-visions.com/ed-distlrn.html

[12] www.lmuaut.demon.co.uk/trc/edissues/ptgloss.htm

[13] http://alt.uno.edu/glossary.html

BIOGRAPHY

Bayan Abu Shawar received her B.Sc. degree in Computer Science from University of Jordan, Jordan, in 1995, and M.Sc degree in Computer Science from University of Jordan, in 1998. She worked as a teaching assistant at University of Petra from 1995-1998, then she worked as a lecturer at Computer Department at University of Jordan. Then she got a school bursary from the School of Computing at university of Leeds to continue PhD in computer Science at University of Leeds 2002-2005. Since finishing the PhD till now she works as an assistant professor in Information and Computing Department at Arab Open University in Jordan. Her research interests are: natural language processing, information retrieval, e-learning and learning management systems.

TEACHING OBJECT-ORIENTED CONCEPTS WITH THE HELP OF ALICE

Khalid Al-Tahat
Information Technology and Computing Department
Arab Open University - Jordan
Email: k_tahat@aou.edu.jo

KEYWORDS
Teaching Object-Oriented Concepts, Object-oriented paradigm, Visualisation, Alice

ABSTRACT

Object-oriented concepts are considered hard to teach and learn in introductory courses in computer science and software engineering. Animated program visualisation can be significantly used to support teaching object-orientation for beginners. However, there is a lack in instructional methods that supports such approach. This paper introduces a new instructional method for teaching object-oriented concepts with the help of a visualised 3D programming environment; Alice. The basic approach is to extract the object-oriented concept related components from a 3D Alice world and then use these components to explain the concept. This approach shows that animated program visualisation has the potential to help in teaching other computing aspects besides programming including methodological concepts.

INTRODUCTION

Formal presentations of object-oriented concepts are mainly theoretical; object-oriented programming books and resources are extremely wordy with minimum diagrams and illustrations. As a result, students find it hard to understand these concepts without being mapped to diagrams drawn with the instructor in the class.

Educators [1, 2] suggest that the main requirements for teaching today's students are:
- Students need to be actively involved in learning
- The ideas must be visually presented
- Students need to apply a concept while it is being taught
- The students must see the relevance of what is being taught

Our experience shows that beginners find it hard to understand object-oriented objects based on theatrical explanations in the class. We believe that introducing these concepts to beginners through visual representations relates the concepts to the real life experiences which help them understand it better. Interactive animated program visualisation can be the starting point for teaching object orientation as we propose in this work.

Interactive visualization has been employed in computer science education since the 1980s and has been successfully used to actively engage learners in different types of activities (Naps et al. 2003). To support teaching object-orientation for beginners, we suggest an instructional method that benefits from the use of animated visualisation tool, Alice. Using visualisation technologies help students understand object-orientation by keeping the focus on the visual representation of the objects while learning about them and their relationships.

Alice is an incredibly fun 3D programming environment created and distributed by Carnegie Mellon University (www.alice.org). Alice has an object-oriented flavour and has recently gained attention as a gentle introduction to object-oriented programming (Mullins 2009). The special characteristics of Alice such as 3D animation, real-life experience of dealing with objects, a direct-manipulation editor for instantiating objects, simple event handling mechanism, multi-media capabilities, and "syntax-free" IDE that engages students and raises their interest in learning programming (Mullins 2009) and (Al-Tahat 2010). Several researchers and computing scholars such as (Al-Tahat 2010; Bishop-Clark et al. 2006; Cooper et al. 2003) report the benefits of using Alice as an object-first tool; students confidence had increased in their programming ability and understanding of basic programming concepts.

Inspired by the benefits of using Alice in teaching programming, we investigate using Alice in teaching object-orientation concepts. We believe that students who enjoyed creating Alice worlds by choosing their objects and designing their interactions will find it interesting as well as easy to understand the methodological concepts behind these worlds.

The structure of the rest of paper is as follows. Section 2 explores Alice worlds and its main components. The main concepts in Object-orientation showing how to teach those using the suggested approach are covered in section 3. Section 4 includes our observations on using the proposed approach. Concluding remarks are given in section 5.

EXPLORING ALICE WORLDS

This section explores Alice worlds; introducing the Alice interface components and describing a simple Alice world that was created to demonstrate the proposed approach of teaching object-oriented concepts with the help of Alice.

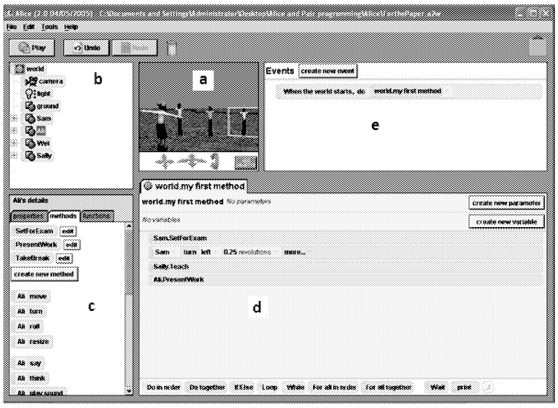


Figure 1: A Screenshot of the Alice Interface.

In the screenshot of the Alice interface, figure 1:

a. The world window provides a view of the virtual world that a students' program will control.
b. The object tree contains a list of the 3D objects in the virtual world.
c. The details area shows the properties, methods and functions for the object selected in the object tree.
d. The methods editor shows the code that defines a method a student is working on.
e. The events area allows students to call methods based on events in the world, such as mouse clicks or changes in the value of a variable

Alice programmer can chose to add objects from the Alice gallery which includes more than 350 3D objects. These objects are organised into groups including people and animals. Alice programmer can access the Alice gallery by pressing the "Add Objects" button underneath the world window (see figure 1.a). This causes the gallery to expand and allows the programmer to select the object they need. When selecting the object, a dialog with the object's information appears with a button to confirm the adding of this object to Alice world.

To build an Alice program, students choose the object they want to program from the object tree (figure 1.b). This will show all the properties, methods and function of that object as command tiles. Then programmers drag command tiles for that object from the details area (figure 1.c) into the method editor (figure 1.d), they then drop it and select the parameter values from a pop-up menu that appears. Method editors can contain any number of method calls as the programmer desires. These methods perform animations on whole 3D objects (e.g. moving a penguin through space) as well as an object's subparts (e.g. penguin's head).
Alice supports collaborative programming and when students press the play button (see figure 1) they are immediately able to see how their own animated program runs.

In this work, our concern will be regarding the objects created in Alice worlds; their characteristics, their behaviours and the relationships existing among and between them. The rest of this work will show how to use Alice to ease understanding object-oriented concepts refereeing to Alice components and the objects appearing in a snapshot of the Alice, figure 1.a.

UNDERSTANDING OBJECT-ORIENTED CONCEPTS WITH THE HELP OF ALICE

This section covers the main object-oriented concepts. We start by explaining the concept from the object-oriented paradigm point of view and then show how Alice supports understanding this concept.

Object, Attributes, Operations and Encapsulation

In the object-oriented paradigm, objects in a software system are viewed, just as in the real world, as things around us (Burton and Bruhn 2004). These things have certain properties or attributes, and are capable of performing certain behaviours or operations. The collection of these attributes and behaviours identify an object and give it its identity. Objects have relationships with other objects and they may exchange messages based on these relationships.

In Alice, programmers can add objects to their world by pressing the "Add Objects" button underneath the world window (see figure 1.a). Students can add an object to their virtual world by dragging the object into the 3D scene and/or by clicking on the picture of the object and pressing the "Add instance to world" button on the dialog box that appears.
Figure 2 below gives a closer look at the Alice world snapshot in figure 1.a. It shows four objects living in this world. The object tree, figure 1.b, tells

us the names of these objects. They are ‘Sam’, ‘Wie’, ‘Ali’ and ‘Sally’.

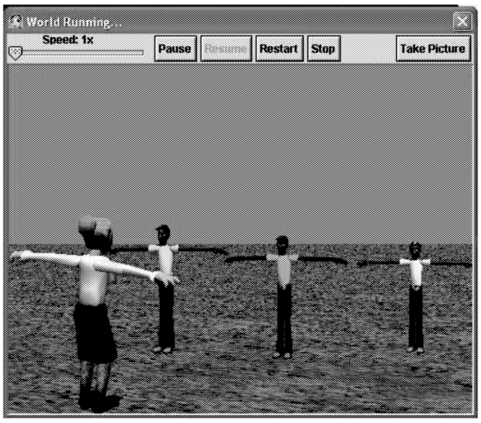


Figure 2: A Closer Look at Alice World Window in Figure 1.a

An attribute, in the object-oriented paradigm, describes a property of an object. The value of an attribute is a piece of information that an object knows. Clicking the properties tab in the Alice world’s details area (figure 1.c) shows the objects’ attributes/properties. Alice defines a number of default properties for its objects and gives Alice programmers the chance to define more properties as part of their program development. Figure 3 shows Sam’s properties; besides others, ‘Sam’ has the properties ‘course’, ‘age’, ’gender’ and ‘color’.

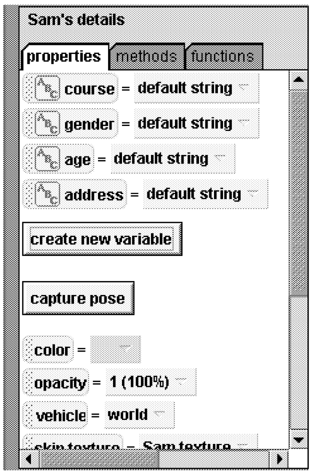


Figure 3: Sam’s Attributes/Properties.

Operations, in the object-oriented paradigm, are procedures that an object knows how to follow to perform its behaviours. Clicking the methods tab in the Alice world’s details area, (figure 1.c) shows the objects’ behaviour/methods. Alice implements a number of default methods for its objects and gives programmers the chance to implement more methods as part of their program development. Figure 4 shows the methods of the Student’s object ‘Sam’; besides others, ‘Sam’ has the methods ‘SetForExam’, ‘PresentWork’, ‘move’ and ‘say’.

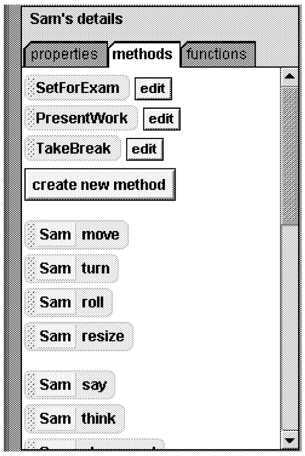


Figure 4: Sam’s Behaviours/Methods.

Packaging the attribute of an object with its behaviours within that object makes the object seen as being an entity and that is how encapsulation is defined in object-oriented paradigm. It makes us see the object as one unit. It also includes hiding the internal structure of an object so other objects can only access the object’s data or change its attributes only through its public methods. Alice supports encapsulation by introducing an object as a 3D entity associated with its attributes and behaviours. In the following section, we introduce class of object and abstraction concepts

Class of Objects and Abstraction

Based on the above definitions of attribute and behaviour, we may define a class of objects on the basis of having common attributes and behaviours. These are common to all the objects in the class. Objects that have the same attributes, operations and relationships belong to the same class. A relationship between two objects results in association among their classes.

Alice programmers are introduced to the class concept when adding new objects to their Alice virtual world. For instance, when Alice programmers choose to add a new ‘Student’ object to their Alice world, they drag the object of the chosen class into the 3D scene. Figure 5 shows adding an object of type ‘Student’ to the Alice world. This method for instantiating, that is creating an instance of an object, Alice objects makes Alice programmers see a class as a template for creating objects, which is another common definition for the class in object-oriented paradigm.

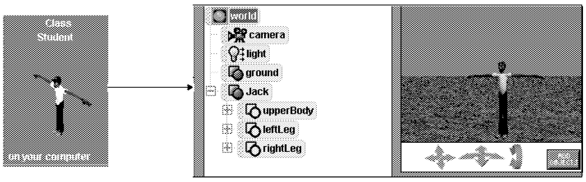


Figure 5: Instantiating an Object From Class Student in Alice

To carry out the abstraction process in the classroom, students are given objects of the same class and asked to extract their common attributes and behaviours to create an abstract data type; the class. As a result the class of objects will naturally evolve from the object details. This helps us introduce the concepts of abstraction and abstract data type (class) in the object-oriented paradigm.

