

**1<sup>ST</sup> INTERNATIONAL  
NORTH-AMERICAN  
SIMULATION TECHNOLOGY CONFERENCE**

**NASTEC'2008**

**EDITED BY**

**Mokhtar Beldjehem**

**Lotfi A. Zadeh**

**Ronald Yager**

**and**

**Madan Gupta**

**AUGUST 13-15, 2008**

**McGILL UNIVERSITY  
MONTREAL, CANADA**

**A Publication of EUROSIS-ETI**

**Printed in Ghent, Belgium**



1<sup>ST</sup> International North-American Conference  
on  
Simulation Technology

MONTREAL, CANADA

AUGUST 13-15, 2008

Organized by  
ETI

And

École Polytechnique de Montreal

Sponsored by  
EUROSIS

Co-Sponsored by

**McGill University**

**Ghent University**

Hosted by  
**McGill University**  
**Montreal, Canada**

## EXECUTIVE EDITOR

**PHILIPPE GERIL  
(BELGIUM)**

## EDITORS

### General Conference Chair

Mokhtar Beldjehem, École Polytechnique de Montréal, Montréal, Canada

### Honorary Chairs

Professor Lotfi A. Zadeh, Berkeley University, CA, USA

Ronald Yager, Iona College, New Rochelle, USA

Madan M. Gupta, University of Saskatchewan, Saskatoon, Saskatchewan, Canada

### Local Programme Committee

Marek Balazinski, Ecole Polytechnique de Montreal, Montréal, Canada

Peter Grogono, Concordia University, Montreal, Canada

Hakim Lounis, UQAM, Montreal, Canada

Houari Sahraoui, University of Montréal, Montréal, Canada

Nematollaah Shiri, Concordia University, Montreal, Canada

### International Programme Committee

Ajith Abraham, Norwegian University of Science and Technology, Norway

Hojjat Adelli, Ohio State University, Columbus, OH, USA

Esma Aimeur, University of Montreal, Montreal, Canada

Troels Andreasen, Roskilde University, Roskilde, Denmark

Riad Assied, Petra University, Amman, Jordan

Bilal M. Ayyub, University of Maryland College Park, MD, USA

Mourad Badri, University of Quebec trois-Rivières, Canada

Linda Badri, University of Quebec trois-Rivières, Canada

Ildar Batyrshin, Kazan State Technological University, Kazan (Tatarstan), Russia

Nabil Belacel, National Research Council, New Brunswick, Canada

Mohamed Bettaz, INI/MESRS, Algiers, Algeria

Prabir Bhattacharya, Concordia University, Montreal, Canada

Loredana Biacino, Università degli Studi di Salerno, Salerno, Italy

Ranjit Biswas, Institute of Technology & Management, Gurgaon, India

Ulrich Bodenhofer, Johannes Kepler University, Linz, Austria

Piero P. Bonissone, General Electric, USA

Gloria Bordogna, Istituto per le Tecnologie Informatiche Multimediali, Milano, Italy

Patrick Bosc, ENSSAT, University of Rennes, France

Boubaker Boufama, University of Windsor, Windsor, Canada

Mounir Boukadoum, UQAM, Montreal, Canada

Ivan Bruha, McMaster University, Hamilton, Ont., Canada

Bill P. Buckles, University of North Texas, USA

Joao P. Carvalho, INESC-ID, Lisboa university, Lisboa, Portugal

Guanrong (Ron) Chen, City University of Hong Kong, China

Alain Colmerauer, University of Aix-Marseille II, Marseille, France

Michel Dagenais, École Polytechnique de Montréal, Montréal, Canada

Mourad Debbabi, Concordia University, Montreal, Canada

Scot Dick, University of Alberta, Canada

Talbi El-Ghazali, Université des Sciences et Technologies de Lille, Lille, France

Jinan Fiaidhi, Lakehead University, Canada

Christian Freksa, University of Bremen, Germany

Claude Frasson, Université de Montréal, Montreal, Canada



## International Programme Committee

Gabriel Gerard, Université de Sherbrooke, Sherbrooke, Canada  
Gianggiacomo Gerla, Università degli Studi di Salerno, Salerno, Italy  
Robert Godin, UQAM, Montréal, Canada  
Madan M. Gupta, University of Saskatchewan, Canada  
Abdelwahab Hamou-Lhadj, Concordia University, Montreal, Canada  
Sami Harari, University of Toulon and the Var, Toulon, France  
Yutaka Hata, University of Hyogo, Japan  
Eyke Hüllermeier, Philipps-Universität Marburg, Marburg, Germany  
Ahmed Ibrahim, RCC Institute of Technology, Concord(Toronto), Canada  
Enso Ikonen, University of Oulu, Finland  
Yao JingTao University of Regina, Sask., Canada  
Janusz Kacprzyk, SRI, Polish Academy of Sciences, Warsaw, Poland  
Okay Kaynak, Bogazici University, Bebek(Istanbul), Turkey  
James M. Keller, University of Missouri, USA  
Bettina Kemme, McGill University, Montreal, Canada  
Etienne E. Kerre, Ghent University, Ghent, Belgium  
Taghi M. Khoshgoftaar, Florida Atlantic University, USA  
Frank Klawonn, University of Applied Sciences Braunschweig/Wolfenbuettel, Wolfenbuettel, Germany  
Erich Peter Klement, Softwarepark Hagenberg, Hagenberg, Austria  
Amit Konar, Jadavpur University, Kolkata, India  
Donald H. Kraft, Louisiana State University, USA  
Vladik Kreinovich, The University of Texas at El Paso, El Paso, Texas, USA  
H. K. Kwan, University of Windsor, Canada  
Guy Lapalme, Université de Montréal, Montreal, Canada  
Frank L. Lewis, University of Texas at Arlington, Worth(TX), USA  
Pawan Lingras, St. Mary University, Hamilton, NS, Canada  
Edwin Lughofer, Softwarepark Hagenberg, Hagenberg, Austria  
Mourad Maouche, University of Philadelphia, Amman, Jordan  
Trevor Martin, University of Bristol, United Kingdom  
Alexander Mehler, University of Bielefeld, Germany  
Jean Meunier, Université de Montréal, Montréal, Canada  
Ali Mili, New Jersey institute of technology, USA  
Guy Mineau, Université de Laval, Laval, Canada  
Zelmat Mimoun, University of M'hamed Bougara - Boumerdès, Algeria  
Gautam Mitra, Brunel University, UK  
Sabah Mohammed, Lakehead University, Canada  
Malek Mouhoub, University of Regina, Sask., Canada  
Sudhir P. Mudur, Concordia University, Montreal, Canada  
Mike Nachtgael, Ghent University, Ghent, Belgium  
Nadia Nedjah, State University of Rio de Janeiro, Rio de Janeiro, Brazil  
Jian-Yun Nie, Université de Montréal, Montréal, Canada  
Abdellatif Obaid, UQAM, Montreal, Canada  
Fakhreddine O. Karray, University of Waterloo, Canada  
Sankar Kumar Pal, Indian Statistical Institute, Kolkata, India  
Costas P. Pappis, University of Piraeus, Piraeus, Greece  
Gabiella Pasi, Istituto per le Tecnologie Informatiche Multimediali, Milano, Italy  
Frederick E. Petry, Naval Research Laboratory, MS, USA  
Hélène Pigot, Université de Sherbrooke, Sherbrooke, Canada  
Bhanu Prasad, Florida A&M University, USA  
Witold Pedrycz, University of Edmonton, Edmonton, Alberta, Canada  
James F. Peters III, University of Manitoba, Canada  
Henri Prade, University of Paul Sabatier, Toulouse, France  
Shahram Rahimi, Southern Illinois University, Carbondale, Il., USA  
Djamal Rebaïne, UQAC, Chicoutoumi, Canada  
Marek Reformat, University of Edmonton, Edmonton, Alberta, Canada  
Burghard B. Rieger, University of Trier, Trier, Germany  
Stuart H. Rubin, Space and Naval Warfare Systems Center, USA  
Imre J. Rudas, Budapest Technical University, Budapest, Hungary  
Aziz Salah, UQAM, Montréal, Canada,

## International Programme Committee

Antoaneta Serguieva, Brunel University, West London, United Kingdom  
Pierre Siegel, University of Aix-Marseille I, Marseille, France  
Constantinos I. Siettos, National Technical University of Athens, Athens, Greece  
James F. Smith, III, Naval Research Laboratory, Washington, DC, USA  
Roman Slowinski, Poznan' University of Technology, Poznan, Poland  
Dutta Sumitra, INSEAD, Fontainebleu, France  
M.N.S. Swamy, Concordia University, Montreal, Canada  
Hamid R. Tizhoosh, University of Waterloo, Canada  
Jose A. B. Tome ,INESC-ID, Lisboa University, Lisboa, Portugal  
Enric Trillas, European Centre for Soft Computing, Mieres(Asturias), Spain  
Edward Tsang, University of Essex, United Kingdom  
Hans Vangheluwe, McGill University, Montreal, Canada  
Athanasios Vasilakos, University of Western, Macedonia, Greece  
Ronald R. Yager, IONA College, New Rochelle, N.Y., USA  
Takeshi Yamakawa, Kyushu Institute of Technology, Kyushu, Japan  
Mustapha Yassine, National University of Amman, Amman, Jordan  
Ting Yu, University of Sydney, Sydney, Australia  
Sahnoun Zaidi, Université de Constantine, Algeria

# **NASTEC 2008**

© 2008 EUROSIS-ETI

Responsibility for the accuracy of all statements in each peer-referenced paper rests solely with the author(s). Statements are not necessarily representative of nor endorsed by the European Simulation Society. Permission is granted to photocopy portions of the publication for personal use and for the use of students providing credit is given to the conference and publication. Permission does not extend to other types of reproduction or to copying for incorporation into commercial advertising nor for any other profit-making purpose. Other publications are encouraged to include 300- to 500-word abstracts or excerpts from any paper contained in this book, provided credits are given to the author and the conference.

All author contact information provided in this Proceedings falls under the European Privacy Law and may not be used in any form, written or electronic, without the written permission of the author and the publisher.

All articles published in these Proceedings have been peer reviewed

EUROSIS-ETI Publications are ISI-Thomson and INSPEC referenced

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  
Final Cover Design by Grafisch Bedrijf Lammaing, Ostend, Belgium

EUROSIS-ETI Publication

**ISBN: 978-90-77381-00-7**

**EAN: 978-90-77381-00-7**

# Preface

**NASTEC** (North-American Simulation Technology Conference) is a series of conferences initiated by Eurosis after in-depth discussions with Dr. Mokhtar Beldjehem and North-American simulationists, addressing issues regarding modeling and simulation (M&S). The first **NASTEC 2008** is being held at Mc Gill University, Montreal, Canada, which is its birth place. It has attracted simulationists, researchers and practitioners, attendees from academic, industry and government agencies in an exchange of ideas and shared experiences. NASTEC aims to be the feast of simulationists in North-America.

The intent of the **NASTEC'2008** event is to nurture the spirit of cooperation and strive to improve the quality of life in this global village through excellence in hybrid soft computing research and education by engineering of next-generation intelligent hybrid soft computing systems for modeling, simulation, software engineering, web computing and virtual reality systems at the service and for the benefits of the humankind.

Computer simulation is being acknowledged as the “third leg” of scientific discovery and analysis, along with theory and experimentation. Simulation technology aims at building the software digital factory. The field of modeling and simulation in general has made significant progress; part of it is reflected in the present proceedings volume. **NASTEC 2008** was able to attract top-level and forefront research; the field itself has brought along a number of new development, unheard of a couple of years ago. The themes to be discussed this year center around novel issues in connection with modeling and simulation: soft computing for modeling and simulation, simulation-based software engineering, web computing and virtual reality systems. The program consists of 15 high-quality papers. Beyond these papers that have undergone a review process, NASTEC 2008 is proud to host three abstracts by Prof. Lotfi A. Zadeh the creator of fuzzy and soft computing, invited talks by Prof. Johann Schumann from NASA Intelligent Systems Division, Prof. JingTao Yao from the University of Regina, Prof. Peter Grogono from Concordia University, and Prof. Brigitte Jumard from Concordia University.

We are grateful to a number of people without whom we would not have been able to put the program together. They include our local program committee and international program committee, which have done an excellent job: We got 4.5 reviews per paper on the average. We would also like to thank many external reviewers who have helped “in the background,” and who made sure that we stuck with our schedule. We are grateful to the large number of authors who have considered NASTEC as the target for their work, and even though we could not accommodate every submission, we hope that the reviews will be helpful to many people. Last, but not least, we are indebted to the staff of Eurosis, École Polytechnique de Montréal and Mc Gill University for making this event a reality.

NASTEC'08 General Conference Chair  
Mokhtar Beldjehem  
Honorary Conference Chairs  
Lotfi A. Zadeh  
Ronald Yager  
Madan Gupta



## CONTENTS

|                                   |            |
|-----------------------------------|------------|
| <b>Preface .....</b>              | <b>IX</b>  |
| <b>Scientific Programme .....</b> | <b>1</b>   |
| <b>Author Listing.....</b>        | <b>119</b> |

### INVITED PRESENTATIONS

|  |          |
|--|----------|
| <b>Granular Computing: A new Paradigm in Information Processing</b><br>JingTaoYao .....  | <b>5</b> |
| <b>Verification and Validation of Neuro-adaptive Aircraft Control Systems</b><br>Johann Schumann, Yan Liu and Pramod Gupta ..... | <b>7</b> |
| <b>Glorious Accidents or Expected Results?</b><br>Peter Grogono.....   | <b>9</b> |

### SOFT COMPUTING THEORY AND PRACTICE

|   |           |
|---|-----------|
| <b>Toward Human Level Machine Intelligence - Is it Achievable? The Need<br/>for a Paradigm Shift</b><br>Lotfi A. Zadeh .....  | <b>13</b> |
| <b>Modeling Numerical and Spatial Uncertainty in Grayscale Image Capture<br/>using Fuzzy Set Theory</b><br>M. Nachtgael, P. Sussner, T. Mélange and E.E. Kerre .....                            | <b>15</b> |
| <b>A Granular Unified Min-Max Fuzzy-Neuro Framework for Learning Fuzzy<br/>Systems</b><br>Mokhtar Beldjehem .....   | <b>23</b> |
| <b>Prediction of Ferrous Bio-oxidation Rate in a Packed Bed Bioreactor<br/>using Artificial Neural Network</b><br>Hasan Yousefi, S. Mohammad Mousavi, Arezou Jafari and Azita Soleymani ..      | <b>31</b> |
| <b>Extension of Rank Based Ant System with Exponential Pheromone<br/>Deposition for Speed-up and Improved Accuracy</b><br>Ayan Acharya, Aritra Banerjee, Amit Konar and Mokhtar Beldjehem ..... | <b>37</b> |

### SIMULATION BASED SOFTWARE ENGINEERING

|   |           |
|---|-----------|
| <b>Computation with Imprecise Probabilities</b><br>Lotfi A. Zadeh ..... | <b>45</b> |
|---|-----------|

## CONTENTS

### **Monte Carlo Validation of Model Stability**

Pierre N. Robillard and Simon Labelle..... 47

### **Standard Error Estimation for EM Applications related to Latent Class Models**

Liberato Camilleri ..... 52

### **A Petri net framework for the modeling, simulation and data analysis of biological models**

Simon Hardy and Pierre N. Robillard..... 57

### **A Granular Unified Framework for a Machine Visual System**

Mokhtar Beldjehem ..... 63

## **WEB COMPUTING AND COGNITIVE SIMULATION FOR PROBLEM SOLVING**

### **A New Frontier in Computation: Computation with Information described in Natural Language**

Lotfi A. Zadeh..... 73

### **Simulation of a Human Machine Interaction: Locate Objects Using a Contextual Assistant**

Chikhaoui Belkacem and Pigot H       ..... 75

### **A Granular Framework for Recognition of Arabic Handwriting: The GOAVMREC System**

Mokhtar Beldjehem ..... 81

### **Behavior Based Predictive Motion Controller for a Mobile Robot**

Krzysztof Skrzypczyk, Krzysztof F           and Adam Galuszka ..... 89

### **Cognitive Modeling of a Cooking Activity: Integration of the Contention Scheduling Theory in the Cognitive Architecture ACT-R**

Pierre-Yves Groussard and H       Pigot..... 94

### **Developing an Ontology Extraction Agent for a Biomedical Learning Social Network**

S. Mohammed, J. Fiaidhi and O. Mohammed ..... 99

## **INDUSTRIAL APPLICATIONS**

### **Moving Containers in Small Terminal as Strips Planning Problem- Preliminary Results**

Adam Galuszka and Krzysztof Skrzypczyk ..... 107



## CONTENTS

|  |            |
|--|------------|
| <b>Introduction to Comparison of Traditional and Virtual Patterns Design in 3D</b> |            |
| Agnieszka Cichocka and Pascal Bruniaux .....                                       | <b>110</b> |



## Message from the Chairs

The organizing committee welcomes you to the 2008 **North-American Simulation Technology Conference (NASTEC 2008)**, held in Montreal, Canada, from August the 13th to August the 15th, 2008. This meeting, co-sponsored by the EUROSIS, is the first of a series of conferences dealing with simulation technology.

The intent of the **NASTEC'2008** event is to nurture the spirit of cooperation and strive to improve the quality of life in this global village through excellence in hybrid soft computing research and education by engineering of next-generation intelligent hybrid soft computing systems for modeling, simulation, software engineering, web computing and virtual reality systems at the service and for the benefits of the humankind.

Computer simulation is being acknowledged as the “third leg” of scientific discovery and analysis, along with theory and experimentation. Ultimately, simulation technology aims at building the cost-effective software *digital factory*.

The three-day program aims to extend and advance the use of modeling and simulation (M&S) technologies in an informal setting arranged to encourage broad discussion about theory, methodologies, best practices and results. Participants will hear, learn and discuss opportunities and problems in using soft computing, simulation-based software engineering, web computing, virtual reality, their synergies and interplays in connection with modeling and simulation breakthroughs to the advancement of the simulation technology and related applications. To promote interaction and discussion in the audience, sufficient time is allotted to presenters not only to introduce their achievements, but also to engage in extended discussions with the participants. Subjects of discussion include, but are not restricted to, examination of approaches and results, the rationale underlying particular methodologies, experimental and theoretical examinations, practical difficulties, insights, and extensions to other application areas.

We believe that **NASTEC 2008** constitutes a seed for the upcoming **NASTEC** series: the quality of accepted papers is still very high, evidencing the real interest and attractiveness of this meeting and the relevance of this scientific and application area in the worldwide scene. This conference will be held in parallel with the North American Simulation and AI in Games Conference (GAMEON-NA). The organizers decided to adopt this structure due to their many common aspects and shared technologies: the parallel organization will allow for more interaction, networking and collaboration among the participants in the two events, and for cross-fertilization of research ideas, more sharing of advanced knowledge and stimulating experiences. We think you will find **NASTEC 2008** a challenging and productive experience.

We hope that you will enjoy the feast of simulationists, the Montreal UNESCO City of Design, the Island of Montreal, the Mc Gill University location, the culture, the food and the Mount Royal: Montreal will be an exciting experience!

NASTEC'08 General Conference Chair  
Mokhtar Beldjehem  
Honorary Conference Chairs  
Lotfi A. Zadeh  
Ronald Yager  
Madan Gupta



# **SCIENTIFIC PROGRAMME**



# **INVITED PRESENTATIONS**





# GRANULAR COMPUTING: A NEW PARADIGM IN INFORMATION PROCESSING

JingTao Yao

Department of Computer Science, University of Regina, Regina, Canada S4S 0A2

Email: [jtyao@cs.uregina.ca](mailto:jtyao@cs.uregina.ca)

URL: <http://www2.cs.uregina.ca/~jtyao>

## INTRODUCTION

Granular computing (GrC) has emerged as one of the fastest growing information processing paradigm in computational intelligence and human-centric systems. It has been gaining popularity in the past ten years. GrC is often loosely defined as an umbrella term to cover any theories, methodologies, techniques, and tools that make use of granules in complex problem solving. As a new paradigm for the problem solving, GrC may be viewed differently from philosophical, methodological and application perspectives. In this talk, I will give a brief introduction to granular computing and discuss its present developments and research directions.

## HISTORICAL VIEW OF GRANULAR COMPUTING

The concept of granular computing was initially called information granularity or information granulation related to the research of fuzzy sets in Zadeh's early paper (Zadeh 1979). The term granular computing first appeared within literature in 1997 (Yao 2007).

Although the term granular computing is new, the basic notions and principles of GrC occur under various forms in many disciplines and fields (Yao 2004, Zadeh 1997). Similar views are shared by research in belief functions, artificial intelligence, cluster analysis, chunking, data compression, databases, decision trees, divide and conquer, fuzzy logic, interval computing, machine learning, structure programming, quantization, quotient space theory, and rough set theory.

## GRANULES, GRANULATION AND GRANULAR RELATIONSHIPS

Granules, granulations and relationships are some of the key issues in the study of GrC (Yao 2005).

A granule can be defined as any subset, class, object, or cluster of a universe. These granules are composed of finer granules that are drawn together by distinguishability, similarity, and functionality (Zadeh 1997).

A group of concepts or objects can be considered as a granule by their spatial neighborhood, closeness, and cohesion. Although granular computing is intended to deal with imprecision, uncertainty and partial truth, the granules may be of crisp or fuzzy format. A granule may have different formats and meanings when used in a particular model. For example, in a set-theoretic setting, such as rough sets and cluster analysis, a granule may be interpreted as a subset of a universal set, while in structured programming, a granule can be a program module (Yao 2004). Granules at the lowest level are composed of elements or basic particles of the particular model that is used. For instance, the finest granules are words in an article universe. They are formed with basic particles, i.e., letters. They may be considered as singleton granules in some special cases.

Granulation involves the process of construction and decomposition of granules (Yao 2005). It is an operation performed on granules. Construction involves the process of forming a larger and higher level granule with smaller and lower level granules that share similarity, indistinguishability, and functionality. Decomposition is the process of dividing a larger granule into smaller and lower level granules. The former is a bottom-up process. The latter is a top-down process. This definition is slightly different with the dictionary definitions of granulation, the act or process of forming something into granules, i.e. decomposition of granulation. Writing an article could be viewed as granulation. The lower level granules, i.e., words, are constructed into an article, a high-level granule. Granulation and computation are two important and related issues of granular computing research. Granulation deals with the construction, interpretation, and representation of granules. Computation deals with the computing and reasoning with granules and granular structures.

Relationships amongst granules may be classified into two types, interrelationship and intrarerelationship. Granulation, regardless of direction, is dealing with relationships between granules. The relationship involved in construction granulation is considered as an interrelationship and the decomposition granulation as an in-

trarelationship. Interrelationship is the basis of grouping small objects together. Granular computing involves structured human thinking. A high-level granule represents a more abstract concept and a low-level a more specific concept. The level of abstraction may be represented in terms of coarse and fine relationships.

## SCHOOLS OF GRANULAR COMPUTING RESEARCH

One of the important developments of granular computing is the triarchic theory of granular computing (Yao 2008). Instead of simply defining what granular computing research is, one may understand the scope of granular computing from the philosophical, methodological and computational perspectives. The philosophical perspective concerns structured thinking. Granular computing combines analytical thinking for decomposing a whole into parts and synthetic thinking for integrating parts into a whole. It is important to consider the conscious effects in thinking with hierarchical structures when using granular computing. The methodological perspective concerns structured problem solving. The techniques for effective human problem-solving, such as systematic approaches to finding a solution, effective problem definition principles, and practical heuristics and strategies to check solutions to a problem, builds major foundations to granular computing. The computational perspective concerns structured information processing. Granular computing also focuses on the application of its theory to knowledge-intensive systems.

Rough sets and fuzzy sets play important roles in GrC development (Yao 2007). An important fuzzy aspect in granular computing is to view granular computing as human-centric intelligent systems. Human-centered information processing was initiated with the introduction of fuzzy sets. The insights have led to the development of the granular computing paradigm (Bargiela and Pedrycz 2008, Zadeh 1997). Shifting from machine-centered approaches to human-centered approaches is considered one of the trends in GrC research. For example, one may integrate different agents in which each pursues its own agenda, exploits its environment, develops its own problem solving strategy and establishes required communication strategies, to form a more effective human-centered information system (Bargiela and Pedrycz 2008).

Another school of thought is rough-granular computing. One may form granules with different criteria from a rough computing point of view (Skowron and Stepaniuk 2007). Granules are constructed in computations aiming at solving optimization tasks. General optimization criterion based on the minimal length principle may be used. In searching for optimal solutions, it

is necessary to construct many compound granules using some specific operations, such as generalization, specification or fusion (Skowron and Stepaniuk 2007). The dominance-based rough set approach is another representation of rough set-based GrC methodology. This approach extends the classical rough set approach by utilizing background knowledge about ordinal evaluations of objects and about monotonic relationships between these evaluations (Slowinski 2008).

Other important granular computing research areas include interval computing, topology, rough logic, quotient space, neural networks, fractal analysis, and quotient space theory (Yao 2009).

## FUTURE OF GRANULAR COMPUTING

In order to broaden and deepen the study of granular computing, one may focus on its foundations and definitions. Important issues, such as the formalization and understanding of granules, granulation, and granular relationships of various granular computing techniques should be emphasized. Applying individual techniques for real applications are essential. Communicating with other disciplines and adopting non-traditional techniques to granular computing research will broaden, enhance, and solidify granular computing research.

## REFERENCES

- A. Bargiela, W. Pedrycz (2008). Toward a theory of granular computing for human-centred information processing. *IEEE Transactions on Fuzzy Systems*. **16**(2), 320-330.
- A. Skowron, J. Stepaniuk (2007). Modeling of high quality granules. In Proc. of the International Conference on Rough Sets and Intelligent Systems Paradigms, Warsaw, Poland, June 28-30, 2007. LNCS **4585**, 300-309.
- R. Slowinski (2008). Dominance-based rough set approach to reasoning about ordinal data: a tutorial. In Proc. of the 3rd International Conference on Rough Sets and Knowledge Technology, Chengdu, China, May 17-19, 2008. LNCS **5009**, 21-22.
- J.T. Yao (2005). Information granulation and granular relationships. In Proc. of the IEEE Conference on Granular Computing, Beijing, China, 326-329.
- J.T. Yao (2007). A ten-year review of granular computing. In Proc. of the IEEE International Conference on Granular Computing. San Jose, USA, 2-4 November 2007, 734-739.
- J.T. Yao (2009). *Novel Developments in Granular Computing: Applications for Advanced Human Reasoning and Soft Computation*. IGI Global.
- Y.Y. Yao (2004). Granular computing. *Computer Science*, **31**(10.A), 1-5.
- Y.Y. Yao (2008). Granular computing: past, present, and future. In Proc. of the 3rd International Conference on Rough Sets and Knowledge Technology, Chengdu, China, May 17-19, 2008, LNCS **5009**, 27-28.
- L.A. Zadeh (1979). Fuzzy sets and information granularity. In *Advances in Fuzzy Set Theory and Applications*. M. Gupta, R.K. Ragade, R.R. Yager (eds). North-Holland Publishing Company. 3-18.
- L.A. Zadeh (1997). Towards a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. *Fuzzy Sets and Systems*. **90**(2), 111-127.

# Verification and Validation of Neuro-adaptive Aircraft Control Systems

Johann Schumann  
RIACS/NASA Ames  
Johann.M.Schumann@nasa.gov

Yan Liu  
Motorola Labs  
yanliu@motorola.com

Pramod Gupta  
UARC/NASA Ames  
Pramod.Gupta@nasa.gov

## KEYWORDS

Neural Networks, Verification, Validation, Control

## ABSTRACT

Traditional fixed-gain control has proven to be unsuccessful to deal with complex changing systems such as a damaged aircraft. Control systems, which use a neural network that can adapt toward changes in the plant, have been actively investigated and test flown as they offer many advantages. We will briefly introduce adaptive flight control and will discuss the specific challenges for the verification and validation (V&V) of such systems. Since performance and safety guarantees cannot be provided at development time, we have developed novel tools and approaches to support V&V and certification, which use a Bayesian approach to monitor sensitivity and performance (confidence) of the neural network during flight.

## INTRODUCTION

Adaptive control systems in aerospace applications have numerous advantages: they can automatically fine-tune system identification and accommodate for slow degradation and catastrophic failures (e.g., a damaged wing or a stuck rudder) alike. A variety of approaches for adaptive controls, based upon self-learning computational models such as neural networks or fuzzy logic, have been developed (e.g., Rysdyk and Calise (1998)). Some are in actual use (e.g., in chemical industry) or have been flight-tested (e.g., the NASA Intelligent Flight Control System (IFCS, Bosworth and Williams-Hayes (2007))). However, the acceptance of adaptive controllers in aircraft and other safety-critical domains is significantly challenged by the fact that methods and tools for analysis and verification of such systems are still in their infancy and no widely accepted V&V approach has been developed. Reliability of learning, performance of convergence and prediction for a nonlinear adaptive controller is hard to guarantee. The analysis of traditional controllers, which have been augmented by adaptive components require technically deep nonlinear analysis methods.

Figure 1 shows the basic architecture of the adaptive controllers developed within NASA's IFCS project: pilot stick commands  $\theta_{cmd}$  are mixed with the feedback,

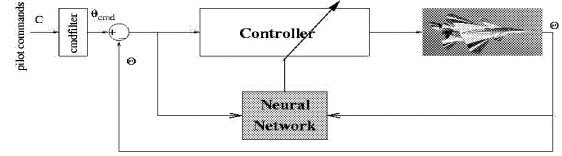


Figure 1: Basic IFCS Adaptive Control Architecture

current sensor readings  $\theta$  (e.g., airspeed, angle of attack, altitude) to form the desired behavior of the aircraft. Then, the controller calculates the necessary movements of the control surfaces (e.g., rudder, ailerons). If the aerodynamics of the aircraft changes radically (e.g., due to a structural damage), there is a deviation between desired and actual behavior of the aircraft. The neural network is trained during flight (“online”) to minimize this deviation. Different types of NNs (DCS, SigmaPi, and MLP) have been investigated within this project.

## V&V AND CERTIFICATION ISSUES

Clearly, an adaptive aircraft controller is a highly safety-critical component of aviation software, and therefore, it has to undergo a rigorous V&V and certification process. Due to the nonlinearity of adaptive controllers, traditional linear analysis techniques and tools cannot be used. Rather, more complex non-linear techniques like Lyapunov stability analysis must be used. In general, adaptive controllers require advanced learning algorithms, which dynamically modify internal parameters (e.g., weights). For such algorithms, no standardized way of performing performance analysis and V&V exists and certification authorities are very reluctant to certify novel components, architectures, and software algorithms.

For such learning algorithms, in general, the convergence time cannot be bounded a priori and there is no guarantee that the global optimum can be reached. The estimation of safety and stability envelopes is strongly related with the performance of the neural network. We therefore have developed a number of tools, which dynamically (i.e., during flight) monitor the performance and sensitivity of the neural network. Using a Bayesian approach, these tools can provide statistical up-to-date evidence on how the neural network is behaving. We also have developed software V&V process guides to support V&V of adaptive control systems.

## PARAMETER SENSITIVITY ANALYSIS

The sensitivity of a controller with respect to input perturbations is an important performance metric for any controller. In a neuro-adaptive system, the internal control parameters are changing while the system is in operation. We are therefore also interested in the *parameter sensitivity* for the neural network. A statistical formulation this provides sensitivity  $s$  and parameter confidence  $\sigma_p^2$ . If we assume a Gaussian probability distributions and the probability of the output of the neural network as  $p(\mathbf{o}|\mathcal{P}, \mathbf{x})$  for output  $\mathbf{o}$  with network parameters  $\mathcal{P}$  and inputs  $\mathbf{x}$ , we can easily calculate sensitivity and parameter confidence for each parameter  $\mathcal{P}$ .

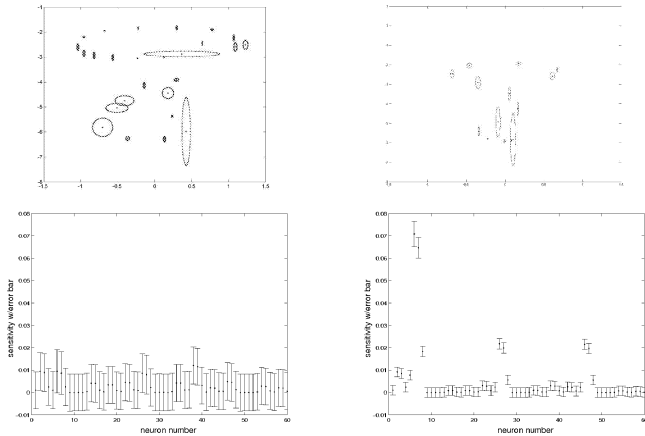


Figure 2: Parameter sensitivity and confidence for DCS (top) and Sigma-Pi (bottom) before and after training.

Fig. 2 (top) shows the sensitivity of the IFCS DCS reference vectors before (l) and after training (r). Small circles correspond to high parameter confidence. The bottom row shows the mean sensitivity and parameter confidence (as error bars) for each of the 60 weights in the Sigma-Pi IFCS network Schumann and Liu (2007). Before training, all weights have similar sensitivity; after training, however, only 7 weights have consistently high sensitivity, i.e., their value really contributes to the output. This observation provides statistical evidence that a dramatic reduction of the network size from 60 to 7 neurons was justified.

## NETWORK CONFIDENCE

The *Confidence Tool* Gupta and Schumann (2004) produces a quality measure (confidence,  $\sigma^2$ ) of the NN output using a Bayesian approach. This tool has been developed for the IFCS Sigma-Pi adaptive controller and successful test flights on a NASA F-15 aircraft have been carried out in early 2006. A similar performance metric (validity index) has been defined for DCS Liu et al. (2005). Figure 3(top) shows the control augmentation signal that the NN produces to compensate for failure (stuck stabilator surface)..  $\sigma^2$  of the network output increases substantially, indicating a large uncertainty in

the network output. Due to the online training of the network, this uncertainty decreases very quickly. A second and third pilot command ( $t = 11s$ ,  $t = 17s$ ) shows that the network has successfully adapted to handle this failure situation (much smaller peaks in  $\sigma^2$ ).

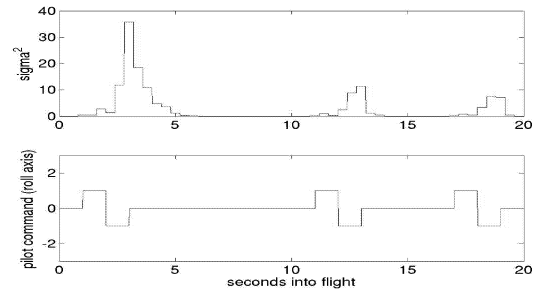


Figure 3: Confidence value  $\sigma^2$  over time (top) and pilot doublet commands (bottom). Failure at  $t = 1.5s$ .

## CONCLUSIONS

Our Bayesian approach allows different models (e.g. networks with different numbers of hidden units, or different network types such as multi-layer perceptrons, Sigma-Pi, RBF, or DCS) to be compared using only the training data. More generally, the Bayesian approach provides an objective and principled framework for dealing with the issues of model complexity.

In aeronautics, the performance of an aircraft is defined in terms of its handling quality (e.g., the Cooper-Harper rating). Current research aims to relate our performance metric with the aircraft handling quality. With the real-time availability of handling quality estimates, our validation tools can be used to alert the pilot and provide assistance/support to decision making.

## REFERENCES

- Bosworth J. and Williams-Hayes P., 2007. *Flight Test Results from the NF-15B Intelligent Flight Control System (IFCS) Project with Adaptation to a Simulated Stabilator Failure*. AIAA 2007-2818.
- Gupta P. and Schumann J., 2004. *A Tool for V&V of Neural Network Based Adaptive Controllers for High Assurance Systems*. In *Proc. HASE*. IEEE.
- Liu Y.; Cukic B.; Jiang M.; and Xu Z., 2005. *Predicting with Confidence - An Improved Dynamic Cell Structure*. In *Advances in Neural Computation*, Springer.
- Rysdyk R. and Calise A., 1998. *Fault tolerant Flight Control via Adaptive Neural Network Augmentation*. AIAA-98-4483, 1722-1728.
- Schumann J. and Liu Y., 2007. *Tools and Methods for the Verification and Validation of Adaptive Aircraft Control Systems*. In *IEEE Aerospace*.

# GLORIOUS ACCIDENTS OR EXPECTED RESULTS?

Peter Grogono

Department of Computer Science and Software Engineering  
Concordia University, Montréal, QC  
grogono@cse.concordia.ca

## ABSTRACT

Evolutionary Computing and Artificial Life apply insights from evolutionary biology to software applications. Our understanding of evolution is itself evolving. Viewed in the light of modern evolutionary theories, the assumptions of Evolutionary Computing and Artificial Life are often naive and sometimes even wrong. How does evolution really work? Can we exploit recent discoveries to expand the possibilities of software? Can software be truly creative?

## BIOLOGY AND COMPUTATION

The idea of developing computer algorithms based on biological evolution has excited researchers since the early days of computing. The first proposals were made in the 1950s, and several streams of evolutionary computing (EC) emerged in the 1960s. Yet it is only in the last few years that EC has demonstrated truly impressive performance. Evolutionary programs have yielded walking robots, high quality amplifiers, sensitive antennas, and a wide range of more esoteric devices. In one field alone — circuit design — EC techniques have reinvented several circuits that were discovered and patented recently. These successes have been due to a confluence of theory, experience, and hardware developments.

In this paper, we use “EC” as an inclusive term for a variety of particular strategies, such as genetic algorithms, genetic programming, evolutionary programming, and techniques based metaphorically on ants, swarms, and weeds. EC is intended to be analogous to Darwinian evolution, which we will refer to as Evolutionary Biology (EB). As EC has evolved during the last few years, so has our understanding of EB. Although the basic principle of evolution — “survival of the fittest”, in Spencer’s well-known aphorism — its implementation is now known to be much more complex and interesting than was previously thought. In summary, evolution requires a *population of individuals* that *reproduce* with *variation* and an *environment* that associates a cost with reproduction.

In nature, fecundity is the *only* criterion for fitness. If absurdly long tail-feathers help males to have more offspring, then males will have absurdly long tail-feathers. EC introduces a fitness function: individuals are first

evaluated and only then allowed to reproduce. This reverses the natural order of events, in which the fitness of an individual can be assessed only at the end of its reproductive life. The fitness function transforms EC from whimsical experimentation to useful engineering.

## STRENGTHS AND WEAKNESSES

Viewed as a form of engineering, EB and EC share strengths and weaknesses. The following quote is often used to argue that even Darwin had doubts about the evolution of eyes (Darwin (1859)):

To suppose that the eye . . . could have been formed by natural selection, seems, I freely confess, absurd in the highest possible degree.

But this sentence is only the introduction to Darwin’s argument showing how evolution should be *expected* to produce eyes, concluding that

. . . the difficulty of believing that a perfect and complex eye could be formed by natural selection . . . should *not* be considered subversive to the theory.

The fossil record provides little evidence of evolving eyes. Yet experiments suggest that the evolution of an eye might require no more than 350,000 years — a blink of an eye, so to speak, in evolutionary time (Nilsson and Pelger (1994)). These experiments assume the existence of a means of detecting light: we might object that discovering a light detector that eyes can use seems to be an even harder problem than evolving the eye. In fact, early single-celled organisms discovered how to detect light, using the gene Pax 6 and the protein rhodopsin. All contemporary eyes are based on this gene and protein and are almost certainly adaptations of the first light detectors.

From eyes and similar examples, it is easy to get the impression that evolution leads to robust and adaptive organisms. This is perhaps the main reason for the interest in EC. It is certainly true that some organisms are robust and adaptive. But evolution also builds quaint, Rube Goldberg contraptions that just happen to work in an environment that changes only slowly.

Even worse, evolution often leads to solutions that would never be accepted by engineers. One example

will suffice (I will mention others in the talk). Again, it concerns eyes. Our retinas are constructed inside-out: the light receptors are behind the retina and the nerve fibres to the brain are in the front. This is a consequence of a simple, basic fact about evolution: change occurs in small increments, and each increment must be better than the previous. Big changes, temporary backward moves, and redesign cannot occur. Squids were luckier: their optical nerves are connected to the back of the retina.

The products of EC sometimes have the same quirkiness as the results of EB. This function (Rooke (2002)), evolved to produce an artistic image, is easily distinguishable from functions designed by people:<sup>1</sup>

```
cos (mul (div (div (pi, dist (mul
(0.703097, dist (div (0.777147, sin
(div (y, minus (-2.19658, cos (x))))),
cos (cos (sin (y))))), sin (y))),
cos (plus (cos (plus (atan (sin (cos
(x))), mul (x, x))), spiral ( (y,
0.418353))))), 0.494688)))
```

One problem that EC must address is how to combine evolutionary strategies with good engineering practice.

## FUTURE DIRECTIONS

Most work in EC is based on a very naive model of EB. For many purposes, this clearly does not matter, since EC has had striking successes. But EB is much more interesting than the simple genotype to phenotype mapping suggests, and I believe that there is much to be gained by studying EB and incorporating its techniques into EC. In this section, I briefly describe aspects of EB and pose challenges (*in italics*) for EC.

When the genetic code was first discovered, biologists assumed that each kind of organism would have its own kind of genes and that a complex organism, such as a person, would have hundreds of thousands of genes. Both assumptions are now known to be quite wrong.

To a first approximation, all organisms have the same genes. The genes that assemble the segments of a fruit fly are the genes that assemble the backbones of vertebrates. The genes that enable a bacterium to derive energy from sugar are the same as the genes that we use for the same purpose. Obviously, there are differences: plants do not need genes for eyes. But the commonalities are striking. EB achieves the software engineer's perennial dream of reuse. The same set of genes are used to assemble an enormous variety of different bodies. *Can EC achieve software reuse by exploiting evolutionary ideas?*

Variation is obtained by gene switching networks that control gene expression using both environmental fac-

<sup>1</sup>This function should not be read as a criticism of Rooke's admirable work but rather as a typical example of the unpredictability of evolution.

tors and other genes. Evolution's motto seems to be "if it ain't broke, don't fix it". Epistasis — interaction between genes — is the means by which EB obtains huge variety from a relatively small set of genes. *Can EC exploit epistasis to build complex systems from simple components?*

Perhaps the biggest difference between EB and most EC is that the genome controls all phases of the organism's life cycle. Our genome's work starts when the fertilized egg splits into two, then four, then eight cells. Later, the same genome controls the varied functions of each of the trillion or so cells in our adult bodies. The genomes of other organisms perform even more astonishing tasks: the genome that causes a larva to eat a leaf later tells the butterfly how to find pollen. *Can EC produce programs that grow, learning as they do so?*

There are similarities as well as differences between EC and EB. Programming languages are sometimes criticized for being fragile by comparison with biological artifacts. A misplaced comma can change the meaning of a program, or make it fail altogether. But biology is no different in this respect: a single error in the genome may be fatal. The techniques that nature has evolved to compensate for fragility follow good engineering practice, including redundant encoding, high fidelity copying followed by error detection and correction, and storing information in a stable and inactive molecule. *Given a simple but unreliable mechanism, can we use EC to evolve an equivalent but highly reliable mechanism?*

Finally, the languages used in EC systems such as Tierra and Avida are based closely on programming languages. *Can we develop a language for EC that is biologically inspired?* I consider this to be the "grand challenge" of EC.