Association Relationship

An association relationship, in the object-oriented paradigm, means that objects of one class are naturally associated with objects of another class in some way (Burton and Bruhn 2004). Association shows how an object of one class communicates with and gains access to an object of another class. Figure 6 shows a single object of type 'Teacher' associated with three objects of type 'Student' with a relationship named 'teaches'. This means that 'Sally' teaches a number of students or these students are taught by the teacher 'Sally'. Besides other communication forms, this association allows 'Sally' to 'giveInstructions' to her students and, on the other hand, allows students to 'presentWork' to their teacher.

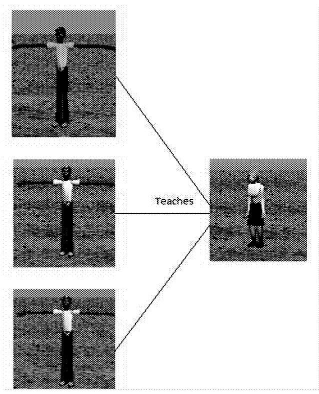


Figure 6: An Association Relationship Between an Object of Class 'Teacher' and Three Objects of Class 'Student'

Composition Relationship

Composition is a type of class association relationship. It represents the whole-part relationship in the object-oriented paradigm. In a whole-part relationship one class is regarded as being made up of one or more other classes (Burton and Bruhn 2004). Composition relationship states that the part cannot exist independently of the whole.

Alice programmers are introduced to the concept of composition when adding new objects to their Alice virtual world. Before conforming the adding of new instance by clicking the button "add instance to the world", an Alice programmer is given some information about the object to be added, including of how many parts it is composed. An object of 'Student' is composed of 16 parts as figure 7 shows, each of which is a smaller object itself.

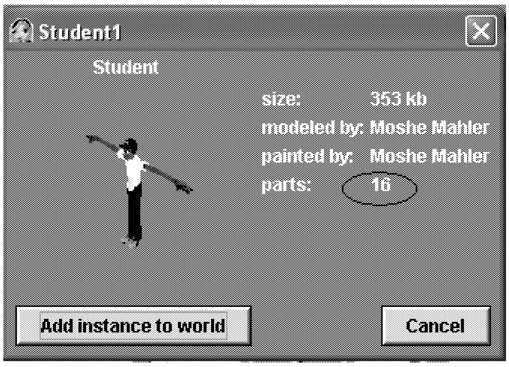


Figure 7: A Student Object is Composed of 16 Smaller Objects / Parts

The Alice object tree for an object of type 'Student', figure 8, shows that a 'Student' object is made up of the objects 'upperBody', 'leftLeg' and 'rightLeg'. The 'leftLeg' object, which is a part of the bigger object 'Student', is made up of the objects 'lowerLeg' and 'foot'. The '+' sign next to the objects 'upperBody' and 'rightLeg' indicates that these objects are also whole objects composed of smaller part objects.

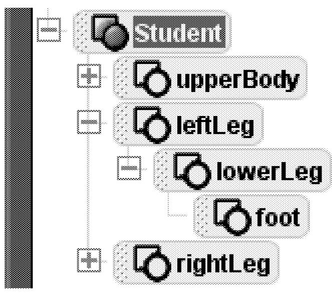


Figure 8: Alice Object Tree for an Object of Type 'Student' Showing the Object Composition

Generalisation/ Inheritance Relationship

Generalisation, in the object-oriented paradigm, is the process of extracting shared characteristics from two or more classes (called child classes or subclasses) (SourceMaking 2011), and combining them into a generalised parent or superclass. Shared characteristics can be attributes, associations or behaviours. According to Fowler (Fowler and Scott, 1997), "generalisation at the implementation perspective is associated with inheritance in programming languages".

Generalisation refers to an 'is-a' or 'substitutability' relationship. Through this relationship, a subclass can make use of the attributes and operations of its superclass. In addition to the attributes and operation they inherit, subclasses can define their own attributes and operations and also redefine the meaning of the inherited operations.

Alice doesn't support generalisation directly, but we can always make use of Alice artefacts and characters to explain this important concepts to our students. To generalise a superclass for classes used in this work namely, 'Teacher' and 'Student' classes, we identify the common attributes and behaviours between these two classes. Some of the attributes common between 'Student' and 'Teacher' classes are 'gender', 'age' and 'address' and so are the behaviours 'move', 'say' and 'think'. Now, we need to find a suitable name for our new class that fulfils the 'is-a' relationship requirements; we can describe a 'Teacher' and a 'Student' 'as-a' type of this new class. Consider class 'Person', we can say that a teacher 'is-a' person and also a student 'is-a' person. So, 'Person' is one possible name for the new superclass. Fig. 9 shows the new superclass 'Person' in a generalisation relationship with the subclasses 'Teacher' and 'Student'.

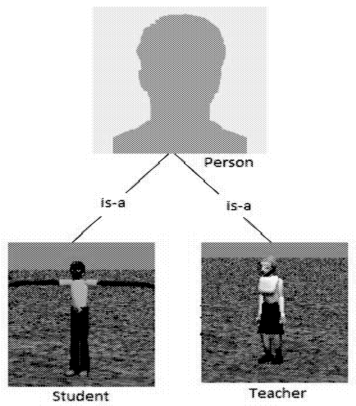


Figure 9: 'Person' is a Superclass for 'Teacher' and 'Student'

In this section, we have demonstrated the proposed instructional approach, of using an animated visualisation tool (Alice) to introduce the main concepts of the object-oriented paradigm.

OBSERVATIONS

The suggested approach was used to teach object-oriented concepts to first year students in their Introduction to programming module. Students were familiar with Alice as the tool was introduced to them during their first two weeks of the same module. The students' response for using Alice, again, to introduce new computing concepts and as part of the teaching tools was very positive. The idea of using their own Alice worlds that they have built before, in learning object oriented concepts, was warmly welcomed. We found that students' understanding of object-oriented concept is better than previous groups who were taught these concepts in traditional ways. We noticed clear improvements in students' in understanding abstractions and abstract data types and identifying real-life objects, their attributes, their behaviours, and means of their interactions. Students skills in using objects in problem solving has also enhanced. Classes' relationships that students used to find hard to grasp and understand, such as inheritance and composition, don't seem to be hard for most of the students any more. Students have included a well-defined attributes and behaviours for objects and classes when using these as part of their suggested solutions. In many occasions, students expressed their appreciations for using Alice to help understanding hard programming concepts and its notations.

CONCLUSION AND FUTURE WORK

In this paper, we have outlined a new approach to teach object-oriented concepts with the help of a 3D visualisation environment; Alice. The importance of this approach lies in its applicability and effectiveness in teaching object-orientation for beginners, where it has been successfully tested with students and approved to be easy to learn and consequently apply. The proposed approach benefits from the visual presentation of the object and the object's details provided by the Alice environment to aid understanding the concepts and explaining them to the students. It starts with students taking a snapshot of a scenario in an Alice world and then identifying the objects, their attributes and behaviours. Using the objects' details, students easily derive the abstract data type (the class) by applying the concept of abstraction. The relationships that may exist between and among these classes was then identified. We found that by using this approach, students do not just learn object-oriented concepts, but they also learn problem solving techniques. The author plans to cover more object-oriented concepts and also is investigating the use of Alice in teaching software modelling.

REFERENCES

Al-Tahat, K. (2010). Using Alice for Teaching Object-oriented Programming for Novice Programmers. In: R. Idrus et al. Eds. 2010. *Electronic Education: Towards an Immersive Learning Environment*. Pearson: Kuala Lumpur, pp. 189-195.

Bishop-Clark, C., Courte, J., & Howard, E. V. (2006) "Programming in pairs with Alice to improve confidence, enjoyment, and achievement", *Journal of Educational Computing Research*, 34:2, pp.213-228.

Burton, P.J. & Bruhn, R.E. (2004). "Using UML to facilitate the teaching of object oriented systems analysis and design". *Journal of Computing Sciences in Colleges*, 19 (3), 278-290.

Cooper, S., Dann, W., & Pausch, R. (2003), "Teaching Objects-first in Introductory Computer Science", *SIGCSE'03* February 19-23, 2003, Reno, Nevada, USA.

Fowler, M. & Scott, K. (1997). *UML Distilled: Applying the Standard Object Modeling Language.* Reading, MA: Addison-Wesley Longman.

Mullins, P., Whitfield, D. & Conlon, M. (2009). "Using Alice 2.0 as a first language". *Journal of Computing Sciences in Colleges*, 24 (3), 136-143.

Naps, T. L., Rößling, G., Almstrum, V., Dann, W., Fleischer, R., Hundhausen, C., Korhonen, A., Malmi, L., McNally, M., Rodger, S., & Velázquez-Iturbide, J. Á. (2003). "Exploring the Role of Visualization and Engagement in Computer Science Education". *ACM SIGCSE Bulliten 35*, 2 (June 20003), 131–152.

WEB REFERENCES

SourceMaking. UML [Online] Available at: http://sourcemaking.com/uml. [Accessed 10 June 2011].

BIOGRAPHY

Dr. Khalid Al-Tahat holds a BSc in Computer Science from Yarmouk University in Jordan, an MSc in Artificial Intelligence from the Malaysian University of Science and Technology, PhD in Computer Science –Software Engineering from the National University of Malaysia. Since 1998, Dr. Al-Tahat worked at eight different educational institutes in Malaysia, Jordan and United Kingdom. His research interests lie in the field of modern software engineering, Artificial Intelligence and teaching programming. He is also a committee member and a reviewer for a number of international conferences in computing and communication technologies.

GAME DESIGN AND AI

MAPSH\D -AIM Framework for Game Design

Russell B. Williams
Ph.D. Master of Cultural & Creative Industries Graduate Studies Division
Higher Colleges of Technology PO Box 25035, Abu Dhabi United Arab Emirates
E-mail: Russell.Williams@hct.ac.ae

KEY WORDS Methodology, Online Gaming, Serious Gaming, Storytelling, Design Process

ABSTRACT

Through design experience and study the author has identified nine interactive components in the process of game design. These are the Mechanic, Aesthetic, Poetic, Hedonic, Schematic, Dynamic, Audience, Intention and Media. Using these components it is possible to establish a comprehensive design process with clear paths of development through iteration and connection. The systematic nature of these components also makes it possible to use them as the basis for theoretical and applied game studies.

INTRODUCTION

Using *MDA*, Mechanics, Dynamics and Aesthetics LeBlanc (2004) as a starting point I have identified nine components in game design: the Mechanic, the Aesthetic, the Poetic, the Schematic, the Hedonic, the Dynamic, the Audience, the Intention, and the Media. These can be grouped as the MAPSH\D and AIM.

MECHANIC

The game mechanic is the part of a game that separates it from other forms of mediated content. Game mechanics involve the rules, actions and behaviors that players encounter when they are operating in the game space. The mechanic is the essence of a game and often the initial focus of game design. Salen and Zimmerman (2004) call this the designed system of the game and say that it is the genesis of meaningful play. Church (1999) talks about it as creating the opportunity for planning, intention and consequence.

Salen & Zimmerman (2004) identify three kinds of rules: Constituative, Operational and Implicit. Constituative rules are at the mathematical core of a game. They are abstract and do not provide an explicit guide for playing the game. Operational rules are the kind found in rulebooks and instruction sheets for games. These are the "rules of play." They constitute the objective or legal understanding of game play. Implicit rules are unwritten and vary from one context to another. These are the rules for conduct during play and will certainly apply to many different games. Implicit rules are provided by the culture in which the game is played and the specific group with whom the game is played rather than the game designer.

The mechanic is key in establishing the identity of a game and separating it from other games. Some argue that the mechanic in the form of the constituative and operational rules should be "exact and unambiguous" while others argue that some ambiguity in the specific rules can create higher levels of player engagement in the process of play. There needs to be enough specificity that the game can be played with little effort but enough latitude to make players a part of an emergent experience that can vary each time the game is played.

AESTHETIC

The Aesthetic is the interface between the internal aspects of the game and the players. It is the experiential form of the game. This is the point where designers discuss the visual, aural and haptic form of the game.

Some basic rules of interface design are that it needs to be simple, usable, consistent, responsive and transparent. Some make the argument for utterly simple and convention-based interfaces that reduce learning of the interface to zero. While this is a commendable goal and might make every game easier to play it would leave players with numbed senses and apathetic responses to game spaces and play. A little learning and surprise will always add some excitement and interest and will be successful when it is well managed. "It must be granted that there is some value in mystification, labyrinth, or surprise in the environment…[but] complete chaos without hint of connection is never pleasurable" (Lynch, 1960).

POETIC

The Poetic is the narrative that connects the content with the Audience. It "provides the players/protagonists with goals, conflict and an element of uncertainty...dramatic tension, an experience that is sought and encountered in computer games and as well as fiction films" (Rusch, 2008, p. 24).

A good story will not only increase interest in the game it will also make it easier to understand the Mechanic and improve the Dynamic of the game. "Not only does the narrative give extra meaning to the rules, but the rules help to perceive the narrative" (Salen & Zimmerman, 2004, p. 387). A good story will connect a game with a player through their interests, aspirations and culture. "There is the embedded narrative that consists of the pre-scripted moments and structures that are relatively fixed in the game. But the true strength of the games on the narrative level lies on the emergent narrative, the narrative that arises during play often in unexpected ways, as a result of the individual moment-to-moment game-play" (Rusch, 2008, 24).

SCHEMATIC

Schematics are representations of the relationships between objects and players in the game space. Every game has a space in which people play. Schematics can be a product of synthetic world, or they can be derived from the abstraction of an actual process, system, or environment (*PSE*).

Schematics involve the representation of relationships rather than reality. The represented relationships can be spatial, social, cultural or perceptual. Perhaps the most enduring image of a schematic representation is the subway map for London. Designed by Harry Beck, the map tells the user about the relationship between the stations rather than the actual topography of London. The important information for the user is getting to their destination and not their position in the physical world. The system in the schematic is abstracted and easier to comprehend than a physical reality.

A schematic is created by abstraction, systematically removing objects and representations that exist in reality but do not explain the underlying functions, relationships, goals, barriers, and affordances which exist in the PSE. Barriers are the obstacles to action within a PSE. Affordances refer to the elements that enable people to take desired actions. Abstraction leads to the conceptual model on which a PSE operates. This conceptual model is what Harry Beck visualized in his original subway map for London. The abstracted PSE becomes the foundation of the Schematic.

HEDONIC

The Hedonic is the pleasure that players get from the game. It is the emotional response of the player to the game, the consequence of the player's experience. This is what some people call "fun" and clearly, the type of fun experienced with *Risk* is not the same kind of fun we experience with *Solitaire* or *Resident Evil*. LeBlanc (2004) has identified eight kinds of fun in relation to playing games:

1. 1. Sensation: game as sense-pleasure
2. 2. Fantasy: game as make-believe
3. 3. Narrative: game as drama
4. 4. Challenge: game as obstacle course
5. 5. Fellowship: game as social framework
6. 6. Discovery: game as uncharted territory
7. 7. Expression: game as self-discovery
8. 8. Submission: game as pastime

A game experience is likely to provide more than one kind of fun. Game designers must ask what kind of fun they expect players to have and design it into the experience. People won't play if they don't think it's fun.

DYNAMIC

The Dynamic is the categorization of the game. It is what the various components look like in practice. "The rules never solely determine the play of a game. The rules are always set into motion within an experiential context that includes particular players with their own levels of desire, skill, and expectation" (Salen & Zimmerman, 2004, p. 449).

Examples of Dynamic categories are strategy, chance, trading, trivia, role-playing, and first-person shooter. While these do not provide a specific description of how the game will be played they provide a quick and simple means of identification through association. Potential users will imagine they have an immediate understanding of the game through this association.

AUDIENCE

A game should be designed with an audience in mind. The identity of these players will impact many aspects of the game design. You cannot employ the same design, communicate the same ideas, or expect players to enjoy an experience in the same way when you are dealing with people who vary by age, education, experience or culture. It is imperative to know who will be playing the game.

There are several ways to define an audience and each one can provide important information about player identity. Demographics and Psychographics are common ways to understand an audience. But both of these schemes leave out the cognitive and physical skills of intended players. It is important to know what your players are able to do before you set them a task. One person may not have the fine motor skills required. Another person may not have the cognitive ability to understand the game mechanic and its expression in play. Therefore it is important to identify the entry behaviors and abilities players must possess before they are able to play a game, grasp its meaning or enjoy the experience.

A very practical way to think about audiences is to use Personas. Cooper (1999) introduced this idea in talking about software application development. A persona involves identifying an archetypical user in a very detailed way, giving her a name and a face and a life. Building this persona will make it possible for everyone involved in the design process to know the members of their target audience in the form of a real person.

INTENTION

A designer can basically have one or more of seven intentions when they create a game: to inform, to connect, to attract, to study, to persuade, to entertain or to express. It is rare that a design will involve only one of these intentions. They work in concert within the communication process.

When learning is an explicit goal it must be thought out and developed clearly otherwise it will fail. When a game attempts to train players in the use or functioning of a process, system or environment then "…the accuracy of the process or effect being simulated for training is of primary importance. In addition to the accuracy of what's being taught, the serious game must also be concerned with whether – and what – the game is actually teaching the player. If the player learns to beat the game but can't usefully apply what he's learned in the real world, then the serious game has failed its mission" (Michael & Chen, 2006, p. 43).

The Intention of a design must be clear and work in concert with the other elements of the game. It must take into consideration the general nature of the game as well as the particular attributes and specifications. It is up to the designer to control the experience of their intentions. People often play with different objectives than those intended.

MEDIA

The design of a game involves the media that are used to deliver it to the target audience. Differences in media forms are based on the dimensions of experience they enable. There are at least nine forms of media with human agency at their core:

1) Graphic
2) Discrete
3) Textual
4) Integrated
5) Temporal
6) Interactive
7) Analogical
8) Virtual
9) Connected

Each form builds on the previous and creates the opportunity for different experiences. In other words, Temporal Media still rely on the dimensions of Graphic Media and Connected Media can take advantage of all of the forms that precede it. Adding a dimension makes each form more complex than the previous and changes the potential in the user experience. The dimensions at work in these forms are line, plane, space, sequence, synergy, time, control, intelligence, presence and distance. Only intelligence and presence specifically depend on digital technology for their existence while connected experiences and the dimension of distance are older than the first telephone.

THE FRAMEWORK

The game design process can begin with any one of the elements identified here. Design for meaningful and serious games will often begin with the establishment of the Schematic as this involves the abstraction of the Process, System or Environment that is at the core of the game. The consideration of Audience, Intention and Media is important to each component and should happen before their design but can happen afterward. Expect major revisions if accessing these components is done too late in the process. Within the AIM components the process is iterative as a clearer definition of one will lead to a clearer definition of the others. Intention and Audience can often be fixed but understanding the best media will result in a careful consideration of these along with the content of the game.

There are natural connections between these components that will dictate some design processes. For example, when Audience is the most important component in the AIM sector, then the MAPSH sector will be engaged at the Hedonic. When Intention is the most important component in the AIM sector, the MAPSH sector will be engaged at the Poetic. When Media is the most important component in the AIM sector, the MAPSH sector will be engaged at the Mechanic. The Mechanic, Aesthetic, Poetic and Hedonic are then processed iteratively with the Schematic remaining relatively stable when it is connected to a real world PSE. The developments at this point may require a return to the AIM sector for further refinement that will then lead to a re-evaluation of the MAPSH components. When the process is complete the final step is the evaluation of the game and the emergence of the Dynamic.

There are other starting points in the design process but each component needs to be developed fully and in an integrative way before the process is complete.

CONCLUSION

Games are as old as any known medium and a framework for game design needs to be developed without myopic devotion to a single mediated form. A view of game design that includes a broader range of media implementations will improve the possibility of developing stronger computer-based games because the medium will not be the beginning and end of the process. Hopefully this framework will provide game designers with a vocabulary, concepts and process that can be used to discover new ways to connect with players and create compelling game experiences.

The components in this framework can go beyond design and serve as a basis for analysis and research. Researchers can use the nine components as variables in their explorations of design and effects. Using what has been offered and further refining the definitions of each component will give researchers points of control and manipulation in laboratory-based effects studies. The components can be used as points of comparison in game content studies. They can also be used in developing survey instruments and identifying the most important aspects of games in the minds of players. Currently we are focused on game features and player-perspective as the basis of analysis, e.g., the Dynamics. Using a more detailed and varied structure will enable us to learn more about how games are created and ultimately inform our understanding of how they are received.

The framework involves nine parts of the game design process: the Mechanic, the Aesthetic, the Poetic, the Hedonic, the Schematic, the Audience, the Intention, the Media, and the Dynamic. Considering these components and the procedural framework they form will inform the design process and provide a consistent basis for game analysis and study. It is important to isolate factors within games when we want to analyze the impact of a game on people. It is also important to isolate and analyze these components in the attempt to create games that are significant, interesting, engaging, meaningful and enduring.

Learning more about this framework will make it possible to diversify and invigorate game design. Putting the focus on the framework will open the door to innovation that goes beyond the quality of the rendering and the number of cool weapons available on the second level of the game.

REFERENCES

Cooper, Alan. (1999). The Inmates are Running the Asylum. Indianapolis, IN: SAMS.

Gibson, J. J. (1977). The theory of affordances. In R. E. Shaw & J. Bransford (Eds.), Perceiving, Acting, and Knowing. Lawrence Erlbaum Associates.

Hunicke, R., LeBlanc, M., & Zubek, R. (2004). MDA: A formal approach to game design and game research. Paper presented at the AAAI th National Congress on Artificial Intelligence in San Jose, CA, July 2004.

Jenkins, Henry. (2004). "Game Design as narrative architecture." In Noah Waldrip-Fruin & Pat Harrigan, editors. (2004). First Person: New Media as Story, Performance, and Game. Cambridge, MA: The MIT Press. pp. 118-130.

Lynch, Kevin. (1960). The Image of the City. Boston: MIT Press.

Michael, David & Chen, Sande. (2006). Serious Games: Games that Educate, Train, and Inform. Boston: Thomson.

Rusch, Doris C. (2008). "Emotional Design of Computer Games and Feature Films," in Andreas Jahn-Sudmann & Ralf Stockman, Eds., *Computer Games as a Sociocultural Phenomenon: Games Without Frontiers and War Without Tears*. New York: Palgrave Macmillan. pp. 22-31.

Salen, Katie, & Zimmerman, Eric. (2004). *Rules of Play: Game Design Fundamentals*. Cambridge, MA: The MIT Press.

Zimmerman, Eric. (2004). "Narrative, interactivity, play, and games: Four naughty concepts in need of discipline." In Noah Waldrip-Fruin & Pat Harrigan, editors. (2004). *First Person: New Media as Story, Performance, and Game.* Cambridge, MA: The MIT Press. pp. 154-164.

WEB REFERENCES

Church, D. (1999). "Formal Abstract Design Tools." Gamasutra. Available online at: http://www.gamasutra.com/features/19990716/design_tools_01.htm

LeBlanc, M. 2004. Mechanics, Dynamics, Aesthetics: A Formal Approach to Game Design. Lecture at Northwestern University, April 2004. Available online at: http://algorithmancy.8kindsoffun.com/MDAnwu.ppt

Lindley, C. A. 2003. Game taxonomies: A high level framework for game analysis and design. Gamasutra. Available online at: http://www.gamasutra.com/features/20031003/lindley_01.shtml

BIOGRAPHY

RUSSELL B. WILLIAMS was born in Los Angeles, California and went to Linfield College where he studied communication. He then worked as a television journalist before pursuing a Master's in Communication at SWBTS followed by a Ph.D. in Mass Communication from Indiana University. He taught the first digital media production courses in the United States at Baylor University and Penn State in the early 1990s and then moved to Hong Kong where he led the development of a digital graphic communication program for 10 years. He is now in Abu Dhabi where he is building a Master's program in Cultural and Creative Industries at the Higher Colleges of Technology. He has been published in international journals, presented in international conferences and advised governments on communication technology and digital entrepreneurship in Asia, Africa, Europe and North America. He also possesses a U.S. patent for a game design and method of play while being academically interested in applied media processes and theory as well as the subjective experience of game play.

Sanctions, Punishment, and Game Design:

Designing MMORPGs for Fair Treatment of Players

Georgios Christou
Department of Computer Science & Engineering,
European University Cyprus
6 Diogenes St., Engomi, 1516 Nicosia, Cyprus
Email: christoug@acm.org

KEYWORDS

Online gaming, design.

ABSTRACT

This paper discusses the question of how game designers can keep their games fair when adding collaborative content and rewards. The paper presents findings from World of Warcraft in-game chat transcripts and data gathered through an online survey that was placed online for approximately 9 months during 2010. From the chat transcripts, we find that players sometimes do abuse mechanisms for collaborative content, which displeases the other players involved. We also find that players believe that there needs to be some kind of sanctioning mechanism to players that abuse grouping mechanisms. Finally we present suggestions to game designers on how to create sanctioning mechanisms based on human cooperation research.