## CONCLUSION

Evolutionary computing is an exciting and growing field with many significant successes to its credit. But its biological foundations have plenty of features waiting to be understood and exploited, ensuring dramatic developments in the near future.

## REFERENCES

- Darwin C., 1859. *The Origin of Species*. John Murray, first ed.
- Nilsson D.E. and Pelger S., 1994. *A Pessimistic Estimate of the Time Required for an Eye to Evolve*. *Proceedings of the Royal Society of London: Biological Sciences*, 256, no. 1345, 53–58.
- Rooke S., 2002. *Eons of Genetically Evolved Algorithmic Images*. In P. Bentley and D. Corne (Eds.), *Creative Evolutionary Systems*, Morgan-Kaufmann, chap. 13. 339–365.

# **SOFT COMPUTING THEORY AND PRACTICE**





# TOWARD HUMAN LEVEL MACHINE INTELLIGENCE -- IS IT ACHIEVABLE? THE NEED FOR A PARADIGM SHIFT

Lotfi A. Zadeh<sup>1\*</sup>

In the fifties of last century, the question "Can machines think?" was an object of many spirited discussions and debates. Exaggerated expectations were the norm, with no exceptions. In an article "Thinking machines—a new field in electrical engineering," published in January 1950, I began with a sample of headlines of articles which appeared in the popular press in the late forties. One of them read "Electric brain capable of translating foreign languages is being built." Today, half a century later, we have translation software, but nothing that approaches the level of human translation. In 1948, on the occasion of inauguration of IBM's Mark I relay computer, Howard Aiken, Director of Harvard's Computation Laboratory, said "There is no problem in applied mathematics that this computer cannot solve." Today, there is no dearth of problems which cannot be solved by any supercomputer. Exaggerated expectations should be forgiven. As Jules Verne said at the turn of last century, "Scientific progress is driven by exaggerated expectations."

Where do we stand today? What can we expect in the future?

AI was born in 1956. Today, half a century later, there is much that AI can be proud of—but not in the realm of human level machine intelligence. A telling benchmark is summarization. We have software that can passably summarize a class of documents but nothing that can summarize miscellaneous articles, much less books. We have humanoid robots but nothing that can compare in agility with that of a four year old child. We can automate driving a car in very light city traffic but there is nothing on the horizon that could automate driving in Istanbul. Far too often, we have to struggle with a dumb automated customer service system which we are forced to use. Such experiences make us keenly aware that human level machine intelligence is an objective rather than reality.

In an article "A new direction in AI—toward a computational theory of perceptions," AI Magazine, 2001, I argued that, in large measure, the lack of significant progress in many realms of human level machine intelligence is attributable to AI's failure to develop a machinery for dealing with perceptions. Underlying human level machine intelligence are two remarkable human capabilities. First, the capability to perform a wide variety of physical and mental tasks, such as driving a car in heavy city traffic, without any measurements and any computations. And second, the

capability to reason, converse and make rational decisions in an environment of imprecision, uncertainty, incompleteness of information, partiality of truth and partiality of possibility. A principal objective of human level intelligence is mechanization of these remarkable human capabilities.

What is widely unrecognized is that mechanization of these capabilities is beyond the reach of classical, Aristotelian, bivalent logic. What is needed for this purpose is fuzzy logic. AI's deep commitment to bivalent logic has impeded its acceptance of fuzzy logic. In my view, achievement of human level machine intelligence is infeasible without the use of fuzzy logic.

What is fuzzy logic? What does it have to offer? There are many misconceptions about fuzzy logic. The following précis of fuzzy logic is intended to correct the misconceptions. Fuzzy logic is not fuzzy. Basically, fuzzy logic is a precise logic of imprecision and approximate reasoning. In fact, fuzzy logic is much more than a logical system. It has many facets. The principal facets are logical, fuzzy-set-theoretic, epistemic and relational. Most of the applications of fuzzy logic involve the concept of a linguistic variable and the machinery of fuzzy if-then rules. The formalism of linguistic variables and fuzzy if-then rules is associated with the relational facet. The cornerstones of fuzzy logic are graduation, granulation, precisation and the concept of a generalized constraint. Graduation should be understood as an association of a concept with grades or degrees.

In fuzzy logic, everything is or is allowed to be a matter of degree or, equivalently, fuzzy. Furthermore, in fuzzy logic everything is or is allowed to be granulated, with a granule being a clump of attribute values drawn together by indistinguishability, equivalence, proximity or functionality. Graduated granulation or, equivalently, fuzzy granulation is inspired by what humans employ to deal with complexity, imprecision and uncertainty. Graduated granulation underlies the concept of a linguistic variable. When Age, for example, is treated as a linguistic variable, its granular values may be young, middle-aged and old. The granular values of Age are labels of fuzzy sets.

A concept which plays a pivotal role in fuzzy logic is that of a generalized constraint, represented as  $X \text{ isr } R$ , where  $X$  is the constrained variable,  $R$  is the constraining relation and  $r$  is an indexical variable which defines the modality of the constraint, that is, its semantics. The principal generalized

<sup>1</sup> Dedicated to Peter Walley.

\* Department of EECS, University of California, Berkeley, CA 94720-1776; Telephone: 510-642-4959; Fax: 510-642-1712; E-Mail: [zadeh@eeecs.berkeley.edu](mailto:zadeh@eeecs.berkeley.edu). Research supported in part by ONR N00014-02-1-0294, BT Grant CT1080028046, Omron Grant, Tekes Grant, Chevron Texaco Grant and the BISC Program of UC Berkeley.

constraints are possibilistic, probabilistic and veristic. The fundamental thesis of fuzzy logic is that information may be represented as a generalized constraint. A consequence of the fundamental thesis is that the meaning of a proposition,  $p$ , may likewise be represented as a generalized constraint. The concept of a generalized constraint serves as a basis for representation of and computation with propositions drawn from a natural language. This is the province of NL-Computation—computation with information described in natural language.

NL-Computation opens the gate to achievement of human level machine intelligence. The validity of this assertion rests on two basic facts. First, much of human knowledge, and especially world knowledge, is described in natural language. And second, a natural language is basically a system for describing perceptions. What this implies is that NL-Computation serves two major functions: (a) providing a

conceptual framework and techniques for precisiation of natural language in the context of human level machine intelligence; and (b) providing a capability to compute with natural language descriptions of perceptions. These capabilities play essential roles in progression toward human level machine intelligence.

In summary, achievement of human level machine intelligence is beyond the reach of bivalent-logic-based tools which AI has in its possession. What is needed for this purpose is addition of concepts and techniques drawn from fuzzy logic to AI's armamentarium. However, what should be stressed is that fuzzy logic is merely one of many tools which are needed to achieve human level machine intelligence. What is obvious is that achievement of human level machine intelligence is a major challenge which will be very hard to meet.

# MODELLING NUMERICAL AND SPATIAL UNCERTAINTY IN GRAYSCALE IMAGE CAPTURE USING FUZZY SET THEORY

M. Nachtegael<sup>a</sup>, P. Sussner<sup>b</sup>, T. Mélangé<sup>a</sup>, E.E. Kerre<sup>a</sup>

<sup>a</sup>Ghent University, Dept. of Applied Mathematics and Computer Science  
Fuzziness and Uncertainty Modelling Research Unit  
Krijgslaan 281 - S9, 9000 Gent, Belgium  
email: Mike.Nachtegael@UGent.be

<sup>b</sup>University of Campinas, Dept. of Applied Mathematics  
Campinas, SP 13083 859, Brazil  
email: sussner@ime.unicamp.br

## KEYWORDS

Uncertainty, Grayscale image, Fuzzy set theory, Interval-valued, Intuitionistic

## ABSTRACT

In this paper, we will discuss interval-valued and intuitionistic fuzzy sets as a model for grayscale images, taking into account the uncertainty regarding the measured grayscale values, which in some cases is also related to the uncertainty regarding the spatial position of an object in an image. We will demonstrate the practical potential of this image model by introducing an interval-valued morphological theory and by illustrating its application with some examples. The results show that the uncertainty that is present during the image capture not only can be modelled, but can also be propagated such that the information regarding the uncertainty is never lost.

## INTRODUCTION

Images are among the most important information carriers in today's world. This importance is not only due to the simple fact that an image can contain an enormous amount of relevant data, but also to the scientific and technological achievements of the last decades. The wide availability of image capturing devices and the easy way to develop images and to make them public (e.g. using the internet) has even enhanced this evolution. Since their introduction, fuzzy set theory [30] and fuzzy logic have given rise to many applications, also in image processing. This is not a surprise, since uncertainty and imprecision are encountered in many image processing applications, e.g. to determine whether a pixel is an edge-pixel or not or whether a pixel is contaminated with noise or not [21, 22], or when measuring the degree to which two images are similar to each other [26]. In other cases, the theory is used as a *tool* to construct

image processing operators. The latter typically occurs in the field of mathematical morphology. The basic morphological operators dilation, erosion, opening and closing constitute the fundamentals of this theory [23], and transform an image into another image, using a structuring element. As an extension from binary to grayscale morphology, different fuzzy morphological models have been introduced [9, 18, 24]. These models were based on the observation that, from a formal point of view, grayscale images and fuzzy sets are modelled in the same way, and consequently tools from fuzzy set theory could be applied in the context of image processing. However, it is only until quite recently that (extended) fuzzy set theory has been used to *model* the uncertainty that occurs with the image capture itself. In particular the extensions based on interval-valued and intuitionistic fuzzy set theory have very nice interpretations in the context of image processing [3, 4, 19, 20].

The goal of this paper is to extensively discuss the potential of extended fuzzy set theory – in particular interval-valued and intuitionistic fuzzy set theory – to model numerical and spatial uncertainty, due to image capture, in grayscale images (Section 2). In order to demonstrate the applicability of this image model we will introduce an interval-valued morphological theory, and we will illustrate this theory with some examples (Section 3). We end our paper with concluding remarks and directions for future research (Section 4).

## MODELLING THE UNCERTAINTY OF IMAGE CAPTURE

### The interval-valued approach

The grayscale value of a pixel in a grayscale image indicates the amount of black or white present at that specific location in the image. All approaches to mathematical morphology use these values to transform the original image. However, one always assumes that these

grayscale values are *certain*, although in practice, due to the circumstances in which images are sometimes captured, the measured values might be uncertain and merely indicate a *likely* value of the image at a specific position. The uncertainty regarding the grayscale value is an immediate fact if one takes into account that any device will round captured values up or down to the finite set of allowed values. The uncertainty grows if several takes of an image reveal different grayscale values for some pixels. This might be the case under identical recording circumstances, and will surely arise when these circumstances change (e.g. a scenery that is illuminated by either a sunny or a cloudy sky; see Figure 1). Not only the recording circumstances can play a role here. Indeed, pixels that belong to the edge of an object might slightly shift position in different takes (e.g. when the camera slightly shifts position; see also Figure 1). This could result in large differences in the measured grayscale value of a specific pixel, and consequently in a large uncertainty regarding the real value of that pixel, i.e. for that specific spatial position in the image.

For all these reasons, it can be useful not to work with grayscale *values* but with grayscale *intervals*, where the interval represents the set to which the actual grayscale value belongs. Such an interval will be small for a pixel that belongs to a larger object in the image and that was captured under more or less identical circumstances, but will be large for a pixel that was captured under different circumstances or that belongs to the edge of a larger object in the image. In this way, the approach of using intervals not only models uncertainty regarding the measurement of *values*, but also regarding the measurement of *spatial positions*.

Specifically in mathematical morphology, also regarding the values of the pixels in the structuring element some uncertainty might exist, even though it is chosen by the user. Indeed, if one wants the structuring element to reflect the importance or weight that is associated with a pixel at a certain position w.r.t. the center of the structuring element, one might not be completely sure how to estimate that weight. The use of an interval with likely values might be a solution in that case.

In the above context, grayscale images and/or structuring elements are actually characterized by interval-valued fuzzy sets. An interval-valued fuzzy set (IVFS) corresponds to a mapping  $A$  from a universe  $\mathcal{U}$  into the class of closed intervals  $[\mu_1, \mu_2] \subseteq [0, 1]$ . Thus,  $A(u) = [\mu_1(u), \mu_2(u)]$  for every  $u \in \mathcal{U}$ . If  $\mu_1(u) = \mu_2(u)$  for all  $u \in \mathcal{U}$  then the interval-valued fuzzy set reduces to a classical fuzzy set. Interval-valued fuzzy sets have been used successfully to implement Zadeh’s paradigm of computing with words [17] and have become increasingly important in applications of rule-based systems and approximate reasoning [7, 10, 25].

The important thing here is that interval-valued fuzzy set theory allows us to model the uncertainty regarding the grayscale values. In the evolution of fuzzy mor-

phology this is quite an important step, since we are making the transfer from *tool* to *model*. Techniques and tools from interval-valued fuzzy set theory can then be used to construct a corresponding morphological model and to define morphological operators that can process interval-valued images. The potential of fuzzy set theory is then fully used, i.e., the theory is employed both as tool and as model.

Note that interval-valued representations also occur naturally in several other image processing problems, e.g. in inverse halftoning [6], and that they also occur in the context of wavelets [5]. In a different context, not as a model but rather as a tool, they have also proven to be useful in edge detection applications [2].

To visualize the place of interval-valued (and intuitionistic) fuzzy morphologies in the field of mathematical morphology, we have summarized several approaches in Table 1, depending on the nature of the image and the structuring element.

Table 1: Approaches to mathematical morphology.

| Image     | Structuring element | Approach   |
|-----------|---------------------|--|
| binary    | binary              | binary morphology  |
| grayscale | binary              | grayscale morphology (threshold approach)                  |
| grayscale | grayscale           | grayscale morphology (umbra approach) + fuzzy morphologies |
| interval  | interval            | interval-valued fuzzy morphologies                         |

To visualize the use of interval-valued fuzzy sets as an image model, we discuss an example in Figures 1 and 2. Figure 1 shows three different takes of the cameraman image: a take with a cloudy sky, a take with a sunny sky and a slightly distorted take. These different takes reveal that the measured grayscale value of several pixels are uncertain. For the cloudy/sunny take the uncertainty is due to the recording circumstances (resulting in different grayscale values for the same pixel); for the cloudy/distorted take the uncertainty is due to the unclear spatial position of the objects in the image (resulting in different grayscale values for mainly these pixels that are on the edge of an object).

To take this uncertainty regarding the grayscale values into account, we construct interval-valued representations of the cameraman image. Starting from the three different takes (cloudy/sunny/distorted), we select for every pixel the lowest grayscale value from the images, resulting in the lower bounds of the grayscale intervals. Similarly, we select for every pixel the highest grayscale value from the images, resulting in the upper bounds of the grayscale intervals. The image with the lower bound

values and the image with the upper bound values are shown in Figure 2. The difference between the lower bound and upper bound images is also shown. The interpretation of this difference is quite nice: the higher the difference for a certain pixel (i.e., the higher the width of the corresponding interval, and the brighter the pixel in the corresponding image), the higher the uncertainty regarding that pixel. In this case, the interval-valued representation takes both numerical and spatial uncertainty into account. This example illustrates the natural way in which the interval-valued approach makes sense in image processing.

### The intuitionistic approach

Other considerations can lead to other extensions of fuzzy mathematical morphology. For example, given a grayscale image, one can assign two separate  $[0,1]$ -valued degrees (to the grayscale value of) every pixel, the first one indicating the belief that the pixel has this specific grayscale value and the second one reflecting the degree of certainty that it differs from the given grayscale value. Consequently, in such a model a pixel gets a specific grayscale value and is associated with two values. In contrast to the previous model, the grayscale values are fixed (i.e. no intervals occur), but the uncertainty regarding the measured value is associated with a couple of  $[0,1]$ -valued degrees.

The approach we just explained starts from the numerical point of view, i.e. we are trying to model the uncertainty regarding measured grayscale values. However, just as in the previous case, this approach can also be looked at from a spatial point of view. Indeed, for a pixel that belongs to the edge of an object there might be quite a lot of uncertainty regarding its grayscale value, depending on the fact whether the pixel actually is part of the object or just belongs to the background of that object. This spatial uncertainty, which is strongly connected to the numerical uncertainty (i.e. the uncertainty regarding the measured value), can also be modelled by the two separate  $[0,1]$ -valued degrees mentioned above. The latter of both points of view is the basis for the works [3, 4], while we used the first point of view in [19]. Regardless from which viewpoint is chosen, using the above approach we are actually considering intuitionistic fuzzy sets [1]. Intuitionistic fuzzy sets (IFS) generalize Zadeh's original definition by defying the law of the excluded middle which claims that if  $u$  belongs to a degree  $\mu$  to a fuzzy set  $A$  then  $u$  does not belong to  $A$  to the extent  $\nu = 1 - \mu$ . In IFS theory, the degrees of membership and non-membership do not have to add up to 1. Instead, IFS theory only requires that this pair satisfies the inequality  $\mu + \nu \leq 1$ .

Formally, a grayscale image  $A$  is in this context represented as a mapping from the universe  $\mathcal{U}$  to the set  $\{(\mu, \nu) | \mu + \nu \leq 1\}$ . Thus,  $A(u) = (\mu(u), \nu(u))$  for every  $u \in \mathcal{U}$ . If  $\nu(u) = 1 - \mu(u)$  for all  $u \in \mathcal{U}$  then the intu-

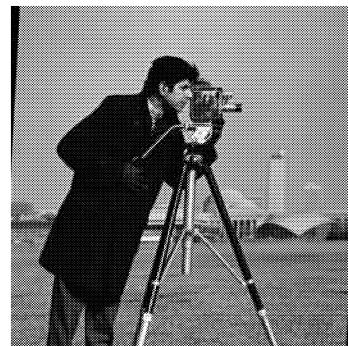
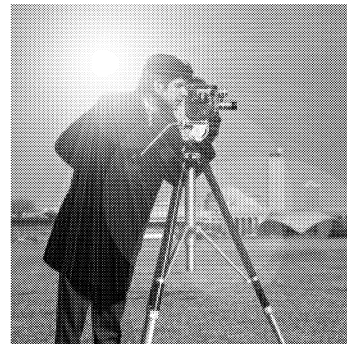


Figure 1: Different captures of the cameraman image: top = take with cloudy sky, middle = take with sunny sky, bottom = take with distortion.

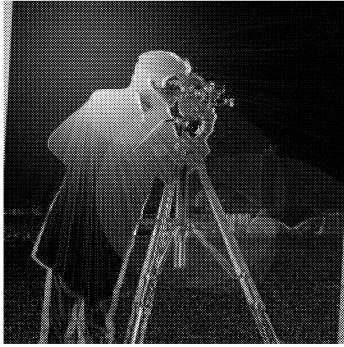


Figure 2: Interval-valued representation of the cloudy/sunny/distorted take of the cameraman image: top = lower bounds, middle = upper bounds, bottom = representation of the interval width. The uncertainty is due to both numerical and spatial uncertainty.

intuitionistic fuzzy set reduces to a classical fuzzy set.

Intuitionistic fuzzy set theory also arises in image processing from a different perspective, just as this is the case for interval-valued fuzzy set theory (see the discussion in the previous subsection). We refer to the works [27, 28, 29] for some applications in which intuitionistic fuzzy sets are used in the context of enhancement and pattern recognition.

### The equivalence between both approaches

The class of IVFS as well as the class of IFS can be regarded as  $\mathcal{L}$ -fuzzy sets (in some universe) where  $\mathcal{L} = (L, \leq_L)$  represents a complete lattice; a complete lattice is a partially ordered set in which every family of elements has a supremum and infimum. An  $\mathcal{L}$ -fuzzy set  $A$  in  $\mathcal{U}$  is characterized by an  $\mathcal{U} - L$  mapping [14]. When  $L = [0, 1]$ ,  $\mathcal{L}$ -fuzzy set theory reduces to classical fuzzy set theory.

In case of IVFS the corresponding complete lattice  $(L^I, \leq_{L^I})$  is defined by:

$$L^I = \{[x, y] | [x, y] \subseteq [0, 1]\},$$

$$[x_1, y_1] \leq_{L^I} [x_2, y_2] \Leftrightarrow x_1 \leq x_2 \text{ and } y_1 \leq y_2.$$

Infimum and supremum of a set  $S = \{[x_s, y_s] \subseteq [0, 1] | s \in I_S \subseteq \mathbb{N}\}$  are given by  $\bigwedge S = [\inf_s x_s, \inf_s y_s]$  and  $\bigvee S = [\sup_s x_s, \sup_s y_s]$ .

In case of IFS the corresponding complete lattice  $(L^*, \leq_{L^*})$  is defined by:

$$L^* = \{(x, y) | x + y \leq 1\},$$

$$(x_1, y_1) \leq_{L^*} (x_2, y_2) \Leftrightarrow x_1 \leq x_2 \text{ and } y_1 \geq y_2.$$

Infimum and supremum of a set  $S = \{(x_s, y_s) \in [0, 1] \times [0, 1] | s \in I_S \subseteq \mathbb{N}\}$  are given by  $\bigwedge S = [\inf_s x_s, \sup_s y_s]$  and  $\bigvee S = [\sup_s x_s, \inf_s y_s]$ .

Although the interval-valued approach (using grayscale intervals) and the intuitionistic approach (using fixed grayscale values, but associating them with additional values expressing some confidence) seem to be completely different in the context of image processing, they actually are exactly the same from a formal point of view [10]. The correspondence between these two extensions of fuzzy set theory is given by:

$$[x, y] \equiv (x, 1 - y),$$

where  $[x, y]$  represents a closed interval in interval-valued fuzzy set theory, and  $(x, 1 - y)$  represents a couple of membership and non-membership degrees in intuitionistic fuzzy set theory.

Since both approaches are identical, we can choose one model for further exploration and development. In general, there is a preference for the interval-valued model because of its very natural interpretation, and because of the fact that the input for this model, i.e. the intervals of grayscale values, can directly result from the image capture process.

# APPLICATION OF INTERVAL-VALUED FUZZY SET THEORY IN MATHEMATICAL MORPHOLOGY

Since we have established interval-valued fuzzy set theory (or, equivalently, intuitionistic fuzzy set theory) as a model for grayscale images, the next challenge is to construct other building stones to develop image processing theories. In this paper, we focus on mathematical morphology to illustrate this process.

## Interval-valued fuzzy morphology

Binary morphology was developed to process binary images, and quite soon extended to grayscale morphology by using the threshold approach [23] and the umbra approach [16]. Fuzzy morphology was an alternative extension, based on the extension of the underlying logical framework of the morphological model, i.e. using fuzzy logical operators as extensions of their binary counterparts [9, 18, 24].

This extension can also be realized in the case of interval-valued fuzzy sets. The logical aspect of interval-valued fuzzy set theory has already been largely investigated [8, 11, 12]. The richness of interval-valued fuzzy logical operators immediately leads to a wide variety of morphological models, depending on the choice of the underlying conjunctive and implicative. In [19, 20] we have developed a specific interval-valued morphological model. The corresponding dilation and erosion are based on the following interval-valued Łukasiewicz-operators: the so-called pessimistic t-norm

$$\mathcal{T}_W^p(x, y) = [\max(0, x_1 + y_1 - 1), \max(0, x_1 + y_2 - 1, x_2 + y_1 - 1)],$$

and the so-called optimistic implicator

$$\mathcal{I}_W^o(x, y) = [\min(1, 1 - x_1 + y_1, 1 - x_2 + y_2), \min(1, 1 - x_1 + y_2)],$$

with  $x = [x_1, x_2]$  and  $y = [y_1, y_2]$ . These operators are adjoint and lead to several interesting properties in mathematical morphology.

In order to simplify the expressions for the interval-valued dilation and erosion, we will make the following identifications regarding the grayscale image  $A$  and the structuring element  $B$ , both modelled as interval-valued fuzzy sets:

$$\begin{aligned} B(u - v) &= [b_1(u - v), b_2(u - v)] \equiv [b_1^u, b_2^u], \\ A(u) &= [a_1(u), a_2(u)] \equiv [a_1^u, a_2^u]. \end{aligned}$$

The following expressions can be derived for the interval-valued fuzzy dilation  $D_W^I$  and erosion  $E_W^I$  corresponding with the above mentioned Łukasiewicz-operators, for all  $v$  in  $\mathcal{U}$ :

$$\bigvee_{u \in \mathcal{U}} [\max(0, a_1^u + b_1^u - 1), \max(0, a_2^u + b_1^u - 1, a_1^u + b_2^u - 1)]$$

$$E_W^I(A, B)(v) = \bigwedge_{u \in \mathcal{U}} [\min(1, 1 + a_1^u - b_1^u, 1 + a_2^u - b_2^u), \min(1, 1 + a_2^u - b_1^u)].$$

## An edge detection application

Consider the following grayscale structuring element  $B$ :

$$B = \begin{pmatrix} 0.5 & 0.8 & 0.5 \\ 0.8 & 1 & 0.8 \\ 0.5 & 0.8 & 0.5 \end{pmatrix}.$$

Note that, for simplicity, we use a *certain* structuring element, i.e., a structuring element that can be represented as a classical fuzzy set; also note that the underlying element corresponds to the center of the structuring element.

Consider the interval-valued representation of the cloudy/sunny/distorted take of the cameraman image as shown in Figure 2. Using the above structuring element, we can perform the interval-valued dilation  $D_W^I$  and the interval-valued erosion  $E_W^I$ . These morphological operators result in new interval-valued images, of which we can display the lower bound image, the upper bound image, and the difference between them (indicating the uncertainty for every pixel). This is done in Figure 3 for the dilation and in Figure 4 for the erosion. We can take it one step further by taking the difference between the dilated and eroded images. This difference should result in an edge-image, just as in the case for regular grayscale images. Note that the difference between two intervals  $[x_1, x_2]$  and  $[y_1, y_2]$  is defined as the interval  $([12, 13])$ :

$$[x_1 - y_2, \max(x_1 - y_1, x_2 - y_2)].$$

The results are displayed in Figure 5, again together with the difference between the lower bound edge-image and the upper bound edge-image to visualize the uncertainty regarding these results. One can see that the lower bound edge-image contains nearly no information (this image results from the difference between the lower bound dilated image and the upper bound eroded image), while the upper bound edge-image produces a more interpretable image.

More specifically, we can make the following observations and conclusions regarding the edge-images. The upper bound edge-image contains real edges and false edges. In this case the false edges are mainly due to the distorted take of the cameraman image. These false edges are situated near the left and right border of the image (lines) and near the lower part of the cameraman contours (edges appear double, slightly shifted). At the same time, we observe a high uncertainty regarding these (real and false) edges. Knowing the “real” cameraman image, we know that the uncertainty for the edges near the left and right border and near the lower part of the cameraman contours actually should result in a rejection of the edges, while this is not the case for (most

of) the other detected edges. One of the future challenges will be to make an automated decision about the nature of the uncertainty, i.e., either due to numerical uncertainty or either due to spatial uncertainty.

In any case, and this is an important conclusion, one can observe that the uncertainty that was present in the original representation of the cameraman image is propagated through the (interval-valued) morphological operators and the edge detection application. This means that the information regarding the uncertainty is not lost. On the contrary, it is fully taken into account and can be used and exploited in further processing.

## CONCLUDING REMARKS AND FUTURE RESEARCH

In this paper we have discussed the important evolution of fuzzy set theory in the context of mathematical morphology. Fuzzy set theory was introduced in this field as a *tool*, used to construct alternative extensions of binary morphology to grayscale morphology. It was however until quite recently that extensions of fuzzy set theory have allowed us to actually *model* uncertainty that comes along with image capture – and modelling uncertainty, that’s what fuzzy set theory is all about. In particular, we have extensively discussed the interval-valued and the intuitionistic approach. We have shown that these theories cannot only be used as image models but also allow the construction of corresponding mathematical morphologies, which lead to specific morphological operators and related applications such as edge detection. Both aspects – the modelling and the construction of morphologies – are important, since we need operators to process grayscale images modelled using interval-valued or intuitionistic fuzzy sets.

Future research will have to focus on the further development of the morphological models. A thorough theoretical study is a must, and should be accompanied with a deep exploration of practical applications such as edge detection, segmentation, etc. In particular, it will be interesting to see how the uncertainty regarding the measured grayscale values is propagated, and how these results can be interpreted and used in practice. Of course, the specific choice of the underlying morphological model will be quite important, and should be clearly motivated.

## REFERENCES

- [1] Atanassov K., Intuitionistic Fuzzy Sets, Physica Verlag, Heidelberg (Germany), 1999.
- [2] Barrenechea E., Image Processing with interval-valued Fuzzy Sets - Edge Detection - Contrast, Ph.D. thesis, Public University of Navarra, 2005.
- [3] Bloch I., Mathematical Morphology on Bipolar Fuzzy Sets, in: Proceedings of ISMM 2007 (International

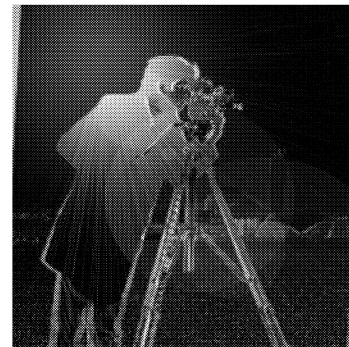


Figure 3: Interval-valued dilation of the cloudy/sunny/distorted take of the cameraman image (Figure 2): top = lower bounds of the dilated image, middle = upper bounds of the dilated image, bottom = representation of the interval width.



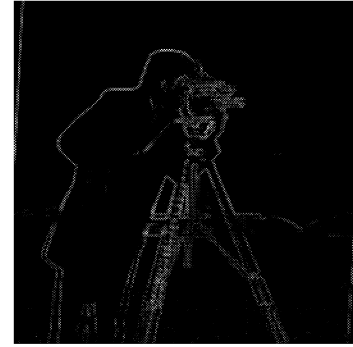
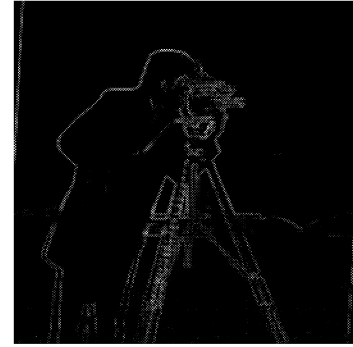
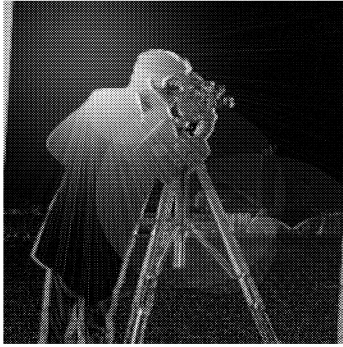
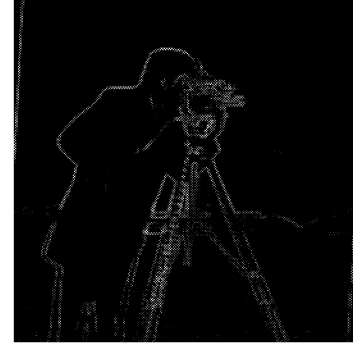


Figure 4: Interval-valued erosion of the cloudy/sunny/distorted take of the cameraman image (Figure 2): top = lower bounds of the eroded image, middle = upper bounds of the eroded image, bottom = representation of the interval width.

Figure 5: Interval-valued edge-image of the cloudy/sunny/distorted take of the cameraman image (Figure 2): top = lower bounds of the edge-image, middle = upper bounds of the edge-image, bottom = representation of the interval width.

- Symposium on Mathematical Morphology), 2007, pp. 3-4.
- [4] Bloch I., Dilation and Erosion of Spatial Bipolar Fuzzy Sets, in: *Lecture Notes in Artificial Intelligence*, Vol. 4578 (Proceedings of WILF 2007), 2007, pp. 385-393.
  - [5] Brito A.E. & Kosheleva O., Interval + Image = Wavelet: For Image Processing under Interval Uncertainty, Wavelets are Optimal, in: *Reliable Computing*, Vol. 4, No. 3, 1998, pp. 291-301.
  - [6] Cabrera S.D., Iyer K., Xiang G. & Kreinovich V., On Inverse Halftoning: Computational Complexity and Interval Computations, in: *Proceedings of CISS 2005* (39th Conference on Information Sciences and Systems), The John Hopkins University, paper 164, 2005.
  - [7] Castillo O. & Melin P., Intelligent Systems with Interval Type-2 Fuzzy Logic, in: *International Journal of Innovative Computing, Information and Control*, Vol. 4, No. 4, 2008, pp. 771-783.
  - [8] Cornelis C., Deschrijver G. & Kerre E.E., Implication in Intuitionistic and Interval-valued Fuzzy Set Theory: Construction, Classification, Application, in: *International Journal of Approximate Reasoning*, Vol. 35, 2004, pp. 55-95.
  - [9] De Baets B., Fuzzy morphology: a logical approach, in: *Uncertainty Analysis in Engineering and Sciences: Fuzzy Logic, Statistics, and Neural Network Approach* (Ayyub B.M. & Gupta M.M., editors), Kluwer Academic Publishers, Boston, 1997, pp. 53-67.
  - [10] Deschrijver G. & Kerre E.E., On the Relationship Between some Extensions of Fuzzy Set Theory, in: *Fuzzy Sets and Systems*, Vol. 133, 2003, pp. 227-235.
  - [11] Deschrijver G., Cornelis C. & Kerre E.E., On the Representation of Intuitionistic Fuzzy t-norms and t-conorms, in: *IEEE Transactions on Fuzzy Systems*, Vol. 12, No. 1, 2004, pp. 45-61.
  - [12] Deschrijver G. & Cornelis C., Representability in Interval-valued Fuzzy Set Theory, in: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, Vol. 15, No. 3, 2007, pp. 345-361.
  - [13] Deschrijver G., Arithmetic Operators in Interval-valued Fuzzy Set Theory, in: *Information Sciences*, Vol. 177, No. 14, 2007, pp. 2906-2924.
  - [14] Goguen J., L-fuzzy sets, in: *Journal of Mathematical Analysis and Applications*, Vol. 18, 1967, pp. 145-174.
  - [15] Hajek P., *Metamathematics of Fuzzy Logic*, Kluwer Academic Publishers, Dordrecht, 1998.
  - [16] Haralick R.M., Sternberg S.R. & Zhuang X., Image analysis using mathematical morphology, in: *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 9, No. 4, 1987, pp. 532-550.
  - [17] Mendel J.M., Computing With Words: Zadeh, Turing, Popper and Occam, in: *IEEE Computational Intelligence Magazine*, Vol. 2, No. 4, 2007, pp. 10-17.
  - [18] Nachtegaele M. & Kerre E.E., Connections between binary, gray-scale and fuzzy mathematical morphologies, in: *Fuzzy Sets and Systems*, Vol. 124, No. 1, 2001, pp. 73-86.
  - [19] Nachtegaele M., Sussner P., Mélangé T. & Kerre E.E., Some Aspects of Interval-valued and Intuitionistic Fuzzy Mathematical Morphology, accepted for IPCV 2008 - International Conference on Image Processing, Computer Vision and Pattern Recognition (July 14-17, 2008, Las Vegas, US).
  - [20] Nachtegaele M., Sussner P., Mélangé T. & Kerre E.E., An Interval-valued Fuzzy Morphological Model based on Lukasiewicz-Operators, accepted for ACIVS 2008 - International Conference on Advanced Concepts for Intelligent Vision Systems (October 21-24, 2008, Juan-les-Pins, France).
  - [21] Schulte S., Nachtegaele M., De Witte V., Van der Weken D. & Kerre E.E., Fuzzy impulse noise reduction methods for color images, In: *Proceedings of FUZZY DAYS 2006* (International Conference on Computational Intelligence), Dortmund (Germany), pp. 711-720.
  - [22] Schulte S., De Witte V., Nachtegaele M., Mélangé T. & E.E. Kerre, A new fuzzy additive noise reduction method, In: *Lecture Notes in Computer Science*, Vol. 4633 (Image Analysis and Recognition - Proceedings of ICIAR 2007), 2007, pp. 12 - 23, ISBN 978-3-540-74258-6.
  - [23] Serra J., *Image analysis and mathematical morphology*, Academic Press Inc, London, 1982.
  - [24] Sussner P. & Valle M.E., Classification of Fuzzy Mathematical Morphologies Based on Concepts of Inclusion Measure and Duality, in: *Journal of Mathematical Imaging and Vision*, accepted for publication, 2008.
  - [25] Turksen I.B. & Zhong Z., An Approximate Analogical Reasoning Schema based on Similarity Measures and Interval-Valued Fuzzy Sets, in: *Fuzzy Sets and Systems*, Vol. 34, No. 3, 1990, pp. 323-346.
  - [26] Vansteenkiste E., Van der Weken D., Philips W. & Kerre E.E., Evaluation of the perceptual performance of fuzzy image quality measures, In: *Lecture Notes in Computer Science*, Vol. 4251, 2006, pp. 623-630.
  - [27] Vlachos I.K. & Sergiadis D.G., Towards Intuitionistic Fuzzy Image Processing, In: *Proceedings of the International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce*, 2005, pp. 2-7.
  - [28] Vlachos I.K. & Sergiadis D.G., Intuitionistic Fuzzy Information - Applications to Pattern Recognition, In: *Pattern Recognition Letters*, Vol. 28, 2006, pp 197-206.
  - [29] Vlachos I.K. & Sergiadis D.G., Hesitancy Histogram Equalization, In: *Proceedings of FUZZ-IEEE 2007*, 2007, pp. 1-6.
  - [30] Zadeh L., Fuzzy Sets, in: *Information Control*, vol. 8, 1965, pp. 338-353.

# A GRANULAR UNIFIED MIN-MAX FUZZY-NEURO FRAMEWORK FOR LEARNING FUZZY SYSTEMS

Mokhtar Beldjehem  
École Polytechnique de Montréal  
C.P. 6079, succ. Centre-Ville  
Montréal QC H3C 3A7, Canada  
E-mail: mokhtar.beldjehem@polymtl.ca

## KEYWORDS

Possibility theory, fuzzy partition, repartitioning operator, if-then fuzzy weighted rules, granularity level, level of details, accuracy level, hybrid fuzzy-neuro possibilistic model, fuzzy hypothesis, fuzzy sequence, approximation of Min-Max relational equations, granulation, abstraction.

## ABSTRACT

We propose a novel computational granular unified framework that is cognitively motivated for learning if-then fuzzy weighted rules by using a hybrid neuro-fuzzy or fuzzy-neuro possibilistic model appropriately crafted as a means to automatically extract or learn fuzzy rules from only input-output examples by integrating some useful concepts from the human cognitive processes and adding some interesting granular functionalities. This learning scheme uses an exhaustive search over the fuzzy partitions of involved variables, automatic fuzzy hypotheses generation, formulation and testing, and approximation procedure of Min-Max relational equations. The main idea is to start learning from coarse fuzzy partitions of the involved variables (both input and output) and proceed progressively toward fine-grained partitions until finding the appropriate partitions that fit the data. According to the complexity of the problem at hand, it learns the whole structure of the fuzzy system, i.e. conjointly appropriate fuzzy partitions, appropriate fuzzy rules, their number and their associated membership functions.

## INTRODUCTION AND MOTIVATIONS

A production system (or a rule-base) the core of a knowledge-based (or a rule-based) system, is basically a formalism for representing knowledge about any area of problem-solving. A program written as a production system is a collection of “production rules,” which takes the form of “If Left-Hand-Side Then Right-Hand-Side.” Where Left-Hand-Side corresponds to the condition part and Right-Hand-Side corresponds to the action (or consequent) part. Such a representation is highly modular, is uniform i.e. all the knowledge of the system is expressed in the same format. It is mentioned that the domain expertise is organized in chunks, or equivalently granules of knowledge, and subsequently each chunk can be learned as a production rule which represents the expert’s answer to a “what if” situation. This way of organizing knowledge in discrete chunks which interact with each other for drawing conclusions by

inference is very natural way of modeling human cognitive processes (Newell and Simon 1972). In addition, the flexibility, the compactness, the approximation capacity, the expressiveness power, the non-linearity, and explicit embedded management of uncertainty provided by a fuzzy production rule make fuzzy production systems or fuzzy rule based systems privileged and very attractive candidates when compared with conventional rule-based systems. Fuzzy rules attempts to capture the “rules-of-thumb” approach generally used by domain experts for decision-making in complex environments. However to determine the required appropriate number of (fuzzy) rules and to elicit these rules from the domain expert remains a knowledge engineering challenge, especially for complex large scale problems for conventional and fuzzy systems alike.

In solving problems the human starts from a coarse description but if needed iterates and goes gradually to a fine-grained description or in-depth details enabling more understanding of the underlying problem until reaching a point where one can effectively find a solution and so stops and does not need any more details. At this point, an excess of precision is not needed (is not necessary) because a certain satisfying trade-off between precision (level of details) and generality of description has been reached and is sufficient and enough for finding a satisfactory approximate solution to the specified problem.

Those problem-solving mechanisms are frequently used by humans in modeling of and/or dealing with complex real world problems. When dealing with practical real world problems, there is an acute need for representing and manipulating imprecision, as it seems that the human mind is indeed conceptually somewhat fuzzy. Most of the time, humans do not use precisely (or crisply) defined terms but at the same time they do manage to communicate and resolve problems effectively. Zadeh has proposed the adoption of Approximate Reasoning (Zadeh 1971, 1979, 1984) and recently Zadeh has suggested that the challenge now facing AI is to produce systems exhibiting “common sense” reasoning, rather than purely logical deduction only, and he points to soft computing (Zadeh 2001).

When using the rich concept of fuzzy sets (Zadeh 1965, 1971) as a basis for possibility theory (Zadeh 1973, 1978) the automatic learning of fuzzy systems using a data-driven approach becomes a problem worth solving because a solution would enable us to build faithfully reliable systems in a more ergonomic convenient cost-effective fashion in various ICTs ranging from diagnosis, modeling, simulation,

vision, pattern recognition, information retrieval, process control to software engineering and so on.