INTRODUCTION

Massively Multi-Player Online Games (MMOGs) are rapidly becoming a favorite pastime online activity (Nielsen Wire, 2010). It has been proposed that the draw to these games is the "social factor" (Ducheneaut & Moore, 2005; Jakobson & Taylor, 2003), players not only playing to fulfill the game's goals, but also to socialize with other players. Sociability in certain MMOGs, such as massively multi-player role playing games (MMORPGs), manifests through the creation of ad-hoc groups so that players can overcome the game's collaborative content (Eklund & Johansson, 2010). The rewards from completing collaborative content are usually not enough for all group participants, hence the challenge and recurrent appeal to completing this type of content when playing MMORPGs. This way of keeping the players' interest also creates problems, as players do not know with whom they become grouped through the game interface, and cannot know if the group participants will treat each other fairly. Thus, it is up to the game designer to ensure that group members behave in a way that builds a positive and pleasant experience, rather than an aggravating one. In this article I examine player perceptions about leaving from an ad-hoc group, and discuss consequences for game design of MMORPGs.

Williams (2006) states that as humans are social creatures, that want to have social ties with other humans. These ties may be instrumental (Habermas, 1981), where one human requires the help of others towards his/her gain, and merely views others as resources, or they could be social (Habermas, 1981), where the actor takes into account the actions and the reactions of others. Virtual communities (Rheingold, 2000) offer the same kinds of social binds as the real world, with MMORPGs being another virtual community, but with the added complication that they also allow socializing through competing and cooperating to achieve game objectives. This is a common way of socializing in World of Warcraft (WoW), one of the most successful MMORPGs to date in terms of user base, with at least 11 million users (Gray, 2008).

Players group together in groups that vary in size from 5 to 25 players, and go against the game's dungeon instances. During these instance runs, both social and instrumental behavior is expected by players, as the point of entering a group and going against the game's collaborative content is to receive rewards as well as have fun playing a game. But, as in any community, problems arise through abuse of societal structures leading to requirements of punishment and sanctioning (Fehr & Fischbacher, 2004).

AD-HOC GROUPS IN WORLD OF WARCRAFT

WoW players experience collaborative content throughout the game, while leveling their characters and when they reach the end-game (currently at level 85). However, to experience end-game content, players need to acquire appropriate equipment for their avatars (called gear), as end-level content has more stringent demands on the capabilities of the players' avatars. These capabilities are augmented with gear that is gathered from dungeon instances (5-player groups) and raids (10 or 25-player groups). As players tackle these harder instances and raids, they have the chance to gather better gear for their avatars. This process keeps the game interesting when players reach the end-level, and it also creates a hierarchy of players in terms of which raids and instances they can tackle, because of gear limitations. It is here where power gamers have an edge (Taylor, 2003), because the more time spent playing instances and heroics, the faster one gathers the required gear.

There are various types of groups of any number of players that may be created in WoW, but in this article we focus on groups of 5 random players (Pick-Up Groups - PUGs) that are created through the Looking For Group (LFG) tool that is offered by the game.

When a group is formed to tackle a 5-player instance, it requires three types of player avatars. A tank, a player whose avatar can withstand a great deal of damage and that can attract the instance monsters (mobs). A healer, a player whose avatar can heal other players by using healing spells. And three "DPS" (Damage-Per-Second) players. These are players whose avatars can affect damage on enemies in the instances. When the LFG tool is used, the players who are in queue are matched by the class that they declare when registering in the queue. While DPS class players are very easy to find, healers and tanks are hard to come by, with tanks making up the smallest amount of WoW's player population (Zardoz, 2010).

A problem that occurs during PUG instance runs is that sometimes a player may decide to leave the group for various reasons. However, when one player leaves the group, the rest of the group members must wait until a new player is available to substitute the player that left. This behavior exemplifies a problem that game designers are faced with when building collaboration structures in a social game. To examine how such problems may be solved and how players themselves view these problems, I created an online survey, and participated in 31 PUG instance runs where I logged the conversation and actions of the other players in the group.

1. What is your age?
2. Where are you from (continent)?
3. How long have you been playing WoW (in months)?
4. Faction you play most (Horde or Alliance)?
5. Which class type do you use the most?
6. When playing in a PUG, how easily do you quit from an instance without finishing it, given that you have each of the roles below? If you never played one of the roles, choose N/A. [Tank class]
7. When playing in a PUG, how easily do you quit from an instance without finishing it, given that you have each of the roles below? If you never played one of the roles, choose N/A. [Healer class]
8. When playing in a PUG, how easily do you quit from an instance without finishing it, given that you have each of the roles below? If you never played one of the roles, choose N/A. [DPS class]
9. What are legitimate reasons to quit a group before finishing the instance?
10. Do you chat when in a PUG?
11. Would you accept someone in the group if you knew that he or she quits instances easily?
12. If a player quits an instance run, he or she is not allowed to enter the LFG queue for 15 minutes. Do you think that the debuff when one quits an instance run is enough of a "punishment"?
13. What, in your opinion, would be a good way to punish people who quit instance runs?

Table 1 The survey questions

RESEARCH METHOD

The data for this study was gathered by means of chat transcript analysis, with records kept through chat logs that were analyzed using the Weft QDA tool (Pressure.to, 2011). The chat logs were kept using the chat logging facility offered by WoW. The observations were taken from playing two Horde characters (one a tank, the other a DPS) on two different European servers, Hellscream and Eonar. However, as the LFG tool allows cross-server group formation, the server where one is logged on may be irrelevant to group formation.

Data was also gathered using an online survey that was placed online and publicized through various game-related forums. It was left active for approximately 9 months during 2010. The survey consisted of 13 questions (Table 1). Five of these were demographics questions, and the rest were multiple choice questions or open-text questions. For some of the multiple choice questions the option was given to not choose from any of the given choices, but to write a short answer. 155 players responded to the survey. The respondents had median age 25, with minimum age 14 and maximum 50. 69% of the participants came from Europe, 24% from North America, and 7% from other continents. Respondents played WoW an average of 43 months (sd = 18.28), with minimum 2 and maximum 68. As the survey was publicized heavily in tank-oriented forums, approximately 46% of the respondents declared that they play mostly as tanks, with healers constituting 17%, and DPS constituting 37% of the sample.

The data from the chat logs were read through multiple times and cross-referenced with notes taken during each instance run, to examine situations where players left the group. Three reasons were found for leaving the group: real-life problem, instrumental behavior, and problematic group. A real life problem occurs when a player cannot continue because he or she must attend to a situation in real life. Instrumental behavior is displayed when a group member decides to leave the group because their own personal aspirations from the instance run have been met. Finally, a problematic group occurs when the players as a group do not have enough power to defeat the collaborative content, resulting in subsequent failures (called wipes) where every member in the group dies, and the group cannot proceed further in the instance.

Instance excursions and player behavior

When PUGs are formed, players are expected to know how to play their class, to know the content of the instance, and to know the tactics necessary to bring down the instance bosses (high level monsters which are significantly more difficult to bring down than mobs). Bosses are the ones that when defeated will give high level rewards in the form of gear, as well as points that players may use to purchase gear for their avatars. Thus, playing instances provides rewards for all players involved, not just for the ones that manage to take the gear (usually two pieces per boss). Especially for well-geared players, instance runs are a way to gather points that will lead to higher level gear than those that can be gathered in any 5-man instance. Therefore, the goal of well-geared players is to go through as many instances as possible as fast as possible, to gather points fast.

Once players have been grouped together by the LFG tool, the players are teleported to the instance, and usually with only a customary salute to the other players (many times not even that) (Eklund & Johansson, 2010), the players start tackling the instance content. An example of a typical, complete run interaction is shown below.

Tank: lo
DPS1: hey
Healer: need all ok
DPS1: ty all

The above fragment is the complete transcript of the chat that occurred during one of my instance runs. Notice how only two players greet the group (including the researcher, signified as DPS1 in the above chat), and that only the researcher thanks the group in the end. The only other utterance is that of Healer, saying/advising that he/she will select 'Need' for the loot that dropped from one of the bosses. As I intended to observe the behavior of the other players, I did not initiate any social interaction, but participated in interactions that were initiated by other players, and always saluted the other players when I entered and left an instance.

When running a PUG instance players are expected to stay until the instance is "cleared", which means that either all the bosses of the instance have been killed, or the last boss has been killed (in some instances not all the bosses have to be killed to reach the final boss). This can be thought of as a social norm in WoW (Fehr & Fischbacher, 2004). In fact, in some instances, players debate whether to compete against all the bosses in the instance, as in the following example:

Answer	Percentage
Bad healing, but without wipes	4
Bad tanking, but without wipes	6
DPS is low	8
A couple of wipes on a mob or boss	27
More than 5 wipes on a mob or boss	60
I got what I needed (points, gear, etc.) from the run, so I can quit.	6
Bad player behavior	82

Table 2. Answers to the question "What are legitimate reasons to quit a group before finishing the instance?"

Tank: skip or do all?
DPS1(Researcher): dont mind
DPS2: do all
DPS2: Points

In this brief interaction, the tank asks if the group wants to do all of the bosses in the instance, with one of the DPS players saying that s/he did not mind, and the other preferring to down all the bosses to get more points.

Of course it is justified to deviate from this norm, and quit a PUG run. To gauge the players' feelings about quitting a PUG, the respondents were asked to name the acceptable reasons to quit a PUG run. The results are shown in Table 2. This question asked participants to choose all of the reasons that applied for them. Thus, the percentages shown do not add up to 100%. As can be seen, the most well accepted reason for leaving is due to unacceptable behavior by other players. The second reason is if there are a lot of wipes during the run. Bad playing by other players does not seem to be a justifiable reason. The questionnaire does not include answers that discuss real life disruptions, as WoW is a game, and it is understood that real life takes precedence, as is exhibited by the following excerpt:

DPS1: guys apologies but i have to leave
DPS1: my kid just woke up
DPS2: :(
TANK: nw :)

This excerpt shows a situation where one of the players must permanently leave the group. The tank answers with "nw", meaning "no worries", sympathizing and forgiving DPS1 for having to leave. DPS2 shows a sad face because DPS1 has to leave. The sad face signifies sympathy to DPS1 who must now disrupt his entertainment to deal with a real world problem.

From Table 2, only 6% of the survey respondents think that it is OK to leave the instance as soon as they have acquired enough points to purchase gear, or because one of the bosses has dropped the loot that they wanted, and they do not want to continue any further. This type of behavior is instrumental, and creates problems to the group, especially if the player that leaves is the tank or the healer. In these cases, the group may need to wait for some time. Also, the replacement player is robbed of their chance to earn the entire number of points given during a run, or roll on the loot that was taken by the deserter, a situation that is exemplified in the following chat excerpt:

Tank1 leaves the party.
Healer: ?
You are now queued in the Dungeon Finder.
DPS1: he wanted a badge
Healer: hate tanks
Healer: lol
Tank2 joins the party.
DPS2: hey
DPS1: welcome
Tank2: great
Tank2: less 1 emblem
Tank2: wht happened to the other tank?
DPS1: sorry the previous tank only wanted one
Healer: he just left:(
DPS1: so he dropped after he got it

In WoW the repercussion to the player leaving the group is a ban on attending another PUG for 15 minutes. This ban, called the "Deserter Debuff", runs its course in 15 minutes of real time whether the player is logged on or not.

The two situations exemplify two different reasons for leaving the group, one clearly accepted by the rest of the group, and another where the group is evidently angry. To examine the ease with which respondents would leave a group the question was posed to them, given the role they play in a group. When compared together, there was no statistical effect of role to the ease with which players decide to leave a group.

The respondents were also asked whether they would accept someone in a PUG whom they knew quits easily. 75.34% of the respondents answered 'No', while 24.66% of the respondents answered 'Yes'.

Finally, the respondents were asked whether the current sanctioning of players leaving a group (15 minutes debuff) is enough. 70.55% of the respondents answered 'Yes', while 29.45% answered 'No'.

DISCUSSION

Leaving a group is a topic that is hotly debated on many WoW forums (World of Warcraft Forums, 2009, 2010a, 2010b), with two reasons for leaving being the most prominent, agreeing with the survey participants: it is acceptable to leave a group if the group cannot complete the instance, or if players behave badly, but not because they *play* badly. The way this action is sanctioned in WoW now is to apply the deserter debuff to everyone, regardless of the reason they leave. However, as shown by the in-game excerpts and from the survey responses, there are legitimate reasons for leaving a group, with a large consensus by the community on the matter. Also, the respondents' answer to whether they would accept a player in their group if they knew the player's behavior, points to the need for a new way of sanctioning. Respondents suggest that to hinder people from deserting the group the debuff's time should be increased, its duration should be counted in play time and not real time, or any loot and points gained by the deserter during the particular instance run should be taken away.

Recently, tools that allow players to rate other players have emerged (i.e. "Playerscore," 2011), because players want to know more about other players than what is disclosed in the game. It is suggested that stable cooperation cannot be attained in groups that allow defection (Dawes & Thaler, 1988). Rather, there is cooperation for the first few periods, and then cooperation degrades to instrumental, behavior. However, providing a mechanism for sanctioning selfish behavior (such as deserting the group), induces a sense of security that the group members will cooperate at high levels voluntarily (Fehr & Fischbacher, 2004).

It is further suggested that human cooperation is based on conditional cooperation, that "prescribes cooperation if the other group members also cooperate, whereas the defection of others is a legitimate excuse for individual defection" (Fehr & Fischbacher, 2004). This is the exact behavior that is exhibited by WoW players in instances, with a minority being the exception to this social norm. Once a group member deserts the group, and a suitable replacement is not soon found, then the other members feel that they can leave as well, thus the group is disbanded. The tool's quick replacement of deserters did indeed rectify this situation, where before the tool's update finding a replacement for any role was an involved process.

As theory on conditional cooperation suggests that the absence of sanctions is likely to cause decreasing contributions over time from team members (Fehr & Fischbacher, 2004), if the sanctions imposed do not impact the sanctioned players, then the sanctions might as well not exist. In other words, with no sanctions, or weak sanctions, the behavior of norm-obeying players will change over time, and the norm may eventually become the exception.

The option of voting a player off the group is a sanctioning mechanism that has been implemented to keep players from exhibiting behavior that goes against the social norms, which agrees with conditional cooperation theory. However, this same sanctioning mechanism works against inexperienced players, effectively shunning them from cooperative content because of inadequacy, which in turn can only be mitigated by instance runs to gather both gear and experience. This creates a vicious circle that can only be broken through an existing social network in the game (Jakobson & Taylor, 2003).

Thus, game designers need to understand that adding formal rules that sanction "quitters" without giving the rest of the group some power over those sanctions, may not be as effective as allowing the group to decide on the punishment of the quitter. Combining sanctions on "quitters" based on the decision of the rest of the group members, together with sanctions that may be imposed on group members, such as reduced payoff depending on the type of sanctions they choose to apply to a group member, the system becomes fairer. As shown by Anderson and Putterman (2006) and Ones and Putterman (2007) if the people who may impose punishments on others also incur some form of reduced payoff, then they punish others less.

CONCLUSION

This article asks the question of how should game designers sanction players that decide to leave an in-game ad-hoc group. Data was presented from observations taken from a chat transcript analysis study that began since the introduction of the LFG tool in patch 3.3.0 of WoW, and 155 responses to an online questionnaire posted during the same period. Two examples of this behavior where discussed from excerpts taken directly from WoW, showing both justified and unjustified deserting of the group by players. Also, the feeling of WoW players towards this behavior was captured in responses to the survey.

The results from the survey show that players find the current sanctions imposed to deserters largely satisfactory, although they also say that they would not accept a player in a group if they knew that s/he left groups before the completion of the instance with ease. Taken together with the popularity of tools written by players - not by the company that created WoW - that allow ratings of players in instance runs, it is legitimate to ask whether better ways of sanctioning this type of behavior exist.

Theory on conditional cooperation provides certain answers that were discussed here, such as allowing the remaining group members to sanction the deserting member, together with a counterbalancing method of sanctioning the remaining members if they impose harsh sanctions on the deserting group member.

REFERENCES

Anderson, C. M., & Putterman, L. (2006). Do non-strategic sanctions obey the law of demand? The demand for punishment in the voluntary contribution mechanism. *Games and Economic Behavior, 54*(1), 1-24.

Dawes, R. M., & Thaler, R. H. (1988). Anomalies Cooperation. *The Journal of Economic Perspectives, 2*(3), 187-197.

Ducheneaut, N., & Moore, R. J. (2005). More that just 'XP': Learning social skills in massively multiplayer online games. *Interactive Technology and Smart Education, 2*(2), 89-100.

Eklund, L., & Johansson, M. (2010). *Social play? A study of social interaction in temporary group formation (PUG) in World of Warcraft.* Paper presented at the Nordic DiGRA 2010, Stockholm, Sweden.

Fehr, E., & Fischbacher, U. (2004). Social norms and human cooperation. *Trends in Cognitive Sciences, 8*(4), 185-190.

Gray, M. (2008). World of Warcraft hits 11 million subscribers worldwide Retrieved 26th July 2010, from http://www.wow.com/2008/10/28/world-of-warcraft-hits-11-million-subscribers-worldwide/

Habermas, J. (1981). *The theory of communicative action (Volume 1): Reason and the realization of society* (T. McCarthy, Trans.). Boston, MA: Beacon Press.

Jakobson, M., & Taylor, T. L. (2003). *The Sopranos meets Everquest: Social Networking in MMOGs.* Paper presented at the Proceedings of DAC2003, Melbourne, Australia.

Nielsen Wire. (2010). What Americans Do Online: Social Media And Games Dominate Activity Retrieved 4th August 2010, from http://blog.nielsen.com/nielsenwire/online_mobile/what-americans-do-online-social-media-and-games-dominate-activity/

Ones, U., & Putterman, L. (2007). The ecology of collective action: A public goods and sanctions experiment with controlled group formation *Journal of Economic Behavior & Organization, 62*(4), 495-521.

. Playerscore. (2011) Retrieved 24th August 2011, from http://www.tentonhammer.com/playerscore

Pressure.to. (2011). Weft QDA: A free qualitative analysis software application Retrieved 18th August 2011, from http://www.pressure.to/qda/

Rheingold, H. (2000). *The Virtual Community: Homesteading on the Electronic Frontier*. London: MIT Press.

Taylor, T. L. (2003). *Power gamers just want to have fun?: Instrumental play in a mmog.* Paper presented at the Level Up conference of the Digital Games Research Association, Utrecht.

Williams, D. (2006). Why game studies now? Gamers don't bowl alone. *Games and Culture, 1*(1), 13-16.

World of Warcraft Forums. (2009). Why do you leave on first wipe? Retrieved 23rd September 2010, from http://forums.wow-europe.com/thread.html?topicId=11824664775&sid=1

World of Warcraft Forums. (2010a). Penalty for leaving group early Retrieved 23rd September 2010, from http://forums.wow-europe.com/thread.html?topicId=14250267742&sid=1

World of Warcraft Forums. (2010b). PUG Deaths - Why do you leave? Retrieved 23rd September 2010, from http://forums.worldofwarcraft.com/thread.html?topicId=23767543427&sid=1

Zardoz. (2010). Armory Data Mining Retrieved 17th August 2010, from http://armorydatamining.appspot.com/

FUZZY Q-LEARNING FOR FIRST PERSON SHOOTERS

Youssef S. G. Nashed
Department of Information Engineering
University of Parma
43124, Parma - Italy
nashed@ce.unipr.it

Darryl N. Davis
Department of Computer Science
University of Hull
HU6 7RX, UK
d.n.davis@hull.ac.uk

KEYWORDS
Reinforcement Learning, Q-Learning, Fuzzy Systems, First Person Shooters, Quake-II.

ABSTRACT

Here machine learning techniques in the context of their application within computer games is examined. The scope of the study was that of reinforcement learning [RL] algorithms as applied to the control of enemies in a video game dynamic environment thus providing interesting new experiences for different game players.
The project proved reinforcement learning algorithms are suitable and useful for simulating intelligence of agents within a game. The implemented learning bot exhibited interesting capabilities to adapt to a human player playing style in addition to outperforming other implemented and downloaded bots. Test results showed that incorporating learning into a game can reduce the extensive scripting and tuning phase, while retaining the ability to guide the NPCs not to exhibit totally unrealistic or unexpected behaviors. This gives the designers the scope to explore new strategies.
The effect of the exploration-exploitation policy on the learning convergence was studied and tested in depth. The combination of the two most popular reinforcement algorithms [Q-Learning and TD(λ)] resulted in faster learning rates and realistic behavior through the exploration period.

INTRODUCTION

Recently the video game industry has been growing exponentially and AI is playing an important role in the success of a game. Nowadays video games are required to deliver stunning graphics, real world physics, impressive sound, believable AI, and gripping storylines to be able to survive in this fierce market. The lifetime of a game title on retailers' shelves depends on how much fun the game is, which in turn depends on many aspects including the repeatability of in-game situations; the less predictable, the more fun the game. First person shooters [FPS] games can be grouped under the action genre which generally involves the player guiding a character through a set of virtual environments while encountering several types of enemies. These games depend on quick responses and eye-hand coordination (Johnson and Wiles 2001).
These games usually provide the player with various weapons to defeat increasingly challenging opponents.
This study requires aspects of Champandard's AI extensions to Quake-II (Champandard 2003) to be built into the Quake-II demonstrator code and extended to allow learning in the game. This will enable studying the effect of applying machine learning algorithms to Non-player characters [NPCs] in terms of emerging behaviors, diversity, optimality, and believability. Although the platform chosen is relatively old when compared to current commercial video games, the modest computational requirements for the game engine leaves room to experiment with more demanding machine learning techniques. Also the freely available source code for Quake-II makes it easier to interface with the game than with most current closed-source proprietary game engines. Finally, it is worth noting that the method described and implemented in this paper can be extrapolated and used in current games with minor optimizations and/or taking advantage of the parallel processing power offered by graphical processing units [GPUs].

RELATED WORK

There has always been a gap between AI research in academia and AI applied in video games. In the past game AI programmers often complained about being given insufficient computational resources compared to graphics which often refrained them from employing any significant AI techniques. With recent breakthroughs in computer and console hardware, and with the advance of dedicated processors for graphics and physics, game AI programmers have all the CPU processing power to borrow well established academic research like neural networks and genetic algorithms (Zanetti and El Rhalibi 2004) and incorporate machine learning into their games.
Game AI is all about believability. If the NPCs in a game can convince the player that they are not artificially controlled characters then the AI is said to be successful. FPS games have been around for many years and their AI systems are well established. However, players still prefer the multiplayer and online versions of these games which means the AI in FPS is still falling short from mimicking the skills of human players.
The following subsection contains an overview of the AI methods that are widely used in FPS games giving examples from commercial games.

AI in Commercial First Person Shooters

Finite State Machines
Finite state machines [FSMs] are a natural way of thinking about synthetic creatures, consisting of a set of states, inputs, and outputs. An agent is always in a given state [e.g. patrol]; when this state receives a specific input [e.g. player sighted],

the state changes [e.g. attack] and an action is performed [e.g. fire]. Games like Doom, Descent, and Quake use FSMs to model the enemies (Bourg and Seeman 2004), while behavior trees [a variation of hierarchical finite state machines] are used in the popular Halo series (Isla 2005).

Fuzzy Logic
Fuzzy logic replaces the standard Boolean logic concept of absolute True or False with degrees of membership to a specified fuzzy set (Zadeh 1965). Fuzzy logic maps perceived input values [crisp input] to a real number, usually between 0 and 1, describing how close this input is to a certain fuzzy variable. For example, if we want to assess how tall a person is using Boolean logic, we might have a rule which states: "If height is more than 180cms, then the person is tall". While in fuzzy logic, we *fuzzify* this input number "height" to degrees of membership in linguistic variables "short-average-tall". Thus, in fuzzy logic a person can be 0.1 short, 0.7 average, and 0.2 tall.
The game Unreal by Epic Games uses fuzzy logic to implement Fuzzy State Machines [FuSM]. These fuzzy machines provide smoother state transition than normal FSMs. 'S.W.A.T. 2' uses fuzzy logic to give each enemy a personality by parameterizing different fuzzy sets like aggression, courage, intelligence and cooperation (Johnson and Wiles 2001).

Planning
In planning systems, an agent is represented in terms of goals and actions. the agent thinks how to reach its closest goal given the current state of the system and available actions at each time step of the simulation or game. Decoupling actions and goals allows more modularity in the design and enables AI designers to produce more complex behaviors that would require a considerable amount of effort and computational memory to be expressed using FSMs.
The AI in the game F.E.A.R by Monolith Productions uses a method very similar to STRIPS (Nilsson 1998) to describe agent goals and actions (Orkin 2006). Similarily, hierarchical task network planning [HTN] is used in Killzone 2 bots (Champandard et al 2009).

Game designers use a mixture of careful level design, behavior scripts, character animation, and above all, extensive fine tuning of several parameters to simulate intelligence in commercial game NPCs. The use of machine learning techniques, like reinforcement learning, can dramatically reduce the time required to design intelligence and increase believability.