On one hand, soft computing has been proposed by Zadeh (Zadeh 1994) and according to him (Zadeh 2001) "It may be argued that is soft computing rather than hard computing that should be viewed as the foundation of Artificial Intelligence (AI)." What is important to note is that soft computing not just a mixture. Rather, it is a synergistic partnership or a forum in which each of the partners contributes a distinct methodology for addressing problems in its domain. In this perspective, the principal constituent methodologies in soft computing are complementary rather than competitive; in particular synergy through hybridization ensures the emergence of desirable properties. The possibility of making fusion of the merits of each one for improved quality is feasible. Since 1990, hybrid soft computing and in particular hybrid fuzzy-neuro or neuro-fuzzy systems have invaded the computer world and constitutes one of the most exciting current topics of research (Beldjehem 1993; Yager and Zadeh 1994; Sinha and Gupta 1999; Pal and Ghosh 2000; Gupta et al. 2002), the advances are also spectacular due to its newness, perspectives and power. Numerical multi-layered networks as well as fuzzy models have been proved to be universal approximators. This has motivated their development and adoption in a large spectrum of successful industrial applications.

On the other hand, the concepts of granulation and abstraction in a fuzzy set theory setting have long been suggested by Zadeh (Zadeh 1976), his co-authors (Bellman et al. 1966) and advocated by others in an AI (Hobs 1985; Giumchglia et al. 1992) setting, in vision engineering (Marr 1982) setting, and in algorithm design (Foster 1992). It is attracting intensive research too and has led to the development of granular computing as an emerging computing paradigm (Yao 2000; Pedrycz 2001; Liu et al. 2002). It has been recently revisited by Zadeh himself (Zadeh 1998) who proposes retargeting it as a design paradigm and/or a methodology in connection with and under the "umbrella" of soft computing.

Bearing in mind that any workable model either mental (human) or computational (machine) is necessarily only abstraction and approximation of the reality, triangular and/or trapezoidal membership functions (MFs) might be used as they are in fact only approximation means to represent data, concepts, objects, entities, relationships, classes and even relations of the real world problems. Bell-shaped and even free-form membership functions may be used too. We consider that the granularity of a fuzzy partition for a variable is of utmost importance as it reflects the level of details (or resolution) required in describing such a variable, whereas the overlapping is connected to the inherent fuzziness in defining the boundaries between classes (granules) of such a variable. Of course a granule is also defined by a fuzzy set represented by a MF. Thus it reflects too a gradual rather than abrupt membership of an object to the class (granule).

The structure of the rest of the paper is as follows; in section II we introduce a new model based on a novel learning design methodology. Section III describes the statement of the learning problem including representation issues, hypothesis generation, formulation and testing, learning algorithm, learning by hybrid fuzzy-neuro Min-Max

networks. The section IV is devoted to the formulation of the learning problem including algorithmic issues of the learning problem, resolution and approximation of a Min-Max equations system, and the presentation of an abstract computational model of a learning session. In section V we conclude and give some perspectives for our future work.

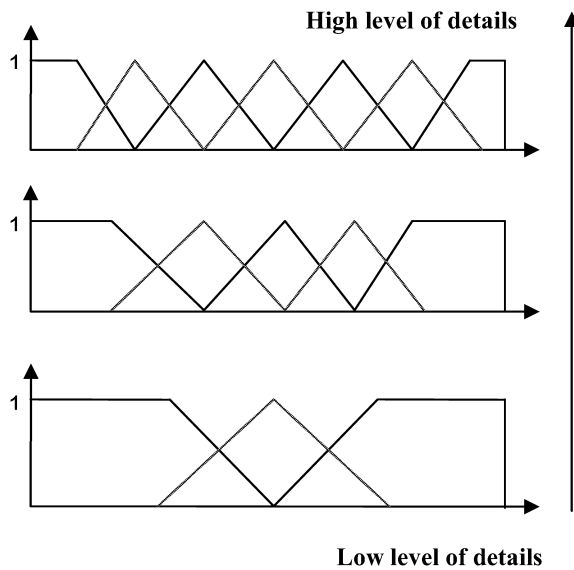
## A NOVEL LEARNING METHODOLOGY

### Motivations for our learning methodology

Fuzzy logic (Zadeh 19965, 1971, 1973, 1979) may be considered as a basis for knowledge and meaning representation and is particularly suited for dealing with natural language. We believe that it is the concept of possibility/necessity distributions (Zadeh 1978), rather than the truth, that will play the primary role in manipulating such knowledge for the perspective of drawing conclusions. Possibility theory (Zadeh 1978; Yager 1986; Dubois & Prade 1988; Olaf 1998) provides a formal framework for representing and dealing with ignorance, and uncertainties prevalent in modeling real world problems in a flexible computerized manner straightforwardly. However it is well accepted that crafting manually fuzzy systems to resolve complex large scale real-world problems is a difficult task that is not always obvious for both the designer (the knowledge-engineer) and the domain expert. This is due partly to the cognitive limits of the human being (Miller 1956), but also to the difficulty of understanding the intricacies of dimensionality and inherent complexities and peculiarities of large scale real world problems. Not to mention the lack of precision in the human-human interaction and communication that affects significantly the knowledge acquisition process during the tandem knowledge-engineer/domain expert relationship. Furthermore once it is undertaken it is labour-intensive, costly, error prone, time-consuming, and done on a trial-and-error basis in an adhoc manner and hence need to be totally or partly automated. This is known as the knowledge acquisition bottleneck problem or the Feigenbaum bottleneck and is a common problem for all AI approaches. Soft computing as an automated knowledge acquisition methodology aims at remedying such a problem.

Various soft computing (SC) techniques have been used to tackle this learning problem from various points of views. However they are based on some idealizing assumptions and no one adopts a holistic approach to resolve such a problem globally, i.e, finding conjointly appropriate fuzzy partitions, fine tuning the membership functions of the labels used in the rules as well as identifying the structure of the fuzzy system (both the required number of rules and rules themselves explicitly) simultaneously. In practice the required number of rules of the system is not known in advance. Indeed learning fuzzy if-then rules is a difficult multi-parameter optimization problem! We have previously devised, developed, formally validated and deployed a hybrid fuzzy-neuro system called Fennec (Beldjehem 1993, 1994, 2002, 2004, 2006, 2008) that was successfully applied to a difficult problem of biomedical diagnosis on Proteins/Biological Inflammatory Syndromes (B.I.S) as well as to a complex handwriting pattern recognition problem. Based on our previous work, we propose herein an integrated

framework to modify the model and extend its ability and scope of applicability by integrating some useful concepts from the human cognitive processes and adding some interesting granular functionalities. The rationale behind using levels of granularity is obvious for the reader.



**Figure 1 From a coarse fuzzy partition to a fine-grained fuzzy partition**

The basic ideas underlying our framework stems from the following interesting remarks about human cognition: Let us first focus our attention on the human problem solving process. In solving problems the human starts from a coarse description but if needed iterates and goes gradually to a fine-grained description or in-depth details enabling more understanding of the underlying problem until reaching a point where one can effectively find a solution and so stops and does not need any more details. At this point, an excess of precision is not needed (is not necessary) because a certain satisfying trade-offs between precision (level of details) and generality of description has been reached and is sufficient and enough for finding a satisfactory approximate solution to the specified problem. Thus after each iteration (increment) a gain of information is obtained enabling more in-depth and more understanding of the underlying situation. Thus, the human converges to a solution gradually by leveraging the level of details. See Figure 1 for more details in connections with a granular soft computing (GrSC) setting. Low levels of details allow coarse or general descriptions reflecting crude approximations whereas high levels of details allow specific descriptions reflecting more or less relatively precise approximations (crisps at the extreme). It is appealing and convenient to mimic mechanically or to emulate computationally such a cognitive process in order to automatically build faithfully by learning an appropriate “good” fuzzy system that exhibits both a high accuracy and a good performance for any problem at hand. This motivates us in building a learning system able to use such abstraction and granulation mechanisms in a fashion that is akin to the way humans achieve problem solving process. In general the required level of details necessary in

describing rules as well as the required number of rules for solving a problem depends to the degree of complexity of the problem at hand and are unknown and hence we propose to detect and determine them by learning within our framework.

## THE STATEMENT OF THE LEARNING PROBLEM

### Description of the Learning Process

The learning is parametric as well as structural. It has to deal with the complexity of the problem and to discover appropriate knowledge chunks, and approximation heuristics for the problem at hand. Taking into account the degree of complexity of the problem at hand as well as the empirical knowledge contained in the training set, the learning subsystem:

- Identify explicitly the appropriate fuzzy partition for each variable by learning. They are used only as references to generate fuzzy hypotheses. For each variable the appropriate number of granules and the slopes of which will be determined during learning. This information could be either kept or thrown away once the learning is completed without loss of information for the system. As they constitute only means for generating appropriate membership functions of fuzzy rules and are not used during inference.
- Find the appropriate membership functions for both the antecedents and consequents of every potential rule that is needed to model the problem at hand.
- Ultimately, build the appropriate “good” collection of if-then fuzzy rules (the rule base or knowledge base that consists of a set of linguistic rules), that fits “best” the data that consists of I/O pairs of the training set.

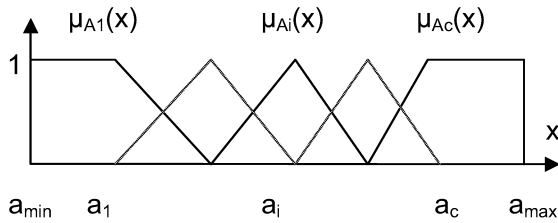
In order to build an automatic workable computational multi-pass learning model some design assumptions are made:

- At each cycle for each input variable  $X_i$  the system generates dynamically a fuzzy partition of  $c$  granules (starting with  $c=2$ , and incrementing  $c$  by 1 or 2 at each cycle until reaching a satisfying point). This point constitutes the stopping criterion of our learning mechanism and it reflects too the accuracy level required for the system. It is worth mentioning that increasing  $c$  alone does not affect the algorithmic computational complexity of the learning process! It is the number of input variables ( $n$ ) of the system when it is very large that affects it significantly. We assume to have a reasonable value for  $n$  which is almost the case in most classes of real world problems.
- An output variable may be dealt with as an input one, but for the sake of simplicity and programmability we assume that a fuzzy partition is given (known a priori for each output variable) and prepared cautiously by the domain expert. As the domain expert is more faced with the difficult problem of capturing relationships between the combinations of inputs variable in relation with a given output variable. In general, for a given output variable the actions (or classes) are well categorized (the number and names of granules are known) by the domain expert even though the slopes of associated MFs have to be questioned during learning.

## FORMULATION OF THE LEARNING PROBLEM

### Hypothesis Generation, Formulation and Testing

How to characterize and to represent a fuzzy partition? What operators are needed in manipulating a fuzzy partition? During learning-time, only one operator is needed to create a fuzzy partition having the required known granularity  $c$ . It is the repartitioning operator. It consists to divide dynamically during learning-time the universe of discourse into  $c$  overlapping granules. It works from scratch, i.e., there is no need for splitting, or fusion or expanding. A partition is used as reference only and its granules do not necessarily constitute MFs for actual rules as they are only used for formulation of initial fuzzy hypotheses during the generation by the systematic exhaustive search algorithm and they are both scale-dependents and context-dependents. We have no other assumption about the fuzzy partition and we are not interested to argue in such matters like “good” partition. The learning will be done at the rule level rather than at the partition level and hence learning a “good” rule is indeed a crucial issue of utmost importance. A fuzzy partition is illustrated in Figure 2 (observe how the rightmost and the leftmost granules are shaped); it is a parameterized family (sequence) of membership functions that cover the universe of discourse for every variable either input or output. It is created dynamically by the execution of the repartitioning operator of granularity equals to  $c$  during learning-time. In fact, it is obtained by superposition of two wave functions defined over the same universe of discourse  $X$  ranging in the interval  $[a_{\min}, a_{\max}]$ . Thus, it is straightforward to extract parameters of granules (MFs) from a given fuzzy partition, as each granule may be considered as an indexed term of the family (or sequence).



**Figure 2 A fuzzy partition of granularity  $c=5$  that is a superposition of two wave functions.**

A fuzzy partition is represented by vector of  $c$  parameters, where  $c$  is the granularity level. A fuzzy partition might be thought of as a sequence of granules, each of which is represented by an indexed term. This makes sense as they are computed and manipulated easily like ordinary terms during learning-time. In general as illustrated in Figure 2, every value  $x$  of the universe of discourse corresponds to at most two granules.  $A_1, A_2, \dots, A_i, \dots, A_c$  are just synthetic linguistic labels interpreted by fuzzy sets of normalized MFs. A fuzzy partition might be thought of as a synthetic alphabet that the system create by learning for future hypotheses generation. Thanks to this flexible scale-dependent representation, regardless the range of the universe of discourse of an input variable, the terms of the fuzzy

partition sequence are explicitly expressed straightforwardly as follows:

The first term (or granule)

$$\mu_{A1}(x) = \begin{cases} (a_2 - x)/(a_2 - a_1), & \text{if } a_1 \leq x \leq a_2 \\ 1 & \text{otherwise} \end{cases}$$

For  $i=2, 3, \dots, c-1$ , where  $c$  is the granularity of the partition or the  $i$ -th term

$$\mu_{Ai}(x) = \begin{cases} (x - a_{i-1})/(a_i - a_{i-1}), & \text{if } a_{i-1} \leq x \leq a_i \\ (a_{i+1} - x)/(a_{i+1} - a_i), & \text{if } a_i < x \leq a_{i+1} \\ 0 & \text{otherwise} \end{cases}$$

And finally the last term

$$\mu_{Ac}(x) = \begin{cases} (x - a_{c-1})/(a_c - a_{c-1}), & \text{if } a_{c-1} \leq x \leq a_c \\ 1 & \text{otherwise} \end{cases}$$

### Learning by Hybrid Min-Max Fuzzy-Neuro Network

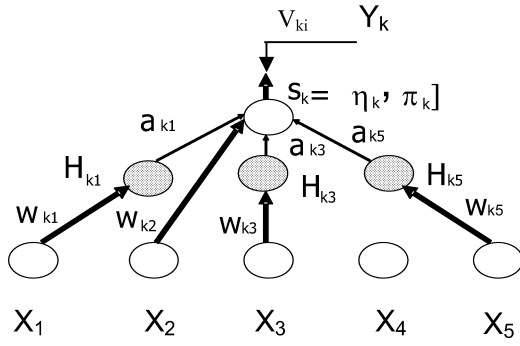
Fuzzy rules attempts to capture the “rules-of-thumb” approach generally used by domain experts for decision-making. However it is well accepted that crafting manually fuzzy systems to resolve complex large scale real-world problems is a difficult task that is not always obvious for both the designer (the knowledge-engineer) and the domain expert. Fuzzy (weighted) rules have been advocated, used, studied and interpreted by many authors (Zadeh 1971; Cayrol et al. 1982; Dubois et al. 1988; Beldjehem 1993; Yager 1996) and machine learned by Beldjehem (Beldjehem 1993). We will focus in dealing with a multi-input single-output (MISO) system as any multiple-input multiple-output (MIMO) system could be converted to a certain number of MISO systems. Let us start with a model overview: As in Beldjehem (Beldjehem 1993) we consider herein to design a fuzzy-neural possibilistic network according to the scheme Fuzzy to Neural (or to switch from fuzzy systems to neural networks). We use fuzzy if-then weighted rules that are herein of the control type instead of the classification type as in (Beldjehem 1993, 1994, 2002, 2004, 2006, 2008) and such a rule looks like:

**If** ( $X_1$  is  $w_{k1}, c_{k1}$ ) and ( $X_2$  is  $w_{k2}, c_{k2}$ ) and ( $X_3$  is  $w_{k3}, c_{k3}$ )

and ( $X_5$  is  $w_{k5}, c_{k6}$ ) **Then**  $Y_k$  is  $V_{ki}$

$c_{kj}$  is a weight that represents the grade of importance of “ $X_j$  is  $w_{kj}$ ” in relation with the output  $Y_k$ . Thus, conversely the weight  $a_{kj} = 1 - c_{kj}$  represents the grade of unimportance of “ $X_j$  is  $w_{kj}$ ” in relation with the same output  $Y_k$ .

Referring to Figure 3, we propose herein a feed-forward fuzzy-neural possibilistic network. We begin with a brief description of the model: two types of weights are associated with the connections.



**Figure 3 Schematic representation of the hybrid fuzzy-neuro possibilistic Min-Max model used.**

Type 1: Direct connections between input cells ( $X_j$ ) and output cell ( $s_k$ ) with only synthetic linguistic weights ( $w_{kj}$ ), interpreted as labels of fuzzy sets, characterizing the variations of the input cells (" $X_j$  is  $w_{kj}$ ") with the output cell ( $s_k$ ), in this case we have  $a_{kj} = [0,0]=0$ . Thus  $(\Pi(X_j; w_{kj}) \vee 0) = \Pi(X_j; w_{kj})$ . Thus the connection between a hidden cell and output cell simply disappears from the graph allowing direct connection.

Type 2: Connections between input cells ( $X_j$ ) and output cells ( $s_k$ ) via intermediate cells ( $H_{kj}$ ), weights associated to connections between input cells ( $X_j$ ) and intermediate cells ( $H_{kj}$ ), are herein artificial or synthetic linguistic ( $w_{kj}$ ), weights associated to connections between intermediate cells ( $H_{kj}$ ), and output cell ( $s_k$ ) are herein numerical intervals ( $a_{kj} \subseteq [0,1]$ ), instead of a scalar value ranging in the interval  $[0,1]$  ( $a_{kj} \in [0,1]$ ).

$w_{kj}$  are unknown artificial or synthetic linguistic weights and  $a_{kj}$  are unknown confidence interval that reflects a domain of possible values of unimportance for the corresponding connections. Thus providing much more flexibility for the network.

A learning session starts with a "blank" fully connected hybrid fuzzy-neuro network without a priori information concerning the weights, i.e. the weights might be thought of as "placeholders" only. Learning is parametric as well as structural. Let us consider now cell activation for an arbitrary output cell ( $s_k$ ), as illustrated in Figure 3, where only connections used in activation of  $s_k$  appear. From the semantic point of view, such a figure reflects a neural representation of an if-then fuzzy weighted rule of control type. Let  $\Pi(X_j; w_{kj}) = \text{Sup}[w_{kj} \cap X_j]$  be possibility measure associated to fuzzy sets  $w_{kj}$  and  $X_j$ . And let  $N(X_j; w_{kj}) = \text{Inf}[w_{kj} \cap \text{Not } X_j]$  be necessity measure associated to fuzzy sets  $w_{kj}$  and  $X_j$ . In general our model is governed by the three abstract fuzzy approximate equations as shown below.

$$\pi_k = \bigwedge_{j \in \{1,2,3,5\}} (\Pi(X_j; w_{kj}) \vee a_{kj}) \quad (1)$$

$$\eta_k = \bigwedge_{j \in \{1,2,3,5\}} (N(X_j; w_{kj}) \vee a_{kj}) \quad (2)$$

$$s_k = [\eta_k, \pi_k] \quad (3)$$

Observe that Maximum ( $\vee$ ) limits lower amplitudes of inputs, we have  $(\Pi(X_j; w_{kj}) \vee a_{kj}) = a_{kj}$  if  $\Pi(X_j; w_{kj}) \leq a_{kj}$ , and amplifies higher ones  $(\Pi(X_j; w_{kj}) \vee a_{kj}) = \Pi(X_j; w_{kj})$ , if

$\Pi(X_j; w_{kj}) \geq a_{kj}$ , so the Min-Max composition indicates a somewhat excitatory character. It is worthwhile to notice that Min-Max composition as containing Min and Max operations is strongly nonlinear. Furthermore, such model has been formally validated and it has been shown recently (Beldjehem 2006, 2008) that Min-Max composition preserves the value approximation property. Observe that when  $a_{kj} = 1$ , the term  $\Pi(X_j; w_{kj}) \vee a_{kj}$  (respectively  $N(X_j; w_{kj}) \vee a_{kj}$ ) is deleted in the application of Minimum ( $\wedge$ ). Thus ensuring the interpretability and transparency of the model. It is now clear that  $a_{kj}$  reflects a notion of unimportance, we point out herein that it is strongly hard if not impossible to make values assignment to grades of unimportance in practical applications, we will propose a mechanism to learn such grades of unimportance. Thus the fuzzy-neuro possibilistic network might be thought of as a transparent learning device of any non-linear mapping of inputs into an output. Its has been proved too that Max-Min composition preserves the value approximation property (Papis 1991) in connections with fuzzy systems setting.

## RESOLUTION OF THE LEARNING PROBLEM

### The Learning Algorithm and Implementation Issues

During a learning session the same learning algorithm is used for each output variable  $Y_j$ . Let us briefly describe the learning algorithm that is composed of many cycles, each of which is executed as follows: For each output variable  $Y_j$  and for each granule belonging to the fuzzy partition that corresponds to  $Y_j$ . Iteratively, an initial fuzzy hypothesis corresponds to a combination of certain number of MFs (each of which corresponds to granule of an input variable) is created (formed) by a systematic exhaustive search procedure. Once a fuzzy hypothesis is formed it is loaded or incorporated in the hybrid fuzzy-neuro network weights for test purposes, its components (elements) will be adjusted to fit the training data. Such hypothesis is considered as a potential candidate to be a rule and then is questioned and adjusted during learning by the means of a hybrid fuzzy-neuro possibilistic network using a successive approximation algorithm of systems of Min-Max relational equations. This adjustment is repeated until finding the ones that minimize the signal error. Hence another new combination is then generated and we repeat the same procedure. Thus the obtained adjusted hypotheses that minimize the cost over all possible combinations and that were embedded in the weights of the hybrid fuzzy-neuro possibilistic network are kept in a temporary learning table.

The algorithm proceeds by increasing the granularity and repeats the same cycle, until reaching a satisfying point. In general the learning is stopped when either a certain level of accuracy has been reached or it is impossible or it is computationally worthless to seek minimizing the error much more, i.e. this situation means that increasing the granularity is no more interesting. In general this point constitutes a trade-offs between tractability and low cost solution. Learning need to find an approximate solution that is not necessarily precise (or crisp) optimal one but at the same time it builds a model that do manage to resolve the problem at hand effectively. At the end one or more of the

obtained adjusted hypotheses that minimize the cost (over all considered granularity levels) constitutes a valid hypothesis and is transferred and stored in a knowledge base (KB) of the system as it consists effectively of a new learned rule. The system check whether or not a rule is new, i. e. whether or not it is already included the KB, and if necessary, transmits it to the KB, in an intelligible form for the storage (Hash table data structure). Assume the system get two or more valid hypotheses, after checking each one, each one is eventually added to the KB as a new rule. The advantage is that by construction (learning) we build a production system with no contradictory rules and thus giving a high satisfactory performance. This is in fact a built-in quality attribute.

Thanks to these granular functionalities, this novel learning algorithm constitutes a departure from the conventional ones, in that it conjointly determine dynamically during the learning-time the required satisfying number of rules necessary to model the problem as well as the rules themselves explicitly. Intuitively, this number is proportional to the degree of complexity of the problem at hand.

The resolution of fuzzy relations equations constitutes a good tool in fuzzy modeling especially for dealing with inverse problems. The fuzzy relational calculus theory (Dinola et al. 1989; Beldjehem 1993) provides us with a set of analytic formulas expressing solutions for some types of equations and their systems. However, the existence of solutions of the system is not known in advance. This makes any preliminary analysis rather tedious if not impossible. We reformulate the problem of solving a system of Min-Max from interpolation-like format to approximation-like one. This means that instead of trying to find exact solution, we try to find the best approximate solution. Any scalar and any element of vectors or matrices are assumed to have its value in the interval [0, 1]. Formally; our problem can be stated as follows: "Given an  $m \times n$  matrix  $R$  and an  $n$  vector  $b$ , find an  $m$  vector  $a$  such that  $(a \Delta R \supseteq b)$  where  $\Delta$  is the Min-Max composition and  $\supseteq$  denotes the fuzzy inclusion operation. Let us consider the case when there is no solution for the system (it does not satisfy the necessary condition, i.e.  $a \Delta R \supset b$ ).

This can be also reflected by only computing a distance. Let  $A, A'$  be fuzzy subsets of  $U$  and  $\alpha, \alpha'$  be the corresponding grades of membership vectors. By  $\|\alpha - \alpha'\|$  we denote the number  $\max_i (|\alpha_i - \alpha'_i|)$ , i.e. the maximum of the absolute

values of the differences between all element of  $\alpha$  and  $\alpha'$ . It might be interpreted as the signal error subject to be minimized. Equivalently by using this distance rather than the fuzzy inclusion concept we get the same results; and for this reason we use such a distance  $\|a \Delta R - b\|$  in our implementation of the system. It corresponds to minimal distance, hence  $a$  is the best approximator. Thus, since our algorithm is valid for both interpolation-like and approximation-like formats, it allows to resolve the more general following problems: "Given an  $m \times n$  matrix  $R$  and an  $n$  vector  $b$ , find all  $m$  vectors such that  $a \Delta R \supseteq b$ ". This algorithm is used as approximation procedure by the learning algorithm in our system. The learning consists mainly in crunching (approximating) systems of Min-Max equations while manipulating abstract synthetic linguistic concepts

(labels, hypotheses). It can be shown that the best approximator (from the fuzzy inclusion point of view) corresponds to the lower bound  $a$  of the inf-semi-lattice. It can be computed straightforwardly using the  $\varepsilon$  resolution operator only. It has been shown by a worst-case analysis that our computing algorithm has a linear complexity of  $\Theta(m \times n)$  (Beldjehem 1993). In order to illustrate the functioning and the behavior of our approximation algorithm let us hand-execute it on the following example,  $R$  and  $b$  are known. The  $\varepsilon$  operator is defined as follows (Beldjehem 1993)

$$x \varepsilon y = \begin{cases} y & \text{if } x < y \\ 0 & \text{otherwise} \end{cases}$$

Firstly, we compute the lower bound  $a$  of the inf-semi-lattice

$$R = \begin{bmatrix} 0.5 & 0.6 & 0.1 & 0.3 & 0.6 \\ 0.7 & 0 & 0.8 & 0.4 & 0.7 \\ 0.8 & 0.3 & 0.5 & 0.7 & 0.6 \\ 0.4 & 0.8 & 0.6 & 0.8 & 0.7 \\ 0.4 & 0.4 & 0.7 & 1 & 0.6 \\ 0.9 & 1 & 1 & 1 & 0.8 \end{bmatrix}$$

$$b = [0.3 \quad 0.3 \quad 0.5 \quad 0.4 \quad 0.5]$$

$$a = \vee (R \varepsilon b), \text{ where } \vee \text{ stands for MAX}$$

$$a = [0.5 \quad 0.3 \quad 0 \quad 0 \quad 0]$$

By performing the Min-Max composition, we have

$$b = [0.3 \quad \underline{0.3} \quad \underline{0.5} \quad \underline{0.4} \quad 0.5] \text{ (the target vector)}$$

$$a \Delta R = [0.4 \quad \underline{0.3} \quad \underline{0.5} \quad \underline{0.4} \quad 0.6]$$

$$\|a \Delta R - b\| = 0.1$$

Observe the surprising remarkable approximating power of  $a$

### Abstract Computational Model of a Learning Session

We are interested herein by establishing the computational abstract model of learning, learning implements a kind of successive approximation of Min-Max system process, and find weights of the hybrid fuzzy-neuro networks that fits "best" the data that consists of pairs I/O of the training set. Formally, from the computational point of view, for each output ( $s_k$ ), a learning session consists to resolve or to approximate  $(r + 1)$  systems of Min-Max equations, as follows:

$$a \Delta R^{(0)} \supseteq b$$

$$a \Delta R^{(1)} \supseteq b$$

$$\vdots$$

$$a \Delta R^{(r)} \supseteq b$$



Learning consists to prefer (validate) the configuration (the fuzzy hypothesis) of the best approximate solution (from the fuzzy inclusion point of view), i.e. which minimizes the local cost function and hence the corresponding deep structure. In other terms the learning process finds incrementally the "best" deep structure which corresponds to the following matrix  $R^{(j)} : j \in [0, r]$  such that:

$$a \Delta R^{(j)} \supseteq a \Delta R^{(j-1)} \supseteq b, \forall j=0 \dots r$$

Or equivalently,

$$\| a \Delta R^{(j)} - b \| \geq \| a \Delta R^{(j-1)} - b \|, \forall j=0 \dots r$$

Learning tries progressively by successive approximation to minimize the local cost function by the generation and the approximation of a new system. Thus, this approximation algorithm constitutes the mathematical machinery of learning. It has been shown that this system is a universal approximator (Beldjehem 2006, 2008), furthermore it is now clear that the ultimate aim of learning is to generate a consistent system which correspond to exact solution (or to establish a universal interpolator), however it seems that is not always the case in practical applications. In general the value of the local cost function may be seen as a quality index for a learning session or a performance index for the system. Learning has high speed due to its simplicity and analytic nature. The learning consists mainly in crunching (approximating) systems of Min-Max equations while manipulating abstract synthetic linguistic concepts (labels, hypotheses). Indeed the fuzzy learning process may be thought of as a new kind of algorithmic fuzzy optimization or rather algorithmic fuzzy approximation.

## CONCLUDING REMARKS

We have developed a cognitively motivated granular computational framework for learning fuzzy systems. This allows the automatic learning of fuzzy if-then knowledge bases (or rule bases) of systems which are large scale, too complex or too ill-defined to admit of precise quantitative analysis, description or control strategy. It may be thought of as an automatic means or a learning device for capturing the description of ill-defined concepts, relations and decisions rules. Such a framework integrates conjointly both the perceptual and the cognitive aspects of the human problem-solving process and ensure a granular processing of the underlying input from different granularity levels. It is the first attempt in the field. Implementation of a system called Neofennec (that is an upgraded or refined version of Fennec) working under the proposed framework is underway. The "good" rule-base (RB) is obtained automatically from training examples. Its inference engine has the inherent ability to generalize, which permit it to classify unseen examples accurately. During learning-time the system finds automatically the adequate levels of details (granularities) for the problem at hand. It is possible using a linguistic approximation to build automatically a true linguistic fuzzy system by learning. We believe that hybrid soft computing, machine learning, knowledge-based systems, performance evaluation have to learn from each other, and could be integrated or fused synergistically (not competitively) in order to build next generation of intelligent computational

systems. Such systems exhibit performance-accuracy trade-off, adaptability, transparency, interpretability, robustness, tractability, tolerance for uncertainty, categorization abilities, value approximation and therefore ensuring evolvability and generalization capacities.

On one hand we have used a systematic exhaustive search algorithm that explores and generates all possible fuzzy hypotheses available in the space-state representation of fuzzy partitions. Even though the algorithmic complexity worst-case analysis showed that it is of exponential complexity  $\Theta(c^n)$ , for practical problems when the number of input variables ( $n$ ) is reasonable and regardless of the granularity ( $c$ ) the learning quickly converges. However when  $n$  becomes very large, regardless of  $c$ , the learning will be faced by combinatory explosion as the number of possible hypotheses to generate risks to increase exponentially, this is known as the curse of dimensionality, and is a common problem for all AI approaches. To remedy such a problem we propose three solutions to attempt to optimize the code and reduce the computational learning complexity, the first solution consists to use pruning techniques either by cutting the search tree once the learning reach a satisfying solution (a level of accuracy) or by using some heuristics. Good heuristics that implements intelligent search (not exhaustive) are indeed cognitively motivated too. The second solution is to exploit the parallelism of the learning algorithm; there is a room for parallel implementation. In order to speed up the learning process, the learning may be easily implemented in a parallel machine of SIMD (single instruction multiple data) type, as the same algorithm is used for each output variable and for every granule of its fuzzy partition. In SIMD machines, all processors execute the same instruction stream on a different piece of data. This approach can reduce both hardware and software complexity.

Another promising alternative that constitutes the third solution is to use an evolutionary algorithms (EA) as in (Pedrycz 1997; Cordon et al. 2001), EAs are optimization techniques based on the mechanics of natural selection and natural genetics. EAs has a great power for global optimization and do not need to know the model previously. EAs also do not require the continuity of the parameters. Therefore EAs can easily handle the multi-parameter problems and for this reason it seems appealing and convenient to use EAs too in our framework. Thus an EA may replace the generator of hypotheses subsystem in our framework. Instead of using an exhaustive search to generate all possible fuzzy hypotheses to test, it may be possible to use an EA that converges to the "best" hypothesis by evolution rather than trying all possibilities. EAs can effectively contribute significantly in our framework thanks to their learning and optimization capabilities. In particular to try to fuzzify concepts used by EAs to obtain and use fuzzy fitness functions (or fuzzy cost), fuzzy crossover, fuzzy mutation and so on to ensure smooth evolvability during learning.

Even though we are more interested in (soft) computation rather than (natural) cognition, i.e. in developing new, powerful and useful tools that learn for resolving real-world problems, we believe that as we understand better how to build these computational systems we'll start to have theories that are powerful enough to explain some aspects of the human cognition.

## REFERENCES

- Beldjehem M. 1993. "Un apport à la conception des systèmes hybrides neuro-flous par algorithmes d'approximation d'équations de relations floues en MIN-MAX: le système Fennec." Ph.D. Thesis in Computer Science and Software Engineering, Université de la Méditerranée, Aix-Marseille II.
- Beldjehem M. 1993. "Fennec, un générateur de systèmes neuro-flous." in Proc. les Actes des Applications des Ensembles Flous, Nîmes, France, 209-218 (in French).
- Beldjehem M. 1993. "Le système fennec." in Electronic BUSEFAL 55, 95-104 (in French).
- Beldjehem M. 1994. "The fennec system." in Proc. ACM Symposium on Applied Computing (SAC), Track on fuzzy logic in Applications, 126-130, Phoenix, AZ (March).
- Beldjehem M. 2002. "Machine Learning based on the possibilistic-neuro hybrid approach: design and implementation." in Electronic BUSEFAL 87, 95-104.
- Beldjehem M. 2002. "Learning IF-THEN Fuzzy Weighted Rules." in Proc. International Conference of Computational intelligence, Nicosia, North Cyprus.
- Beldjehem M. 2006. "Validation of Hybrid MinMax Fuzzy-Neuro Systems." in Proc. International conference of NAFIPS, Montreal.
- Beldjehem M. 2008. "Towards a Validation Theory of Hybrid MinMax Fuzzy-Neuro Systems." in Proc. the WSEAS International Conference, Sofia.
- Beldjehem M. 2008. "A Validation Theory of Hybrid MinMax Fuzzy-Neuro Systems." in Proc. the CIMSA International Conference, Istanbul.
- Bellman R.; Kalaba R.; and Zadeh L. 1966. "Abstraction and Pattern Classification". J. Math. Anal. Appl. 13, 1-7.
- Cayrol M.; Farreny H.; and Prade H. 1982. "Fuzzy Pattern Matching." Kybernetics 11, 103-116.
- Cordon O.; Herrera F.; Hoffman F.; and Magdalena L. 2001. Genetic Fuzzy System, Evolutionary Tuning and Learning of Fuzzy Knowledge Bases, World Scientific.
- Dinola A.; Sessa S.; and Pedrycz W. 1989. Fuzzy Relation Equations and Their Applications To Knowledge Engineering, Kluwer Academic Publishers, Dordrecht, NL.
- Dubois, D. and Prade H. 1988. Possibility Theory: An Approach to Computerized Processing of Uncertainty, Plenum Press, New York, USA.
- Dubois D.; Prade H.; and Testemale C. 1988. Weighted fuzzy pattern matching, Fuzzy sets and systems 28, 313-331.
- Foster C. L. 1992. Algorithms, Abstraction and Implementation: Levels of details in cognitive sciences, Academic Press, London.
- Giunchiglia F. and Walsh T. 1992. "A theory of abstraction." Artificial Intelligence 56, 323-390.
- Gupta M.; Jin L.; and Homma N. 2002. Static and Dynamic Neural Networks: From Fundamentals to Advanced Theory, John Wiley and Son Inc., New York.
- Hobbs J. R. 1985. Granularity, Proceedings of the 9<sup>th</sup> International Joint Conference on Artificial Intelligence, 432-435.
- Liu T. Y.; Yao Y. Y.; and Zadeh L. A. 2002. "Data Mining, Rough Sets and Granular Computing.", Physica-Verlag, Heidelberg.
- Marr D. 1982. Vision, A Computational Investigation into Human Representation and Processing of Visual Information, Freeman W. H. and Company, San Francisco.
- Miller G. 1956. "The magical number seven, plus or minus two." The Psychological Review 63, 81-97.
- Newell A. and Simon, H. 1972. Human Problem Solving, Englewood Cliffs, NJ: Prentice Hall.
- Pappis C. P. 1991. "Value approximation of fuzzy systems variable." Fuzzy sets and systems 39, 111-115.
- Sinha N. K. and Gupta M. 1999( Ed.). Soft Computing and Intelligent Systems: Theory and Applications, Academic Press, New York.
- Yager R. 1986. Fuzzy Set and Possibility Theory: Recent Developments, Pergamon Press: New York.
- Yager, R. and Zadeh, L. A. 1994. Fuzzy Sets, Neural Networks and Soft Computing, Van Nostrand Reinhold: New York.
- Yager R. 1996. "On the interpretation of Fuzzy If-Then Rules," Applied Intelligence 6, 141-151.
- Yao Y. Y. 2000. Granular computing: basic issues and possible solutions, Proceedings of the 5th Joint Conference on Information Sciences, 186-189.
- Olaf W. 1998. Possibility Theory with Application to Data Analysis, John Wiley & Son inc.
- Pal S. K. and Ghosh A. (Ed.). 2000. Soft Computing in Image Processing, Physica-Verlag, Heidelberg.
- Pedrycz W. (Ed.). 1997. Fuzzy Evolutionary Computation, Kluwer Academic Publishers
- Pedrycz W. (Ed.). 2001. Granular Computing: An emerging paradigm, Springer. Series: Studies in Fuzziness and Soft Computing 70.
- Zadeh L. A. 1965. "Fuzzy sets." Info. Control 89, 338-353
- Zadeh L. A. 1971. "Toward a theory of fuzzy systems." in : R. E. Kalman, N. Declaris, Eds., Aspects of Network and System Theory (Holt, Rinehart and Winston, New York), 209-245.
- Zadeh L. A. 1973. "Outline of a new approach to the analysis of complex systems and decision processes." IEEE Trans. Syst. Man Cybernet. 3, 28-44.
- Zadeh L. A. 1976. "Fuzzy sets and information granulation." in Advances in Fuzzy Set Theory and Applications, in M. Gupta, R. K. Ragade, R. R. Yager (Eds.), North-Holland Publishing Company, 3-18.
- Zadeh L. A. 1978. "Fuzzy sets as a basis for a theory of possibility." Fuzzy sets and syst. 1, 3-28.
- Zadeh L. A. 1979. "A theory of approximate reasoning." in Machine Intelligence 9 (J.E. Hayes et al.; Eds). Elsevier, 149-194.
- Zadeh L. A. 1984. "A theory of commonsense knowledge." in Aspect of Vagueness (H.J. Skala, S. Termini and E. Trillas; Eds). Dordrecht: Reidel, 257-295.
- Zadeh L. A. 1994. "Fuzzy logic neural networks, and soft computing." Communications of the ACM 37, 77-84.
- Zadeh L. A. 1997. Towards a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic, Fuzzy Sets and Systems 19, 111-127.
- Zadeh L. A. 1998. "Some reflections on soft computing, granular computing and their roles in the conception, design and utilization of information / intelligent systems." Soft Computing 2, 23-25.
- Zadeh L. A. 1998. "Soft Computing, Fuzzy Logic and Recognition Technology." In Proc. IEEE Int. Conf. Fuzzy Syst., Anchorage, AK, 1678-1679.
- Zadeh L. A. 2001. "A new AI: Toward computational theory of perceptions." AAAI Magazine 22, No. 1, 73-84, Springer.

# PREDICTION OF FERROUS BIOOXIDATION RATE IN A PACKED BED BIOREACTOR USING ARTIFICIAL NEURAL NETWORK

Hasan Yousefi

Department of Mechanical Engineering,  
Lappeenranta University of Technology,  
Lappeenranta, Finland

S. Mohammad Mousavi

Biotechnology Group, Chemical  
Engineering Department, Tarbiat  
Modares University, Tehran, Iran  
Department of Chemical Technology,  
Lappeenranta University of  
Technology, Lappeenranta, Finland

Arezou Jafari

Laboratory of Engineering  
Thermodynamics, Lappeenranta  
University of Technology,  
Lappeenranta, Finland  
National Petrochemical Company,  
Tehran, Iran

Azita Soleymani

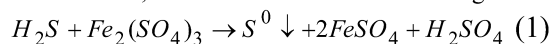
Department of Chemical Technology,  
Lappeenranta University of  
Technology, Lappeenranta, Finland

## ABSTRACT

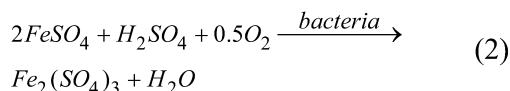
The biological oxidation of ferrous ion by iron oxidizing bacteria is potentially a useful industrial process for the removal of  $H_2S$  from industrial gases, desulphurization of coal, removal of sulfur dioxide from flue gas, the treatment of acid mine drainage and the regeneration of an oxidant agent in hydrometallurgical leaching operations. The main purpose of this study is to predict the ferrous biooxidation rate by immobilization of a native *Sulfobacillus* species on the LDPE particles in a packed-bed bioreactor using artificial neural network (ANN). Five control factors, including temperature, initial pH of feed solution, dilution rate, initial concentration of  $Fe^{+3}$  and aeration rate are considered in the experiments. One of the most powerful optimizers, Differential Evolution (DE) algorithm is used to find the best number of neurons for a hidden layer and their weights. The prediction results by using the proposed ANN are satisfactorily.

## INTRODUCTION

The use of microorganisms capable of oxidizing  $H_2S$  and producing elementary sulfur or sulfate from a complete and/or incomplete metabolism has been considered a potential alternative for the large-scale treatment of this gas (Ebrahimi et al. 2003; Oprime et al. 2001). In the bioprocess of  $H_2S$  removal an aqueous  $Fe_2(SO_4)_3$  solution is used as an absorbent.  $H_2S$  is absorbed and oxidized to elemental sulfur. At the same time,  $Fe^{3+}$  is reduced to  $Fe^{2+}$  according to

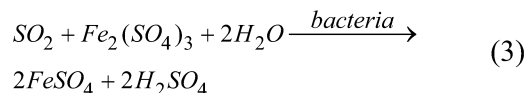


Elemental sulfur is removed from the solution by a separator, and the reactant  $Fe^{3+}$  is regenerated from  $Fe^{2+}$  by biological oxidation in an aerated bioreactor according to the following reaction:



Biological removal of sulfur dioxide from flue gas has also been reported in the literature (Gasiorek 1994). This process

is based on the wet scrubbing of the gas stream with a ferric sulfate solution:



The resultant ferrous sulfate solution is deoxidized to the ferric state, using iron oxidizing bacteria. The ferric sulfate solution produced is then recycled to the wet scrubbing tower to repeat the cycle.

The process of microbiological desulphurization have been applied for quality improvement of coals used as a fuel or a raw material in the chemical industry. In the nature the pyritic sulfur oxidation of coal is a process that happens quite slowly. This process can be accelerated in the presence of certain microorganisms. Biological desulphurization has attractions because it operates at close to ambient temperatures and involves no associated loss of coal carbon (Rubiera et al. 1997).