Machine Learning and First Person Shooters

Quake-II has been used as a research environment for developing autonomous agents for some time since the source code was released by Id software. Anticipation bot (Laird 2001) developed the SOAR engine (Laird et al. 1987) with Quake-II to demonstrate dynamical hierarchical task decomposition and prediction. Emobot (Hooley et al. 2004) was developed to prove that simple emotions like fear and anger can have a great effect on agents' actions and believability. While (Parker and Bryant 2009) used the game to train a neuroevolution controller for agent shooting based solely on raw visual input data.

In (Patel et al. 2011) the standard version of Q-Learning was used for agent combat tuning in an abstracted version of the game Counter-Strike, which achieved good policy results. However, the environment was a scaled down game with a state space of known size, which is not always applicable in a highly dynamic continuous simulation. This paper uses fuzzy inference to overcome the problem of continuous state space. (Westra and Dignum 2009) implemented an offline evolutionary neural learner for weapon and item selection in Quake III, while here we are concerned with online learning of the complete bot behavior.

REINFORCEMENT LEARNING

The standard RL model is depicted in Figure 1. At each time step $t+1$ the agent receives s as the state of the environment, which can be separated to i as input to the agent, and r as the scalar reward signal, usually a floating point number between -1 and 1, for the action a chosen at the previous time step t. The agent's behavior B should choose the action that is expected to increase the long-run sum of rewards. The input state s is perceived by the agent as input according to the input function I, which is the way the agent views the environment state. Typically I is the identity function [that is the agent perceives the exact state of the environment], however this function can change in the case of partially observable environments (Kaelbling et al. 1996).

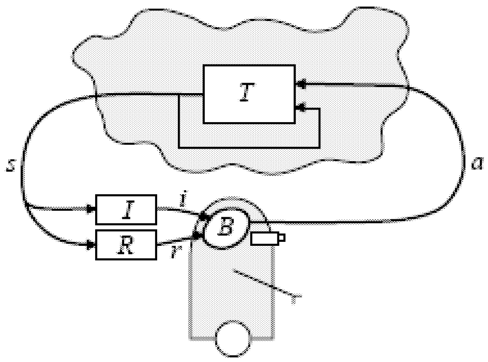


Figure 1: The Standard Reinforcement Learning model [taken from (Kaelbling et al. 1996)]

Exploration Versus Exploitation

The exploration-exploitation dilemma is very popular in the field of RL algorithms. This problem is studied in depth within statistical decision theory and the k-armed bandit problems (Berry and Fristedt 1985).
The balance between the exploration period and the exploitation period is called the exploration policy. It is extremely difficult to find an optimal exploration policy, nevertheless various researches proposed reasonable policies that produce good results.
Some approaches to this problem use dynamic programming incorporating basic Bayesian reasoning (Berry and Fristedt 1985), or learning automata techniques (Kaelbling et al. 1996). In (Gu et al. 2003) a genetic algorithm was used for exploration, while Jouffe proposed a heuristic based method based on the frequencies and success of the choices (Jouffe 1998). However, the most successful and used approaches are that of ad-hoc strategies.

Ad-hoc strategies are simple undirected [randomized] algorithms. They are far from optimal but are proven reasonable, traceable, and computationally efficient for most applications. The two main exploration stratigies that are commonly used are the ε-greedy exploration and the Boltzmann exploration.
In the ε-greedy exploration algorithm, actions with the highest estimated payoff are always chosen. The downside of this approach is that the algorithm can get stuck in suboptimal actions while the optimal solution is starved and never found. One useful heuristic is called "optimism in the face of uncertainty", in which actions are initialized with high optimistic estimations so that negative feedback is needed to remove the action from selection.
The Boltzmann exploration algorithm uses the Boltzmann distribution to assign action selection probabilities.

$$P(a) = \frac{e^{\frac{E(a)}{T}}}{\sum_{a' \epsilon A} e^{\frac{E(a')}{T}}} \qquad (1)$$

Where A is the set of all actions, $P(a)$ is the selection probability of action a, $E(a)$ is the estimated payoff of action a, and T is the temperature of the system.
The temperature parameter T is decreased over time to decrease exploration and allow exploitation of higher estimated actions. T values need careful tuning as they are used to control convergence.

Delayed Reward

The two main problems facing RL methods are temporal credit assignment and the curse of dimensionality.
In the standard RL model the agent performs an action then receives an immediate reward from the environment; this is not always applicable in the case of complex dynamic environments where the agent has to execute a series of actions to reach a state of the environment with a high reward signal [goal]. This kind of reward functions is called a sparse reward function.
Sparse reward functions are zero everywhere, except for a few places. Contrastly, dense reward functions give non-zero rewards most of the time.
The agent must learn from delayed rewards to cope with such environments. When a reinforcement signal is received, the value function of all the previous actions leading to the current state should be updated with a discounted version of the reward. Temporal credit assignment [delayed reward] problems are well modeled as Markov Decision Processes [MDPs].
MDPs consist of a set of actions A, a set of states S, a reward function R, and a state transition function T with which $T(s, a, s')$ means the probability of the environment being in state s' after executing action a in state s.
Some RL algorithms use a model of the environment, these use MDPs and learn the state transition function or construct the model while learning.
To form a model of a video game environment is extremely difficult; both sets of states and actions would be huge. Consequently this study is concerned with model free techniques like Q-Learning and Temporal difference [TD] algorithms.

Temporal Difference TD(λ)

TD methods introduced in (Sutton 1988) do not wait until the end of an episode like Monte Carlo methods to update the value function or estimate. Instead, it uses the temporal difference Δ_t in the expected return $V(s_t)$ value of a state s_t to the previous one(s).
The parameter λ controls how many steps or states to look back when adjusting the state expected return values, when $\lambda=0$ the algorithm only looks one step back. The update rule for a $TD(0)$ algorithm is:

$$\Delta_t = r_{t+1} + \gamma V(s_{t+1}) - V(s_t)$$

Where r_{t+1} is the reward signal received at state s_{t+1} and γ is a discount factor. Therefore:

$$V(s_t) = V(s_t) + \alpha\Delta_t$$

Where α is the learning rate.
In addition to the state expected return value V, (Sutton 1988) defines **eligibility traces** of a state $e(s)$ as the degree to which a state has been visited in the recent past. These traces are used when updating the expected value function of a state each according to its eligibility. The update rule for $e(s)$ is as follows:

$$e(s) = \begin{cases} \lambda\gamma e(s) + 1, & s = current\ state \\ \lambda\gamma e(s), & otherwise \end{cases} \qquad (2)$$

Q-Learning

Q-Learning (Watkins and Dayan 1992) is probably the most used RL algorithm. Q-Learning can be considered a $TD(0)$ method but instead of estimating a value function for each state, Q-Learning maintains a value function $Q(s,a)$ for each state-action pair.
Let's say the agent has performed an action a, received a reward signal r, and moved from state s to s'. The Q value for the stat-action pair (s, a) is updated using an equation very similar to the TD update equation, except that the highest Q value for the pair (s', a') is used. Where a' is a possible action that can be executed in state s'.

$$Q(s,a) = Q(s,a) + \alpha(r + \gamma \max_{a'} Q(s',a') - Q(s,a)) \qquad (3)$$

Where α is the learning rate and γ is a discount factor.

Generalization

In a simple discrete state space environment, the state and action spaces can be enumerated and the Q values stored in a lookup table. The use of such a data structure in a large continuous state space environment is impractical. This problem is called the curse of dimensionality, and to address it some sort of input generalization should be employed. Generalization uses compact storage of learned experience and applies it to similar states and actions.
Input generalization is achieved by the use of function approximators such as Cerebellar Model Articulation Controller [CMAC] (Sutton 1996), neural networks (Lin 1992), and fuzzy inference systems (Berenji 1996) (Jouffe 1998).

DESIGN AND IMPLEMENTATION

Being a FPS game, Quake-II players move in a 3D virtual world viewing it from the eyes of the main character. The main task of a player is staying alive and eliminating his/her opponents.

The research involved the design and implementation of three agent types. Each agent [bot] was designed and implemented in an incremental fashion, its action was built and tested individually and every bot contributed in the implementation of its successor. The navigation module used a reactive architecture and is divided into two levels: *Low level actions* [short term goals] like obstacle avoidance and turning; and *high level actions* [long term goals] like taking cover, gathering, investigating, chasing, fleeing, and dodging. There are only two effectors for the navigation module: *Step*, which takes a 2D vector specifying the direction as input, and *Turn*, which takes three floating point values as input, specifying the yaw, pitch, and roll angles for the bot orientation.

Rule Based FSM Bot "Tom"

Tom is the first and simplest bot, implemented using scripts. Tom uses a FSM to control his actions and decisions. Here FSMs are designed as a form of Rule Based Systems [RBS]. The implemented RBS is a *forward chaining* RBS with a *first applicable* conflict resolution mechanism (Bourg et al. 2004). This system can be described in terms of three mutually exclusive sets: states, actions, and inputs; and a rule base.

States

States dictate the action that is chosen at a given point in time. Tom can be in one and only one of the following states:

Searching: looking around exploring as much possible of the level map.
Wandering: randomly moving in the level map.
Attacking: firing the held weapon at an enemy.
Hunting: pursuing the enemy's current position.
Gathering: collecting health, armor, weapons, or ammo.
Fleeing: running away from the enemy to a safe position.

Actions

Actions are the outputs of an AI system. They are considered the bot's effectors to interact with the environment. Tom's effectors are outlined below:

Turn: keep yawing with a certain angle.
Move: step in a given direction.
Shoot: fire the held weapon.
Seek: move to a given location on the map.
Avoid: move away from a given location on the map.

Inputs

They represent the values and observations that are fed into the system through sensors. Tom has the following Boolean sensor inputs:

Enemy sighted: another player or bot is in the field of view.
Item sighted: an item health, weapon, or ammo is in the field of view.
Is confident: health is above 20%.
Is hit: suffering from a projectile or weapon damage.

Rules

The rule base consists of production rules. These rules are basically in the form of "*If ... then"* statements. They are crafted by an expert to transfer human knowledge to the system. Table 1 illustrates the rules at each state.

Table 1: Rules for the FSM Scripted Bot

Rank	*State*	*Input*	*New State*	*Action*
1	Searching	Enemy Sighted	Attacking	Shoot
2	Searching	Is Hit	Searching	Turn
(Default)	Searching	Any Other	Wandering	Move
1	Wandering	Enemy Sighted	Attacking	Shoot
2	Wandering	Item Sighted & ¬Is Confident	Gathering	Seek
3	Wandering	Is Hit	Searching	Turn
(Default)	Wandering	Any Other	Wandering	Move
1	Attacking	Enemy Sighted & Is Confident	Hunting	Seek
2	Attacking	Enemy Sighted & ¬Is Confident	Fleeing	Avoid
(Default)	Attacking	Any Other	Wandering	Move
1	Hunting	Enemy Sighted	Attacking	Shoot
(Default)	Hunting	Any Other	Wandering	Move
1	Gathering	Enemy Sighted	Attacking	Shoot
2	Gathering	Item Sighted & ¬Is Confident	Gathering	Seek
3	Gathering	Is Hit	Searching	Turn
(Default)	Gathering	Any Other	Wandering	Move
1	Fleeing	Enemy Sighted	Attacking	Shoot
(Default)	Fleeing	Any Other	Wandering	Move

Fuzzy Rule Based Agent "Yianni"

Yianni uses Tom's RBS redesigned as a fuzzy rule based system (FRBS). Yianni uses fuzzy logic instead of Boolean sensors to classify sensor crisp values as degrees of membership to a specific fuzzy set.

There are two types of FRBSs; the Mamdani FRBS and the Takagi-Sugeno-Kang (TSK) FRBS. The difference between the two is the types of inputs [antecedents] and outputs [consequents] used (Alcalá et al. 2000).

The Mamdani system rules are constructed in the following form:

$$\text{IF } X_1 \text{ is } A_1 \text{and } \dots \text{ and } X_n \text{ is } A_n \text{ THEN Y is B}$$

Where $X = \{X_1, X_2\ \dots, X_n\}$ is the set of real value input variables, $A = \{A_1,\ A_2,\ \dots,\ A_n\}$ is the set of fuzzy linguistic variables for the inputs, Y is an output variable, and B is also a fuzzy linguistic variable.

The TSK consequents are a function of the input variables, taking the form:

$$\text{IF } X_1 \text{ is } A_1 \text{and } \dots \text{ and } X_n \text{ is } A_n \text{ THEN Y is } (\mu_1 \wedge \dots \wedge \mu_n)$$

Where μ_1 is the truth value for input variable X_1 in the fuzzy set defined by A_1, and $\wedge$ is the disjunction [min] operator.