Mostly, in these biological processes the iron-sulfur bacteria species such as *A. ferrooxidans*, *A. thiooxidans* and *Sulfobacillus* are used (Fecko et al. 1991; Mousavi et al. 2006a; Mousavi et al. 2006b). These bacteria belong to the chemoautotrophs whose unique feature is the ability to receive energy required for life processes from inorganic sulfur compounds and carbon atoms required for cell build up from assimilated carbon dioxide.

In recent years, most studies (Long et al. 2004; Mesa et al. 2004; Mousavi et al. 2007) have been aimed at improving the rate of biooxidation of  $Fe^{2+}$ . Many types of reactors operating under both batch and continuous regimes have been studied in order to obtain better results using *A. ferrooxidans* and less attention has been paid to *Sulfobacillus* species, on the other hand, there is no scientific literature about the application of ANN to predict the ferrous iron biooxidation rate in the bioreactors.

A number of publications quote the successful application of ANN in various research fields such as process control and estimation (Jules et al. 1990; Yousefi and Handroos 2006; Yousefi et al. 2007), pattern recognition (Shuta et al. 2002), fault detection and property prediction (Jayshankar and Bhagwat 2002). Several authors have successfully implemented ANN in various areas of food and fermentation technology such as modelling and control (Ferriera et al. 2001; Kulkarni et al. 2004), dynamic modelling and state estimation (de Assis and Filho 2000; James et al. 2002), and optimization (Tholudur and Ramirez Fred 1996).

This paper describes a case study investigating the parameters that influence biooxidation rate of ferrous iron using a native *Sulfobacillus* species in a packed-bed bioreactor. Concentration of  $\text{Fe}^{+3}$  in effluent of bioreactor and rate of ferrous biooxidation are the key factors for evaluating the performance of bioreactor. Factors such as temperature, initial pH, dilution rate, initial  $\text{Fe}^{+3}$  concentration and rate of aeration affect the biooxidation rate of ferrous ion. The main objective was using of ANN approach to predict the biological oxidation rate of ferrous iron.

## MATERIALS AND METHODS

The details of materials and methods used in this study, were described in the previous work (Mousavi et al. 2007). Below a brief description has been provided.

The microorganism used in this study was originally isolated from the sphalerite concentrate of Kooshk lead and zinc mine. The bacterium was determined as *Sulfobacillus* species, which may be distinguished by its morphology, chromosomal DNA base compositions and by its abilities to grow autotrophically on reduced sulfur. These species are iron- and sulfur-oxidizing, gram-positive and sporulating rods (Norris et al. 1996). The composition of the medium for growth and maintenance of cells, was as follows:  $\text{FeSO}_4 \cdot 7\text{H}_2\text{O}$ : 44.2 g,  $(\text{NH}_4)_2\text{SO}_4$ : 3 g,  $\text{MgSO}_4 \cdot 7\text{H}_2\text{O}$ : 0.5 g,  $\text{K}_2\text{HPO}_4$ : 0.5g, KCl: 0.1 g,  $\text{Ca}(\text{NO}_3)_2$ : 0.01 g and Yeast Extract: 0.2 g in 1020 mL solution (Atlas, 1997). To culture the bacteria, 200 mL of the medium was transferred into a 500 mL Erlenmeyer flask and was incubated with *Sulfobacillus* culture, 10% (v/v), on a rotary shaker at 180 rpm and 60 °C. The initial pH was set to 1.5 with 1N  $\text{H}_2\text{SO}_4$  solution.

The biological oxidation was studied in a bioreactor shown in figure 1. Bioreactor design was based on a glass column with inlet for air and outlet for effluent at the bottom. The main part of bioreactor was biocatalyst bed with 7 and 45 cm in diameter and length, respectively. Total operating volume of bioreactor was about 2 L. The temperature of bioreactor was controlled using an external jacket. The reactor was aerated at different aeration rates and the flow rates for fresh media were regulated with a peristaltic pump during the experiments. To provide a uniform temperature inside the bioreactor and to increase the residence time of the reactant in the biocatalyst bed, part of the liquid collected in the collection container was re-circulated to the top of the bioreactor using a peristaltic pump at a flow rate of  $1.2 \text{ Lh}^{-1}$ .

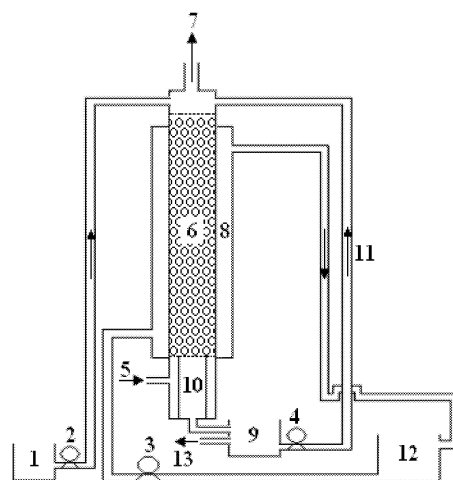


Figure 1: Schematic of packed-bed bioreactor used in this study. 1- fresh feed; 2,3,4- peristaltic pumps; 5- influent air; 6- packed bed; 7- effluent air; 8- jacket; 9- reservoir; 10- Particles support; 11- recycling stream; 12- water bath; 13- effluent solution.

*Sulfobacillus* cells were immobilized on LDPE particles of 3 mm diameter as support. The particles had a density of about  $930 \text{ kg m}^{-3}$ . Batch culture for the immobilization of cells was performed in 1000 mL Erlenmeyer flask containing 400 mL of mineral medium and 600 biomass support particles. The medium was inoculated with cell suspension, 10% (v/v), and incubated on a rotary shaker for 72 h at 150 rpm and 60 °C. Before complete consumption of ferrous iron had occurred, the spent medium was replaced by fresh medium followed by three consecutive runs without inoculation. After immobilization of cells in batch culture has been achieved to a constant level, support particles were placed in the bioreactor. The bioreactor influent solution contained ferrous sulfate, which was converted to ferric sulfate by bacteria present on the surfaces of particles. The bacteria were inoculated to the column while it was operated as a batch reactor. Once more than 95%  $\text{Fe}^{2+}$  oxidation was established, the reactor column was changed to a continuous mode of operation. Steady-state conditions were used at each flow rate for estimating the rate of ferrous iron oxidation. After a change in the flow rate, steady-state conditions were achieved when no further change occurred in the iron oxidation rate. The time required to achieve steady-state conditions at each flow rate varied depending on the flow rate. Experiments were performed in four different levels of dilution rate (based on the total volume of the liquid in the bioreactor). It should be mentioned that the concentration of ferrous iron in the bioreactor influent solution was adjusted to 12 g/L for all of experiments.

Determination of ferric and total iron concentration in bacterial solutions was based on the method described by Karamanev et al. (2002). Difference between concentrations of total iron and ferric iron led to obtain ferrous concentration in the solution. The observation of free bacteria in the solution was done by visual count, using a Thoma chamber with an optical microscope. The pH of the cultural suspensions was monitored at room temperature with a pH meter calibrated with a low pH buffer.

## NEURAL NETWORK DESIGN

An artificial neural network is used to improve the reliability of the bioreactor behaviour's prediction. The obtained neural network can be used to optimize the conditions for bioreactor's inputs to get the best performance. ANNs provide an approximate model structure to fit the experimental data.

An ANN consists of massively interconnected nonlinear memoryless processing elements known as neurons or nodes. The strength of the connections between the neurons is called the weight. Each neuron accepts a weighted set of inputs with a bias given by (Rumelhart et al. 1986)

$$n = \sum_{i=1}^P w_i x_i + b \quad (4)$$

where  $P$  and  $w_i$  are the number of elements and the interconnection weights of the input vector  $x_i$ , respectively, and  $b$  is the bias for the neuron. Note that the knowledge is stored as a set of connection weights and biases, which have to be adjusted in order to allow the network to perform a required task. Then, the neuron responds with an output. For this aim, the sum of the weighted inputs is processed through an activation function, represented by  $f$ , and the output that it computes is

$$f(n) = f\left(\sum_{i=1}^P w_i x_i + b\right) \quad (5)$$

Basically, the neuron model emulates the biological neuron that fires when its inputs are significantly excited, i.e.  $n$  is big enough. There are many ways to define the activation function, such as a threshold function, log-sigmoid function and hyperbolic tangent sigmoid function. One of the most commonly used functions satisfying these requirements is the hyperbolic tangent function as follow;

$$f(x) = (e^x - e^{-x}) / (e^x + e^{-x}) \quad (6)$$

Using a suitable learning method, ANNs can be trained to perform a particular function by adjusting the values of connections, i.e. weighting coefficients, between the processing nodes. The training process continues until the network output matches the target. The error between the output of the network and the desired output is minimized by modifying the weights. When the error falls below a predetermined value or the maximum number of epochs is exceeded, the training process is terminated. Then, this trained network can be used for simulating the system outputs for the inputs that have not been introduced before. The architecture of an ANN is usually divided into three parts: an input layer, hidden layers and an output layer. The information contained in the input layer is mapped to the output layer through the hidden layers. Each neuron can receive its input only from the lower layer and send its output to the neurons only on the higher layer.

The performance of the ANN based prediction is evaluated by a regression analysis between the network outputs, i.e. predicted parameters, and the corresponding targets, i.e. experimental values. The criteria used for measuring the network performance are the root mean square error and absolute fraction of variance. The root mean square error is given by

$$RMSE = \sqrt{(1/NP) \sum_{i=1}^{NP} (a_i - p_i)^2}, \quad (7)$$

where,  $RMSE$  is the root mean square error,  $NP$  is the population size of both sets  $a$  and  $p$ . Finally, the absolute fraction of variance (regression constant), a statistical indicator that can be applied to multiple regression analysis, is determined from

$$R^2 = 1 - \left( \frac{\sum_{i=1}^{NP} (a_i - p_i)^2}{\sum_{i=1}^{NP} p_i^2} \right), \quad (8)$$

The regression constant ranges between 0 and 1. A very good fit yields an  $R^2$  value of 1, whereas a poor fit results in a value near 0.

Utilizing a standard back propagation algorithm the input vectors with five variables and the corresponding target vector with variables from the training set were introduced to the network for training. The training procedure adjusted the weighting coefficients using the Differential Evolution algorithm for Global optimization. This algorithm is a widely used iterative optimization technique that locates the global minimum of a function. The output of the network was compared to the desired output at each presentation, and an error was computed. The training process is terminated when the maximum number of epochs is exceeded or the performance goal is met. Figure 2 shows the schematic structure of neural network. In the figure the biases are not shown. The five inputs are temperature, initial pH of feed solution, dilution rate, initial concentration of  $Fe^{+3}$  and aeration rate gotten from experimental tests that are shown as A, B, C, D and E respectively. The output of system is biooxidation rate of ferrous iron shown as F. Table 1 shows the maximum and minimum values for the inputs. The hidden layer consists of 15 neurons, so the total number of proposed neural network has 106 weights and biases.

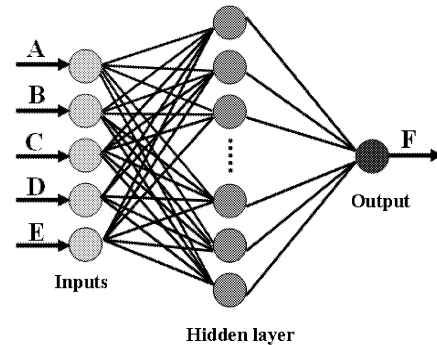


Figure 2: The structure of the ANN for modelling the experimental bioreactor

Table 1: Maximum and minimum values of inputs

| Parameter | Units                             | Min Value | Max Value |
|-----------|-----------------------------------|-----------|-----------|
| A         | °C                                | 40        | 65        |
| B         | -                                 | 0.5       | 3         |
| C         | h <sup>-1</sup>                   | 0.1       | 0.5       |
| D         | g L <sup>-1</sup>                 | 1         | 15        |
| E         | mL min <sup>-1</sup>              | 50        | 300       |
| F         | g L <sup>-1</sup> h <sup>-1</sup> | 1.1       | 7.8       |

The performance of an ANN is affected by the characteristics of the network, such as the number of hidden layers and the number of nodes in each hidden layer. For

instance, too few neurons may yield underfitting, but too many neurons may result in overfitting, which means that all the training data fit well, but the neural network does not satisfactorily predict new test data that have not been presented in the training process. On the other hand, there are no definite methods to determine the optimal number of hidden layers and the optimal number of neurons on each hidden layer. Therefore, by trial and error with different ANN configurations, the network was selected to consist of one hidden layer with 15 neurons along with input and output layers. Choosing the weights is important in the ANN. To avoid the local minimum problem, the differential evolution (DE) algorithm is used to find the offline weights (Storn and Price 1995).

Differential evolution is a simple and powerful population based, direct-search algorithm for globally optimizing functions defined especially on functions with real-valued parameters. In this paper a class of DE presented by Come et al. (1999) is used.

## RESULTS AND DISCUSSION

In order to develop the ANN, the available data set from the experimental work was divided into training and test sets. The data set consists of 59 input-output pairs. While 84% of the data (50 input-output pair) set was randomly assigned as the training set, the remaining (9 input-output pair) 16% was employed for testing the network.

The activation function in the hidden layer was chosen as the tangent sigmoid function. All the input and output values were normalized by pre-processing so that they fall in the interval  $[-1, 1]$  using the following equation:

$$\rho_n = 2(\rho - \min(\rho)) / (\max(\rho) - \min(\rho)) - 1, \quad (9)$$

where  $p_n$  is the normalized form of the vector  $p$ .

After the training process has been finished, the network is ready for prediction. Then, the input vectors from the test data set were presented to the trained network. The responses of the network, i.e. the predicted output, were compared with the experimental ones for the performance measurement. The computer code solving the back propagation algorithm and measuring the network performance was implemented under the Matlab/ Simulink environment. Figure 3 shows the schematic diagram of the procedure.

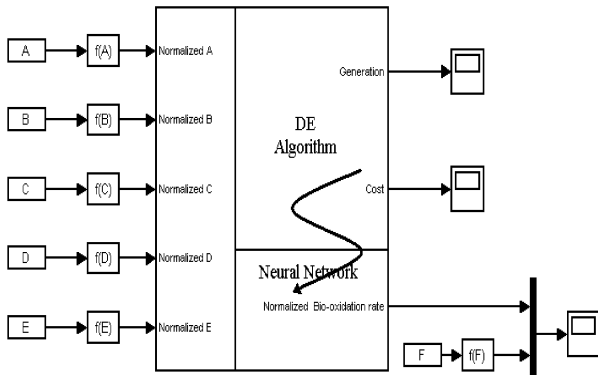


Figure 3: Schematic diagram of training utilizing differential evolution algorithm

The initial upper and lower bounds are  $[1, -1]$ , respectively. Note that generally the number of the population ( $NP$ ) is five times the number of unknown parameters ( $D$ ). Here, because of a computer memory saturation problem, the maximum  $NP$  was 250. The results show that  $DE$  can find the global minimum cost for the system. Depending on the expected value of the cost and its absolute fraction of variance, the  $DE$  finds the proper weights. The cost of the system is defined as follows:

$$F_{Co}(z_{i,G}) = \sum_{k=1}^{k=50} e(k)^2, \quad (10)$$

$$e(k) = F_{Target}(k) - F_{ANN}(k), \quad (11)$$

where  $F_{Target}(k)$  and  $F_{ANN}(k)$  are the outputs of real test and ANN for each training data.

The number of neurons for hidden layer was ascertained by trial and error, and the most suitable number of neurons was chosen in such a way that the training results were converge out to the experimental data. Figure 4 shows that the best number of neurons for ANN which is a network with 15 neurons in hidden layer. The proposed ANN gives results in good agreement with the experimental data.

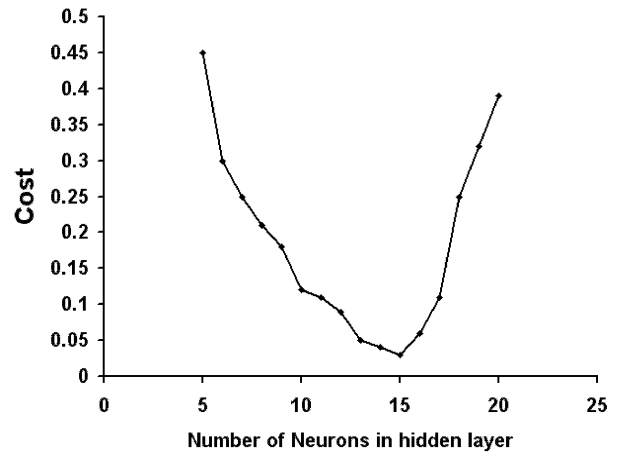


Figure 4: Costs for training data with various numbers of neurons in the hidden layer of ANN

Figures 5 and 6 show the training set and test set results, respectively. There are very good agreements between the normalized predicted data by ANN and normalized experimental data. On the basis of such good trainings, the resulting ANN are capable enough to simulate the other half of data by their application to the relevant networks gave such simulated data as depicted. In these figures, normalized simulated and experimental data are compared. An excellent agreement can be seen in between them. The equation in the form of  $Y=AX+B$  appearing in figure 5 and 6 is the equation of the adapted least regression line with best state as  $Y=X$  happening when all the points fall exactly on a line at  $45^\circ$ , i.e. the network predicts results exactly the same as the experimental ones. Constants of the equation however, show the deviation of the state from the ideal one. In addition to the equation, the regression constants ( $R^2$ -value) which are also appeared in these figures show the agreement of trained and simulated data with experimental data. In an ideal situation, when these parameters are exactly similar,  $R^2 = 1$ .

Figure 7 shows all the predicted data (59 input-output pairs) and their experimental data. The figure shows a very good agreement for all data set.

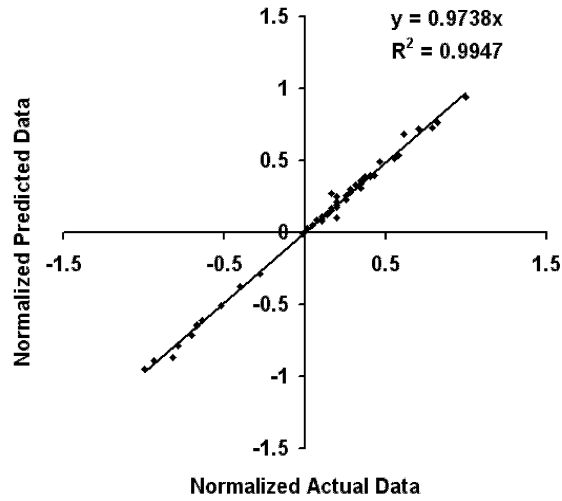


Figure 5: The proposed ANN training result

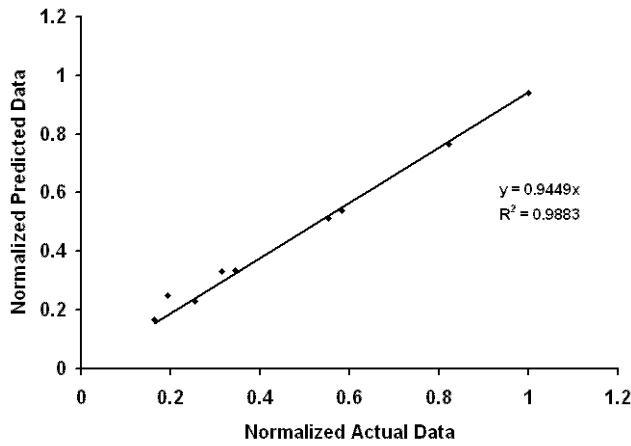


Figure 6: The proposed ANN test result

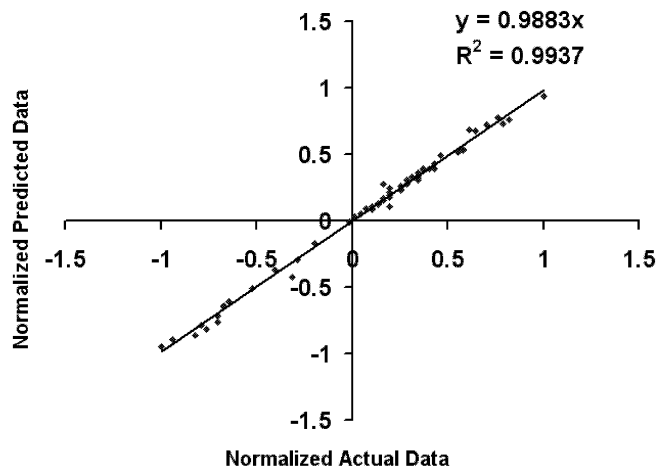


Figure 7: The ANN all data result

The Global minimum cost for the proposed ANN was 0.034 and the related weights are used as final values of weights

for predicting the biooxidation rate by ANN. Figure 8 is the plot of the cost when the differential evolution algorithm is used to search for the global minimum of the defined cost. The figure shows that the total cost after 28732 generations converges and remains constant, so the related weights are the best value for the proposed ANN.

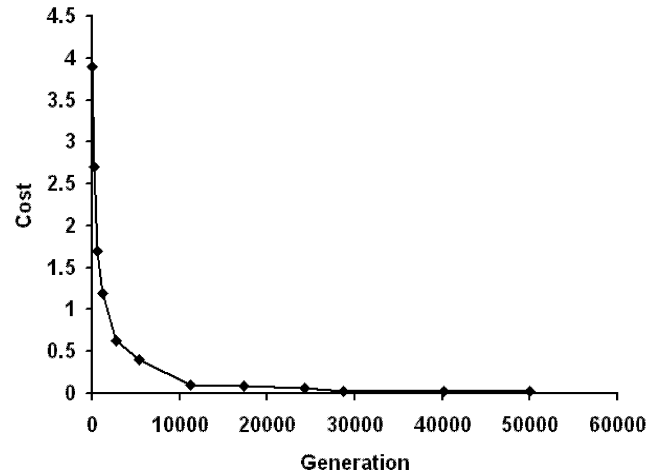


Figure 8: Training costs of Proposed ANN using Differential Evolution algorithm

Figure 9 is used to verify the robustness of the obtained weight. The figure show a sample weight of neural network against the number of generations. It is clear that the amount of the sample weight converges to its final value (0.86) and remains constant.

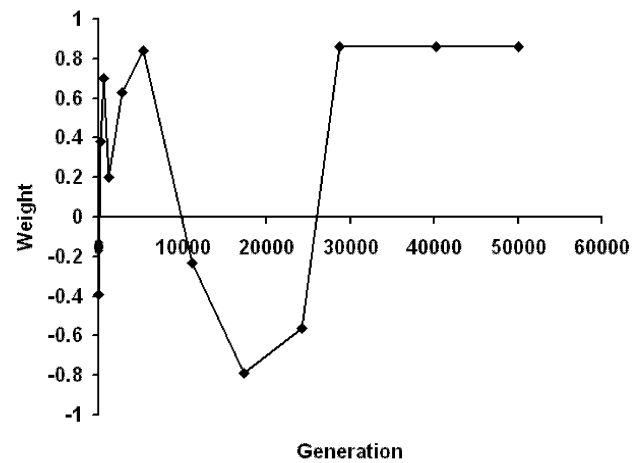


Figure 9: Converging of a sample weight of Proposed ANN using Differential Evolution algorithm

## CONCLUSION

The applicability and capability of the ANN modelling approach for predicting the biooxidation rate of ferrous iron in the presence of *Sulfobacillus* species in a packed-bed bioreactor was investigated. For this aim, totally 59 test runs covering a wide range of operating conditions were performed on an experimental system. Then, an ANN model based on an artificial neural network was developed to predict various performance parameters. Based on the five input parameters, the trained ANN model was used for

predicting the performance of the system in terms of the temperature, initial pH of feed solution, dilution rate, initial concentration of  $\text{Fe}^{+3}$  and aeration rate. The performance of the ANN predictions was measured using the root mean square error and absolute fraction of variance. The ANN model usually demonstrated a good statistical performance with the absolute fractions of variance in the range of 0.9883–0.9947. The best number of hidden layer neurons and their related values for the weights of neural network were investigated by using differential evolution algorithm. The results show a good agreement between predicted biooxidation rates of ferrous iron and experimental data. The robustness of proposed weights of the neural network was examined. This study reveals that biooxidation rate can alternatively be modelled using ANN within a high degree of accuracy.

## REFERENCES

- Atlas, R.M. 1997. "Handbook of Microbiological Media". Second ed., CRC Press, Boca Raton.
- Corne, D.; M. Dorigo; and F. Glover. 1999. "New Ideas in Optimization", McGraw-Hill.
- de Assis A.J. and R.M. Filho. 2000. Soft Sensors Development for On-Line Bioreactor State Estimation." *Comp. Chem. Eng.* 24, 109–110.
- Ebrahimi, S.; R. Kleerebezem; M.C.M. van Loosdrecht; and J.J. Heijnen. 2003. "Kinetics of the Reactive Absorption of Hydrogen Sulfide into Aqueous Ferric Sulfate Solutions." *Chem. Eng. Sci.* 58, 417–427.
- Fecko, P.; H. Raclavska; and V. Malysiak. 1991. "Desulphurisation of Coal from Northern Bohemian Brown Coal Basin by Bacterial Leaching." *Fuel* 70, 1187–1191.
- Ferreira, L.S.; M.B. De Souza Jr; and R.O.M. Folly. 2001. "Development of an Alcohol Fermentation Control System Based on Biosensor Measurement Interpreted by Neural Networks." *Sens. Actuators* 75, 166–171.
- Gasiorek, J. 1994. "Microbial Removal of Sulfur Dioxide from a Gas Stream." *Fuel Processing Technol.* 40, 129–138.
- James, S.; R. Legge; and H. Buddman. 2002. "Comparative Study of Black Box and Hybrid Estimation Methods in Fed Batch Fermentation." *J. Process Control* 12, 113–121.
- Jayshankar, P.Y. and S.S. Bhagwat. 2002. "Simple Neural Network Model for Prediction of Physical Properties of Organic Compounds." *Chem. Eng. Technol.* 25, No.11, 1041–1046.
- Jules, T.; V.B. Vincet; and C. Arlette. 1990. "On Line Prediction of Fermentation Variable Using Neural Network." *Biotechnol. Bioeng.* 36, 1041–1048.
- Karamanev, D.G.; L.N. Nikolov; and V. Mamartarkova. 2002. "Rapid Simultaneous Quantitative Determination of Ferric and Ferrous Ions in Drainage Waters and Similar Solutions." *Minerals Engineering* 15, 341–346.
- Kulkarni, S.G.; A.K. Chaudhary; S. Nandi; S.S. Tambe; and B.D. Kulkarni. 2004. "Modeling and Monitoring of Batch Processes Using Principal Component Analysis (PCA) Assisted Generalized Regression Neural Networks (GRNN)." *Biochem Eng. J.* 18, 193–201.
- Long, Z.; Y. Huang; Z. Cai; W. Cong; and F. Ouyang. 2004. "Immobilization of *Acidithiobacillus Ferrooxidans* by a PVA–Boric Acid Method for Ferrous Sulphate Oxidation." *Process Biochem.* 39, 2129–2133.
- Mesa, M.M.; J.A. Andrades; M. Mac'ias; and D. Cantero. 2004. "Biological Oxidation of Ferrous Iron: Study of Bioreactor Efficiency." *J. Chem. Technol. Biotechnol.* 79, 163–170.
- Mousavi, S.M.; S. Yaghmaei; and A. Jafari. 2007. "Influence of Process Variables on Biooxidation of Ferrous Sulfate by an Indigenous *Acidithiobacillus Ferrooxidans*. Part II: Bioreactor Experiments." *Fuel* 86, No.7-8, 993–999.
- Mousavi, S.M.; S. Yaghmaei; F. Salimi; and A. Jafari. 2006a. "Influence of Process Variables on Biooxidation of Ferrous Sulfate by an Indigenous *Acidithiobacillus Ferrooxidans*. Part I: Flask Experiments." *Fuel* 85, No.17-18, 2555–2560.
- Mousavi, S.M.; S. Yaghmaei; M. Vossoughi; A. Jafari; and R. Roostaazad. 2006b. "Zinc Extraction from Iranian Low-Grade Complex Zinc–Lead Ore by Two Native Microorganisms: *Acidithiobacillus Ferrooxidans* and *Sulfobacillus*." *Int. J. Miner. Process.* 80, 238–243.
- Norris, P.R.; D.A. Clark; J.P. Owen; and S. Waterhouse. 1996. "Characteristics of *Sulfobacillus Acidophilus* sp. nov. and other Moderately Thermophilic Mineral Sulphide-Oxidizing Bacteria." *Microbiology* 141, 775–783.
- Oprime, M.E.A.G.; O. Garcia Jr.; and A.A. Cardoso. 2001. "Oxidation of  $\text{H}_2\text{S}$  in Acid Solution by *Thiobacillus Ferrooxidans* and *Thiobacillus Thiooxidans*." *Process Biochem.* 37, 111–114.
- Rubiera, F.; A. Moran; O. Martinez; E. Fuente; and J. Pis. 1997. "Influence of Biological Desulphurisation on Coal Combustion Performance." *Fuel Processing Technol.* 52, 165–173.
- Rumelhart, D.E.; G.E. Hinton; and R.J. Williams. 1986. "Learning Internal Representations: by Error Propagation." *Parallel Distributed Processing: Explorations in the Microstructures of Cognition.* 1, 318–362.
- Storn, R. and K. Price. 1995. "Differential Evolution - A Simple and Efficient Adaptive Scheme for Global Optimization over Continuous Spaces". Technical Report TR-95-012, Berkeley, CA.
- Shuta, T.; H. Taizo; K. Naoki; S. Youichi; K. Takeshi; and H. Hiroyuki. 2002. "Artificial Neural Network Predictive Models for Allelic Disease Using Single Nucleotide Polymorphism Data." *J. Biosci. Bioeng.* 93, No.5, 470–478.
- Tholudur A. and W. Ramirez Fred. 1996. "Optimization of Fed-Batch Bioreactor Using Neural Network Parameter Function Models." *Biotechnol. Progress.* 12, 302–309.
- Yousefi, H.; M. Hirvonen; H. Handroos; and A. Soleymani. 2007. "Application of Neural Network in Suppressing Mechanical Vibrations of a Permanent Magnet Linear Motor." In press, *Control Engineering Practice*, (doi:10.1016/j.conengprac.2007.08.003).
- Yousefi, H. and H. Handroos. 2006. "Experimental and Simulation Study on Control of a Flexible Servo-Hydraulic System Using Adaptive Neural Network and Differential Evolution Strategy". *8th Biennial ASME Conference on Engineering Systems Design and Analysis (ESDA 2006)*. Torino, Italy.



# Extension of Rank Based Ant System with Exponential Pheromone Deposition for Speed-up and Improved Accuracy

Ayan Acharya  
Aritra Banerjee  
Amit Konar

Department of Electronics and Telecommunication Engineering  
Jadavpur University  
Kolkata: 700032  
India

E-mail: {masterayan| aritraetce}@gmail.com, konaramit@yahoo.co.in

Mokhtar Beldjehem  
École Polytechnique de Montréal  
Campus de l'Université de Montréal  
Montréal QC H3C 3A7,  
Canada  
Email: mokhtar.beldjehem@polymtl.ca

## KEYWORDS

Ant Colony Optimization, Ant System, Rank Based Ant System, Stability Analysis, Exponential Pheromone Deposition

## ABSTRACT

The paper introduces a novel pheromone deposition approach to improve the performance of traditional ant system algorithm that employ constant deposition rule. We select the rank based ant system model, an extension of basic ant system algorithm, to compare the proposed deposition rule with the standard one. A simplified analysis of basic ant system dynamics is carried out to find the parameter range for system stability in both kind of deposition approach. A roadmap of connected cities where the shortest route between two given cities is to be found out is chosen as a problem environment and extensive simulations are performed to find the ranges of major controlling parameters for best performance of the proposed approach. Experiments reveal that our method, with this empirically obtained optimum parameter set, outstrips its traditional counterpart by a large margin. Finally, we attempt to establish an algebraic relationship between the parameter set of the algorithm and the feature set of the problem environment.

## INTRODUCTION

**Ant Colony Optimization (ACO)** is a paradigm for designing metaheuristic algorithms for combinatorial optimization problems. While roaming from food sources to the nest and vice versa, ants deposit on the ground a substance called *pheromone*. Ants can smell pheromone and choose, in probability, paths marked by stronger pheromone concentration. Hence, the pheromone trail allows the ants to find their way back to the food source or to the nest. ACO algorithms simulate this behavior of ant colony to solve difficult NP hard optimization problems.

**Ant System (AS)** is the earliest form of ant colony optimization algorithm that has been modified by numerous researchers till date. **Rank Based Ant System (AS<sub>rank</sub>)** model is one such improved model. Our paper extends the AS model by introducing an exponential pheromone deposition approach, contrary to the uniform deposition approach used in classical AS algorithms. We attempt to solve the deterministic AS dynamics using differential equation. The analysis helps in determining the range of parameters in the exponential pheromone deposition rule to confirm stability in pheromone trails. The deterministic solution to the AS dynamics undertaken here does not violate the stochastic nature of the AS because a segment of trajectory here is always selected probabilistically.

A uniform pheromone deposition by an ant cannot ensure subsequent ants to follow the same trajectory. A non-uniform non-decreasing time function, however, ensures that subsequent ants close enough to a previously selected trial solution will follow the trajectory, as it can examine gradually thicker deposition of pheromones over the trajectory. Naturally, *deception probability* (D.Merkle and M. Middendorf 2002) of the ants will be less, consequently improving expected convergence time and final solution.

The paper is structured in 6 sections. In next section, a brief introduction of AS and AS<sub>rank</sub> is provided. We formulate a scheme for the general solution of the Ant System in third section. Stability analysis with complete solution to ant system dynamics for different pheromone deposition rules is undertaken in fourth section. Comparative study of the proposed and classical AS<sub>rank</sub> is carried out in the penultimate section. Conclusions are listed in final section.

## ANT SYSTEM AND RANK BASED ANT SYSTEM: A REVIEW

The theory of ant system can best be explained in the context of TSP (M.Dorigo and L.M. Gambardella. 1997). The basic ACO algorithm for TSP can be described as follows:

**procedure** ACO algorithm for TSPs

- Set parameters, initialize pheromone and ants' memory
- while** (termination condition not met)
- Construct Solution
- Apply Local Search ( optional)
- Best Tour check
- Update Trails

**end**

**end** ACO algorithm for TSPs

**Ant System (AS)** (C. Blum and M. Dorigo 2005; M.Dorigo *et al* 1996) was the earliest implementation of the ACO algorithm. Basically it consists of two levels:

1. **Initialization:** 1.Any initial parameters are loaded. 2. Edges are set with an initial pheromone value. 3. Each ant is individually placed on a random city.
2. **Main Loop:**

**Construct Solution:** Each ant constructs a tour by successively applying the probabilistic choice function which can be described as follows:

$$P_i^k(j) = \begin{cases} (\tau_{ij}^\alpha)(\eta_{ij}^\beta) / \sum_{k: k \in N_i^k} (\tau_{ik}^\alpha)(\eta_{ik}^\beta) & \text{if } q < q_0 \\ 1 & \text{if } (\tau_{ij}^\alpha)(\eta_{ij}^\beta) = \max \{ (\tau_{ik}^\alpha)(\eta_{ik}^\beta) : k \in N_i^k \} \text{ with } q > q_0 \\ 0 & \text{if } (\tau_{ij}^\alpha)(\eta_{ij}^\beta) \neq \max \{ (\tau_{ik}^\alpha)(\eta_{ik}^\beta) : k \in N_i^k \} \text{ with } q > q_0 \end{cases} \quad (1)$$

where  $P_i^k(j)$  is the probability of selecting node  $j$  after node  $i$  for ant  $k$ . A node  $j \in N_i^k$  ( $N_i^k$  being the neighborhood of ant  $k$  when it is at node  $i$ ) if  $j$  is not already visited.  $\eta_{ik}$  is the

visibility information generally taken as the inverse of the length of link  $(i,k)$  and  $\tau_{ik}$  is the pheromone concentration associated with link  $(i,k)$ .  $q_0$  is a pseudo random factor deliberately introduced for path exploration.  $\alpha, \beta$  are the weights for pheromone concentration and visibility.

- **Best Tour check:** Calculate the lengths of the ants' tours and compare with best tour length so far. If there is an improvement, update it.
- **Update Trails:** 1. Evaporate a fixed proportion of the pheromone on each edge. 2. For each ant perform the 'Ant Cycle' pheromone update.

Now, let us consider a small segment of the tour by an ant. Let  $i$  and  $j$  be two successive nodes, on the tour of an ant and  $\tau_{ij}(t)$  be the pheromone concentration at time  $t$  associated with the edge of the graph joining the nodes  $i$  and  $j$ .

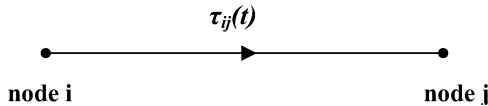


Fig. 1: Defining  $\tau_{ij}(t)$

Let  $\rho > 0$  be the pheromone evaporation rate and  $\Delta\tau_{ij}^k(t)$  be the pheromone deposited by ant  $k$  at time  $t$ . The basic pheromone updating rule in AS is then given by,

$$\tau_{ij}(t) = (1-\rho)\tau_{ij}(t-1) + \sum_{k=1}^m \Delta\tau_{ij}^k(t) \quad (2)$$

In  $AS_{rank}$  algorithm, introduced by Bullnheimer et al in 1999, each ant deposits an amount of pheromone that decreases with its rank. Additionally, the best-so-far ant is allowed to deposit the largest amount of pheromone. The pheromone updating rule (2) is therefore modified as,

$$\tau_{ij}(t) = (1-\rho)\tau_{ij}(t-1) + \sum_{r=1}^{w-1} (w-r)\Delta\tau_{ij}^r(t) + w\Delta\tau_{ij}^{bs} \quad (3)$$

where  $\Delta\tau_{ij}^r$  is the amount of pheromone deposited by ant of rank  $r$  on the arcs it has visited and is defined as follows:

$$\Delta\tau_{ij}^r = \begin{cases} 1/C^r, & \text{if arc (i,j) belongs to } T^r \\ 0, & \text{otherwise} \end{cases}, \quad C^r \text{ being the length}$$

of the tour  $T^r$  constructed by ant of rank  $r$ .  $\Delta\tau_{ij}^{bs}$  in (3) is defined as  $\Delta\tau_{ij}^{bs} = \begin{cases} 1/C^{bs}, & \text{if arc (i,j) belongs to } T^{bs} \\ 0, & \text{otherwise} \end{cases}$ , where

$C^{bs}$  is the tour length of the best-so-far tour  $T^{bs}$ . In each iteration, only  $(w-1)$  best ranked ants and best-so-far ant is allowed to deposit pheromone.

The other two algorithms which achieve superior performance compared to AS are **Elitist Ant System** (M.Dorigo et al 1991; M.Dorigo et al 1996) and **Max-Min Ant System** (T. Stützle and H. H. Hoos. 2000). **Ant Colony System** (ACS) (M.Dorigo and L.M. Gambardella. 1997), **Approximate Non-deterministic Tree Search** (ANTS) (V. Maniezzo. 1999) and **Hyper Cube Framework for ACO** (C. Blum et al. 2001), on the other hand, achieve improvement over AS by modifying its basic structure.

There is another set of algorithm of ACO where amelioration is achieved by introducing newer kind of pheromone evaluation strategy. In general, it is assumed that pheromone evaluation is done only locally by ants. Michels and Middendorf (1999) first extended this local view of ants by introducing a look-forward strategy. Here, every ant takes also the quality of the next possible decisions into account. The first global pheromone evaluation strategy was proposed

later on by Merkle and Middendorf (2000). They termed the strategy as summation evaluation which takes into account all the former and next possible decisions along with the local one. Merkle et al (2000) later showed how a combination of local pheromone evaluation and weighted summation evaluation strategy can help improve the performance even further. We, however, exclude detailed discussion of these algorithms and will concentrate only on AS and  $AS_{rank}$  algorithm in particular.

## DETERMINISTIC FRAMEWORK FOR SOLUTION OF BASIC ANT SYSTEM DYNAMICS

This section focuses on the development of deterministic framework using differential equation. It follows from (2),

$$\begin{aligned} \tau_{ij}(t) - \tau_{ij}(t-1) &= -\rho\tau_{ij}(t-1) + \sum_{k=1}^m \Delta\tau_{ij}^k(t) \\ \Rightarrow \frac{d\tau_{ij}}{dt} &= -\rho\tau_{ij} + \sum_{k=1}^m \Delta\tau_{ij}^k(t) \quad \therefore (D+\rho)\tau_{ij} = \sum_{k=1}^m \Delta\tau_{ij}^k(t) \quad (4) \end{aligned}$$

where,  $D \equiv d/dt$  is the differential operator.

Evidently, (4) gives the solution for the ant dynamics. Now, to solve (4), we have to separate the complimentary function and the particular integral. We consider two different forms of  $\Delta\tau_{ij}^k(t)$  corresponding to both forms of deposition rule and try to determine the complete solution of  $\tau_{ij}(t)$ .

### Evaluation of Complimentary Function (CF):

The complimentary function of (4) is obtained by setting  $\sum_{k=1}^m \Delta\tau_{ij}^k(t)$  to zero. This gives only the transient behavior of the ant system dynamics. Therefore, from (4),

$$(D+\rho)\tau_{ij} = 0, \Rightarrow D = -\rho$$

Thus, the transient behavior of the Ant System is given by

$$CF: \tau_{ij}(t) = Ae^{-\rho t} \quad (5)$$

where  $A$  is a constant which is to be found out from initial condition.

### Evaluation of Particular Integral for Both Forms of Deposition Rule:

The steady state solution of the ant system dynamics is obtained by computing particular integral of (4).

Case I:  $\Delta\tau_{ij}^k(t) = C_k$ , constant, where  $C_k > 0$ ,

Case II:  $\Delta\tau_{ij}^k(t) = C_k(1 - e^{-t/T})$ , where  $C_k > 0$  and  $T > 0$ .

The particular integral (PI) for the Ant System can be obtained from (4). This is given by,

$$\tau_{ij} = \frac{1}{D+\rho} \sum_{k=1}^m \Delta\tau_{ij}^k(t) \quad (6)$$

Case I: When  $\Delta\tau_{ij}^k(t) = C_k$ , we obtain from (6)

$$\begin{aligned} PI &= \frac{1}{D+\rho} \sum_{k=1}^m C_k = \frac{1}{\rho} \cdot (1 + D/\rho)^{-1} \sum_{k=1}^m C_k \\ &= \frac{1}{\rho} \left(1 - \frac{D}{\rho} + \frac{D^2}{\rho^2} - \dots\right) \sum_{k=1}^m C_k = \frac{1}{\rho} (1) \sum_{k=1}^m C_k = \sum_{k=1}^m C_k / \rho \quad (7) \end{aligned}$$

Case II: When  $\Delta\tau_{ij}^k(t) = C_k(1 - e^{-t/T})$ , we obtain from (6),

$$PI = \frac{1}{D+\rho} \sum_{k=1}^m C_k (1-e^{-t/T}) = \frac{1}{D+\rho} \sum_{k=1}^m C_k \frac{1}{D+\rho} \sum_{k=1}^m C_k e^{-t/T}$$

$$= \sum_{k=1}^m C_k / \rho - \frac{1}{D+\rho} \sum_{k=1}^m C_k e^{-t/T} = \sum_{k=1}^m C_k / \rho - \sum_{k=1}^m C_k e^{-t/T} / (\rho - \frac{1}{T}) \quad (8)$$

## STABILITY ANALYSIS OF ANT SYSTEM DYNAMICS WITH COMPLETE SOLUTION

In this section, we obtain the complete solution of the ant system dynamics for determining the condition for stability of the dynamics.

**Case I:** For constant deposition rule, the complete solution can be obtained by adding CF and PI from (5) and (7)

respectively and is given by,  $\tau_{ij}(t) = Ae^{-\rho t} + \sum_{k=1}^m C_k / \rho$ .

$$\text{At } t=0, \tau_{ij}(0) = A + \sum_{k=1}^m C_k / \rho, \Rightarrow A = \tau_{ij}(0) - \sum_{k=1}^m C_k / \rho$$

Therefore, the complete solution is,

$$\tau_{ij}(t) = [\tau_{ij}(0) - \sum_{k=1}^m C_k / \rho] e^{-\rho t} + \sum_{k=1}^m C_k / \rho \quad (9)$$

It follows from (9) that the system is stable for  $\rho > 0$  and converges to steady state value  $\sum_{k=1}^m C_k / \rho$  as time increases.

The plot below supports the above observation.

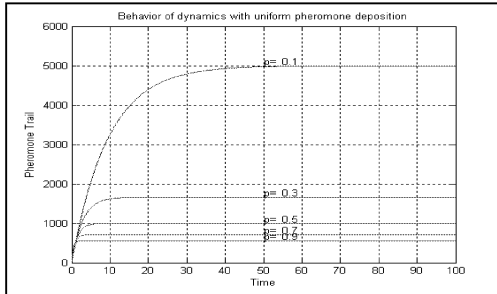


Fig.2.  $\tau_{ij}(t)$  versus  $t$  for constant pheromone deposition

**Case II:** For exponentially increasing pheromone deposition, the complete solution is,

$$\tau_{ij}(t) = Ae^{-\rho t} + \sum_{k=1}^m C_k / \rho - \sum_{k=1}^m C_k e^{-t/T} / (\rho - \frac{1}{T})$$

$$\text{Now, at } t=0, \tau_{ij}(0) = A + \sum_{k=1}^m C_k / \rho - \sum_{k=1}^m C_k / (\rho - \frac{1}{T})$$

$$\therefore A = \tau_{ij}(0) - \sum_{k=1}^m C_k / \rho + \sum_{k=1}^m C_k / (\rho - \frac{1}{T})$$

$$\therefore \tau_{ij}(t) = [\tau_{ij}(0) - \sum_{k=1}^m \frac{C_k}{\rho} + \sum_{k=1}^m \frac{C_k}{(\rho - \frac{1}{T})}] e^{-\rho t} + \sum_{k=1}^m \frac{C_k}{\rho} - \sum_{k=1}^m \frac{C_k e^{-t/T}}{(\rho - \frac{1}{T})} \quad (10)$$

Clearly, the system is stable for positive values of  $\rho$  and  $T$  and converges to  $\sum_{k=1}^m C_k / \rho$  in its steady state.

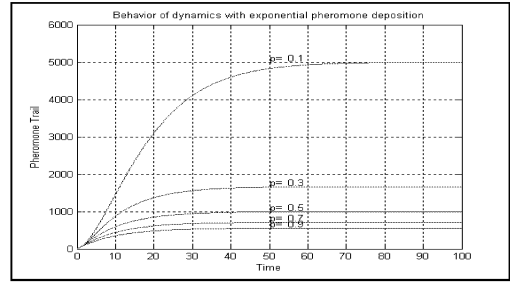


Fig.3.  $\tau_{ij}(t)$  versus  $t$  for exponential pheromone deposition with  $T=10$

## SIMULATION RESULTS

As our problem environment, we choose a roadmap of connected cities where the shortest route between two given cities is to be found out. We represent the cities as nodes and the paths connecting these cities as edges. Therefore, in effect, the problem environment takes the form of a connected graph. Implementation of ACO algorithm in such problem is little different from that of TSP. All ants, in this problem, are placed on the source node and they are allowed to construct tour to reach the destination node in the shortest possible route. At each intermediate step, ant decides its next position by the same probability based selection approach as given in (1). The interpretation of the terms in equation (1) in context of this problem is also exactly the same as given earlier in context of TSP. The only difference lies in the termination condition of the algorithm. In TSP, ant stops moving if it finds a dead end or reaches the beginning node. In this problem, ant terminates its tour on occurrence of a dead end or on reaching the destination node.

In constant deposition approach, ant deposits uniform pheromone on all edges belonging to the tour of the ant. But in exponential deposition approach, pheromone deposited by ant increases as it moves closer to the destination node. This implies that edges lying closer to the destination node receive greater amount of pheromone than those lying closer to the source node. To find the optimum parameter range of ACO algorithm with such approach we performed simulations on seven different node distributions with 200,220,235,250,265,280,300 number of nodes. We, however, owing to space constraint, present results for only two such distributions with 200 and 300 number of nodes.

### Result for Roadmap I:

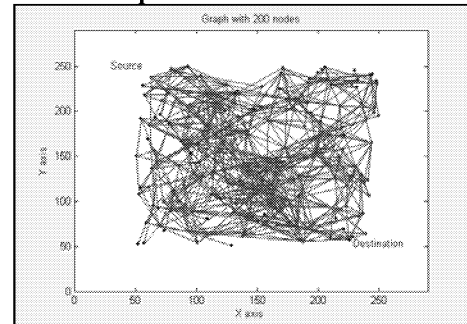


Fig 4: Graph with 200 nodes

The bold black line in above figure shows the theoretical minimum path as found by Dijkstra's algorithm. In most optimal solutions for this problem environment with 200

nodes, number of edges belonging to a tour is in between 8 and 10.  $T$ , therefore, is set at a value 6.0 following the philosophy of the proposed deposition rule. Both  $\alpha$  and  $\beta$  are varied in the range of 0.5 to 5.0 in steps of 0.5 and the solution accuracy and convergence time of the proposed method are observed. The results with  $w=6$  and 25 number of ants are presented below. Proposed method works best for  $\alpha=1.0$  and  $\beta=3.5$  as obvious from figures (5) and (6). A comparative study of the two kinds of deposition rule is also presented in figure 7. We use  $\alpha=1.0$  and  $\beta=5.0$  (Bullnheimer *et al* 1999; M. Dorigo and T. Stutzle. 2005) for simulating the standard  $AS_{rank}$  algorithm. Also,  $\rho$  is set at a value 0.1 in both deposition rules. The plots of figure 7 reveal that the proposed method outperforms the traditional one in terms of both solution quality and convergence time.

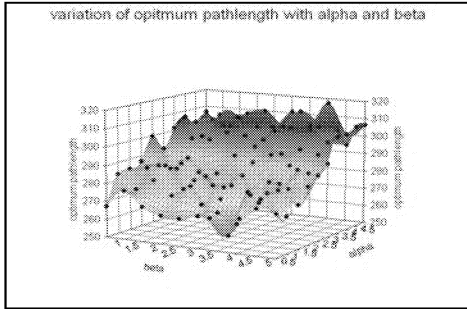


Fig 5: Variation of optimum path length with  $\alpha$  and  $\beta$

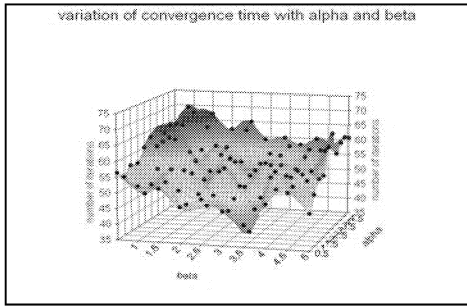


Fig 6: Variation of convergence time with  $\alpha$  and  $\beta$

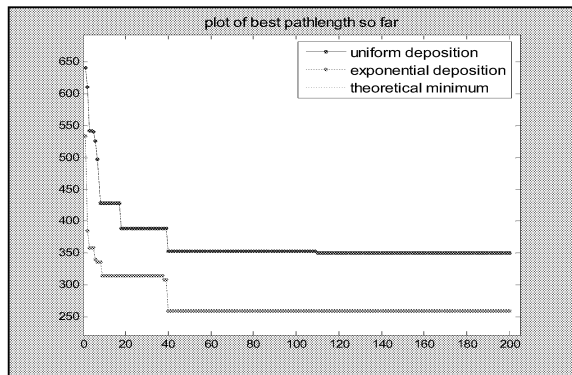


Fig 7: Comparative study of two deposition rules with 200 nodes

## Result for Roadmap II:

Here we consider a graph with 300 nodes. The proposed algorithm works best with  $\alpha=1.0$  and  $\beta=4.0$ .  $T$  here is set at 8.0 as number of edges in most optimal solutions lie between 10 and 15 for such environment. A comparative study of two

deposition rules over this environment is presented in figure 11. The proposed method again outshines its traditional counterpart. The optimal solution found by the proposed method almost matches the one found by Dijkstra's algorithm.

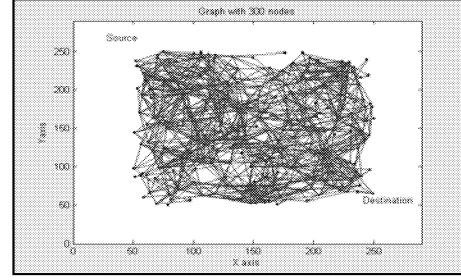


Fig 8: Graph with 300 nodes

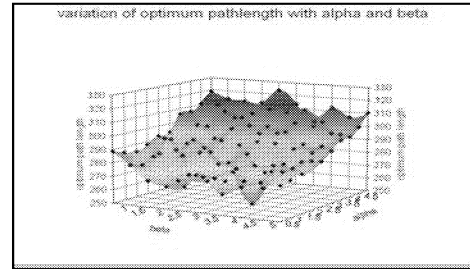


Fig 9: Variation of optimum path length with  $\alpha$  and  $\beta$

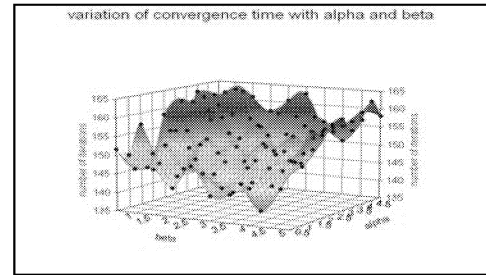


Fig 10: Variation of convergence time with  $\alpha$  and  $\beta$

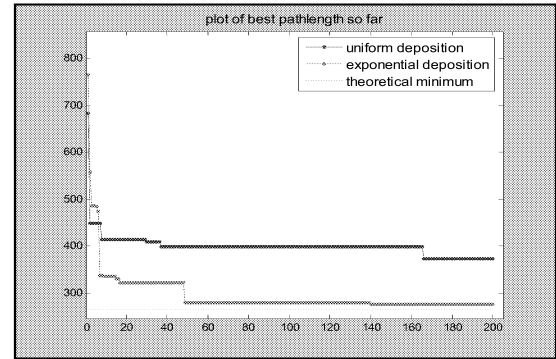


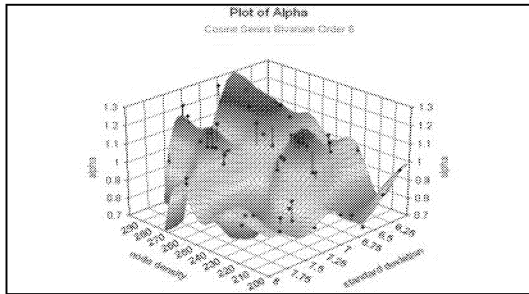
Fig 11: Comparative study of two deposition rules with 300 nodes

For other 5 environments, optimum performance was achieved at  $\alpha=1.0$  and  $\beta=3.5$  or 4.0. Thus, it can be estimated that the proposed method works best for values of  $\alpha$  and  $\beta$  lying in the neighborhood of these values.

## Establishment of algebraic relationship:

We now attempt to establish a relationship between the features of problem environment and optimum values of  $\alpha$

and  $\beta$ . The two parameters which we select as problem environment feature are i) **Node density**: number of nodes scattered per unit area ii) **Standard deviation of nearest neighbor distance**: which is the standard deviation of length of smallest edge associated with all nodes. We vary  $\alpha$  over the range 0.7 to 1.3 and  $\beta$  over the range 3.2 to 4.3 in steps of 0.1 ( i.e. over the range that we have obtained from previous set of experiments) and run the proposed algorithm on 49 different node distributions to record the optimum values of  $\alpha$  and  $\beta$  for each such distribution. The plots along with the algebraic relationship are presented below. The surface fitting through the data points was performed using a tool **TableCurve3D V 4.0**. The algebraic equations discovered help in finding almost accurately the optimum values of  $\alpha$  and  $\beta$  when problem feature set is known in advance.



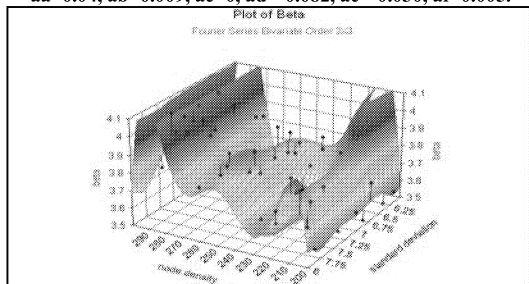
**Fig 12 : Surface Fitting for variation of  $\alpha$**

**Function: Cosine Series Bivariate Polynomial Order 6**

(  $x$ :  $x$  scaled 0 to  $\pi$   $y$ :  $y$  scaled 0 to  $\pi$ ,  $x \equiv$  no of nodes in 200 sq units,  $y \equiv$  standard deviation)

$$\begin{aligned} \alpha = & a + b\cos(x) + c\cos(y) + d\cos(2x) + e\cos(x)\cos(y) + f\cos(2y) + \\ & g\cos(3x) + h\cos(2x)\cos(y) + i\cos(x)\cos(2y) + j\cos(3y) + \\ & k\cos(4x) + l\cos(3x)\cos(y) + m\cos(2x)\cos(2y) + \\ & n\cos(x)\cos(3y) + o\cos(4y) + p\cos(5x) + q\cos(4x)\cos(y) + \\ & r\cos(3x)\cos(2y) + s\cos(2x)\cos(3y) + t\cos(x)\cos(4y) + \\ & u\cos(5y) + v\cos(6x) + a\cos(5x)\cos(y) + ab\cos(4x)\cos(2y) + \\ & accos(3x)\cos(3y) + adc\cos(2x)\cos(4y) + aecos(x) + cos(5y) + \\ & afcos(6y). \end{aligned}$$

$$\begin{aligned} a=0.93, b=-0.052, c=0.12, d=-0.042, e=-0.16, f=-0.09, g=0.027, h=0.037, \\ i=0.093, j=0.014, k=-0.050, l=-0.092, m=-0.065, n=-0.073, \\ o=-0.029, p=0.012, q=0.059, r=0.098, s=0.103, t=0.142, u=0.004, v=0.04, \\ aa=0.04, ab=0.009, ac=0, ad=-0.082, ae=-0.030, af=0.005. \end{aligned}$$



**Fig 13 : Surface Fitting for variation of  $\beta$**

**Function: Fourier Series Bivariate Polynomial Order 2\*3**

(  $x$ :  $x$  scaled 0 to  $\pi$   $y$ :  $y$  scaled 0 to  $\pi$ )

$$\begin{aligned} \beta = & a + b\cos(x) + c\cos(y) + d\sin(x) + e\sin(y) + f\cos(2x) + g\cos(2y) + h\sin(2x) \\ & + i\sin(2y) + j\cos(3x) + k\cos(3y) + l\sin(3x) + m\sin(3y) + n\cos(x)\cos(y) \\ & + o\cos(x)\sin(y) + p\sin(x)\cos(y) + \\ & q\cos(x)\cos(2y) + r\cos(2x)\cos(y) + s\cos(x)\sin(2y) + \\ & t\sin(2x)\cos(y) + u\sin(x)\sin(y) + v\sin(x)\cos(2y) + \\ & aa\cos(2x)\sin(y) + ab\sin(x)\sin(2y) + ac\sin(2x)\sin(y). \end{aligned}$$

$$\begin{aligned} a=-0.116, b=-0.086, c=-0.124, d=7.134, e=-0.264, f=4.270, \\ g=-0.450, h=-0.019, i=-0.003, j=-0.089, k=0, l=1.210, m=0.129, \\ n=-0.085, o=0.044, p=0.179, q=0.043, r=0.197, s=0.023, t=0.180, \\ u=-0.455, v=0.160, aa=-0.448, ab=0.054, ac=-0.075. \end{aligned}$$

## CONCLUSIONS

The paper presents a novel approach of stability analysis as well as a new kind of pheromone deposition rule which outperforms the traditional approach of pheromone deposition used so far in all variants of ant system algorithms. Our works are on progress and we are trying to compare the two kinds of deposition rule using other variants of ant system algorithms like **MMAS** and **EAS** and find optimum parameter setting of ACO algorithm with proposed deposition rule for such models also.

## REFERENCES

- B. Bullnheimer, R.F. Hartle and C. Strauss. 1999. "A New Rank Based Version of The Ant System-A Computational Study". *Central European Journal for Operations Research and Economics*
- C. Blum, A. Roli and M. Dorigo, 2001 "HC-ACO: The Hyper Cube Framework for Ant Colony Optimization" In proceedings of Metaheuristic International Conference, vol. 2, pp. 399-403
- C. Blum and M. Dorigo. 2005. "Search bias in ant colony: On the role of competition balanced systems" *IEEE Transactions on Evolutionary Computation*, vol. 9, no.2, pp. 159-174.
- D. Merkle and M. Middendorf. 2000 "An Ant Algorithm with a New Pheromone Evaluation Rule for Total Tardiness Problem" In S.Cagnoni *et al.* *Real Word Applications of Evolutionary Computing Proceedings of EvoWorkshops 200*, Edinburgh 17, April 2000, Springer Verlag, LNCS, 1803 pp. 287-296
- D. Merkle, M. Middendorf and H. Schmeck. 2000 "Pheromone Evaluation in Ant Colony Optimization", In Proceedings of 26<sup>th</sup> International Conference of the IEEE Electronics Society, Piscataway, NJ, IEEE Press, pp. 2726-2731
- D. Merkle and M. Middendorf. 2002 "Modeling the dynamics of ant colony optimization algorithms," *Evolutionary Computation*, vol.10, no. 3, pp. 235-262.
- M. Dorigo and C. Blum. 2005 "Ant colony optimization theory: A survey" *Theoretical Computer Science* 344, pp 243 – 278
- M. Dorigo and G. Di Caro. 2004 "The Ant Colony Optimization meta-heuristic," *IEEE Transactions on System, Man, and Cybernetics-Part B*, vol. 34, no. 2, pp. 1161-1172.
- M. Dorigo and L.M. Gambardella. 1997 "Ant Colony System: A Co-operative Learning Approach to Travelling Salesman Problem", *IEEE Transactions on Evolutionary Computation*, vol. 1, no. 1, pp. 53-66
- M. Dorigo and T. Stutzle. 2005 *Ant Colony Optimization*, Prentice-Hall of India Private Limited ISBN-81-203-2684-9
- M. Dorigo, V. Maniezzo, and A. Colomi. 1991 "Positive feedback as a search strategy", Dipartimento di Electronica, Politecnico di Milano, Italy, Tech Rep. 91-016.
- M. Dorigo, V. Maniezzo and A. Colomi. 1996 "The Ant System: Optimization by a colony of cooperating agents" *IEEE Transactions on Systems, Man, and Cybernetics-Part B*, Vol.26, No.1, pp.1-13
- R. Michels and M. Middendorf. 1999 "An Ant System for The Shortest Common Supersequence Problem", In D. Corne, M. Dorigo, F.Glover(Eds), *New Ideas in Optimization* pp: 692-791
- T. Stützle and H. H. Hoos. 2000 "Max-Min Ant system", *Future Generation Computer Systems*, Vol.16, Issue 8, pp.: 889 - 914
- V. Maniezzo. 1999 "Exact and Approximate Nondeterministic Tree Search Procedures for the Quadratic Assignment Problem", *INFORMS Journal on Computing*, vol. 11, no. 4, pp. 358-369
- W.J.Gutjahr. 2006 "On the finite-time dynamics of ant colony optimization," *Methodology and Computing in Applied Probability*, vol. 8, no. 1, pp. 105-133



# **SIMULATION BASED SOFTWARE ENGINEERING**





# COMPUTATION WITH IMPRECISE PROBABILITIES

Lotfi A. Zadeh<sup>1\*</sup>

An imprecise probability distribution is an instance of second-order uncertainty, that is, uncertainty about uncertainty, or uncertainty<sup>2</sup> for short. Another instance is an imprecise possibility distribution. Computation with imprecise probabilities is not an academic exercise—it is a bridge to reality. In the real world, imprecise probabilities are the norm rather than exception. In large measure, real-world probabilities are perceptions of likelihood. Perceptions are intrinsically imprecise, reflecting the bounded ability of human sensory organs, and ultimately the brain, to resolve detail and store information. Imprecision of perceptions is passed on to perceived probabilities. This is why real-world probabilities are, for the most part, imprecise.

What is important to note is that in applications of probability theory in such fields as risk assessment, forecasting, planning, assessment of causality and fault diagnosis, it is a common practice to ignore imprecision of probabilities. The problem with this practice is that it leads to results whose validity is in doubt. This underscores the need for approaches in which imprecise probabilities are treated as imprecise probabilities rather than as precise probabilities.

Peter Walley's seminal work "Statistical Reasoning with Imprecise Probabilities," published in 1991, sparked a rapid growth of interest in imprecise probabilities. Today, we see a substantive literature, conferences, workshops and summer schools. An exposition of mainstream approaches to imprecise probabilities may be found in the 2002 special issue of the Journal of Statistical Planning and Inference (JSPI), edited by Jean-Marc Bernard. My paper "[A perception-based theory of probabilistic reasoning with imprecise probabilities](#)," is contained in this issue but is not a part of the mainstream. A mathematically rigorous treatment of elicitation of imprecise probabilities may be found in "A behavioural model for vague probability assessments," by Bert de Cooman, Fuzzy Sets and Systems, 2005.

The approach which is outlined in the following is rooted in my 1975 paper "[The concept of a linguistic variable and its application to approximate reasoning](#)," Information Sciences, but in spirit it is close to my 2002 JSPI paper. The approach is a radical departure from the mainstream. Its principal distinguishing features are: (a) imprecise probabilities are dealt with not in isolation, as in the mainstream approaches, but in an environment of imprecision of events, relations and constraints; (b)

imprecise probabilities are assumed to be described in a natural language. This assumption is consistent with the fact that a natural language is basically a system for describing perceptions.

The capability to compute with information described in a natural language opens the door to consideration of problems which are not well-posed mathematically. Following are very simple examples of such problems.

1.  $X$  is a real-valued random variable. What is known about  $X$  is: (a) usually  $X$  is much larger than approximately  $a$ ; and (b) usually  $X$  is much smaller than approximately  $b$ , with  $a < b$ . What is the expected value of  $X$ ?
2.  $X$  is a real-valued random variable. What is known is that  $\text{Prob}(X \text{ is small})$  is low;  $\text{Prob}(X \text{ is medium})$  is high; and  $\text{Prob}(X \text{ is large})$  is low. What is the expected value of  $X$ ?
3. A box contains approximately twenty balls of various sizes. Most are small. There are many more small balls than large balls. What is the probability that a ball drawn at random is neither large nor small?
4. I am checking-in for my flight. I ask the ticket agent: What is the probability that my flight will be delayed. He tells me: Usually most flights leave on time. Rarely most flights are delayed. How should I use this information to assess the probability that my flight may be delayed?

To compute with information described in natural language we employ the formalism of [Computing with Words](#) (CW) (Zadeh 1999) or, more generally, NL-Computation (Zadeh 2006). The formalism of Computing with Words, in application to computation with information described in a natural language, involves two basic steps: (a) precisiation of meaning of propositions expressed in natural language; and (b) computation with precisiated propositions. Precisiating of meaning is achieved through the use of generalized-constraint-based semantics, or GCS for short. The concept of a generalized constraint is the centerpiece of GCS. Importantly, generalized constraints, in contrast to standard constraints, have elasticity. What this implies is that in GCS everything is or is allowed to be graduated, that is, be a matter of degree. Furthermore, in GCS everything is or is allowed to be granulated. Granulation involves partitioning

<sup>1</sup> Dedicated to Peter Walley.

\* Department of EECS, University of California, Berkeley, CA 94720-1776; Telephone: 510-642-4959; Fax: 510-642-1712; E-Mail: [zadeh@eecs.berkeley.edu](mailto:zadeh@eecs.berkeley.edu). Research supported in part by ONR N00014-02-1-0294, BT Grant CT1080028046, Omron Grant, Tekes Grant, Chevron Texaco Grant and the BISC Program of UC Berkeley.

of an object into granules, with a granule being a clump of elements drawn together by indistinguishability, equivalence, similarity, proximity or functionality.

A generalized constraint is an expression of the form  $X \text{ is } R$ , where  $X$  is the constrained variable,  $R$  is the constraining relation and  $r$  is an indexical variable which defines the modality of the constraint, that is, its semantics. The principal modalities are: possibilistic ( $r = \text{blank}$ ), probabilistic ( $r = p$ ), veristic ( $r = v$ ), usuality ( $r = u$ ) and group ( $r = g$ ). The primary constraints are possibilistic, probabilistic and veristic. The standard constraints are bivalent possibilistic, probabilistic and bivalent veristic. In large measure, scientific theories are based on standard constraints.

Generalized constraints may be combined, projected, qualified, propagated and counterpropagated. The set of all generalized constraints, together with the rules which govern generation of generalized constraints from other generalized constraints, constitute the Generalized Constraint Language (GCL). Actually, GCL is more than a language—it is a language system. A language has descriptive capability. A language system has descriptive capability as well as deductive capability. GCL has both capabilities.

The concept of a generalized constraint plays a key role in GCS. Specifically, it serves two major functions. First, as a means of representing the meaning of a proposition,  $p$ , as a generalized constraint; and second, through representation of  $p$  as a generalized constraint it serves as a means of dealing with  $p$  as an object of computation. It should be noted that representing the meaning of  $p$  as a generalized constraint is equivalent to precisiation of  $p$  through translation into GCL. In this sense, GCL plays the role of a meaning precisiation language. More importantly, GCL provides a basis for computation with information described in a natural language. This is the province of CW or, more generally, NL-Computation.

A concept which plays an important role in computation with information described in a natural language is that of a granular value. Specifically, let  $X$  be a variable taking values in a space  $U$ . A granular value of  $X$ ,  $*u$ , is defined by a proposition,  $p$ , or more generally by a system of propositions drawn from a natural language. Assume that the meaning of  $p$  is precisiated by representing it as a generalized constraint,  $GC(p)$ .  $GC(p)$  may be viewed as a definition of the granular value,  $*u$ . For example, granular

values of probability may be defined as approximately 0.1, ..., approximately 0.9, approximately 1. A granular variable is a variable which takes granular values. For example, young, middle-aged and old are granular values of the granular variable Age. The probability distribution in Example 2 is an instance of a granular probability distribution. In effect, computation with imprecise probability distributions may be viewed as an instance of computation with granular probability distributions.

In the CW-based approach to computation with imprecise probabilities, computation with imprecise probabilities reduces to computation with generalized constraints. What is used for this purpose is the machinery of GCL. More specifically, computation is carried out through the use of rules which govern propagation and counterpropagation of generalized constraints. The principal rule is the extension principle (Zadeh 1965, 1975). In its general form, the extension principle is a computational schema which relates to the following problem. Assume that  $Y$  is a given function of  $X$ ,  $Y = g(X)$ . Let  $*g$  and  $*X$  be granular values of  $g$  and  $X$ , respectively. Compute  $*g(*X)$ .

In most computations involving imprecise probabilities what is sufficient is a special form of the extension principle which relates to possibilistic constraints. More specifically, assume that  $f$  is a given function and  $f(X)$  is constrained by a possibility distribution,  $A$ . Assume that  $g$  is a given function,  $g(X)$ . The problem is to compute the possibility distribution of  $g(X)$  given the possibility distribution of  $f(X)$ . In this case, the extension principle reduces the solution of the problem in question to the solution of a variational problem (Zadeh 2006).

In summary, the CW-based approach to computation with imprecise probabilities opens the door to computation with probabilities, events, relations and constraints which are described in a natural language. Progression from computation with precise probabilities, precise events, precise relations and precise constraints to computation with imprecise probabilities, imprecise events, imprecise relations and imprecise constraints is an important step forward—a step which has the potential for a significant enhancement of the role of natural languages in human-centric fields such as economics, decision analysis, operations research, law and medicine, among others.

# MONTE CARLO VALIDATION OF MODEL STABILITY

Pierre N. Robillard, Simon Labelle  
Département de génie informatique et de génie logiciel  
École Polytechnique de Montréal,  
Montréal, Qc, Canada  
E-mail: Pierre-n.robillard@polymtl.ca

## KEYWORDS

Monte Carlo simulation, protocol analysis, stability of model, model validation

## ABSTRACT

Monte Carlo Validation of Protocol Analysis is an original computer simulation approach used to validate qualitative and subjective coding performed in protocol analysis. The coding of the protocol is a subjective task that is very time-consuming. Subjectivity generates most variability in the resulting coding. The variability from the coders is parameterized and new coding data are simulated using Monte Carlo approach. These numerous simulated coded sequences are used to validate the stability of the models derived from the protocol analysis. The simulation generates codes with random variations of the qualitative interpretation within an observed range of variability. This enables the researcher to determine under which qualitative parameter ranges the resulting model remains invariable, or to evaluate the sensitivity of the model to these parameters. This new approach makes a contribution to the need for evaluating the validity of a model derived from a unique experiment based on time-consuming protocol analysis.

## INTRODUCTION

Cognitive sciences use protocol analysis as a suitable approach for studying cognitive behaviors. Protocol analysis, which is based on a transcript of observational data, is often used as the basic approach in empirical studies (Ericsson and Simon, 1993). Human behavior is characterized by coding schemes, and statistical tools are used to derive models of the behavioral patterns emerging from the tasks under study.

The Monte Carlo simulation approach has been developed to understand the impact of the coders' subjectivity on a model derived from a protocol analysis of design meetings. The observational approach used was to videotape the meetings, and then have a specially trained

typist transcribe the videos to produce a document called a *protocol transcript*.

A transcript entry is called a (*verbal*) *move*. The protocol transcript is an accurate written representation of all the moves performed by the participants during the meeting. The protocol must be encoded to enable efficient statistical analysis. The coding scheme defines the formalism to encode the protocol, and should be capable of encoding the moves adequately and yet be formal enough to support quantitative analysis.

The coding of the video sequences is a subjective task that is very time-consuming. Consequently the coding activities are the weak link in any study based on protocol analysis, in that two coders who perform the same move characterizations may have a different interpretation of the moves and may not replicate exactly the same coding scheme.

A coder who replicates his previous coding session later on is likely to end up with different code sequences. This behavior is natural, and it is unrealistic to expect exactly the same coding patterns from two coders, or even with the same coder. The reasons for discrepancies between codes are numerous and natural. Coding is a subjective activity based on the assignment of qualitative codes to natural moves, and many statistical tools have been developed to evaluate the reliability of coding activities.

The traditional scientific approach to validating models derived from experimental studies is to replicate the experiments. Experimental replication is multipurpose. It validates the appropriateness and reliability of the observed data, the correctness of the statistical analysis and the significance of the results. However, observational empirical studies, unlike exact sciences such as physics or chemistry, cannot be replicated for validation purposes. Difficulties come from the impossibility of using the same people, in the same environment, carrying out the same project. Thus, replications of observational empirical studies involving human behaviors are impossible, since the teammates will have learned from previous experiments, and different individuals are likely to exhibit different behaviors.

What is often recommended is to analyze the reliability of an empirical study before deriving a model based on these studies. In the case of protocol analysis, reliability is associated with stability between coders (Baer, 1977), such stability being obtained by measuring the degree of similarity between the resulting codes. Stability could also mean that a coder codes in the same way during the entire experiment, to verify that the coder will not introduce bias in one way or another during the various coding sessions (Medley and Mitzel, 1963).

Measurement validity is defined by its representativeness and adequacy (Kerlinger, 1973, Curtis, 1980), validity often being confirmed by human experts (Herbert and Attridge, 1975). Indeed, a detailed analysis by experts of the experimental setup may be necessary to judge the validity of the experiment. Another approach would be to measure the same phenomena in various ways and then compare the results (Campbell and Fiske, D. 1959).

There is also another problem with observational empirical studies, which is the nature of the subjective and qualitative components of the coding that generate variability in the protocol analysis. This problem is of general interest, since it occurs for any observational empirical study involving the measurement of human behavior.

This paper presents a novel approach based on Monte Carlo Coding Replication (MCCR), which enables the validation by simulation of the coding performed in the protocol analysis. The approach consists in measuring the variability of protocol analysis on a representative sample of the data, and then simulating the coding with the same variability for all the transcripts. The impact of the coding variability obtained from various simulated coders is analyzed on the model resulting from the protocol analysis.

The idea behind MCCR is to simulate the replication of the coding in the protocol analysis. The simulation generates codes with random variation of the qualitative interpretation within an observed range of variability. This enables the researcher to determine under which qualitative parameter ranges the resulting model remains invariable, or to evaluate the sensitivity of the model to these parameters. This approach is different from statistical analysis, where the volume of data is related to the statistical significance of the results. For example, in a protocol analysis, the number of coded events determines the statistical significance of the analysis, but says little on the impact of the different coding on the analysis.

The purpose of MCCR is to study, based on computer simulation of qualitative parameter variability, the stability of the model derived from the observational study initially coded by the human coder.

## CASE-STUDY

The use of the MCCR approach and the impact of its application are illustrated on data from protocol analysis

and resulting models that have already been published. (D'Astous et al, 2005). The model at the time it was published did not take into account the results obtained subsequently from the MCCR approach presented in this paper.

The reliability of the coding of the transcript moves into a sequence of codes based on the coding scheme was validated according to accepted practices, such as the kappa coefficient, to compare codes from two different coders. This paper presents an approach to measure the impact of coding variability between reliable coders on the resulting models.

For the purposes of this research, the variability of coding has to be characterized. Coder variability was measured from session 2, which was judged the most representative of the seven sessions that were coded. Session 2 was coded three times. Coder A, who collaborated in defining the coding scheme, did the first coding, called *Version 1*. One year later, at the beginning of this project, coder B, trained by coder A, did a second coding of session 2, called *Version 2*. Coder B continued coding in the other sessions, and six months later did a third coding of session 2, called *Version 3*.

This approach provides three coded versions of the same session with different coder characteristics. Versions 1 and 3 provide the information for the usual inter-coder agreement studies, while Versions 2 and 3 provide some information on the stability of the code according to the experience of the coder. Statistical studies were performed on the three versions.

The codes were shown to be reliable, according to accepted practices (Cone, 1977, Cooil and Rust, 1994). Table 1 presents the Perreault and Leigh reliability index (1989) for the three versions. These indices, obtained from two different statistical approaches, are above a threshold estimated to signify good agreement between coders. It is interesting to note that the best agreement is between Versions 1 and 3, which were coded by experienced coders. Versions 1 and 3 were used in this paper for demonstrating the MCCR approach.

Table 1: Perreault and Leigh Reliability Index

| Versions        | Reliability Index | Std dev |
|-----------------|-------------------|---------|
| Version 1 and 2 | 0.73              | 0.024   |
| Version 1 and 3 | 0.83              | 0.020   |
| Version 2 and 3 | 0.79              | 0.022   |

The reliability and appropriateness of the coding scheme constitute a major issue. Much iteration was required to define the coding scheme. Once an acceptable set of activities had been defined, two coders were trained to measure the reliability of the coding scheme. The problem addressed in this paper was to determine how the variations in coder practices affect the resulting models.

Table 2 presents the number of occurrences of each code for Versions 1 and 3. Codes are labelled A,B,C,D and E. For example, the first element in the first line (A) and first column (A) indicates that 11 moves were coded as A in both Versions 1 and 3. The element in the second column indicates that three moves coded A in Version 1 were coded B in the Version 3. Column 3 shows that no A move in Version 1 was coded C in Version 3. The last column gives the total number of moves with a particular code in Version 1. The last line gives the total number of moves with a particular code in Version 3. For example, moves were coded A 20 times in Version 1 and 16 times in Version 3, with 11 moves coded A in the two versions. The values on the diagonal are the exact coded agreements for the two versions. There are similar tables of code occurrences for Versions 1 and 2, and Versions 2 and 3.

Table 2: Number of Codes from Version 1 and 3

| VERSION<br>1/3 | A  | B  | C  | D  | E  | SUM |
|----------------|----|----|----|----|----|-----|
| A              | 11 | 3  |    | 2  | 4  | 20  |
| B              |    | 15 |    | 2  | 4  | 21  |
| C              |    | 2  | 11 | 1  | 5  | 19  |
| D              | 5  | 1  |    | 19 | 3  | 28  |
| E              |    |    | 2  | 1  | 50 | 53  |
| SUM            | 16 | 21 | 13 | 25 | 66 |     |

The reliability indices presented in Table 1 show that all three coded versions of session 2 are reliable for discourse analysis. With the present state of the art, it is not possible to claim that one version is better than another. The main challenge is to understand and quantify the impact of this variability on the models derived from this protocol analysis, where only one coder coded all the sessions.

## SIMULATION PROCEDURE

The following presents a new approach to replicating coding activities based on Monte Carlo simulation. MCCR is based on the following three hypotheses:

1. The reference session, which is session 2 in this case, is representative of all sessions, and consequently the variations between coders observed for this session are representative of the variations for any of the sessions.
2. The simulated data will not generate hidden dependencies between the data, and the generated data will be similar to real data.
3. The coders are stable, and their coding is homogeneous throughout all the sessions.

The first hypothesis is necessary to reduce the effort required to measure variability, which is quite time-

consuming. In theory, all the sessions could be unique and variability could be measured on each of them. Pragmatic considerations and the homogeneity of the seven sessions, in terms of reviewing activities, make the first hypothesis reasonable. The second hypothesis is guaranteed by the parameters of the normal distributions used for the simulations, which are specific to, and derived from, each session. The third hypothesis is self-evident, and is verified by statistical analysis of the resulting codes. MCCR is based on two parameters. The first takes into account the qualitative and subjective differences observed between coders in the reference session. The second takes into account the profile of moves within the simulated session.

Table 2, presented previously, shows the differences in the qualitative coding of moves between the coder of Version 1 and the coder of Version 3. To enable meaningful simulation, we also need the intrinsic characteristics of each session, since each session is unique. The move duration is an intrinsic property of moves within a session. Table 3 shows the average duration of a move respectively for the moves of sessions 1 and 3. For example, the A moves, which were coded A in the two versions, have an average duration of 21.3 seconds with a standard deviation of 11.8 sec.

Table 3: Average Move Duration

| Duration<br>Version<br>1/3 | A    | B    | C    | D    | E    |
|----------------------------|------|------|------|------|------|
| A                          | 21.3 | 8    |      | 19   | 10.3 |
| B                          |      | 10.4 |      | 9.5  | 5    |
| C                          |      | 20   | 20.6 | 5    | 8.5  |
| D                          | 8    | 4    |      | 15.7 | 19.7 |
| E                          |      |      | 9.5  | 3    | 15.1 |

The A moves of version 1 that were coded B in Version 3 have an average duration of 8 seconds, with a standard deviation of 1.7 seconds.

## MCCR SIMULATION EXAMPLE

The simulation procedure is illustrated, on a small sample of codes. It is easy to follow and can be readily generalized to the whole sessions. All simulated sessions have first to be coded by one coder for which an agreement matrix exists. Within a session, each code is simulated individually, and then all the codes are reassembled to compose a new simulated session.

To summarize, MCCR is performed on complete protocol analysis performed by a well-trained coder. The coding is redone in a representative session, called the *reference session*, by another well-trained coder or by the same coder later on. The reliability of the two coded protocols is measured with known tools, such as agreement indices or kappa coefficients. Qualitative

differences between the coders, which are interpreted as the qualitative variations in the coding scheme, are characterized by a probability matrix derived from the two codings of the reference session. Profiles of coded moves are characterized by the average and standard deviation of their durations. These parameters are used to specify the parameters of the normal distributions of the random generator. All codes are simulated according to the distribution specifications provided by the reference session. Sequences of simulated codes are built according to the specific code duration within a given session. A simulated session of exactly the same duration is obtained with a coding variability that is within the range shown for the reference session. The kappa coefficient is computed on the simulated sessions to validate their reliability. As many simulations as needed could be performed to obtain a statistically significant and stable model.

MCCR provides a new version of the human-coded version that is likely to have, on average, the same characteristics, as a version obtained by another coder. It will have the same number of codes with a variability that will be within the characteristics measured in the reference session. MCCR is efficient and could be performed hundreds of times in any given session, which means that the whole spectrum of coding variability can be explored.

### COMPARING SINGLE-CODER AND SIMULATED MODELS

This section shows the impact of the MCCR approach, as applied to review meetings, in comparing the exchange pattern model derived from protocol analysis performed by a single coder and the model derived from multi-simulated coding.

All the observed meetings that composed the empirical study were coded by experienced coders, and an intercoder variability study on one of the representative sessions and 10 series of simulations of the meetings were performed. This is equivalent to having recoded the transcript ten times with ten different coders who have statistically the same behaviors within reliable variability. More simulations did not add any new information to the model. This approach enables us to see whether or not the resulting model would have been different if different coders had been used, or, put differently, to see the core of the resulting model that is invariant under the qualitative components of the coding.

Figure 1 shows the patterns resulting from the single-coder model on the left-hand side and patterns from the simulated Monte-Carlo model on the right-hand side.

The models show the links between the exchanges. The Monte-Carlo model obtained from 10 simulations of the coders variability illustrates three with full dot punched links that were not observed from the single coder model analysis.

The simulated model confirmed the salient features of the single-coder model. MCCR may resolve ambiguities that arise from the qualitative components of the coding.

### CONCLUDING REMARKS

MCCR enables us to analyze the impact of individual qualitative variability on coding transcripts from protocol analysis. It is much easier to perform than real replication of the protocol analysis by different coders, and requires far less resources. Moreover, it provides information on the components of the resulting models which are sensitive to the qualitative components of the coding.