This implementation uses the TSK FRBS which can be expressed in terms of its input and output variables along with the rule base.

Fuzzy Input Variables [Antecedents]

Sensor Input fuzzification provides smooth transition between states and emulates the way humans think. Crisp floating point values are fed into the fuzzification component to be mapped to a degree of membership of a linguistic variable (e.g. how close an enemy is).

Yianni has the following list of sensor inputs, some with an accompanying fuzzy membership function and others are Boolean sensors.

Projectile distance: the length of the vector from the bot to the closest projectile observed. The game engine provides sensor data at each time step which in turn is fuzzified into three linguistic variables: *near*, *close*, and *far* according to the membership functions in Figure 2.

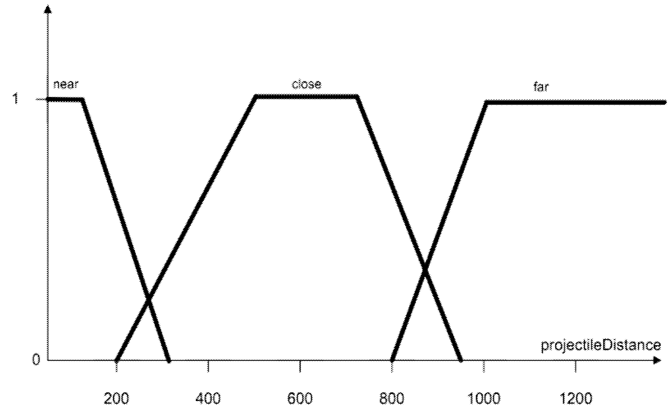


Figure 2: Fuzzy sets for input variable projectile distance

Enemy distance: the length of the vector from the bot to the closest enemy observed. The crisp input is obtained from the game engine sensors and assigned a degree of membership to the *near*, *close*, and *far* linguistic variables based on the functions depicted in Figure 3.

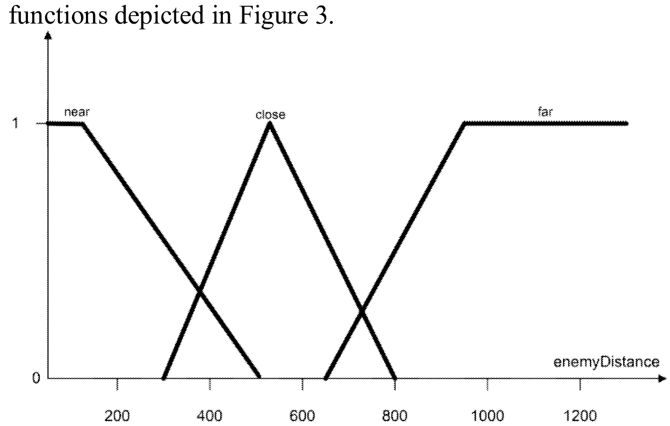


Figure 3: Fuzzy sets for input variable enemy distance

Item distance: the length of the vector from the bot to the closest item observed. Using the same membership functions as the enemy distance in Figure 3, item distance is also fuzzified into *near*, *close*, and *far*.

Confidence: the crisp value for this fuzzy variable is gathered from the agent's personal status and calculated as follows:

Confidence (out of 400) = *health* (%) + *armor* (%) + (*ammo*/2) + (*weapon rank* * 10)

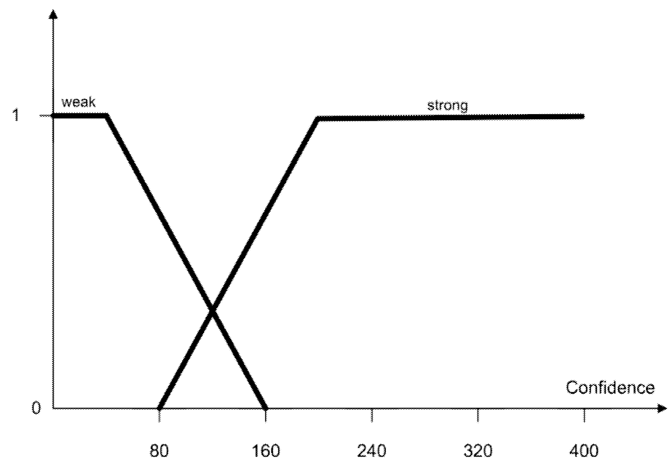


Figure 4: Fuzzy sets for input variable confidence

Four different game variables form the crisp value for the fuzzy confidence input variable. Each variable is clamped to 100 then added together to give the final value out of 400, which is then fuzzified into either *weak* or *strong* confidence based on the membership functions defined in Figure 4.

In Quake-II, maximum ammo for a weapon is 200, hence the division by 2. Weapons rank from 1-10 with increasing powers, so the multiplication by 10 gives the weapon strength out of 100 [a bot switches automatically to a higher rank weapon when acquired].

Got shot: a Boolean variable indicating that the bot has been hit by a projectile or any enemy weapon.

Enemy present: a Boolean variable indicating that the bot can see an enemy.

Enemy disappeared: a Boolean variable indicating that the bot was seeing an enemy but then the enemy ran away or took cover.

Fuzzy Output Variables [Consequents]

TSK consequents are assigned the value of the minimum [sometimes the product] of the truth values computed for the antecedents in a fuzzy rule. If more than one rule has the same consequent, the output variable will have multiple confidence values to choose from. This can be solved in various ways; the most popular techniques are bounded sum [sum and bound to 1], and maximum value [equivalent to OR-ing the confidences together].

The maximum value method is used here to resolve the multiple confidences per output conflict. Yianni has the following output variables each representing a tendency to a certain action:

Dodge: try to avoid incoming visible projectiles [bullets, rockets, etc...]

Flee: moving away [opposite direction] from an attacking enemy.

Attack: aiming and shooting the held weapon at a visible enemy.

Chase: moving towards or pursuing an enemy.

Gather: moving towards a visible item [health kit, armor, weapon, or ammo]

Search: turning around looking for enemies.

Investigate: moving towards a location in the level and searching the area.

Wander: random movement and turning while avoiding obstacles.

Actions

The agent's overall action is a weighted vector of a movement direction, and a weighted orientation angle for turning.

Each output fuzzy variable contributes to the action with the value calculated after firing the fuzzy rule where this variable is a consequent. This provides smooth control implementing the architecture proposed by Saffiotti for controlling autonomous robots (Saffiotti et al. 1993).

Fuzzy Rules

TSK rules are used to control Yianni's behavior and can be seen in Table 2. It is noticeable the amount of rules required for Yianni is far less than that for Tom. This proves to be more efficient for the inference system as fewer rules have to be matched at every time step.

Table 2: Fuzzy Agent Behavior Rules

Output	*Condition*
Dodge	Projectile Distance is *near*
Flee	(Enemy Distance is *near* OR Enemy Distance is *close*) AND Confidence is *weak*
Attack	Enemy Present
Chase	(Enemy Distance is *near* OR Enemy Distance is *close*) AND Confidence is *strong*
Gather	(Item Distance is *near* OR Item Distance is *close*) AND Confidence is *weak*
Search	Got Shot AND ¬Enemy Present
Investigate	Enemy Disappear
Wander	Always [Default Behavior if none of the above Rules Fire]

Fuzzy Q-Learning Agent "Babis"

Babis is a learning agent and is built upon Yianni's fuzzy system. The learning algorithm implemented is a fuzzy Q-Learning method (Glorennec and Jouffe 1997) used by many researchers in the robotics field (Deng et al. 2003)
Using fuzzy logic for input state generalization in reinforcement learning methods has been used for a while now (Berenji 1996). While most fuzzy reinforcement learning algorithms used actor-critic architectures (Berenji et al. 1992) (Wang et al. 2007), the method implemented here has the advantage of being exploration insensitive because of the use of eligibility traces of (Sutton 1988).
Fuzzy Q-Learning is specifically designed to work with TSK FRBSs, which makes it suitable for adding learning to the existing system built for Yianni. Firstly, the TSK rule base has to be amended to allow *J* competing actions defined for every combination of the state [input] vector. Then a q value is associated with each of the competing action as follows:

$$\begin{aligned} &\text{IF x is } \mathrm{S_i} \text{ then } \mathrm{a[i,1]} \text{ with } \mathrm{q[i,1]} \\ &\qquad \text{OR } \mathrm{a[i,2]} \text{ with } \mathrm{q[i,2]} \\ &\qquad \dots \\ &\qquad \text{OR } \mathrm{a[i,J]} \text{ with } \mathrm{q[i,J]} \end{aligned}$$

Where x is the vector of input variables $\{x_1, x_2, \ldots, x_n\}$, S_i is the state vector defined "x_1 is $S_{i,1}$ AND x_2 is $S_{i,2}$ AND … AND x_n is $S_{i,n}$", $(S_{i,j})_{j=1}^{n}$ are fuzzy sets, $a_i \in A_i$, and A_i is the set of possible actions that can be chosen in state S_i.
The learning agent has to find the best consequent [action] for each rule using the quality value q. The q values are initialized to zeros at the beginning. Secondly, at each time step an exploration/exploitation policy [EEP] is used to choose an action for each rule using the Boltzmann probability assignment from equation (1).
Let i^+ be the index of the selected action in rule i using the EEP, and i^* is the index of the action having the maximum q value in rule i. Therefore the overall quality Q of the inferred global action a in state x is given by:

$$\mathrm{Q(x,a)} = \frac{\sum_{i=1}^{N} \alpha_i(x) * q[i,i^+]}{\sum_{i=1}^{N} \alpha_i(x)} \qquad (4)$$

Where N is the total number of rules, and $\alpha_i(x)$ is the inferred truth value for input vector x in rule i.
Thus, the value V for state x is given by:

$$\mathrm{V(x,a)} = \frac{\sum_{i=1}^{N} \alpha_i(x) * q[i,i^*]}{\sum_{i=1}^{N} \alpha_i(x)} \qquad (5)$$

To update the action q values, we have to calculate the error signal ΔQ. Let x be a state, a the action taken by the agent, y the new state and r the reinforcement signal received from the environment. By substituting in equation (3) we get:

$$\Delta Q = r + \gamma V(y) - Q(x,a) \qquad (6)$$

By using an ordinary gradient descent we get:

$$\Delta q[i,i^+] = \alpha * \Delta Q * \frac{\alpha_i(x)}{\sum_{i=1}^{N} \alpha_i(x)} \qquad (7)$$

Where Δq is the amount added to the q value of action i^+, and α is the learning rate.
We use the eligibility traces to speed up the learning process and make it exploration insensitive, changing equation (2) to the form:

$$e(i,j) = \begin{cases} \lambda\gamma e(i,j) + \frac{\alpha_i(x)}{\sum_{i=1}^{N} \alpha_i(x)}, & j = i^+ \\ \lambda\gamma e(i,j), & otherwise \end{cases} \qquad (8)$$

Where λ is the discount parameter of TD(λ)
After introducing eligibility traces equation (7) becomes:

$$\Delta q[i,j] = \alpha * \Delta Q * e(i,j) \qquad (9)$$

Finally, the learning algorithm steps are:

1. Collect sensor information and fuzzify them to state *x*.
2. For each rule, choose the consequent action using Boltzmann EEP.
3. Apply the global action *a* from blending all rules actions and receive reinforcement.
4. Compute the overall quality of the system *Q(x,a)* using equation (4).
5. Observe the new state *y*.
6. Compute the value of the new state *V(y)* using (5).
7. Calculate the error signal ΔQ using formula (6).
8. Update the eligibility traces for all actions using (8).
9. Use the ΔQ and the eligibility traces to update all the action *q* values using (9).

It is worth noting that the convergence of this algorithm depends on many parameters:
The number of input variables forming the state vector and the number of linguistic variables describing each input decide how many rules are specified. The temperature parameter in the Boltzmann probability equation used for EEP decides how long the exploration period is, and when to start choosing only actions with maximum q values. The discount and learning rate factors affect the speed of convergence of the learning algorithm.
It is worth noting that action *q* values can be initialized based on prior knowledge, giving some sort of bias towards certain actions in specific state input configurations. This provides more human control on the learning process but might prevent the agent from exploring new solutions.
Babis has the same fuzzy input and output variables as Yianni. However, additional rules were added to cover the remaining input space configurations. Three input variables were considered in the learning process: enemy distance, item distance, and confidence; making chase, flee, and gather the three competing actions for the bot to choose from.