Observational studies are difficult to realize in an industrial environment, and analysis of the data is very time-consuming. MCCR is a new method developed to extract from the observed model the components related to the subjective and qualitative components of the observations. This method is of general interest and can be applied to any observational study.

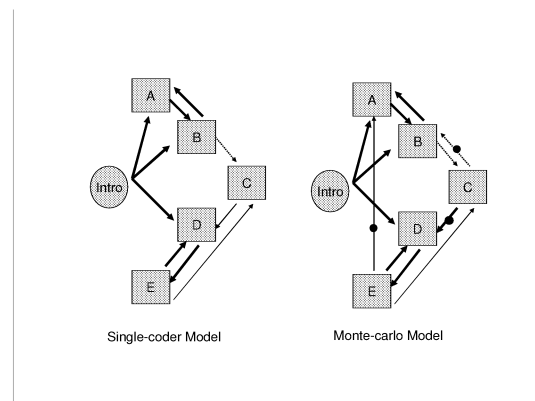


Figure 1: Single-Coder Model (left-hand side) and the Simulated Model (right-hand side).

### REFERENCES

- Baer, D. M. 1977. "Reviewer's Comment: Just Because It's Reliable Doesn't Mean That You Can Use It". *Journal of Applied Behavior Analysis*, Vol. 10, No. 1, 117-119.
- Campbell, D. T. and D. Fiske. 1959. "Convergent and Discriminant Validation by the Multi-Trait, Multi-Method Matrix". *Psychological Bulletin*, Vol. 56, 81-105.
- Cone, J. D. 1977. "The relevance of Reliability and Validity for Behavioral Assessment". *Behavior Therapy*, Vol. 8, 411-426.
- Cool, B. and R. T. Rust. 1994. "Reliability and Expected Loss: A Unifying Principle". *Psychometrika*, Vol. 59, No 2, 203-216.
- Curtis, B. 1980. "Measurement and Experimentation in Software Engineering". *Proceedings of the IEEE*, Vol. 68, No 9, 1144-1157.
- D'Astous, P., Détienné F., Visser W., Robillard, P.N., 2004. "Changing our view on design evaluation meetings methodology: a study of software technical review meetings". *Design Studies*, Vol 25, issue 6, Nov, 625-655.

- D'Astous, P., and P.N Robillard. 2002. "Empirical Study of Exchange Patterns during Software Peer Review Meetings". *Information and Software Technology*, 44, 639-648.
- Ericsson, K. A. and H. A. Simon, 1993. *Protocol Analysis: Verbal Reports as Data*, Revised Edition. MIT Press, Cambridge.
- Herbert, J. and & C. Attridge. 1975. "A Guide for Developers and Users of Observation Systems and Manuals". *American Educational Research Journal*, Vol. 12, No 1, 1-20.
- Kerlinger, F. N. 1973. *Foundations of Behavioral Research*, 2nd edition. Holt, Rinehart and Winston, New York,
- Medley, D. M. and H. E.Mitzel. 1963. "Measuring Classroom Behavior by Systematic Observation". In Gage, N. L. (ed.) *Handbook of Research on Teaching*. Rand McNally, Chicago, 247-328.
- Perreault, W.D.J., and L.E. Leigh. 1989. "Reliability of Nominal Data Based on Qualitative Judgments", *Journal of Marketing Research*, Vol. 26, 135-148.

# STANDARD ERROR ESTIMATION FOR EM APPLICATIONS RELATED TO LATENT CLASS MODELS

Liberato Camilleri  
Department of Statistics and Operations Research  
University of Malta  
Msida (MSD 06)  
Malta  
E-mail: liberato.camilleri@um.edu.mt

## KEYWORDS

EM algorithm; Numerical differentiation; Incomplete data; Maximum likelihood estimation; Proportional Odds Model; Latent Class Model.

## ABSTRACT

The EM algorithm is a popular method for computing maximum likelihood estimates. It tends to be numerically stable, reduces execution time compared to other estimation procedures and is easy to implement in latent class models. However, the EM algorithm fails to provide a consistent estimator of the standard errors of maximum likelihood estimates in incomplete data applications. Correct standard errors can be obtained by numerical differentiation. The technique requires computation of a complete-data gradient vector and Hessian matrix, but not those associated with the incomplete data likelihood. Obtaining first and second derivatives numerically is computationally very intensive and execution time may become very expensive when fitting Latent class models using a Newton-type algorithm. When the execution time is too high one is motivated to use the EM algorithm solution to initialize the Newton Raphson algorithm. We also investigate the effect on the execution time when a final Newton-Raphson step follows the EM algorithm after convergence. In this paper we compare the standard errors provided by the EM and Newton-Raphson algorithms for two models and analyze how this bias is affected by the number of parameters in the model fit.

## 1. INTRODUCTION

A limitation of the EM algorithm is that the estimated information matrix, in contrast to the case for gradient methods such as Newton-Raphson, is not a direct by-product of maximization. Procedures for obtaining the information matrix within the EM algorithm have been suggested by several authors.

An approach for computing the Fisher information matrix within the EM framework was suggested by (Louis 1982). His methodology is based on a result by (Fisher 1925) that showed that, given the incomplete data, incomplete data

scores are conditional expectations of the complete data scores. The author derives a procedure for extracting the observed information matrix when the EM algorithm is used to find maximum likelihood estimates in incomplete data problems. The technique requires the computation of the complete data gradient vector and the Hessian matrix but does not require those associated with the incomplete data log-likelihood function. A criticism of this approach is that the procedure is often computationally demanding and hard to implement because it requires the computation of both a complete-data score vector and second derivative matrix.

An alternative approach for computing the Fisher information matrix using gradients only was suggested by (Meilijson 1989). Methods that only require gradients are easier to compute analytically and less demanding to compute numerically. An appealing advantage of this procedure, in contrast to the approach suggested by (Louis 1982), is that once the individual scores have been identified there is no additional analysis to perform. Meilijson's methodology is based on a result by (Fisher 1925) in which the evaluation of individual score vectors of the incomplete data is a by-product of the application of the E-step of the EM algorithm. The Fisher information matrix may be consistently estimated by the empirical variance-covariance matrix of these individual score vectors and the M step may be replaced by a Newton-type step. This permits a unification of EM methodology and Newton methods. A demerit of Meilijson's technique is that it applies only to specialized cases in which the observed data are independent and identically distributed samples.

Another approach for computing the observed information matrix is the well-known supplemented EM (SEM) algorithm, suggested by (Meng and Rubin 1991). The SEM algorithm numerically differentiates the EM operator  $M(\boldsymbol{\varphi})$  and uses a result by (Dempster, Laird and Rubin 1977) that relates the Jacobian of  $M(\boldsymbol{\varphi})$  to the Hessian matrix  $H(\boldsymbol{\varphi})$ , both evaluated at  $\hat{\boldsymbol{\varphi}}$ . The authors claim that their algorithm can be applied to any problem to which EM has been applied, assuming that one has access to the complete-data asymptotic variance-covariance matrix. (Segal, Bacchetti and Jewell 1994) point out that the SEM



algorithm requires very accurate estimates of  $\hat{\boldsymbol{\phi}}$  and so they can be much more expensive to obtain than the EM estimates. (McCulloch 1998) remarks that for many problems the method of obtaining standard errors using the SEM algorithm can be numerically unstable. (Jamshidian and Jennrich 2000) point out that, algorithms that numerically differentiate  $M(\boldsymbol{\phi})$  may suffer from the error magnification problem when the EM algorithm is slow. The authors remark that algorithms that numerically differentiate the score vector  $\mathbf{g}(\boldsymbol{\phi})$  are appropriate for all maximum likelihood applications and they do not suffer from the error magnification problem.

The variance-covariance matrix can be obtained by other techniques that do not use numerical differentiation. Bootstrapping uses computer intensive resampling and treats a given sample as the population. An empirical probability distribution is constructed from the sample of size  $n$  in which the probability of each observation is  $1/n$ .  $K$  random samples each of size  $n$  are drawn with replacement from this empirical distribution where some of the observations in a sample may be duplicated. The EM algorithm is then performed on each sample to calculate the vector of parameters  $\hat{\boldsymbol{\phi}}_k$ . Hence a probability distribution is constructed from all the resampled parameter estimates in which the probability of each  $\hat{\boldsymbol{\phi}}_k$  is  $1/K$ . This distribution is the bootstrapped estimate of the sampling distribution of  $\hat{\boldsymbol{\phi}}$  which can be used to provide estimates for the standard errors. The primary advantage of bootstrapping is that no assumptions about the shape of the sampling distribution are made. Jackknifing is a different resampling technique in which a single observation is omitted at a time. Thus, each sample consists of  $n-1$  observations formed by deleting a different observation from the sample. A jackknifed estimate of the sampling distribution of  $\hat{\boldsymbol{\phi}}$  can be obtained in a similar way to the bootstrap procedure. (Agresti 2002) remarks that bootstrap and jackknife procedures are useful tools for estimating standard errors when samples are small or data is sparse.

## 2. A GENERAL MODEL

A latent class model relates a set of observed multivariate categorical variables to a latent variable which is discrete. Latent class analysis, unlike cluster analysis, uses a model-based approach that combines conventional statistical estimation methods to classical clustering techniques. In this methodology latent classes are defined by the criterion of conditional independence where the observed variables within each segment are statistically independent. The assumption of conditional independence has been widely used in latent class modelling. It is directly analogous to the assumption, in the factor analysis model, that observed variables are conditionally independent given the factors. This implies that the observed correlations between the items are due to the clustered nature of the population, whereas within a cluster, the items are independent.

To illustrate the procedure, we fit a latent class model to a data set as suggested by (Camilleri and Green 2004) using the EM algorithm and a Newton-type algorithm. The aim is to assess the bias of the standard errors between these maximization procedures. The EM algorithm for fitting latent class models is implemented using GLIM software (Generalized linear interactive models). The Newton-type algorithm is implemented using the facilities of GLLAMM (Generalized linear latent and mixed models). GLLAMM software uses numerical first and second derivatives of the log-likelihood and produce standard errors by maximizing the marginal log-likelihood using Raphson algorithm. The GLLAMM framework accommodates a large class of models including structural equation, multilevel, latent class and longitudinal models.

Let  $\boldsymbol{\phi} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\pi})$  be the vector comprising the parameters of the latent class model with  $K$  segments. The  $n^{\text{th}}$  density function is of the form

$$P(\mathbf{Y}_n = \mathbf{y}_n | \boldsymbol{\phi}) = \sum_{k=1}^K \pi_k \cdot P(\mathbf{Y}_n = \mathbf{y}_n | \boldsymbol{\alpha}, \boldsymbol{\beta}_k) \quad (1)$$

$\pi_k$  are the unconditional probabilities that sum to 1 and represent the proportion of respondents that are allocated to each segment. The marginal or conditional probability  $P(y_{jn} = r | \boldsymbol{\alpha}, \boldsymbol{\beta}_k)$  follows the Proportional Odds model suggested by (McCullagh 1980)

$$P(y_{jn} = r | \boldsymbol{\alpha}, \boldsymbol{\beta}) = F(\alpha_r + \mathbf{x}_j' \boldsymbol{\beta}) - F(\alpha_{r-1} + \mathbf{x}_j' \boldsymbol{\beta}) \quad (2)$$

In this model  $y_{jn}$  is a rating response elicited by the  $n^{\text{th}}$  respondent for the  $j^{\text{th}}$  item;  $\boldsymbol{\alpha}$  is a vector of threshold parameters;  $\boldsymbol{\beta}$  is a vector of regression parameters and  $\mathbf{x}_j$  are item covariates. The choice of  $F(\cdot)$  is the Logistic distribution which leads to the logit link.

The likelihood function of the data set is obtained by taking the product of the  $N$  density functions.

$$L(\boldsymbol{\phi}) = \prod_{n=1}^N \sum_{k=1}^K \pi_k \cdot P(\mathbf{Y}_n = \mathbf{y}_n | \boldsymbol{\alpha}, \boldsymbol{\beta}_k) \quad (3)$$

The log-likelihood function is given by:

$$l(\boldsymbol{\phi}) = \sum_{n=1}^N \ln \sum_{k=1}^K \pi_k \cdot P(\mathbf{Y}_n = \mathbf{y}_n | \boldsymbol{\alpha}, \boldsymbol{\beta}_k) \quad (4)$$

Maximum likelihood estimation can be carried out via standard numerical optimization routines such as the Newton Raphson method or alternatively using the EM algorithm. The popularity of the EM algorithm arises from its computational elegance, particularly for latent class models. The idea behind the EM algorithm is to augment the observed data by introducing unobserved data,  $\lambda_{nk}$  indicating whether the  $n^{\text{th}}$  respondent belongs to the  $k^{\text{th}}$  segment.

An effective procedure to fit a latent class model with  $K$  segments is to maximize the expected complete log-likelihood function using the iterative EM algorithm.

$$L(\boldsymbol{\varphi} | \boldsymbol{\Lambda}) = \prod_{n=1}^N \prod_{k=1}^K [\pi_k \cdot P(\mathbf{Y}_n = \mathbf{y}_n | \boldsymbol{\alpha}, \boldsymbol{\beta}_k)]^{\lambda_{nk}} \quad (5)$$

The complete log likelihood  $l(\boldsymbol{\varphi} | \boldsymbol{\Lambda})$  is given by:

$$l(\boldsymbol{\varphi} | \boldsymbol{\Lambda}) = \sum_{n=1}^N \sum_{k=1}^K [\lambda_{nk} \cdot \ln P(\mathbf{Y}_n = \mathbf{y}_n | \boldsymbol{\alpha}, \boldsymbol{\beta}_k) + \lambda_{nk} \cdot \ln(\pi_k)] \quad (6)$$

The complete log-likelihood function  $l(\boldsymbol{\varphi} | \boldsymbol{\Lambda})$  has a simpler form compared to  $l(\boldsymbol{\varphi})$  given by (4) and the derivatives are easier to compute.

Each iteration is composed of two steps: an E-step and an M-step. In the E-step,  $E[l(\boldsymbol{\varphi} | \boldsymbol{\Lambda})]$  is calculated with respect to the conditional distribution of the unobserved data  $\boldsymbol{\Lambda} = (\lambda_1, \lambda_2, \dots, \lambda_N)$  given the vector of observed responses  $\mathbf{y}_n$  and using the provisional parameter estimates  $\boldsymbol{\varphi}$ . This is achieved by using Bayes' theorem to estimate  $\lambda_{nk}$ .

$$E(\lambda_{nk}) = \frac{\pi_k \cdot P(\mathbf{Y}_n = \mathbf{y}_n | \boldsymbol{\alpha}, \boldsymbol{\beta}_k)}{\sum_{k=1}^K \pi_k \cdot P(\mathbf{Y}_n = \mathbf{y}_n | \boldsymbol{\alpha}, \boldsymbol{\beta}_k)} = p_{nk} \quad (7)$$

In the M-step,  $E[l(\boldsymbol{\varphi} | \boldsymbol{\Lambda})]$  is maximized with respect to  $\boldsymbol{\varphi}$ . This is achieved by replacing  $\lambda_{nk}$  by their expected posterior probabilities  $p_{nk}$ . So

$$E[l(\boldsymbol{\varphi} | \boldsymbol{\Lambda})] = \sum_{n=1}^N \sum_{k=1}^K [p_{nk} \cdot \ln P(\mathbf{Y}_n = \mathbf{y}_n | \boldsymbol{\alpha}, \boldsymbol{\beta}_k) + p_{nk} \cdot \ln(\pi_k)] \quad (8)$$

The two terms on the right hand side of the expression can be maximized separately. The maximization of  $E[l(\boldsymbol{\varphi} | \boldsymbol{\Lambda})]$  with respect to  $\pi_k$  is straightforward and can be worked directly by differentiation. The maximum of  $E[l(\boldsymbol{\varphi} | \boldsymbol{\Lambda})]$  with respect to  $\pi_k$ , subject to the constraint  $\sum_{k=1}^K \pi_k = 1$ , is obtained by maximizing the augmented function.

$$\sum_{k=1}^K \sum_{n=1}^N p_{nk} \ln \pi_k - \delta \left( \sum_{k=1}^K \pi_k - 1 \right) \quad (9)$$

$\delta$  is the Lagrange multiplier. Setting the derivative with respect to  $\pi_k$  equal to zero yields

$$\hat{\pi}_k = \frac{1}{N} \sum_{n=1}^N p_{nk} \quad \text{for } k = 1, 2, \dots, K \quad (10)$$

The maximization of  $E[l(\boldsymbol{\varphi} | \boldsymbol{\Lambda})]$  with respect to  $\boldsymbol{\alpha}$  and  $\boldsymbol{\beta}_k$  in GLIM is performed by transforming the polychotomous responses as a vector of 0-1 indicators. This allows the

use of Poisson likelihood in model fitting by considering each term of  $\sum_{n=1}^N \sum_{k=1}^K p_{nk} \cdot \ln P(\mathbf{Y}_n = \mathbf{y}_n | \boldsymbol{\alpha}, \boldsymbol{\beta}_k)$  a weighted Poisson log-likelihood function. This maximization step can be accommodated using the OWN model facilities of GLIM4.

Since the probabilities,  $p_{nk}$  are unknown then the iterative procedure is initiated by setting random assignment to these probabilities. The algorithm alternately updates the parameters  $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\pi}$  and the prior weights,  $p_{nk}$  until the process converges.

Maximum likelihood estimation in GLLAMM is carried out via a Newton-Raphson algorithm. The algorithm uses numerical first and second derivatives of the likelihood function, which is computationally demanding and time expensive even with few model parameters. The Newton-Raphson algorithm can be derived by considering an approximation of  $\partial l(\boldsymbol{\varphi}) / \partial \boldsymbol{\varphi}$  using a first order Taylor series expansion around the parameter  $\boldsymbol{\varphi}^m$  evaluated at the  $m^{\text{th}}$  iteration.

$$\frac{\partial l(\boldsymbol{\varphi})}{\partial \boldsymbol{\varphi}} \approx \frac{\partial l(\boldsymbol{\varphi}^m)}{\partial \boldsymbol{\varphi}} + \frac{\partial^2 l(\boldsymbol{\varphi}^m)}{\partial \boldsymbol{\varphi} \partial \boldsymbol{\varphi}'} (\boldsymbol{\varphi} - \boldsymbol{\varphi}^m) \quad (11)$$

Gradient methods are iterative and updated parameters can be evaluated by setting  $\partial l(\boldsymbol{\varphi}) / \partial \boldsymbol{\varphi}$  to zero. Denoting the gradient vector and Hessian matrix by  $\mathbf{g}(\boldsymbol{\varphi}^m)$  and  $\mathbf{H}(\boldsymbol{\varphi}^m)$ , the updated parameters are given by:

$$\boldsymbol{\varphi}^{m+1} = \boldsymbol{\varphi}^m - \mathbf{H}(\boldsymbol{\varphi}^m)^{-1} \mathbf{g}(\boldsymbol{\varphi}^m) \quad (12)$$

If the log-likelihood is quadratic in the parameters, as in the case of linear regression models, the equations can be solved in closed form and maximum likelihood estimates  $\hat{\boldsymbol{\varphi}}$  are found in a single iteration.

### 3. RESULTS OF THE STUDY

Although the EM algorithm yields maximum likelihood estimates of the parameters it fails to provide standard errors of these parameter estimates as a by-product of the iterative algorithm. On the other hand, a Newton-type algorithm provides correct standard errors; however, there is a computing cost associated with our patience in waiting for an output. It is well known that Newton-type methods require good starting values and a fast convergence is only guaranteed if these starting values are near the solution. Another problem is that obtaining first and second derivatives numerically is computationally intensive and a Newton-type algorithm may become very expensive particularly when fitting models with a considerable number of parameters. This paper compares the standard errors of the parameters provided by the EM and Newton-Raphson algorithms for two models and contrast execution times when using GLIM and GLLAMM software.

GLLAMM software can fit proportional odds models by specifying the **family** to be binomial and the **link** to be ologit. This link corresponds to the logit link functions appropriate for ordinal data. The syntax **nrf** specifies the number of latent variables; the syntax **nip** specifies the number of latent classes (segments) and the syntax **ip(fn)** yields non-centred latent classes. Some of the terms in the GLIM output were intrinsically aliased. In order to get a similar solution using GLLAMM we had to constrain these parameters to zero using the **constraint define** command in GLLAMM.

It was noted that estimation with GLLAMM using a Newton-type algorithm took about fifty times longer compared to GLIM using an EM algorithm. For problems with large numbers of parameters and latent variables, Newton-type methods can become infeasible and computationally demanding. When the computer cost is too high one is motivated to use GLIM's EM algorithm solution to initialize GLLAMM's Newton Raphson algorithm. This reduces considerably the execution time for GLLAMM. It was noted that when a final Newton-Raphson step was applied to GLIM's EM solution after convergence the algorithm always converged in at most three iterations yielding a solution which was concave. In spite of this improvement, estimation with GLLAMM still took about five times longer compared to GLIM.

In the first illustration a Latent class model was fitted to a data set (Camilleri and Green 2004) that provided rating scores to a number of items (profiles) described by three car-attributes. The linear predictor included brand as a sole main effect with four categories. The latent variable, segment, was interacted with each level of brand and the model was estimated with two latent classes, four latent variables and a logit link function. A 7-point scale was used for the rating scores yielding 6 threshold (cut-point) parameters. The GLIM solution required 34 iterations and took 3 minutes to converge. The log-likelihood of this solution was 9807.98. The parameter estimates elicited from the EM algorithm were then used as starting values for the Newton-Raphson algorithm. GLLAMM required three iterations and took 9 minutes to converge. The log-likelihood of the GLLAMM solution was 9807.62.

| Term            | GLIM Output |          | GLLAMM Output |          |
|-----------------|-------------|----------|---------------|----------|
|                 | Estimate    | St Error | Estimate      | St Error |
| Cutp1           | -4.061      | 0.134    | -4.063        | 0.177    |
| Cutp2           | -2.816      | 0.127    | -2.814        | 0.171    |
| Cutp3           | -1.858      | 0.124    | -1.856        | 0.169    |
| Cutp4           | -0.927      | 0.122    | -0.925        | 0.168    |
| Cutp5           | 0.118       | 0.121    | 0.119         | 0.167    |
| Cutp6           | 1.362       | 0.126    | 1.364         | 0.168    |
| Brand(1).Seg(1) | -2.871      | 0.177    | -2.870        | 0.274    |
| Brand(1).Seg(2) | -1.149      | 0.140    | -1.148        | 0.191    |
| Brand(2).Seg(1) | -0.636      | 0.174    | -0.636        | 0.270    |
| Brand(2).Seg(2) | -0.603      | 0.139    | -0.603        | 0.189    |
| Brand(3).Seg(1) | -2.628      | 0.176    | -2.629        | 0.332    |
| Brand(3).Seg(2) | -1.360      | 0.140    | -1.360        | 0.190    |
| Brand(4).Seg(1) | -2.541      | 0.177    | -2.541        | 0.273    |
| Brand(4).Seg(2) | Aliased     | Aliased  | Aliased       | Aliased  |

**Table 1:** Parameter estimates and standard errors elicited the EM and EM+NR algorithms.

Another interesting observation is that GLIM provided deflated standard errors where the deflation for each standard error varied from 24% to 47%. The cause for this deflation is that the EM algorithm has to estimate  $KN$  missing or unobserved values  $\lambda_{nk}$  together with the model parameters.

In the second illustration another Latent class model was fitted to the same data set. The linear predictor includes brand and a two-level door attribute as main effects and the interaction of brand with a quadratic function of price. The latent variable, segment, was again interacted with each term. The model was estimated with two latent classes, thirteen latent variables and a logit link function. The GLIM solution required 34 iterations and took 10 minutes to converge. The log-likelihood of this solution was 9004.64. Using GLIM's parameter estimates as initial values, GLLAMM required 3 iterations that took 36 minutes to converge. The log-likelihood of the GLLAMM solution was 9003.24 and the amount of deflation of GLIM's standard errors compared to GLLAMM's varied from 0% to 19%.

| Term                    | GLIM Output |          | GLLAMM Output |          |
|-------------------------|-------------|----------|---------------|----------|
|                         | Estimate    | St Error | Estimate      | St Error |
| Cutp1                   | -0.631      | 0.843    | -0.634        | 0.877    |
| Cutp2                   | 0.043       | 0.843    | 0.045         | 0.877    |
| Cutp3                   | 0.604       | 0.843    | 0.602         | 0.877    |
| Cutp4                   | 1.181       | 0.843    | 1.180         | 0.877    |
| Cutp5                   | 1.802       | 0.843    | 1.803         | 0.877    |
| Cutp6                   | 2.513       | 0.843    | 2.513         | 0.877    |
| Door(1).Seg(1)          | -1.295      | 1.135    | -1.297        | 1.214    |
| Door(1).Seg(2)          | -0.314      | 0.044    | -0.312        | 0.053    |
| Door(2).Seg(1)          | -0.799      | 1.135    | -0.798        | 1.214    |
| Door(2).Seg(2)          | Aliased     | Aliased  | Aliased       | Aliased  |
| Brand(2).Seg(1)         | -0.436      | 1.079    | -0.434        | 1.090    |
| Brand(2).Seg(2)         | 1.082       | 1.188    | 1.080         | 1.213    |
| Brand(3).Seg(1)         | -0.275      | 1.078    | -0.273        | 1.090    |
| Brand(3).Seg(2)         | 0.625       | 1.189    | 0.623         | 1.215    |
| Brand(4).Seg(1)         | -0.569      | 1.083    | -0.567        | 1.104    |
| Brand(4).Seg(2)         | 1.597       | 1.186    | 1.597         | 1.233    |
| Brand(1).Price.Seg(1)   | 0.410       | 0.213    | 0.411         | 0.218    |
| Brand(1).Price.Seg(2)   | 0.406       | 0.234    | 0.405         | 0.244    |
| Brand(2).Price.Seg(1)   | 0.598       | 0.212    | 0.598         | 0.218    |
| Brand(2).Price.Seg(2)   | 0.319       | 0.233    | 0.317         | 0.244    |
| Brand(3).Price.Seg(1)   | 0.515       | 0.212    | 0.515         | 0.216    |
| Brand(3).Price.Seg(2)   | 0.246       | 0.234    | 0.246         | 0.241    |
| Brand(4).Price.Seg(1)   | 0.494       | 0.212    | 0.494         | 0.218    |
| Brand(4).Price.Seg(2)   | 0.133       | 0.233    | 0.131         | 0.244    |
| Brand(1).PriceSq.Seg(1) | -0.017      | 0.014    | -0.017        | 0.014    |
| Brand(1).PriceSq.Seg(2) | -0.043      | 0.016    | -0.043        | 0.016    |
| Brand(2).PriceSq.Seg(1) | -0.030      | 0.014    | -0.030        | 0.014    |
| Brand(2).PriceSq.Seg(2) | -0.037      | 0.015    | -0.037        | 0.016    |
| Brand(3).PriceSq.Seg(1) | -0.026      | 0.014    | -0.026        | 0.014    |
| Brand(3).PriceSq.Seg(2) | -0.033      | 0.016    | -0.033        | 0.016    |
| Brand(4).PriceSq.Seg(1) | -0.023      | 0.014    | -0.023        | 0.014    |
| Brand(4).PriceSq.Seg(2) | -0.023      | 0.015    | -0.023        | 0.016    |

**Table 2:** Parameter estimates and standard errors elicited the EM and EM+NR algorithms.

An interesting observation is that when complex models are fitted the discrepancy between GLIM's standard errors compared to GLLAMM's was smaller. An explanation for this occurrence is that the proportion of model parameters

compared to the proportion of missing values increases when more terms are included in the model fit. It was also noted that when complex models are fitted a higher proportion of the posterior probabilities approach 0 or 1. This is due to the fact that complex models explain the heterogeneity in the data better than simple models.

#### 4 CONCLUSIONS

Newton-type algorithms are essential to elicit correct standard errors for the parameter estimates; however, these algorithms are extremely slow since they use numerical first and second derivatives of the log-likelihood. This execution time problem becomes more severe when the number of latent variables in the latent class model is increased. Estimation with a Newton-type algorithm may take fifty times longer compared to estimation with an EM algorithm. The study proposes using the EM algorithm solution as an initialization step. Equipped with very good starting values the final Newton-Raphson step converges quickly. This procedure guarantees correct standard errors of the parameters estimates and reduces execution times considerably. Another interesting finding is that the bias between the correct and incorrect standard errors obtained respectively by Newton-type and EM algorithms becomes less conspicuous as the model complexity increases.

#### REFERENCES

- Aitkin, M., Anderson, D., Francis, B. and Hinde, J. (1994), *Statistical Modelling in GLIM*, Oxford Science Publications.
- Camilleri, L. and Green, M. (2004), Statistical Models for Market Segmentation, *Proceedings of the 19<sup>th</sup> International Workshop Statistical Modelling, Florence*. 120-124.
- Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977), Maximum Likelihood from Incomplete Data via the EM algorithm, *Journal of the Royal Statistical Society, B*, 39, 1-38.
- Fisher, R.A. (1925), Theory of Statistical Estimation. *Proc. Camb. Phil. Society.*, 22, 700-725.
- Francis, B., Green, M. and Payne, C. (1993), *The GLIM 4 manual*, Oxford Science Publications.
- Green, M. (2000), Statistical Models for Conjoint Analysis, *Proceedings of the 15<sup>th</sup> International Workshop on Statistical Modelling, Bilbao*. 216-222.
- Green, P.J. (1984), Iteratively Reweighted Least Squares for Maximum Likelihood Estimation *Journal of Royal Statistical Society, B*, 46, 149-192.
- Jamshidian, M. and Jennrich, R.I. (1997), Acceleration of the EM algorithm using quasi-Newton methods, *Journal of the Royal Statistical Society B*, 569-587.
- Jamshidian, M. and Jennrich, R.I. (2000), Standard Errors for EM Estimation, *Journal of the Royal Statistical Society B*, 257-270.
- Louis, T.A. (1982), Finding the Observed Information Matrix when using the EM algorithm, *Journal of the Royal Statistical Society*, 44, 226-233.
- McCullagh P. (1980) Regression Models for Ordinal Data, *J.R.Statist.Soc B*, 42, 109-142.
- McCulloch, C.E. (1998), Maximum Likelihood Variance components estimation for Binary Data, *Journal of the American Statistical Association*, 89, 330-335.
- Meilijson, I. (1989), A Fast Improvement to the EM algorithm on its Own Term, *Journal of the Royal Statistical Society*, 51, 127-138.
- Meng, X.L. and Rubin, D.B. (1991), Using EM to obtain asymptotic variance-covariance matrices: the SEM algorithm, *Journal of American Statistical Association*, 86, 899-909.
- Nelder, J.A and Wedderburn, R.W.M. (1972), Generalized Linear Models, *Journal of the Royal Statistical Society, A*, 135, 370-384.
- Rabe-Hesketh, S., Pickles, A. and Skrondal, A. (2001), GLLAMM: A General Class of Multilevel Models and Stata Program, *Multilevel Modelling Newsletter*, 13, 17-23.
- Segal, M.R., Bacchetti, P. and Jewell, N.P. (1994), Variance for Maximum Penalized Likelihood estimates obtained via the EM algorithm, *Journal of the Royal Statistical Society B*, 56, 345-352.
- Skrondal, A. and Rabe-Hesketh, S. (2004), *Generalized Latent Variable Modelling*, Chapman & Hall/CRC.
- Vermunt, J.K. (2004), An EM algorithm for the estimation of parametric and non-parametric hierarchical non-linear models, *Statistica Neerlandica*, 58, 220-233.

#### AUTHOR BIOGRAPHY

**LIBERATO CAMILLERI** studied Mathematics and Statistics at the University of Malta. He received his PhD degree in Applied Statistics in 2005 from Lancaster University. His research specialization areas are related to statistical models, which include Generalized Linear models, Latent Class models, Multi-Level models and Random Coefficient models. He is presently a lecturer in the Statistics department at the University of Malta.

# A Petri net framework for the modeling, simulation and data analysis of biological models

Simon Hardy

Mount Sinai School of Medicine, New York, USA

Pierre N. Robillard

École Polytechnique de Montréal, Montréal, Canada

simon.hardy@mssm.edu, pierre-n.robillard@polymtl.ca

## KEYWORDS

Petri nets, Biochemical modeling, Model validation, Simulation, Quantitative analyses, Invariants

## Abstract

Petri nets have been used for the modeling and simulation of molecular biology systems for almost two decades. Different methodologies and techniques have been developed; a modeling and simulation software tool designed specifically for biological applications has been implemented. In this paper, we discuss the integration of some of these methodologies and techniques into a unified modeling and simulation framework where Petri net theory is used to validate biological models, simulate them and finally, analyze the simulation data.

## Introduction

Computational modeling is increasingly used in cellular biology for the interpretation of biological data. Models mediate between theory and reality. If a model replicates experimental data, then it is a good approximation of reality until it is disproven. In the last few years, modeling and simulation tools have been developed and adapted to biological and biochemical models. A vast majority of modeling tools is based on differential equations, but other formalisms, like Petri nets, offer additional possibilities and techniques useful for modeling. The Petri net formalism is a formal method in computer science that has been applied in numerous domains to analyze systems composed of concurrent and parallel processes. Computational and communication systems, but also industrial, business and manufacturing processes are examples of fields and applications where the use of Petri nets have been reported. Since the original work of Reddy et al. (1993), this formalism has also been used for the analysis of biochemical systems: several contributions have been made in order to use and adapt Petri net techniques to the specificities of biochemical models. Some of these are qualitative approaches for the analysis and validation of Petri net models and they can be very useful to verify their prop-

erties. Others are quantitative approaches for the study of the dynamics of models through simulation with hybrid and or stochastic processes. Some contributions are cited in this paper but articles on the subject are available for an exhaustive review (Hardy and Robillard, 2004; Matsuno et al., 2006b).

The goal of this paper is to discuss how a unified modeling and simulation methodological framework can be based on Petri nets. Such a framework using a graph formalism complements mathematical methodologies like differential equations. The structure of this framework is shown in Figure 1. Also, recent advances in the application of Petri net techniques for the analysis of the simulation data of biochemical systems are presented in this paper. The two main threads of this article are the adaptation of the invariance properties of Petri nets to different biochemical concepts and the use of different types of Petri nets at each phase of the modeling and simulation process.

The next section introduces the Petri net basics. Section 3 presents the Petri net-based techniques for the validation of metabolic and signaling pathway models. Section 4 reviews the simulation of biochemical models with a hybrid Petri net formalism. The last section presents new adaptations of some Petri net concepts to the analysis of the simulation data of signaling pathway models.

## Basic Petri net concepts

This section introduces some basic concepts of Petri net theory that are useful in the context of biological modeling. These concepts constitute only a subset of the whole theory. For a complete presentation of the Petri net theory, readers should consult David and Alla (2004). Among the Petri net concepts, we present the invariants. Invariants have been linked to different entities in biochemical models and we will show in later sections how this relationship can be used for model validation and simulation data analysis.

The original Petri net formalism is a modeling language depicting systems as directed bipartite graphs, i.e. directed arcs linking nodes from two disjoint sets named

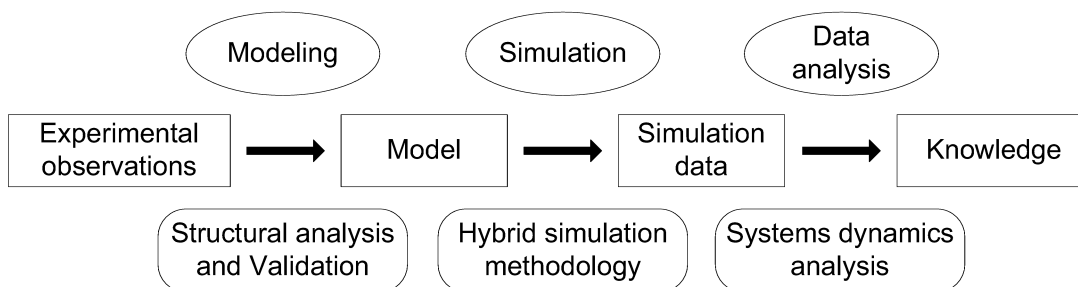


Figure 1: Modeling and simulation process from experimental observations to knowledge using a Petri net framework. The bottom rounded rectangles present the activities that can be performed with Petri nets at each phase of this process when applied to molecular biology.

*places* and *transitions*. Places are the passive elements of the model, like entities and states. Transitions are the active elements of the model, like processes and events. Directed *arcs* represent the causal relations between places and transitions. These arcs are weighted, the default value being 1. In the graphical representation of a Petri net, places are displayed as circles and transitions as rectangles. In Petri net models of a biochemical system, places and transitions most often represent molecular substances and chemical reactions respectively. Also, the weights of the net’s arcs commonly represent the stoichiometric coefficients of chemical reactions.

The dynamic elements of a discrete Petri net are called *tokens*. Places contain an integer number of tokens, called *mark*, and transitions withdraw tokens from or add tokens to places. This process of withdrawing and adding tokens is called transition firing. The firing of a transition can happen only when the transition preconditions are met. If a firing occurs, then the transition postconditions are met. The transition firing has two consequences: tokens are removed from input places, as stated by the preconditions, and tokens are added to output places in a number corresponding to the weights of the output arcs.

The state of a Petri net model is given by the token distribution in its places. The token distribution is called the *marking*. Most of the time, a firing modifies the marking, thus changing the state of the model. In biochemical Petri net models, the tokens are molecules. The molecules change their forms (tokens moving from place to place) as they form molecular complexes, undergo chemical modification, etc. (transition firings). In this kind of Petri net model, the marking indicates the distribution of the molecules between the different molecular substances.

The conceptual framework of Petri nets is usually used to understand “how” a system works. Part of this understanding comes from a mathematical analysis of Petri nets. This is possible because the Petri net formalism is more than just a graphical representation. Petri nets can also be expressed in a linear algebra fashion. The

properties determined with this mathematical representation serve to achieve a qualitative and structural analysis of a system and a validation of a model. Some of these properties are named *invariants*, they are structural properties of Petri nets. In biochemical models, invariants usually have a mass conservation meaning. In the next paragraphs, we present the theoretical background on Petri net invariants. We will elaborate on their biochemical meaning in the next section.

The structure of a Petri net model, i.e. the arrangement of places, transitions and arcs, can be expressed by a matrix. This is the incidence matrix  $\mathbf{W}$ . One dimension of the incidence matrix is the number of places of the model, and the other dimension is the number of transitions. Each element  $w_{ij}$  of the matrix indicates the token change at place  $i$  after firing transition  $j$ . From the incidence matrix, it is possible to determine the invariants. Among all the reachable markings of a model, some quantities do not change, even when transitions are fired. This first type of invariant is the marking invariant (p-invariant). Every p-invariant of a Petri net model is a positive vector  $\mathbf{x}$  that is a solution of the following equation:

$$\mathbf{x}^T \cdot \mathbf{W} = \mathbf{0} \quad (1)$$

A p-invariant characterizes a conservation component of the model, which is a set of places over which the weighted sum of the tokens is constant for every reachable marking.

Among all possible firing sequences of a model, some repetitions are possible. This second type of invariant is a firing invariant (t-invariant). Every t-invariant of a Petri net model is a positive vector  $\mathbf{y}$  that is a solution of the following equation:

$$\mathbf{W} \cdot \mathbf{y} = \mathbf{0} \quad (2)$$

A t-invariant characterizes a repetitive component of the model, which is a firing sequence composed of several transitions causing a return to the model’s initial state. In other words, the transition firings of a repetitive component together have a null effect on the marking of the model.

The invariant properties of Petri nets and different types of Petri nets can be useful at each phase of the modeling and simulation process shown in Figure 1. In this figure, the first layer represents the three activities of this process: modeling, simulation and data analysis. The second layer represents the progression from experimental observations to knowledge. Of course, a feedback loop from simulation to experimental observations is also present to validate the modeling. This loop is not present in the figure but modeling is an iterative process. The third layer represents the techniques that can be implemented with Petri net theory. The next three sections present these techniques.

### Validation of biochemical models with Petri nets

The validation of biochemical models with Petri nets is done by analyzing the structure of the modeled pathways. This means that only the topology of the interconnections between molecular substances, usually specified by stoichiometric coefficients, is considered. Validation approaches do not deal with the kinetic details of the reactions. By validating a model, it is possible to detect inconsistencies. This validation also allows the verification that the model respects some logical and temporal properties. A validation should be performed before any sophisticated questions are asked about the behavior of the model. The Petri net models appropriate for this kind of validation are constructed with the assumption that many chemical reactions can be considered as irreversible (the flux in one direction is negligible in comparison with the flux in the opposite direction). To accomplish a model validation, consistency criteria must be established. The invariants of Petri nets are fine tools to define such criteria. The relationship between the invariant properties and concepts of biochemical models has already been discussed in the literature. This relationship depends on the nature of biochemical models. In metabolic pathways models, the marking invariants express the conservation relations of metabolites and the firing invariants are related to the elementary flux modes (Zevedei-Oancea and Schuster, 2003; Voss et al., 2003; Heiner and Koch, 2004). At steady state, some molecular quantities should remain constant and some enzymatic reactions are essential to maintain this state. These two biochemical concepts correspond to Petri net invariants. The marking invariants can be linked to the constant molecular quantities and the firing invariants can be connected to the steady state flux modes. Consequently, the analysis of a model and the identification of its invariants confirm expected characteristics of the model.

In signal transduction models, marking and firing invariants represent different biochemical concepts. A signaling pathway is considered active when enzymes change state to transmit a signal. The total concentration of all forms of a signaling enzyme is modeled as a constant

quantity. This conserved quantity is a marking invariant in a Petri net model. In signaling pathway models, the firing invariants are associated to the different signal flows of the pathway. The result of these two relationships is that the marking invariants of Petri net signaling models can be used for checking the biological plausibility of the different groups of enzymatic molecules and the firing invariants can be used for checking the biological meaning of certain signal flows. Structural analysis and validation approaches based on these relationships between invariants and biochemical concepts in signaling pathways have been developed (Sackmann et al., 2006; Li et al., 2006). With these approaches, the architecture of the information flow can be decomposed and a simpler picture of the signal networking of these systems can be generated.

Once a model has been validated, the next step is to perform a quantitative analysis using simulation.