TESTING AND RESULTS

The three bots were developed within the FEAR platform (Champandard 2003). The platform is developed to support AI research and teaching through building synthetic

creatures in virtual environments. Theoretically, it can interface with any game engine as a backend, yet its only implementation is with the Quake-II engine.
Bots are written in C++, compiled into dynamic link libraries [DLLs] and are defined in XML files.
The DLL file is the implementation of the control processes, using FEAR components specified in the XML files through interfaces called hooks to interact with the game engine.
The knowledge in the rule base of the implemented bots is based upon experience in the game and trial and error. In the fuzzy Q-Learning algorithm we use $\gamma = 0.9$ as the discount factor and $\alpha = 0.1$ as the learning rate.
The reward function is defined as follows:
On pain: $r = -0.1$, on item pickup: $r = +0.1$, on death: $r = -1.0$, hurting an enemy: $r = +0.1$, and killing an enemy: $r = +1.0$
To evaluate the performance of the implemented bots, a simple map was created comprising of one large room, scattered obstacles, and various power-ups. Performance evaluation is achieved through death matches between implemented bots. The difference between how many times a bot has killed the other is called the frag difference, which is used as the main performance evaluation criteria in testing. Frag difference conveys superiority of a bot's actions over the other in combat.

Tom VS Yianni

The winner from the first two implemented bots will have the chance to compete against Babis to assess whether learning has added any power to the agent's behaviors or not.
Results from ten death matches [each has a time limit of 3 minutes in *accelerated mode*] were collected and presented in Table 3.

Table 3: Death Match Results for Tom VS Yianni

#	*Tom's Kills*	*Yianni's Kills*	*Frag Difference*
1	974	**1028**	54
2	900	**1002**	102
3	832	**928**	96
4	894	**942**	48
5	833	**880**	47
6	900	**1003**	103
7	918	**1018**	100
8	947	**1050**	103
9	815	**853**	38
10	839	**890**	51

Yianni VS Babis

Babis' performance depends on the learning algorithm and the time it takes to converge to an optimum policy, which in turn depends on many parameters.
During the exploration period the agent may behave unrealistically, so the temperature parameter was chosen to show the effect of the exploration period on Babis' performance through ten death matches against Yianni. Results are plotted in Figure 5. T=0.5 yields a good exploration policy but with one loss while a lower exploration time does not find the optimal behavior policy.

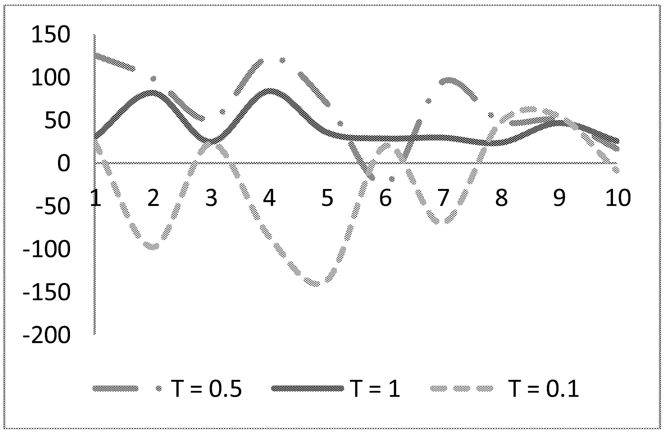


Figure 5: Temperature Effect On The Frag Difference

To speed up learning TD(λ) methods were combined with the fuzzy Q-Learning algorithm by using the eligibility traces for each rule action. Another ten death match simulations were run to show the effect of changing the λ value on Babis' performance. Results are plotted in Figure 6.

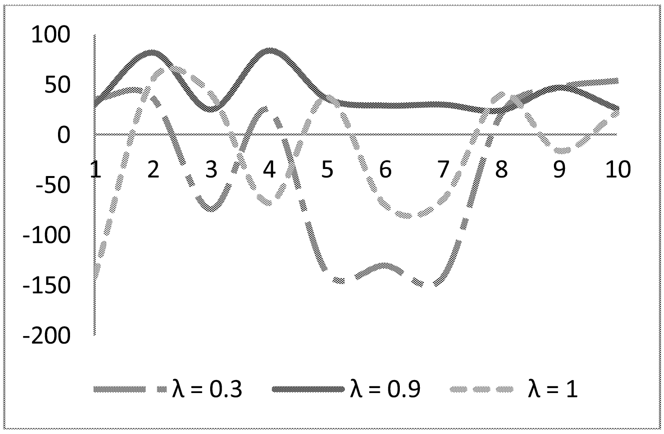


Figure 6: λ Effect On The Frag Difference

CONCLUSION AND FUTURE WORK

From the results presented and from observing the bot actions, the following points are noticed: Babis the learning bot is evidently superior to both of the scripted bots [and Champandard's bots]. The learnt modified rule base presented surprising preferences. For instance, Babis prefers chasing when confidence is weak and flees when confidence is strong; which may seem to be counterintuitive. However, after careful observation, this strategy proved to be very useful. At the beginning of a confrontation Babis is usually cautious attacking while avoiding the opponent, then after inflicting some damage [and getting hurt in the process], Babis becomes aggressive and finishes the badly wounded fleeing enemy.
An advantage of fuzzy Q-Learning is the amount of human control over the learning system. This is a desirable feature in the game AI field, as game designers do not usually wish for completely unexpected behaviors from the NPCs; they demand tools to script or at least guide the NPC in the course of the game.
As for disadvantages, although we show the proposed learning technique works well in adapting to the environment and opponents, the basic fuzzy membership functions still have to be crafted by a designer or an

experienced player, which, while being a desirable feature for game designers, is still a time consuming task. Fuzzy tuning (Alcalá et al. 2000) (Er et al. 2004) addresses this problem, where not only the consequents are learnt, but also the fuzzy membership function boundaries, and even the functions themselves. The reward function is another problem with the implementation. The reward signal is general to the agent's global action, which might be deceiving especially in the case of multiple opponents. Hierarchical reinforcement methods (Ponsen et al. 2006) separate the reward signal for every action, so the agent receives accurate feedback. A combination of the subsumption architecture (Brooks 1986) and hierarchical reinforcement learning can be an interesting experiment within game AI.

Finally, genetic algorithms can be used offline to optimize the fuzzy rule base as in (Bonarini 1996), or to form new generations of bots after each death match. This can be achieved by mating between the fittest bots. Fitness functions might be based on statistical information gathered during fights such as the number of kills, survival time, or gathering capabilities.

AKNOWLEDGEMENTS

Youssef S. G. Nashed is supported by the European Commission MIBISOC grant (Marie Curie Initial Training Network, FP7 PEOPLE-ITN-2008, GA n. 238819).

REFERENCES

Alcalá, R.; J. Casillas; O. Cordón; F. Herrera; and S. J. Zwir. 2000. "Learning and tuning fuzzy rule-based systems for linguistic modeling". *Knowledge-based Systems: Computer techniques* 3 (29), 889–941.

Berenji, H. R. 1996. "Fuzzy Q-Learning for Generalization of Reinforcement Learning". In *Proceedings of the Fifth IEEE International Conference on Fuzzy Systems*, 2208-2214.

Berenji, H. R. and P. Khedkar. 1992. "Learning and Tuning Fuzzy Logic Controllers Through Reinforcements". *IEEE Transactions on Neural Networks*, 3, 724-740.

Berry, D. A. and B. Fristedt. 1985. *Bandit Problems: Sequential Allocation of Experiments*. Prentice-Hall, NJ.

Bonarini, A. 1996. "Evolutionary learning of fuzzy rules: competition and cooperation". In *Fuzzy Modelling: Paradigms and Practice*.

Bourg, D. M. and G. Seeman. 2004. *AI for Game Developers*. O'Reilly , California.

Brooks, R. A. 1986. "A robust layered control system for a mobile robot". *IEEE Journal of Robotics and Automation*, 14-23.

Champandard, A. J. 2003. *AI Game Development: Synthetic Creatures with Learning and Reactive Behaviors*. New Riders Publishing.

Champandard, A.; T. Verweij; and R. Straatman. 2009. "The AI for Killzone 2's multiplayer bots." In *Proceedings of Game Developers Conference*, Paris, France.

Deng, C. and M. J. Er. 2003. "Real-time dynamic fuzzy Q-learning and control of mobile robots". In *2nd WSEAS International Conference on Electronics, Control and Signal Processing. Singapore*.

Er, M. J. and C. Deng. 2004. "Online Tuning of Fuzzy Inference Systems Using Dynamic Fuzzy Q-Learning". In *IEEE Transactions on Systems, Man, and Cybernetics*. 34.

Glorennec, P. Y. and L. Jouffe. 1997. "Fuzzy Q-Learning". In *Proceedings of the 6th IEEE International Conference on Fuzzy Systems*.

Gu, D.; H. Hu; and L. Spacek. 2003. "Learning Fuzzy Logic Controller for Reactive Robot Behaviours". In *Proceedings of IEEE/ASME International Conference on Advanced Intelligent Mechatron*. Kobe, Japan.

Hooley, T.; B. Hunking; M. Henry; and A. Inoue. 2004. "Generation of Emotional Behavior for Non-Player Characters: Development of EmoBot for Quake II". In *Proceedings of the 19th National Conference on Artificial Intelligence (AAAI04)*, San Jose, California, 954-955.

Isla, D. 2005. "Handling Complexity in the Halo 2 AI". In *Proceedings of the Game Developers Conference*.

Johnson, D. and J. Wiles. 2001. "Computer Games with Intelligence". *10th IEEE International Conference on Fuzzy systems*. Melbourne, Australia.

Jouffe, L. 1998. "Fuzzy Inference System Learning by Reinforcement Methods". In *IEEE Transactions On Systems, Man, And Cybernetics*.

Kaelbling, L. P.; M. L. Littman; and A. W. Moore. 1996. "Reinforcement Learning: A Survey". *Journal of Artificial Intelligence Research*, 4 , 237-285.

Laird, J. E. 2001. "It Knows What You're Going To Do: Adding Anticipation to a Quakebot". In *Proceedings of Agents 2001 conference*.

Laird, J. E.; A. Newell; and P. S. Rosenbloom. 1987. "Soar: An architecture for general intelligence". *Artificial Intelligence*, 33, 1-64.

Lin, L. J. 1992. "Self-Improving Reactive Agents Based On Reinforcement Learning, Planning and Teaching". *Machine Learning* , 8, 293-321.

Nilsson, N. J. 1998. "STRIPS Planning Systems". In *Artificial Intelligence: A New Synthesis*, 373-400

Orkin, J. 2006. "3 States and a Plan:The AI of F.E.A.R". *Game Developers Conference GDC*

Patel, P. G.; N. Carver; and S. Rahimi. 2011. "Tuning Computer Gaming Agents using Q-Learning". In *Proceedings of the Federated Conference on Computer Science and Information Systems*. 581-588

Parker, M. and B. D. Bryant. 2009. "Backpropagation without human supervision for visual control in Quake II" *Computational Intelligence and Games. CIG 2009.* 287-293.

Ponsen, M.; P. Spronck; and K. Tuyls. 2006. "Hierarchical Reinforcement Learning in Computer Games". *ALAMAS'06 Adaptive Learning Agents and Multi-Agent Systems*, 49–60. Brussels: Vrije Universiteit.

Saffiotti, A.; E. H. Ruspini; and K. Konolige. 1993. "Blending Reactivity and Goal-Directedness in a Fuzzy Controller". In *Proceedings of the Second IEEE Conference on Fuzzy Systems*, San Francisco, CA, 134-139.

Sutton, R. S. 1988. "Learning to Predict by the Methods of Temporal Differences". *Machine learning*, 3, 9-44.

Sutton, R. S. 1996. "Generalization in reinforcement learning: Successful examples using sparse coarse coding". *Advances in Neural Information*, 8.

Wang, X. S.; Y. H. Cheng; and J. Q. Yi. 2007. "A fuzzy Actor–Critic reinforcement learning network". *Information Sciences*, 177, 3764–3781.

Watkins, C. J. and P. Dayan. 1992. "Q-learning". *Machine Learning*, 8 (3), 279-292.

Westra, J. and F. Dignum. 2009. "Evolutionary Neural Networks for Non-Player Characters in Quake III". *Computational Intelligence and Games. CIG 2009,* 302-309.

Zadeh, L. A. 1965. "Fuzzy sets". *Information and Control*, 8, 338-353.

Zanetti, S. and A. El Rhalibi. 2004. "Machine learning techniques for FPS in Q3". In *Proceedings of the 2004 ACM SIGCHI International Conference on Advances in computer entertainment technology*, Singapore, 239-244

AUTHOR LISTING

AUTHOR LISTING