### Simulation of biochemical models with Petri nets

The Petri net formalism described so far and used in the presented validation approaches is the original form of this formalism, known as the place-transition net. This modeling language can represent discrete, non-deterministic, asynchronous systems. Other types of Petri nets have been developed to enhance the modeling capabilities of the formalism. For example, continuous and hybrid Petri nets are extensions to the original theory enabling the modeling of continuous quantities and timed processes. Continuous Petri nets were developed to model systems involving flows. This type of Petri nets can represent a system of ordinary differential equations using continuous places and transitions (Matsuno et al., 2000). The marking of a continuous place is represented by a real number. A continuous transition has a speed. Hybrid Petri nets were developed to combine elements of discrete and continuous natures. This last Petri net extension is a hybrid simulation methodology and it has been used to model and simulate different biochemical systems. Models of genetic regulation networks (Doi et al., 2006; Matsuno et al., 2006a), metabolic pathways (Chen and Hofestädt, 2003; Matsuno et al., 2003a) and signal transduction pathways (Matsuno et al., 2003b; Koh et al., 2006; Troncale et al., 2006) have been the objects of quantitative analyses using Hybrid Petri nets. To transform a discrete Petri net model that has been validated into a hybrid Petri net model that can be simulated, dynamic information must be added to the model. A discrete or continuous type must be assigned to each place and transition of the model and parameters must be specified. For discrete transitions, this parameter is a timed delay between events. This value can be deterministic or stochastic. For continuous transitions, the parameter is a speed rate equation. Finally, the places are given initial values corresponding to initial concen-

trations.

A hybrid methodology is appropriate for the simulation of many biological systems for two reasons. First, because certain biological systems combine discrete and continuous processes; this is especially true of information flow through signaling networks in various parts of the neuron (Jordan et al., 2000). Second, because some explicit dynamic parameters cannot be obtained experimentally, thus discrete elements are more appropriate for approximating those parameters. Overall, a hybrid model might be more realistic than a continuous model. The simulation data of hybrid Petri net biochemical models are usually quantitatively analyzed with concentration graphs to study the behavior and dynamics of the models. This straightforward analysis approach has been used in the previously cited examples. However, for complex models, Petri net techniques can also be useful for a more sophisticated analysis of the simulation data to achieve a system-level understanding of the models' behaviors. This new idea is explored in the next section.

### **Analysis of the simulation data of biochemical models with Petri nets**

Systems biology can be summarized as the search for "a system-wide perspective on component interactions, [this perspective] is required so that network properties, such as a particular functional state or robustness, can be quantitatively understood and rationally manipulated." (Sauer et al., 2007) A number of systems biologists turned to computational methods and tools in their search for this system-wide perspective of biological systems (Ideker et al., 2001). These methods comprise computational modeling and simulation methodologies, but also computational techniques for data analysis. Recently, we have developed two Petri net-based techniques for the analysis of the simulation data of hybrid Petri nets. Both techniques are tailored for signal transduction models. They use the invariant properties of Petri nets and their relations to biochemical concepts to process the raw simulation data and hopefully highlight some systemic characteristics of models.

The first technique is a visualization method of simulation data (Hardy and Robillard, 2007). This technique can be summarized in two points. First, it uses the graphical representation of Petri nets to map the simulation data onto the topology of models by coloring its places according to their marking. This animated display presents simultaneously the structure of a biochemical system and its dynamics. Looking at these two dimensions all together facilitates the observation of the model systemic behavior. Second, it uses the marking invariants to color the model structure. This novel use of this Petri net attribute enables the visualization of the simulation data of signaling pathway models according to their main feature. This main fea-

ture is that the propagation of signals in pathways is the result of the activation and deactivation of enzymes through conformational changes as they turn "on" and "off", rather than the production and consumption of metabolites. The effect is to show in a simple manner the concentration distribution of the different conformations of an enzyme, thus informing the viewer of the global state of this enzyme. Since using the marking invariant property for data visualization is a way to emphasize the basic mechanism of signal transduction, the result is a more significant display. This type of presentation of the simulation data is not unique to Petri nets since other tools implements the idea of a mapping of simulation data onto a graph (Qeli et al., 2003; Rost and Kummer, 2004), but the system-meaningful display of signal transduction models created with marking invariants is exclusive to a Petri net technique.

The second Petri net-based technique is a method for the analysis of the dynamics of signal propagation (Hardy and Robillard, 2008). This method is a combination of the mathematical identification of the marking and firing invariants from Petri net theory with a graph exploration algorithm. It has three main features. First, the marking invariant analysis serves to generate a simplified graph representation of a signaling network model. Instead of showing the entire Petri net model and the fine details of every enzymatic activity, this simplified graph displays only the interconnections between different enzymes. Second, the firing invariant analysis serves to identify the network components acting as relays of the informational flux. The activity of these relays, computed from the simulation data, provides temporal information about the signal propagation. Third, it is possible to characterize special transduction paths as regulation motifs, such as positive and negative feedback loops, with a formal definition of these motifs. This method can be a key tool to help computational biologists deciphering the information processing capabilities of complex cellular signaling networks

### **The next steps toward a unified Petri net framework for systems biology**

The Petri net methods previously presented have been developed to support different phases of the modeling and simulation process. Despite the apparent continuity in these applications of Petri net theory to computational systems biology, their successive use in a single project is not without obstacles. Developed independently, not all of these methods have been implemented into software tools, and the ones that use available software have different file formats for the model specifications. For the moment, the unified framework is only of a conceptual nature and centered on Petri net theory. This section is a discussion about the next steps to take toward a concrete unified Petri net framework.

Some methodologies presented in this paper are based



on existing Petri net tools. The Integrated Net Analyzer INA (Starke, 2003) has been used to perform invariant analyses. Also, Cell Illustrator (Nagasaki et al., 2003) is a Petri net model editor and simulator developed specifically for molecular biology. On the other hand, some methodologies presented in this paper have just been theoretical contributions up to now. To be more widely used, software implementation is an unavoidable step for the establishment of a unified Petri net framework.

For existing tools, the compatibility between specifications format is still a hurdle. Two solutions can be adopted to solve this problem. Format conversion software is one solution. Such software is not available for the moment but is in preparation. A unique file format could also be developed and adopted. One format has already been proposed (Chen et al., 2002), but the community did not embrace it, probably because the need for a consensual format has not been felt until now.

Regardless of a growing interest for Petri net methods in the last years, the application of this formalism for molecular biology models remains fairly recent and underused. More biologists need to adopt this formalism and develop Petri net-based models of biological systems in order to have more integrated tools be designed and implemented. For this to happen, a greater number of collaborations between biologists and computer scientists must be established.

## Conclusion

The application of the Petri net theory to molecular biology is reaching a turning point. After 15 years of research on this subject and an increasing number of biologists adopting this formalism, it is now time for the basis of a unified modeling and simulation framework to be established. In this paper, we presented various Petri net-based methods and discussed how they can be integrated into a single framework. These methods are used for structural analysis and model validation, for quantitative analysis via simulation and for the analysis of system dynamics from simulation data. Their common point is mostly how their use of Petri nets invariants for their different significations in biochemical models.

Petri nets are a handy and valuable formalism to biologists. It has a simple and clear graphical representation, similar to the representations already familiar to them. It can also model biological systems more realistically by dealing with the inherent uncertainties related to experimental measurements of the dynamic parameters of the cell. Instead of an accurate kinetic modeling, Petri nets can support simpler mathematical functions. Finally, as we have suggested in this paper, Petri nets offer a unified conceptual framework, assisting biologists to transform experimental observations into knowledge.

## Acknowledgement

This work was supported in part by the grant A-0141 from the National Sciences and Engineering Research Council of Canada and a NSERC Postgraduate Scholarship.

## References

- Chen, Ming, Andreas Freier, Jacob Köhler, and Alexander Rüegg. "The biology Petri net markup language." In *Proceedings of Promise'2002*, edited by J. et al. Desel. 2002, volume 21 of *Lecture Notes in Informatics*, 150–161.
- Chen, Ming, and Ralf Hofestädt. "Quantitative Petri net model of gene regulated metabolic networks in the cell." *In Silico Biology* 3, 3: (2003) 347–365.
- David, René, and Hassane Alla. *Discrete, Continuous, and Hybrid Petri Nets*. Berlin: Springer, 2004. 524 p.
- Doi, Astushi, Masao Nagasaki, Hiroshi Matsuno, and Satoru Miyano. "Simulation-based validation of the p53 transcriptional activity with hybrid functional Petri net." *In Silico Biology* 6: (2006) 0001.
- Hardy, Simon, and Pierre N. Robillard. "Modeling and simulation of molecular biology systems using petri nets: modeling goals of various approaches." *Journal of Bioinformatics and Computational Biology* 2, 4: (2004) 595–613.
- . "Petri net-based visualization of signal transduction pathway simulations." *Computational Biology and Chemistry* (submitted).
- . "Petri net-based method for the analysis of the dynamics of signal propagation in signaling pathways." *Bioinformatics* 24, 2: (2008) 209–217.
- Heiner, Monica, and Ina Koch. "Petri Net Based Model Validation in Systems Biology." In *Proceedings Application and Theory of Petri Nets 2004, Bologna, Italy, June*. Berlin: Springer, 2004, volume 2679 of *Lecture Notes in Computer Science*, 216–237.
- Ideker, Trey, Timothy Galitski, and Leroy Hood. "A new approach to decoding life: Systems biology." *Annual Review on Genomics and Human Genetics* 2: (2001) 343–372.
- Jordan, J. Dedrick, Emmanuel M. Landau, and Ravi Iyengar. "Signaling networks: the origins of cellular multitasking." *Cell* 103, 2: (2000) 193–200.
- Koh, Geoffrey, Huey Fern Carol Teong, Marie-Veronique Clement, David Hsu, and P.S. Thiagarajan. "A decomposition approach to parameter estimation in

- pathway modeling: a case study of the Akt and MAPK pathways and their crosstalk.” *Bioinformatics* 22, 14: (2006) e271–280.
- Li, Chen, Shunichi Suzuki, Qi-Qei Ge, Mitsuru Nakata, Hiroshi Matsuno, and Satoru Miyano. “Structural modeling and analysis of signaling pathways Based on Petri nets.” *Journal of Bioinformatics and Computational Biology* 4, 5: (2006) 1119–1140.
- Matsuno, Hiroshi, Atsuchi Doi, Masao Nagasaki, and Satoru Miyano. “Hybrid Petri net representation of gene regulatory network.” *Pacific Symposium on Biocomputing* 341–352.
- Matsuno, Hiroshi, Sachie Fujita, Atsushi Doi, Masao Nagasaki, and Satoru Miyano. “Towards biopathways modeling and simulation.” In *Proceedings of the 24<sup>th</sup> Conference on Applications and Theory of Petri Nets (ICATPN 2003)*, edited by W. M. P. van der Aalst, and E. Best. Springer-Verlag, 2003a, volume 2679 of *Lecture Notes in Computer Science*, 3–22.
- Matsuno, Hiroshi, Shin-Ichi T. Inouye, Yasuki Okitsu, Yasushi Fujii, and Satoru Miyano. “A new regulatory interaction suggested by simulations for circadian genetic control mechanism in mammals.” *Journal of Bioinformatic and Computational Biology* 4, 1: (2006a) 139–153.
- Matsuno, Hiroshi, Chen Li, and Satoru Miyano. “Petri Net Based Descriptions for Systematic Understanding of Biological Pathways.” *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences* E89-A, 11: (2006b) 3166–3174.
- Matsuno, Hiroshi, Yukiko Tanaka, Hitoshi Aoshima, Atsushi Doi, Mika Matsui, and Satoru Miyano. “Biopathways representation and simulation on hybrid functional Petri net.” *In Silico Biology* 3, 3: (2003b) 389–404.
- Nagasaki, Masao, Atsushi Doi, Hiroshi Matsuno, and Satoru Miyano. “Genomic Object Net: I. A platform for modeling and simulating biopathways.” *Applied Bioinformatics* 2, 3: (2003) 181–184.
- Qeli, Emir, Bernd Freisleben, Daniela Degenring, Aljoscha Wahl, and Wolfgang Wiechert. “MetVis: a tool for designing and animating metabolic networks.” *The European Simulation and Modelling Conference 2003* 333 – 338.
- Reddy, Venkatramana N., Michael L. Mavrovouniotis, and Michael N. Liebman. “Petri net representation in metabolic pathways.” In *Proceedings of the First International Conference on Intelligent Systems for Molecular Biology (ISMB)*, edited by L. Hunter, D. Searls, and J. Shavlik. AIII Press, 1993, 328–336.
- Rost, Ursula, and Ursula Kummer. “Visualisation of biochemical network simulations with SimWiz.” *IEE Systems Biology* 1, 1: (2004) 184 – 189.
- Sackmann, Andrea, Monika Heiner, and Ina Koch. “Application of Petri net based analysis techniques to signal transduction pathways.” *BMC Bioinformatics* 7, 1: (2006) 482.
- Sauer, Uwe, Matthias Heinemann, and Nicola Zamboni. “GENETICS: Getting Closer to the Whole Picture.” *Science* 316, 5824: (2007) 550–551.
- Starke, Peter H. “INA – The Integrated Net Analyzer.”, 2003.
- Troncale, Sylvie, Fariza Tahi, David Campard, Jean-Pierre Vannier, and Janine Guespin. “Modeling and Simulation with Hybrid Functional Petri Nets of the Role of Interleukin-6 in Human Early Haematopoiesis.” *Pacific Symposium on Biocomputing* 11: (2006) 427–438.
- Voss, Klaus, Monika Heiner, and Ina Koch. “Steady state analysis of metabolic pathways using Petri nets.” *In Silico Biology* 3, 3: (2003) 367–387.
- Zevedei-Oancea, Ionela, and Stefan Schuster. “Topological analysis of metabolic networks based on Petri net theory.” *In Silico Biology* 3, 3: (2003) 323–345.

# A GRANULAR UNIFIED FRAMEWORK FOR A MACHINE VISUAL SYSTEM

Mokhtar Beldjehem  
École Polytechnique de Montréal  
C.P. 6079, succ. Centre-Ville  
Montréal QC H3C 3A7, Canada  
E-mail: mokhtar.beldjehem@polymtl.ca

## KEYWORDS

Novel layered granular vision architecture, visual front-end module, mid-level module, fuzzy set, fuzzy partition, level of details, decision levels, granular soft vision (GrSV), Min-Max fuzzy-neuro model, granulation, value approximation, abstraction, non-linear digital filters, soft computing, hybridization, X-rays images.

## ABSTRACT

We propose a novel unifying framework for building a novel layered granular architecture for a machine visual system that accommodates a large spectrum of potential vision problems. Thus removing the ad hoc nature of present solutions and providing the basis for a new generation of machine visual systems. Such a framework works by integrating some useful concepts from the human vision and cognitive processes and adding some interesting granular functionalities of human vision. It advocates further hybridization of non-linear digital filters and soft computing in implementing such a next generation of intelligent machine visual systems. Our focus herein will be on the low level and mid-level stages of such a framework. The goal is to build an automatic system that can be used for degraded multi-modal image processing, including X-rays, MRI, Sonar, etc for diagnosis, recognition, registration and information fusion multipurposes. For illustration purposes, an investigation concerning its application to a real world problem is also provided. We are interested by an application to automatic detection and classification of patients' spines affected by Idiopathic Scoliosis from X-rays images.

## INTRODUCTION AND MOTIVATIONS

It is well-accepted that vision problems are ill-posed, ill-defined and computationally intractable. Nevertheless it is possible to find feasible solution for a large class of practical vision problems. Our aim is at building a general unifying framework enabling the building of an effective modular machine visual processing system in a coherent fashion. Consider our case study, detection and classification of patients' spines affected by Idiopathic Scoliosis from X-rays images, thanks to the fuzziness of the human senses (perception) and the accuracy of human visual system (HVS); a radiologist can detect the pedicles and locate them manually. Why a machine (program) could not? This is due may be to the ability of the mechanism of the perception of

the variation in image brightness by the HVS but also it is connected to the concepts of visual logarithmic non-linearity and spatial adaptability of the HVS. Logarithmic non-linearity has led to the so-called *homomorphic* image processing.

Our approach is well motivated and based on psycho-cognitive considerations of the *theory of first global vision* (Marr 1982). This is what motivates our approach, taking into account Marr ideas (Marr 1982) "*...vision is the process of discovering from images what is present in the world and where it is.*" Some principles of the *gestalt theory*, and in particular the active vision systems approach from them we take our guiding design principles and in conformance with the software engineering point of view: rather than building in an ad hoc manner a visual system for every given problem at hand, it is more attractive and advantageous to try to build a flexible machine visual system able to handle and effectively solve a large number of vision problems using visual data/information/knowledge processing.

The concepts of *granulation* and *abstraction* in a fuzzy set theory setting have long been suggested by Zadeh (Zadeh 1976), his co-authors (Bellman and al. 1966) and advocated by others in an AI (Hobs 1985; Giumchglia et al. 1992) setting, in vision engineering (Marr 1982) setting, and in algorithm design (Foster 1992). It is attracting intensive research too and has led to the development of *granular computing* as an emerging computing paradigm (Yao 2000, Liu and al. 2002, Pedrycz 2001). It has been recently revisited by Zadeh himself (Zadeh 1998) who proposes retargeting it as a design paradigm and/or a methodology in connection with and under the "umbrella" of *soft computing* (SC). Bearing in mind that any workable vision model either mental (human) or computational (machine) is necessarily only *abstraction* and *approximation* of the reality, triangular and/or trapezoidal membership functions (MFs) might be used as they are in fact only approximation means to represent image data, concepts, objects, entities, relationships, classes and even relations of the real world vision problems. Bell-shaped and even free-form membership functions may be used too.

We consider that the *granularity* of a *fuzzy partition* for a variable is of utmost importance as it reflects the *level of details* (or *resolution*) required in describing such a variable, whereas the overlapping is connected to the inherent fuzziness in defining the boundaries between classes (*granules*) of such a variable. Of course a granule is also

defined by a fuzzy set represented by a membership function. Thus it reflects too a gradual rather than abrupt membership of an object to the class (granule).

As early pointed out in (Marr 1982) vision is indeed a data/information/knowledge-processing task. In order to meet such design requirements and to put them working in practice we propose to use and integrate the concepts of *modularity, abstraction, granularity, grouping, operator size, scale-space representation*. Thus our current work is on the mainstream of what has been called *vision engineering* (Marr 1982, 1993) and/or *perceptual engineering* (Nevatia 1982; Jain 1988; Zadeh 2001).

## A CASE STUDY FOR ILLUSTRATION PURPOSES

We are interested by an application to automatic detection and classification of patients' spines affected by Idiopathic Scoliosis from X-rays images. Idiopathic Scoliosis is a pathological condition that can induce a deformation of the spine; it is generally diagnosed soon in the infantile, juvenile or adolescence periods. Scoliosis is an ancient disease that remains incompletely understood despite a collective medical experience that approaches 4000 years. In order to understand the disease and to document the shape of the scoliotic spine, various research studies have been conducted about the positions and orientations of vertebrae in scoliotic patients. Idiopathic scoliosis is the most common type of spinal deformity confronting orthopedic surgeons. Its onset can be rather insidious, its progression relentless, and its end results deadly. Proper recognition and treatment of idiopathic scoliosis help to optimize patient outcomes. Once the disease is recognized, effective ways exist to treat it. Our focus herein is on both visual low level and mid-level processing of an X-rays image. See (Mould 1981) for technical details about X-rays images.

Because of the fuzzy nature of conventional X-rays, they are usually examined by two or sometimes three radiologists. Conventional X-rays photography has depended on the absorption of X-rays in different tissues to form an image. Dense tissue, like bones (vertebras), absorbs more X-rays while soft material like flesh, absorbs less, to create an image of light and dark areas, with images of soft tissue (including pedicles) that lack details and appear fuzzy. Furthermore we are faced by various problems inherent to X-rays image such as highly degraded (very) noisy, low contrast, and the superposition of various structures due to the projection on photographic plate technology used. Characterizing the spinal deformation consists on evaluating the Cobb angles.

The aim of this investigation is to devise an algorithm that automatically detect and localize pedicles enabling computing of the Cobb angles. It is pointed out that idiopathic scoliosis is a complex 3-dimensional deformity. Cobb angles are used as clinic indices to reflect the degree of gravity and are used as references points for the 3D reconstruction of the spine. However segmentation of pedicles is a difficult problem as their contours are thin and weakly contrasted. It is worth mentioning that this phase is only a means not a goal as it prepares for next phases of 3D

image reconstruction and registration or diagnosis useful for both visualization and surgical purposes. For example, this will enable to track the curve progression that varies based on time and on the idiopathic scoliosis group in which a patient belongs (i.e., infantile, juvenile, adolescent).

Due to the nature of an X-rays image, conventional methods used for pre-processing and segmentation are seldom useful in dealing with such a situation. As they need to adjust manually several parameters, a task that has to be undertaken frequently on a case-by-case basis in an intuitive manner and that is not always obvious for the clinicians or surgeons for whom the algorithm is only a black box.

## A NOVEL MACHINE VISION METHODOLOGY

The automatic machine visual system that we are interested to design and build will have modular granular three-levels architecture reflecting three layers of abstractions :

- **Level 1 (low level):** Perceptual stage that corresponds to the raw visual data processing layer. This layer is inspired from the basic idea of "*guessing what is important without knowing why.*"
- **Level 2 (intermediate level or mid-level):** Interfaces stage in-between perceptual/reasoning that corresponds to some kind of visual information processing layer! That is somewhat a task such as attempts to attach meaning to visual data when manipulating those data but without reasoning about those data. For example, segmentation features and primitives selection/extraction falls in this layer. Thus, basically it includes procedures that connect to the raw data and relate it to high-level representations. Defining properties of segments and relationships among segments too fall in this layer. This level may be thought of as a mapping and it should be physically and logically independent from the low level stage.
- **Level 3 (high level):** Reasoning stage (*cognition, linguistic*), that correspond to the knowledge representation and reasoning layer (or a problem-solving layer). Thus, it is contextual by nature and consists of semantic manipulation and is by excellence task-oriented, i. e. oriented toward the resolution of the problem at hand and is application-dependent. Its underlying bulk task is basically implementation issues of how to analyze/understand/recognize/interpret/classify the contents of the image. This level may be thought of as decision making ability and should be physically and logically independent from the intermediate level stage. The focus of attention herein is on resolution of the problem at hand including related tasks. Basically it consists of description of scenes and objects in relation to an object. This corresponds in the HVS to directing attention towards some interesting structures in the scene and description of scenes and objects in relation to an object.

Thus to summarize the framework aims at building a system that starts from low-level raw image data to squeeze or grasp gradually a high-level knowledge-based computational model for vision problem-solving. It is worth mentioning that each level is physically and logically independent of levels below it. This logical and physical independence means explicitly, that both logical and physical changes in level 1 do not affect level 2 and level 3. And changes in level 2 do not affect level 3. It is conceptually clear herein within our framework that we distinguish between the concepts of data (raw), information and knowledge. Indeed, intuitively there is a barrier between perceiving (without semantic) and interpreting (with semantic or meaning). As one can perceive objects without assigning meaning to them, whereas interpreting is pure cognition in essence and involves both linguistic and reasoning. Reasoning has to be related to a given task and hence constitutes the high level processing. Drawing conclusions or making inference (by deduction, by induction, by analogy, etc.) about the objects and their relationships falls in this high level stage.

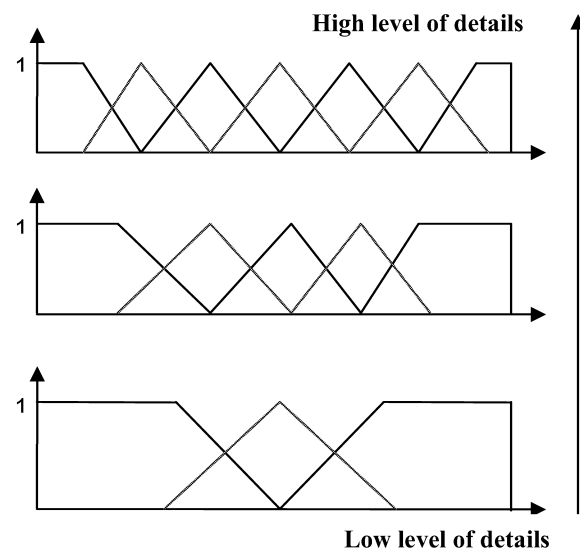
We are interested herein by building two modules: the first module works at the low level as a *visual front-end* computational module that operates on the raw pixel values without any type of pre-processing. It constitutes an interface between the raw image data and the reasoning process. No specific assumption will be made about how higher-level processes are to operate on the output. In other terms it has to be reasoning-independent, therefore this approach will be applicable to a variety of reasoning strategies and finally the system could cope with large number of problems. The second module works at the *mid-level* and plays the role of a mapping in-between the low level and the high level.

*Fuzzy logic* (Zadeh 19965, 1971, 1973, 1979) may be considered as a basis for knowledge and meaning representation and is particularly suited for dealing with Machine Vision problems. We believe that it is a good candidate that constitute the interface between low-level and high-level vision. We believe that it is the concept of *possibility/necessity distributions* (Zadeh 1978), rather than the truth, that will play the primary role in manipulating such vision knowledge for the perspective of drawing conclusions. *Possibility theory* (Zadeh 1978; Yager 1986, Dubois & Prade 1988; Olaf 1998) provides a formal framework for representing and dealing with ignorance, and uncertainties prevalent in modeling real world vision problems.

However it is well accepted that crafting manually fuzzy systems to resolve complex large scale real-world problems is a difficult task that is not always obvious for both the designer (the knowledge-engineer) and the domain expert. This is due partly to the cognitive limits of the human being (Miller 1956), but also to the difficulty of understanding the intricacies of dimensionality and inherent complexities and peculiarities of large scale real world problems. Furthermore once it is undertaken it is labour-intensive, costly, error prone, time-consuming, and done on a trial-and-error basis in an adhoc manner and hence need to be totally or partly automated. This is known as « *the Knowledge aquisition bottleneck* » problem or the Feigenbaum bottleneck and is a

common problem for all AI approaches. Soft computing as automated knowledge acquisition methodology aims at remedying such a problem.

The basic ideas underlying our framework stems from the following interesting remarks about human vision: Let us first focus our attention on the human vision process. In solving vision problems the human starts from a coarse description but if needed iterates and goes gradually to a fine-grained description or in-depth details enabling more understanding of the underlying problem until reaching a point where one can effectively find a solution and so stops and does not need any more details. At this point, an excess of precision is not needed (is not necessary) because a certain satisfying trade-offs between *precision (level of details or resolution)* and *generality* of description has been reached and is sufficient and enough for finding a satisfactory approximate solution to the specified vision problem. Thus after each iteration (increment) a gain of information is obtained enabling more in-depth and more understanding of the underlying situation. Thus, the human converge to a solution gradually by leveraging the level of details.



**Figure 1 From a coarse fuzzy partition to a fine-grained fuzzy partition**

See Figure 1 above for more details in connections with a granular soft vision setting. It is appealing and convenient to mimic mechanically or to emulate computationally such a vision process in order to automatically build faithfully by means of hybrid *Min-Max fuzzy-neuro learning* (Beldjehem 1993) of an appropriate “good” fuzzy visual system that exhibits both a high accuracy and a good performance for any problem at hand. This motivates us in building a learning system able to use such abstraction and granulation mechanisms in a fashion that is akin to the way humans achieve vision problem solving process. In general the required level of details for solving a problem depends to the degree of complexity of the problem at hand and is unknown and hence we propose to detect it during learning-time.

Fuzzy and in particular *hybrid Min-Max Fuzzy-Neuro Systems* (Beldjehem 1993, 1994, 2002, 2004, 2006, 2008) have proved to be more reliable in terms of robustness and effectiveness over conventional ones in resolving a spectrum class of real world vision problems and moreover they have been validated formally (Papis 1991; Beldjehem 2006, 2008) that they preserve the *value approximation topological property* and it is likely that they are too plausible Biologically ! Examples of such map in Biology are : in connection with the human vision, the retinotopic map which takes input from the retina (at the eye) and maps it onto the visual cortex (back of the brain) in a two dimensional map. But also, the somatosensory map which maps our touch centres on the skin to the somatosensory cortex. And also the tonotopic map which maps the responses of our ears to the auditory cortex. Each of these maps is believed to be determined genetically but refined by usage. Thus the MinMax Fuzzy-Neuro possibilistic network might be thought of as a *transparent learning device* of any non-linear mapping of inputs into an output that is being proved formally (mathematically) to be tolerant to small changes in input.

## NON-LINEAR DIGITAL FILTERS

*Non-linear digital filters* (Pitas 1990; Lindeberg 1994) for the segmentation and restoration of images have proved to be more reliable in terms of robustness and effectiveness over conventional linear ones.

In order to tackle this problem appropriately, we propose to perform the following tasks on the X-rays image using non-linear models:

### Low level: Smoothing, image enhancement and restoration (noise reduction and contrast enhancement)

The main interest is to try to suppress and remove unnecessary and disturbing details, such that later stage processing task can be simplified and so as that significant structures (pedicles in our case study) can be extracted from the X-rays image without any prior information. In general image enhancement is necessary to support the human visual perception. Due to the limited authorized weak X-rays dose during acquisition, the radiographies are highly degraded images. Restoration is still a mandatory preprocessing step in order to reduce and remove noise, to eliminate irrelevant details and to enhance relevant structures.

Conventional linear filtering by regularization is unsuited to deal with X-rays image as it tries to more fuzzify the contours and may be to destroy them, whereas anisotropic diffusion tries to keep and enhance them and hence constitute a promising approach. Partial differential equations (PDEs) allow the modeling of such regularization. PDEs belong to one of the most important parts of mathematical analysis, are closely related to physical world. One of the main interests in using PDEs is that the theory is well established. Moreover PDEs enable the combination of the *scale-space theory* with *mathematical morphology*.

We propose to develop new smoothing kernels for highly degraded image using PDEs and diffusion fields. This filter incorporates specific diffusion fields and since each of which account for characteristics of a given application, it brings a new degree of freedom to it, in order to address various problems of greater practical interests. Other alternative that consist to replace a diffusion field by a non-linear tensor will be also investigated.

The real challenge is to keep the shape of pedicles unchanged while reducing the noise we propose to develop a new chock filter for image enhancement while ensuring contrast intensification between ill-defined regions of the X-rays image.

### Intermediate level: Image segmentation with prior knowledge

The problem of image segmentation plays a key role for the later high-level stage. We suggest using the mumpford-shah framework; we propose to develop new methods for object extraction through *the level set method* that is parameter free and can account for topological changes and integrate visual cues. This is achieved by minimizing the Mumford-Shah functional. Better segmentation results are expected by incorporating information. Biplanar radiography of scoliotic patients is routinely performed at North-American hospitals. Thus, a sufficient amount of data is available for analysis.

This motivates us in trying to exploit the a priori information concerning the shape and perimeter of a pedicle. In fact a pedicle has a close to ellipsoidal shape. We propose to develop an algorithm similar to Cremers for segmentation of pedicles. This necessitates building a pedicles database. We will extract manually the pedicles contours from available vertebrae radiographies. Spinal deformation affects neither the shapes of vertebrae nor the shapes of pedicles. It affects only their localization and spatial orientation. Our novel model is one-object driven problem that incorporates an information distance into the popular region-based active contour (the *Mumford-Shah* framework), along with prior knowledge. This method has several advantages: it uses most of the laws of the *Gestalt theory* (pixel grouping by a characteristic, regularity of the border, prior knowledge), and is computationally tractable.

## HYBRIDIZATION OF NON-LINEAR DIGITAL FILTERS AND SOFT COMPUTING

Since 1990, soft computing and in particular hybrid fuzzy-neuro or neuro-fuzzy systems have invaded the computer world and constitutes one of the most exciting current topics of research (Beldjehem 1993; Yager and Zadeh 1994; Sinha and Gupta 1999; Pal and Ghosh 2000; Gupta and al. 2002), the advances are also spectacular due to its newness, perspectives and power. Moreover it has been proved that soft computing (Zadeh 1998; Pal and Ghosh 2000) can also be applied successfully in dealing with grey level images, and for this reason we are exploring the possibility of hybridization of non-linear filters and soft computing

models, in particular synergy through *hybridization* ensures the emergence of unexpected desirable properties. We believe that hybridization is a promising methodology and is technically feasible and there exist some gaps to bridge as they are complementary rather than competitive.

In the sequel; we will focus on the relevance of soft computing to the problem: Soft computing provides non-linear computational models that have human like decision making capabilities for processing and analyzing of visual/image data. It could be effectively incorporated everywhere within the three levels. Consider the problem of object extraction from a graytone image such as an X-rays: (How can one define exactly the target or object region in the image when its boundary is ill-defined?). Any conventional hard thresholding made for the extraction of the object will propagate the associated uncertainty to subsequent stages (e. p., thinning, skeleton extraction, primitive selection, etc.) and this might, in turn, affect feature analysis and recognition. Decision taken at a particular stage will have an impact on the subsequent stages. Grey information is expensive and informative and has to be kept until making a decision at the highest level.

Fuzzy logic algorithms are based on the ground of the rigorous and the formal mathematical of fuzzy sets theory and they could effectively play the numerical-symbolic interface role as well as they could be incorporated in various stage of our framework thanks to their flexibility, in particular extracting fuzzy primitives (features) allows capturing the inherent vagueness of edges and segmented outputs of image regions for recognition purposes. Uncertainty in an image pattern may be explained in terms of grayness ambiguity or spatial (geometrical) ambiguity or both. Regions in an image are not always crisply defined; uncertainty can arise within every phase of the aforementioned tasks. It seems appealing and convenient to use soft computing techniques. Grayness ambiguity means “indefiniteness” in deciding whether a pixel is white or black. Spatial ambiguity refers to “indefiniteness” in the shape and geometry of a region within the image. Grayness ambiguity measures are reflected by index of fuzziness and entropy, whereas spatial ambiguity measures are represented by fuzzy geometrical properties. It is convenient, natural and appropriate to avoid committing our selves to specific (hard) decisions (e. g., segmentation, edge detection, and skeletonization):

- by allowing the segments or skeletons or contours to be fuzzy subsets of the image
- The results of image segmentation should be fuzzy subsets rather than ordinary subsets
- In general, the subsets being characterized by the possibility (degree) to which each pixel belongs to them.

Soft computing provides the machinery for solving such a problem. We will obtain different fuzzy segmented version of the same image. These different outputs correspond to different ambiguity values (or *decision levels*). In fact if the gray levels are scaled to lie in the range [0, 1], a graytone image can be viewed as a fuzzy set. Therefore all fuzzy

operators including, contrast intensification, erosion, dilation,  $\alpha$ -cuts, etc. are straightforward to implement. But now, with soft computing techniques, it is possible to obtain pin sharp pictures, which not only reveals bones ... but muscles, pedicles, blood vessels and cartilage flesh. Soft computing exploits the tolerance for imprecision, uncertainty and partial truth to achieve tractability, robustness and low solution cost

Explicitly hybrid soft computing could be adequately applied in all levels namely at the point level, local level, global level, object level (pedicle). The real income will be to keep the shape of pedicles unchanged while reducing the noise, for deferent decision making purposes. Furthermore the detection algorithm should be invariant to translation and rotation. Thus, the pedicles will be segmented accurately over patients adopting different position and having various deformations. Integrating a priori knowledge about the close to ellipsoidal shape of a pedicle is possible and can guide the segmentation procedure. To detect edge we measure the rate of change of gray level (membership value) at each pel.

On the other hand, an evolutionary algorithm (EA) as in (Pedrycz 1997; Cordon et al. 2001) is an optimization technique based on the mechanics of natural selection and natural genetics. It has a great power for global optimization and does not need to know the model previously. It also does not require the continuity of the parameters. Therefore EA can easily handle the multi-parameter problems and for this reason it seems appealing and convenient to use genetic algorithms too. As non-linear filters involve the tuning and optimization of several parameters, EAs can effectively contribute significantly in image enhancement thanks to their learning and optimization capabilities. In particular to try to fuzzify concepts used by genetic algorithms to obtain and use fuzzy fitness functions, fuzzy crossover, fuzzy mutation and so on, ensuring *smooth evolvability*, using SVMs should also be considered for such a task as another feasible alternative.

Various architectures for integrating non-linear digital filters and soft computing are possible : From loosely-coupled, through tightly coupled, to fully integrated and intelligent cooperation.

A promising perspective should be to try to design new fuzzy non-linear filters, fuzzy-neuro or neuro-fuzzy filters, rule-based filters, and/or new fuzzy morphological operators. This can be achieved either by trying to fuzzify the existing known non-linear filters and morphological operators or by radically building new ones from scratch. Approximate models based on if-then fuzzy rules integrating human-like processing should constitute a promising privileged approach. This is in fact due to their transparency and interpretability.

A segmentation problem might be formulated as a clustering problem and thus using techniques from clustering algorithms such as fuzzy C-means techniques constitute an effective promising approach in segmentation of an X-ray

Image. We may too combine them using cascade non-linear digital filters followed by C-means strategy or conversely.

In practice we might either fuse or embedd symbiotically or combine non-linear digital filters with soft computing models using a cascade strategy. The possibility of making fusion of the merits of each one for improved quality is feasible. Ultimately such hybridization methodology will contribute to the conception and design of next generation of hybrid evolvable soft non-linear fuzzy filters ensuring performance close to the accuracy of human visual system (HVS).

## CONCLUDING REMARKS

This work might be thought of as an attempt to engineer granular soft vision (GrSV) systems that operates at the low level as well as at the mid and high levels. Implementation of the two *visual front-end* and *mid-level* modules working under such a framework is underway. Such convenient easy to develop, easy to debug and easy to maintain *layered granular architecture* accommodates well and overcomes the complexity of the software engineering of the visual system leaving us only focusing on algorithms issues that will allow the implementation of a coherent solution.

We believe that non-linear digital filters and soft computing models have to learn from each other, and could be coupled synergistically (not competitively) in order to build new generation of hybrid models for the segmentation and restoration of grayscale images. A challenge is to build such hybrid models having advantages of performance, non-linearity, accuracy, stability, robustness, adaptability, tractability, tolerance for uncertainty, edge preservation and detection properties. The possibility of making fusion of the merits of each one for improved quality is feasible.

We have addressed major problems of greater practical interest. Algorithms and models that are being developed in the course of our project will have generic applicability in various applications, in particular to mention only a few, in Sonar, in OCR of contents from strongly degraded documents, systems employed in brain surgery on humans and other medical imaging problems.

Even though we are more interested in vision engineering rather than (natural) vision, i.e. in developing new, powerful and useful vision tools that learn for resolving real-world vision problems, we believe that as we understand better how to build these computational vision systems we'll start to have theories that are powerful enough to explain some aspects of the human visual system (HVS).

We hope that this paper will be a starting point for more integration of non-linear digital filters and soft computing models that will definitely allow building flexible machine vision systems that mimic the human visual system (HVS) tackling practical industrial and medical complex real world vision problems of great importance.

The advantages of this hybrid methodology might be summarized as follows:

- To benefit from the synergism of non-linear digital filter and soft computing models (complementarity rather than competitity to ensure capability enhancement)
- To benefit from the advantages of each one and avoid the weaknesses of each one, to cope with the complexity of multifaceted real-world vision problems
- To start from low-level raw data to squeeze or grasp high-level knowledge-based computational models for problem-solving. To obtain clarification and to allow verification and validation
- To overcome the problem of knowledge engineering known as « *the Knowledge aquisition bottleneck* » or the Feigenbaum bottleneck
- To learn a high level representation (expressed in terms of IF/THEN Fuzzy production rules)
- To work at a higher level of abstraction during inference and classification (third level or high level stage)
- To ensure good *performance-interpretability* tradeoffs
- To build physically, biologically and cognitively motivated vision models
- To understand and replicate the human visual system (HVS), the human perception, cognition and intelligence

According to Zadeh (Zadeh 1994) "The exploitation of tolerance for imprecision and uncertainty underlies the remarkable human ability to understand distorted speech, decipher sloppy handwriting, comprehend nuances of natural language, summarize text, *recognize and classify images* and more generally, make rational decision in an environment of uncertainty and imprecision." This ability is what granular hybrid soft computing (GrSC) systems try to capture by learning and emulate computationally in general and what soft computing can bring especially to machine visual systems and non-linear digital filters.

An extension and greater improvement of the framework and the model are worth further consideration. We hope that this will serve as a starting point or a pointer for those interested in pursuing this line of vision engineering research.

## REFERENCES

- Beldjehem M. 1993. "Un apport à la conception des systèmes hybrides neuro-flous par algorithmes d'approximation d'équations de relations floues en MIN-MAX: le système Fennec." Ph.D. Thesis in Computer Science and Software Engineering (Artificial Intelligence), Université de la Méditerranée-Aix-Marseille II, Marseille (in French).
- Beldjehem M. 1993. "Fennec, un générateur de systèmes neuro-flous." in *Proceed. les Actes des Applications des Ensembles Flous*, Nîmes, France, 209 -218 (in French).
- Beldjehem M. 1993. "Le système fennec." in *Electronic BUSEFAL* 55, 95-104 (in French).



- Beldjehem M. 1994. "The fennec system." in *Proceed. ACM Symposium on Applied Computing (SAC), Track on fuzzy logic in Applications*, 126-130, Phoenix, AZ (March).
- Beldjehem M. 2002. "Machine Learning based on the possibilistic-neuro hybrid approach: design and implementation." in *Electronic BUSEFAL* 87, 95-104.
- Beldjehem M. 2002. "Learning IF-THEN Fuzzy Weighted Rules." in *Proceed. International Conference of Computational intelligence*, Nicosia, North Cyprus.
- Beldjehem M. 2006. "Validation of Hybrid MinMax Fuzzy –Neuro Systems." in *Proceed. International conference of NAFIPS*, Montreal.
- Beldjehem M. 2008. "Towards a Validation Theory of Hybrid MinMax Fuzzy–Neuro Systems." in *the Proceed. WSEAS International Conference*, Sofia.
- Beldjehem M. 2008. "Towards a Validation Theory of Hybrid MinMax Fuzzy–Neuro Systems." in *the Proceed. CIMSA International Conference*, Istanbul.
- Bellman R.; R. Kalaba R.; and Zadeh L. 1966. "Abstraction and Pattern Classification". *J. Math. Anal. Appl.* 13, 1-7.
- Cordon O.; Herrera F.; Hoffman F.; and Magdalena L. 2001. Genetic Fuzzy System, Evolutionary Tuning and Learning of Fuzzy Knowledge Bases. World Scientific.
- Dubois D. and Prade H. 1988. *Possibility Theory: An Approach to Computerized Processing of Uncertainty*, Plenum Press, New York, USA.
- Foster C. L. 1992. *Algorithms, Abstraction and Implementation: Levels of details in cognitive sciences*. Academic Press, London.
- Giunchiglia F. and Walsh T. 1992. "A theory of abstraction." *Artificial Intelligence* 56, 323-390.
- Gupta M. M.; Jin L.; and Homma N, 2002. *Static and Dynamic Neural Networks: From Fundamentals to Advanced Theory*. John Wiley and Sons Inc., New York.
- Hobbs J. R. 1985. Granularity, *Proceed. of the 9<sup>th</sup> International Joint Conference on Artificial Intelligence*, 432-435.
- Jain R. 1988. "Perception engineering". *Mach. Vision Appl.* 1, 73-74.
- Lindeberg T. 1994. *Scale-Space Theory in Computer Vision*. Kluwer, Dordrecht.
- Liu, T. Y.; Yao Y. Y.; and Zadeh L. A. 2002. *Data Mining, Rough Sets and Granular Computing*. Physica-Verlag, Heidelberg.
- Marr D. 1982. *Vision, A Computational Investigation into Human Representation and Processing of Visual Information*. Freeman W. H. and Company, San Francisco.
- Marr D. 1993. *Vision Engineering: Designing Computer Vision Systems*, (with A. Rosenfeld) in *Handbook of Pattern Recognition and Computer Vision*, (eds.), C.H. Chen, L.F. Pau and P.S.P. Wang, World Scientific Publishing Company, Singapore, 805-815.
- Mould R F . 1981. *Radiotherapy Treatment Planning*. Adam Hilger Ltd, Bristol and Boston.
- Miller G. 1956. "The magical number seven, plus or minus two." *The Psychological Review* 63, 81-97.
- Nevatia R. 1982. *Machine Perception*. Prentice-hall, Englewood.
- Olaf W. 1998. *Possibility Theory with Application to Data Analysis*. John Wiley & Son inc.
- Pal S. K. and A. Ghosh A. (Ed.). 2000. *Soft Computing in Image Processing*. Physica-Verlag, Heidelberg.
- Pappis C. P. 1991. "Value approximation of fuzzy systems variable," *Fuzzy sets and systems* 39, 111-115.
- Pedrycz W. (Ed.). 1997. *Fuzzy Evolutionary Computation*, Kluwer Academic Publishers.
- Pedrycz W. (Ed.). 2001. *Granular Computing: An emerging paradigm*, Springer. Series: Studies in Fuzziness and Soft Computing 70.
- Pitas I. and Venetsanopoulos A. N. 1990. *Nonlinear Digital Filters, Principles and Applications*, Kluwer Academic Publishers, Boston.
- Sinha N. K. and Gupta M. M. 1999( Ed.). *Soft Computing and Intelligent Systems: Theory and Applications*. Academic Press, New York.
- Yager R. R. 1986. *Fuzzy Set and Possibility Theory: Recent Developments*. Pergamon Press, New York.
- Yager, R. R. and Zadeh L. A. 1994. *Fuzzy Sets, Neural Networks and Soft Computing*. Van Nostrand Reinhold, New York.
- Yao, Y.Y. 2000. "Granular computing: basic issues and possible solutions." *Proceedings of the 5th Joint Conference on Information Sciences*, 186-189.
- Zadeh L. A. 1965. "Fuzzy sets." *Info. Control* 89, 338-353.
- Zadeh L. A. 1971. "Toward a theory of fuzzy systems." in R. E. Kalman, N. Declaris, Eds., *Aspects of Network and System Theory* (Holt, Rinehart and Winston, New York), 209-245.
- Zadeh L. A. 1973. "Outline of a new approach to the analysis of complex systems and decision processes." *IEEE Trans. Syst. Man Cybernet.* 3, 28-44.
- Zadeh L. A. 1976. "Fuzzy sets and information granulation." in *Advances in Fuzzy Set Theory and Applications*, in M. Gupta, R. K. Ragade, R. R. Yager (Eds.), North –Holland Publishing Company, 3-18.
- Zadeh L. A. 1978. "Fuzzy sets as a basis for a theory of possibility." *Fuzzy sets and syst.* 1, 3-28.
- Zadeh L. A. 1979. "A theory of approximate reasoning." in *Machine Intelligence 9* (J.E. Hayes et al.; Eds). Elsevier, 149-194.
- Zadeh L. A. 1984. "A theory of commonsense knowledge." in *Aspect of Vagueness* (H.J. Skala, S. Termini and E. Trillas; Eds). *Dodrecht: Reidel*, 257-295.
- Zadeh L. A. 1994. "Fuzzy logic neural networks, and soft computing." *Communications of the ACM* 37, 77-84.
- Zadeh L. A. 1997. "Towards a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic." *Fuzzy Sets and Systems* 19, 111-127.
- Zadeh L. A. 1998. "Some reflections on soft computing, granular computing and their roles in the conception, design and utilization of information / intelligent systems." *Soft Computing* 2, 23-25.
- Zadeh L. A. 1998. "Soft Computing, Fuzzy Logic and Recognition Technology." In *Proc. IEEE Int. Conf. Fuzzy Syst.*, Anchorage, AK, 1678-1679.
- Zadeh L. A. 2001. "A new AI: Toward computational theory of perceptions." *AAAI Magazine* 22, No. 1, 73-84, Springer.



# **WEB COMPUTING AND COGNITIVE SIMULATION FOR PROBLEM SOLVING**



# A NEW FRONTIER IN COMPUTATION—COMPUTATION WITH INFORMATION DESCRIBED IN NATURAL LANGUAGE

Lotfi A. Zadeh<sup>1\*</sup>

What is meant by Computation with Information Described in Natural Language, or NL-Computation, for short? Does NL-Computation constitute a new frontier in computation? Do existing bivalent-logic-based approaches to natural language processing provide a basis for NL-Computation? What are the basic concepts and ideas which underlie NL-Computation? These are some of the issues which are addressed in the following.

What is computation with information described in natural language? Here are simple examples. I am planning to drive from Berkeley to Santa Barbara, with stopover for lunch in Monterey. It is about 10 am. It will probably take me about two hours to get to Monterey and about an hour to have lunch. From Monterey, it will probably take me about five hours to get to Santa Barbara. What is the probability that I will arrive in Santa Barbara before about six pm? Another simple example: A box contains about twenty balls of various sizes. Most are large. What is the number of small balls? What is the probability that a ball drawn at random is neither small nor large? Another example: A function,  $f$ , from reals to reals is described as: If  $X$  is small then  $Y$  is small; if  $X$  is medium then  $Y$  is large; if  $X$  is large then  $Y$  is small. What is the maximum of  $f$ ? Another example: Usually the temperature is not very low, and usually the temperature is not very high. What is the average temperature? Another example: Usually most United Airlines flights from San Francisco leave on time. What is the probability that my flight will be delayed?

Computation with information described in natural language is closely related to Computing with Words. NL-Computation is of intrinsic importance because much of human knowledge is described in natural language. This is particularly true in such fields as economics, data mining, systems engineering, risk assessment and emergency management. It is safe to predict that as we move further into the age of machine intelligence and mechanized decision-making, NL-Computation will grow in visibility and importance.

Computation with information described in natural language cannot be dealt with through the use of the machinery of natural language processing. The problem is semantic imprecision of natural languages. More specifically, a natural language is basically a system for describing perceptions. Perceptions are intrinsically imprecise, reflecting the bounded ability of sensory organs,

and ultimately the brain, to resolve detail and store information. Semantic imprecision of natural languages is a concomitant of imprecision of perceptions.

Our approach to NL-Computation centers on what is referred to as generalized-constraint-based computation, or GC-Computation for short. A fundamental thesis which underlies NL-Computation is that information may be interpreted as a generalized constraint. A generalized constraint is expressed as  $X \text{ isr } R$ , where  $X$  is the constrained variable,  $R$  is a constraining relation and  $r$  is an indexical variable which defines the way in which  $R$  constrains  $X$ . The principal constraints are possibilistic, veristic, probabilistic, usuality, random set, fuzzy graph and group. Generalized constraints may be combined, qualified, propagated, and counter propagated, generating what is called the Generalized Constraint Language, GCL. The key underlying idea is that information conveyed by a proposition may be represented as a generalized constraint, that is, as an element of GCL.

In our approach, NL-Computation involves three modules: (a) Precision module; (b) Protoform module; and (c) Computation module. The meaning of an element of a natural language, NL, is precisiated through translation into GCL and is expressed as a generalized constraint. An object of precision,  $p$ , is referred to as precisiend, and the result of precision,  $p^*$ , is called a precisiand. Usually, a precisiend is a proposition, a system of propositions or a concept. A precisiend may have many precisiands. Definition is a form of precision. A precisiand may be viewed as a model of meaning. The degree to which the intension (attribute-based meaning) of  $p^*$  approximates to that of  $p$  is referred to as cointension. A precisiand,  $p^*$ , is cointensive if its cointension with  $p$  is high, that is, if  $p^*$  is a good model of meaning of  $p$ .

The Protoform module serves as an interface between Precision and Computation modules. Basically, its function is that of abstraction and summarization.

The Computation module serves to deduce an answer to a query,  $q$ . The first step is precision of  $q$ , with precisiated query,  $q^*$ , expressed as a function of  $n$  variables  $u_1, \dots, u_n$ . The second step involves precision of query-relevant information, leading to a precisiand which is expressed as a generalized constraint on  $u_1, \dots, u_n$ . The third step involves an application of the extension principle, which has the effect of propagating the generalized

<sup>1</sup> Dedicated to Peter Walley.

\* Department of EECS, University of California, Berkeley, CA 94720-1776; Telephone: 510-642-4959; Fax: 510-642-1712; E-Mail: [zadeh@eeecs.berkeley.edu](mailto:zadeh@eeecs.berkeley.edu). Research supported in part by ONR N00014-02-1-0294, BT Grant CT1080028046, Omron Grant, Tekes Grant, Chevron Texaco Grant and the BISC Program of UC Berkeley.

constraint on  $u_1, \dots, u_n$  to a generalized constraint on the precisiated query,  $q^*$ . Finally, the constrained  $q^*$  is interpreted as the answer to the query and is retranslated into natural language.

The generalized-constraint-based computational approach to NL-Computation opens the door to a wide-ranging enlargement of the role of natural languages in

scientific theories. Particularly important application areas are decision-making with information described in natural language, economics, systems engineering, risk assessment, qualitative systems analysis, search, question-answering and theories of evidence.

# SIMULATION OF A HUMAN MACHINE INTERACTION: LOCATE OBJECTS USING A CONTEXTUAL ASSISTANT

Chikhaoui Belkacem  
Pigot Hélène

Domus Laboratory, Computer Science Department, Faculty of Science,  
University of Sherbrooke, Sherbrooke (QC), J1K 2R1 Canada  
email: {belkacem.chikhaoui, helene.pigot}@usherbrooke.ca

## ABSTRACT

The standard development of human machine interfaces needs the respect of ergonomic norms and rigorous approaches, which constitutes a major concern for computer system designers. The increased need on easily accessible and usable interfaces leads researchers in this domain to create methods and models that make it possible to evaluate these interfaces in terms of utility and usability. This paper presents a study about the simulation of a human machine interaction with an interface of a contextual assistant, using the cognitive architecture ACT-R emphasizing on the time execution of tasks. The results of our model were consistent with those obtained by the Fitts Law model which is a powerful analytical method for evaluating human machine interfaces, developed in this study mainly to support our results.

## INTRODUCTION

The evaluation of Human Machine Interfaces (HMI) is becoming increasingly important and constitutes an integral part in the development cycle of computer systems. While the development of interfaces presents some challenges, their evaluation needs rigorous methods to ensure they fulfill the initial specifications and the quality of accessibility, usability and usefulness (Nielsen and Phillips, 1993; Eugenio et al., 2003). Two main approaches for evaluation are currently used, empirical approaches and analytical approaches. Empirical approaches are essentially based on performances or opinions of users gathered in laboratories or other experimental situations. These approaches are user-focused. Unlike the empirical approaches, analytical approaches are not based directly on the user performance, but rather, on the automated examination of interfaces using well-defined structures and rigorous analysis techniques.

The HMI should be resumed by the actions of pushing buttons displayed on a screen. According to this approach the Fitts law estimates the time needed to reach the targets displayed on the interface. Nevertheless, the HMI implies three human components, which must be taken in account. The first component is perceptual.

In our case the human perceives the signal in a visual manner. The second one is cognitive. Here the human retrieves in his memory the object required and reasons to satisfy specific goals. The third one is motor and necessitates pressing on the selected button.

In this study, we aim to evaluate the interaction with an interface of a contextual assistant developed for cognitively impaired people. The aim of this application is to assist people while preparing meals in their kitchen by using cognitive assistance (Pigot et al., 2005). Due to the related population and the kind of errors they commit we need to take in account the cognitive part involved in the HMI. We then use a powerful analytical method based on cognitive models, emphasizing the cognitive analysis of the tasks and the time execution. We choose to base our analytical method on the cognitive architecture ACT-R (Anderson et al., 2004). Thanks to ACT-R the interaction is decomposed in rules simulating the cognitive behavior of a human using the contextual assistant. We first present an overview of the cognitive architecture ACT-R and of the contextual assistant. Once the task simulated is defined, the model, we developed, is introduced and the results of the simulation are compared to the time estimated by the Fitts law to interact with the contextual assistant.

## BACKGROUND

In this section we present an overview of the cognitive architecture ACT-R, and then we introduce the contextual assistant application and the interface to be modeled.

### Cognitive architecture ACT-R

The cognitive architecture ACT-R is built to simulate and understand human cognition (Anderson et al., 2004, 2005). It consists of a set of modules integrated through a central production system. ACT-R is an hybrid architecture that combines two subsystems: symbolic system including semantic and procedural knowledge, and subsymbolic system evaluating knowledge activations. The subsymbolic system assigns activations to chunks (semantic knowledge) and rules (procedural knowledge).

The activation level is one of the criteria to choose the more predominant knowledge available at a specific time. In ACT-R the perceptual and motor modules are used to simulate interfaces between the cognitive modules and the real world (Byrne, 2001; Bothell, 2004).

### Visual and Motor Modules of ACT-R

The visual module that is part of the perceptual modules, has two subsystems, the positional system (where) and the identification system (what) that work together in order to send the specified chunk to the visual module. The positional system is used to find objects. When a new object is detected, the chunk representing the location of that object is placed in the visual-location buffer according to some constraints provided by the production rule (Bothell, 2004). The identification system is used to attend to locations which have been found by the positional system. The chunk representing a visual location will cause the identification system to shift visual attention to that location. The result of an attention operation is a chunk, which will be placed in the visual buffer (Byrne, 2001; Bothell, 2004). The motor module contains only one buffer through which it accepts requests (Bothell, 2004). Two actions are available in ACT-R, to click with the mouse or press a key on the keyboard.

### Contextual Assistant

The Contextual assistant application is developed to assist persons with cognitive disabilities (Pigot et al., 2007a; Lussier-Desrochers et al., 2007). The aim is to foster autonomy in the daily living tasks and particularly during complex cooking tasks such as preparing pancakes, or spaghetti (Pigot et al., 2007b). The cooking task is decomposed of steps displayed on a touch screen. The two first steps consist of gathering the utensils and ingredients necessary to the recipe (Figure 1). The other steps explicit the recipe using photo and video on the screen as well as information dispatched all around the kitchen. The contextual assistant is specifically designed to help people remembering the places where the objects are stored. To do so, the contextual assistant contains an interface called the locate application displaying the objects to search. When an object is pushed in the main interface, the contextual assistant looks for the location of that object in the environment using techniques of pervasive computing and indicates the location by highlighting the appropriate locker containing that object as shown in Figure 2. In this study we simulate the first two steps of the spaghetti recipe. They consist of first knowing the list of objects to gather, either utensils or ingredients, and then to use the locate application in order to find each object.

The contextual assistant interface is displayed on a 1725L 17" LCD Touchscreen, with 13.3" (338 mm) hor-

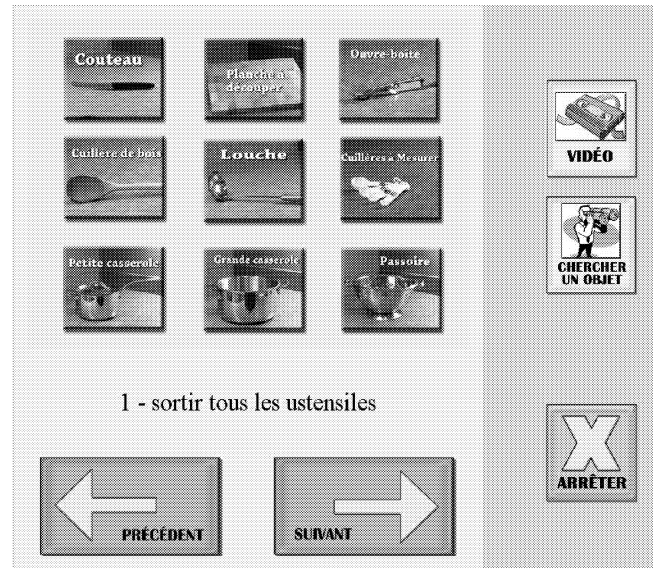


Figure 1: Main interface of the contextual assistant

izontal and 10.6" (270 mm) vertical useful screen area. It is configured to 1024 x 768 optimal native resolution running Macintosh. The screen is fixed under a closet nearby the oven in order to be easily accessible and also protected against the cooking splashes.

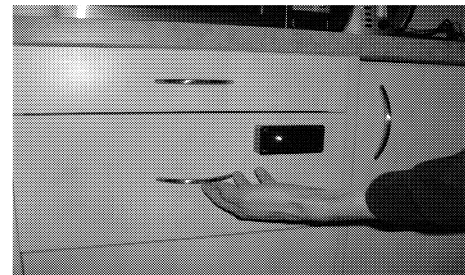


Figure 2: Locker state when an object is pushed

### MODELING THE INTERACTION WITH THE CONTEXTUAL ASSISTANT USING ACT-R

In this section, we present the modeling process of the tasks involved in our study, which are gathering utensils and gathering ingredients, emphasizing on the perceptual and cognitive parts, using the perceptual motor modules of ACT-R.

#### Task analysis: gathering utensils and ingredients

We model the first two steps of the recipe, gathering utensils and gathering ingredients. The interactions with the touch screen are simulated without taking in account the time taken by the subject to pick up the objects in the environment. The two first steps require



three subtasks (Figure 3). The first subtask consists of activating the locate application in order to locate each object required by the recipe. This is done by pushing the button "LOOK-FOR-OBJECT" (in French, "CHERCHER-UN-OBJET"), which is displayed on the main interface of the contextual assistant (Figure 1). The second subtask is to locate each object, either utensils or ingredients, needed in the current step by pushing the button corresponding to the object in the locate application. The third task consists of coming back to the main application in order to know the next step of the recipe. The tree decomposition is presented in figure 3, where the translation in English is available to compare the tasks tree from the interface of Figure 1. The nodes in capital indicate the action to click on the named button, while the other nodes represent tasks to be decomposed.

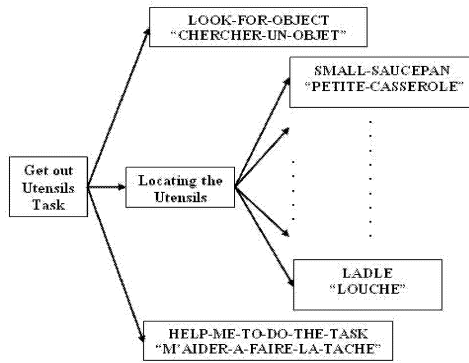


Figure 3: Tree representing the gathering utensils task

### Gathering ingredients and utensils model

The model developed aims to simulate the HMI during the two first steps of the recipe. In that task, three different interfaces are involved, the interface of the locate application and the two of the contextual application displaying the utensils and ingredients needed in the recipe. The model uses ACT-R to emphasize the cognitive processes involved when looking for an object and choosing the button to push. It is decomposed of three phases, the visual phase, the recognizing phase and the motor phase. The visual phase consists of localizing the object to perceive and then identifying it. We consider that all buttons displayed on the screen are objects, either the button used to locate a utensil or ingredient, or the buttons to navigate in the interface. The first one is the button "LOOK-FOR-OBJECT" as described in Figure 3. Then, all the utensils needed in the recipe are presented in the visual interface of ACT-R. Finally, to complete the first step of the recipe, the button "HELP-ME-TO-DO-THE-TASK" is presented in order to come back to the main interface of the contextual assistant and pursue the second step of the recipe. Each object

of the interface is displayed at defined coordinates (x, y) on the screen. These coordinates specify the request made to the visual-location buffer of ACT-R, which creates a chunk representing the location of the specified object. After that, the identification system identifies the name of the object and creates a chunk placed in the visual buffer. The location and identification phases last 185 ms (Bothell, 2004; Byrne, 2001). The objects are presented to the visual module of ACT-R by the mean of a list of all the objects (buttons of the interface) to be pushed on. Figure 4 shows some ACT-R productions responsible of the visual encoding phase.

```

(P start-application
=goal> ISA      begin ; Initializing the model
=>
-imaginal>
+visual-location>
; Making request of the visual-location buffer
ISA      visual-location
:attended nil
+goal> ISA      get-object
state      find-location )
(P attend-utensil
=goal> ISA      get-object
state      find-location
; Move attention to the location
; screen-x 122 and screen-y 250
=visual-location>
ISA      visual-location
screen-x 122
screen-y 250
?visual> state free
=>
+visual> ISA      move-attention
screen-pos =visual-location
=goal> state      attend )

```

Figure 4: Example of some ACT-R productions responsible for the visual encoding phase

The recognizing phase begins when the chunk of the object is placed in the visual module. This phase implies to recover that specific chunk from the declarative memory. The result of this phase is a chunk that represents the object with some characteristics as color, localization on the screen, name, and kind of object. The motor phase consists of activating the motor actions via a request to the motor buffer in order to click on the object. The three phases process is applied for each object displayed in the interface for the two steps of the recipe. The gathering utensils and ingredients model finishes when the last object of the ingredient list is reached.

The ACT-R model is developed using the ACT-R 6 environment. No noise is introduced in the perceptual motor modules. no retrieval error is modeled in the recognizing phase. These restrictions lead to a deterministic model. Figure 5 shows an example of execution traces of the ACT-R model for the visual encoding and the shift attention actions respectively. The visual-location request takes place at time 0.050 seconds and the request to move-attention is made at time 0.100 seconds. The encoding needs still 0.085 seconds to be completed and store the chunk into the visual buffer.

|   |            |   |
|---|------------|---|
| 0.000   | PROCEDURAL | PRODUCTION-SELECTED START-APPLICATION       |
| 0.000   | PROCEDURAL | BUFFER-READ-ACTION GOAL                     |
| 0.050   | PROCEDURAL | PRODUCTION-FIRED START-APPLICATION          |
| THE SUBJECT STARTS TO LOOK FOR NEW OBJECT             |            |   |
| 0.050   | PROCEDURAL | MODULE-REQUEST VISUAL-LOCATION              |
| 0.050   | PROCEDURAL | MODULE-REQUEST GOAL                         |
| 0.050   | PROCEDURAL | CLEAR-BUFFER IMAGINAL                       |
| 0.050   | PROCEDURAL | CLEAR-BUFFER VISUAL-LOCATION                |
| 0.050   | PROCEDURAL | CLEAR-BUFFER GOAL                           |
| 0.050   | VISION     | Find-location                               |
| 0.050   | VISION     | SET-BUFFER-CHUNK VISUAL-LOCATION LOC1       |
| 0.050   | GOAL       | CREATE-NEW-BUFFER-CHUNK GOAL ISA GET-OBJECT |
| 0.050   | GOAL       | SET-BUFFER-CHUNK GOAL GET-OBJECTO           |
| 0.050   | PROCEDURAL | CONFLICT-RESOLUTION                         |
| 0.050   | PROCEDURAL | PRODUCTION-SELECTED ATTEND-UTENSIL          |
| 0.050   | PROCEDURAL | BUFFER-READ-ACTION GOAL                     |
| 0.050   | PROCEDURAL | BUFFER-READ-ACTION VISUAL-LOCATION          |
| 0.050   | PROCEDURAL | QUERY-BUFFER-ACTION VISUAL                  |
| 0.100   | PROCEDURAL | PRODUCTION-FIRED ATTEND-UTENSIL             |
| SHIFT ATTENTION TO A SPECIFIED LOCATION ON THE SCREEN |            |   |
| 0.100   | PROCEDURAL | MOD-BUFFER-CHUNK GOAL                       |
| 0.100   | PROCEDURAL | MODULE-REQUEST VISUAL                       |
| 0.100   | PROCEDURAL | CLEAR-BUFFER VISUAL-LOCATION                |
| 0.100   | PROCEDURAL | CLEAR-BUFFER VISUAL                         |
| 0.100   | VISION     | Move-attention LOC1-0 NIL                   |
| 0.100   | PROCEDURAL | CONFLICT-RESOLUTION                         |
| 0.185   | VISION     | Encoding-complete LOC1-0 NIL                |
| 0.185   | VISION     | SET-BUFFER-CHUNK VISUAL TEXT1               |

Figure 5: Example of execution trace of the ACT-R model for the visual encoding action

## Results of the ACT-R model

Figure 6 shows the progress of time depending on progress in the task of get out utensils and get out ingredients respectively. The first task (get out utensils) lasted 6510 ms and the second task (get out ingredients) lasted 8101 ms, the overall time to complete the whole task equal to the sum of the two previous times:  $7107 + 8101 = 15208$  ms. The time taken to gather the utensils and ingredients is linear depending on the number of objects to search. No differences are observed between the object locations on the screen.

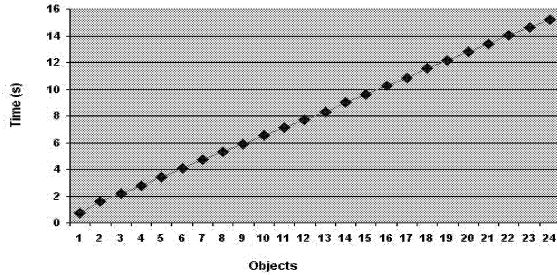


Figure 6: Progress of time depending on progress in the tasks of get out utensils and get out ingredients

## MODELING THE INTERACTION WITH THE CONTEXTUAL ASSISTANT USING FITTS LAW

In order to support and validate our results, we used the Fitts Law model, widely used in the evaluation of human machine interfaces. In the Fitts Law, the movement time is proportional to the target amplitude and

inversely proportional to the target width.

### Fitts Law model

In human machine interfaces, the formulation of Fitts Law (Fitts, 1954) states that the movement time (MT) is function of target amplitude (A) and target width (W). Our model is based on the Mackenzie's [1995] version of Fitts Law in which the movement time (MT) follows the equation:

$$MT = a + b * \log_2\left(\frac{A}{W} + 1\right) \quad (1)$$

The second term of the equation (1):  $\log_2\left(\frac{A}{W} + 1\right)$  is known as the index of difficulty ID, where a and b are constants derived empirically. They can be interpreted by the y-intercept and the slope of a predictive linear regression equation (MacKenzie, 1995) (MacKenzie et al., 1991). In our study, the user- interface interaction is based on the use of a touchscreen, assuming that users remain standing at a distance of 30 cm from the touchscreen, and point directly on the displayed objects by touching them using their index finger. The index finger is held down before starting the interaction, which constitutes the start position. After each pointing action, users returned their index finger to the start position, and the procedure continued like that.

### Results of the Fitts Law model

Table 1 shows the index of difficulty values obtained when applying the formulate  $\log_2\left(\frac{A}{W} + 1\right)$  on some objects displayed in the interface, and the corresponding predicted movement time (MT) obtained by applying the equation (1). The target amplitude (A) remains constant while the button width (W) varies as seen in figure 1.

| Object-Name     | A<br>(cm) | W<br>(cm) | ID<br>(bits) | MT<br>(ms) |
|-----------------|-----------|-----------|--------------|------------|
| BIG-SAUCEPAN    | 30        | 5.8       | 2.625        | 614.125    |
| NEXT-BUTTON     | 30        | 7.6       | 2.306        | 553.834    |
| LOOK-FOR-OBJECT | 30        | 3.8       | 3.152        | 713.728    |
| MUSHROOMS       | 30        | 5.8       | 2.625        | 614.125    |

Table 1: Index of Difficulty values for some Objects in the Interface and the corresponding movement time

The total time of the whole task applying the Fitts Law is estimated using the following equation:

$$MT_{Total} = \sum_{i=1}^n MT_i \quad (2)$$

Where n represents the number of objects used by the user in the interface, and  $MT_i$  the corresponding movement time of each object. The total movement time of

the whole task applying the equation (2) is: 14977 ms (14.977 s).

## COMPARISON OF RESULTS

The results of the predicted time of the task gathering utensils, gathering ingredients and the predicted time of the whole task in both models ACT-R model and Fitts Law model are shown in Table 2.

| Tasks   | ACT-R | Fitts Law |
|---|-------|-----------|
| Predicted time of getting out Utensils Task (ms)    | 7107  | 6954      |
| Predicted time of getting out Ingredients Task (ms) | 8101  | 8023      |
| Predicted time of the whole Task (ms)               | 15208 | 14977     |

Table 2: Time estimation of gathering utensils task, gathering ingredients task and the whole task in both models ACT-R and Fitts Law

The ACT-R results are consistent with the Fitts Law model as shown in Figure 7. The predicted time to point each object is very close in both models ACT-R and Fitts Law.

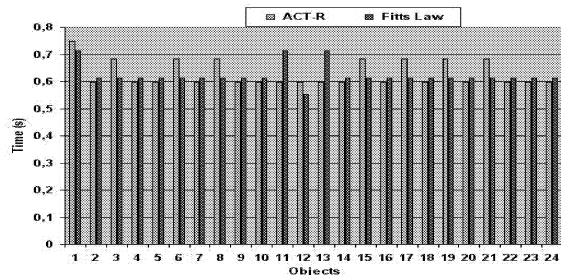


Figure 7: Comparison between the predicted time of each object using ACT-R model and Fitts Law model

## GENERAL DISCUSSION

The ACT-R model we developed is proved robust and efficient in our analysis. In fact, the results obtained by the ACT-R model were consistent with those obtained by the Fitts Law model in terms of the predicted time execution of tasks as mentioned previously; this demonstrates that cognitive models and particularly ACT-R can give good predictions in the evaluation of HMI. The results of the ACT-R model show that, the size of objects in the interface is not taken into consideration, and our model does not make difference in the predicted time of the pushing "HELP-ME-TO-DO-THE-TASK" button for example, and the pushing "WOODEN-SPOON" object; these two actions have the same predicted time which equals to 597 ms. Unlike ACT-R model, the Fitts

Law model takes in account the object's size in the interface. The predicted time for the pushing "HELP-ME-TO-DO-THE-TASK" button using the Fitts Law is 713 ms and the predicted time for the pushing "WOODEN-SPOON" object is 614 ms. However, some differences are noted as presented in Figure 7. The simulation of the HMI with the object number 11 and 13 takes more time with the Fitts Law. It corresponds to buttons representing the action "HELP-ME-TO-DO-THE-TASK" and "LOOK-FOR-OBJECT" respectively. This is due to the smaller size of these buttons (width = 3.8 cm), compared with the size of the other objects. On the other side, the object number 12 necessitates less time to be pushed. It corresponds to the button representing the action "NEXT", which has the largest size (width = 5.8 cm) in the interface. The simulation of the HMI with the object number 1 as shown in Figure 7, takes more times with the ACT-R model, it corresponds to the button representing the action "LOOK-FOR-OBJECT". This is due to the initialization of the model such as the goal buffer, the retrieval buffer and the visual buffers. In the ACT-R model, the focus is essentially on the visual encoding and recognizing of objects and how to interact with the interface using motor actions. This is supported by some scientific literature such as the use of cognitive models in the evaluation of expert cell phone menu interaction (Amant et al., 2007). The results of the ACT-R model are considered suitable and correct comparing them to those obtained by the Fitts Law model. In fact, as shown in Table 2 the estimated time of the whole task in the ACT-R model (15208 ms) is very close to the Fitts Law model time estimated to 14977 ms. We believe nevertheless, that our study lays out new perspectives of research in this domain particularly how to use perceptual motor modules of the ACT-R architecture to simulate the HMI.

## CONCLUSION

The main goal of our study is to evaluate the HMI of a contextual assistant by simulating the HMI, focusing on the time execution of tasks. We used the cognitive architecture ACT-R as a powerful tool to develop our model. Our ACT-R model consists of two parts, the model of the interface of the contextual assistant which represents the environment to interact with, and the model of the cognitive processes required to interact with the interface. The perceptual part of the cognitive processes constitutes the difficult part in our ACT-R model, due to the scarcity in the documentation about the perceptual module in the literature. The results of the ACT-R model were compared with those obtained by the Fitts Law model, developed in this study in order to argue and support our results. The results of our model were consistent with the results of the Fitts Law model. Our model gives a good prediction of user performance, which makes it powerful and realistic.

## FUTURE IMPROVEMENTS

The model we developed constitutes the first step of the evaluation of HMI using a contextual assistant. Three future improvements will add scientific validity to our model. First, the results of our model were compared with those obtained by the Fitts Law model. The results of the Fitts Law model are not always good and exact, but have a certain percentage of errors. It would be interesting to do some experiments with real persons to collect real data and compare them with our results. Second, our model is deterministic and does not make errors. It should be extended to allow errors in the pointing actions. These errors are essentially related to memory problems that may occur in the task modeling (Serna et al., 2005; Dion and Pigot, 2007) and during the interaction with the interface of the contextual assistant. Finally, the action of searching an object is resumed to the HMI with the touch screen. The contextual assistant offers an interaction with the environment to help people recovering utensils and ingredients dispatched in the kitchen. It would be interesting in the future to model this part and simulate the movement of users picking up the objects in the kitchen. Therefore, the extended model should simulate people making a task with contextual assistant and the errors committed by people with cognitive impairments.

## REFERENCES

- Amant R.S.; Horton T.E.; and Ritter F.E., 2007. *Model-based Evaluation of Expert Cell Phone Menu Interaction*. *ACM Transactions on Computer-Human Interaction*, 14(1), 1–24.
- Anderson J.R.; Bothell D.; Byrne M.D.; Douglass S.; Lebiere C.; and Qin Y., 2004. *An Integrated Theory of the Mind*. *Psychological Review*, 111, 136–1060.
- Anderson J.R.; Taatgen N.A.; and Byrne M.D., 2005. *Learning to Achieve Perfect Time Sharing: Architectural Implications of Hazeltine, Teague, Ivory (2002)*. *Journal of Experimental Psychology: Human Perception and Performance*, 31, No. 4, 749–761.
- Bothell D., 2004. *ACT-R 6.0 Reference Manual*. Working draft.
- Byrne M.D., 2001. *ACT-R/PM and menu selection: Applying a cognitive architecture to HCI*. *International Journal of Human-Computer Studies*, 55, 41–84.
- Dion A. and Pigot H., 2007. *Modeling cognitive errors in the realization of an activity of the everyday life*. In *Cognitio 2007*.
- Eugenio B.D.; Haller S.; and Glass M., 2003. *Development and Evaluation of NL interfaces in a Small Shop*. *AAAI Spring Symposium on Natural Language Generation in Spoken and Written Dialogue*, 1–8.
- Fitts P.M., 1954. *The information capacity of the human motor system in controlling the amplitude of movement*. *Journal of Experimental Psychology*, 47, NO. 6, 381–391.
- Lussier-Desrochers D.; Lachapelle Y.; Pigot H.; and Bauchet J., 2007. *Apartments for People with Intellectual Disability: Promoting Innovative Community Living Services*. In *2nd International Conference on Intellectual Disabilities/Mental Retardation*.
- MacKenzie I.S., 1995. *Movement Time Prediction in Human Computer Interfaces*. In *Readings in human-computer interaction*. 2nd, 483–493.
- MacKenzie I.S.; Sellen A.; and Buxton W., 1991. *A comparison of input devices in elemental pointing and dragging tasks*. In *Proceedings of the CHI '91 Conference on Human Factors in Computing Systems*. New York: ACM, 161–166.
- Nielsen J. and Phillips V.L., 1993. *Estimating the relative usability of two interfaces: heuristic, formal, and empirical methods compared*. *Proceedings of the INTERACT '93 and CHI '93 conference on Human factors in computing systems*, 214–221.
- Pigot H.; Bauchet J.; and Giroux S., 2007a. *Assistive devices for people with cognitive impairments*, John Wiley and Sons, chap. 12. *The Engineering Handbook on Smart Technology for Aging, Disability and Independence*.
- Pigot H.; Lussier-Desrochers D.; Bauchet J.; Lachapelle Y.; and Giroux S., 2007b. *A smart home to assist recipes completion*. In *Festival of International Conferences on Caregiving, Disability, Aging and Technology (FICCDAT), 2nd International Conference on Technology and Aging (ICTA)*.
- Pigot H.; Savary J.P.; Metzger J.L.; Rochon A.; and Beaulieu M., 2005. *Advanced technology guidelines to fulfill the needs of the cognitively impaired population*. In *3rd International Conference On Smart Homes and health Telematic (ICOST)*. IOS Press, Assistive Technology Research Series, 25–32.
- Serna A.; Pigot H.; and Rialle V., 2005. *Modeling the performances of persons suffering Alzheimer's disease on an activity of the daily living*. In *18th Congress of the International Association of Gerontology*.

# A GRANULAR FRAMEWORK FOR RECOGNITION OF ARABIC HANDWRITING: THE GOAVMREC SYSTEM

Mokhtar Beldjehem  
École Polytechnique de Montréal  
C.P. 6079, succ. Centre-Ville  
Montréal QC H3C 3A7, Canada  
E-mail: mokhtar.beldjehem@polymtl.ca

## KEYWORDS

Modular Granular Recognition Architecture, Arabic Handwriting Segmentation, Perceptual Features, Global Visual Indices, Morphological Analysis, Syntax Analysis (Parsing), Semantic Analysis, Pragmatics, Fuzzy Logic, Granular Soft Computing (GrSC), Prolog, Cooperative Morphological-Guided Recognition, Off-line Arabic Handwriting Recognition.

## ABSTRACT

We propose a novel cognitively motivated unifying framework for Arabic handwriting recognition that takes into account the nature of the human reading process of Arabic handwriting. This Modular Granular Architecture tackles the problem by observing Arabic handwriting from both perceptual and linguistic points of view and hence analyzes the underlying input signal from different granularity levels. It is based on three levels of abstraction: a low granularity level that uses perceptual features called global visual indices, a medium granularity level that is the conventional recognition stage and a high granularity level (linguistic level) that consists on morphological analysis dedicated to segmentation/ recognition.

The original idea is the effective use of Arabic word's morphology in the recognition not only in post-processing. This architecture carries well around the Arabic word's morphology, as typically in Arabic, the Arabic word's morphology is by excellence the logical structure (even semantic) of a given Arabic word, whereas the visual data constitute the physical geometric (topological) structure of a given word. We need to integrate both of them for an effective cooperative recognition of Arabic Handwriting. This framework subsumes the lexicon-driven approaches; in that it can recognize a word that does not exist within the lexicon. Implementation of a system called the GOAVMRec working under our framework is underway.

## INTRODUCTION AND MOTIVATIONS

If modern technology has caused the "paper explosion", then modern technology by the means of automatic recognition has to resolve the resulting problems. The automatic recognition of text either printed or handwritten on scanned images has enabled various applications in many ICTs fields such as data entry, retrieval of words in

large volume of documents, automatic bank cheques processing, automatic administrative forms analysis, automatic sorting of postal mail, recognition of text from a fax cover-sheet, convenient editing of hardcopy paper contents, automatic building of text databases, recovering and restoring contents of old degraded ancient manuscripts.

The automatic recognition of Arabic handwritten text is a problem worth solving because a solution would enable us to design and use systems with a more ergonomic convenient and flexible human-machine interaction in various ICTs fields. The research study in (Hyder and Khoujah 1988) points out that the Arabic script based languages, over fifty in number, are used by nearly a fifth of the population of the globe. The complex nature of the Arabic language is evident in the cursiveness of the text, character overlapping, various character shapes, "Tashkeel", diacritics and/or dots and the variety of available Calligraphic Arabic styles (more than 12) that exist, not to mention variability induced by different writing styles for multiscripter or omniscripter handwriting recognition. As a result, these specific Arabic language complexities present major technical challenges in the Recognition Technology.

In general the problem of handwriting recognition may be approached in one of the two modes: online or offline. However, offline recognition is more popular because of its pragmatic viability. In this article we focus on unconstrained off-line Arabic handwriting recognition. An intensive research has been devoted to printed rather than handwritten Arabic recognition. A rich research literature exists for Latin, Chinese and Hindi handwriting (Simon 1985; Bozinovic and al. 1989; Tappert et al. 1990; Simon 1992; Suen et al. 1993; Plamondon et al. 2000; Liu et al. 2004) only a few articles deal to Arabic handwriting. That is because Arabic handwritten text yields very complicated shapes and patterns that are among the most difficult to segment and classify accurately. It is worth mentioning that it is not possible to apply directly many of the available recognition algorithms proposed for other languages to Arabic Recognition.

Research in Arabic handwriting recognition can be traced early to works by Amin (Amin et al. 1980; Amin 1984). Recently interesting surveys about Arabic Handwriting

Conventionally in OCR systems, once the recognition is completed, either spell-checking tool-based and/or a manual or semi-manual proofreading process is performed for the verification/correction purposes. This task is labour-intensive, tedious, costly, time-consuming, and prone to errors. Therefore this task needs to be automated either totally or partly and integrated early in the recognition chain by incorporating linguistic information, i. e. high level knowledge. We point out herein the advantages of using the perceptual features and segmentation methods as a mandatory task before recognition, and the importance of considering context and more specifically the integration and contribution of a morpho-lexical analyzer to segmentation/ recognition; this is because the knowledge about the Arabic language is important for improving the Arabic handwriting recognition throughput and accuracy.

Referring to Table 1, the Arabic alphabet comprises 29 basic letters (28 if we exclude the *hamza* "ء") that behaves either like a special letter or like a diacritic. Arabic letters are contextual letters and so their associated multiple shapes are of major importance.

Dots are located either above, below, or inside the letter's body. There are six letters {ا, ذ, ر, ز, و, ي} that can not be connected to their successors within a word or a pseudo-word; they are called "stubborn letters". Several Arabic letters share the same main body (but differing in the number and position of dots). Remark the connectedness, and the greater similarity, and hence the resulting confusability in machine recognition of Arabic letters. In Arabic there is no upper or lower cases problem; this is may be the only one advantage of Arabic over Latin.

adjacent letters can be stacked on top of each other; some letters can be fused to form new shapes. In fact an Arabic word or a pseudo-word either printed or handwritten is a continuous smooth flow of letters. Not to mention different available Arabic Calligraphic styles (more than dozen). Arabic handwritten text may be voweled or unvoweled, this is called “*Tashkeel*” (see Table 1 and Figure 1), the placement of it, is determined by the *rules of grammar*. This adds more complexity to Arabic recognition as any serious OCR system has to cope with “Tashkeel” to fulfill design goals that can lead to a coherent model for Arabic handwriting recognition. This is in fact a serious problem that has been unfortunately overlooked in the past.

- a) EF: End of word shape
- b) MF: Middle of word shape
- c) BF: Beginning of word shape
- d) IF: Isolated shape

| Name  | EF | MF | BF | IF |
|-------|----|----|----|----|
| DĀD   | ض  | ض  | ض  | ض  |
| TĀ    | ط  | ط  | ط  | ط  |
| ZĀ    | ظ  | ظ  | ظ  | ظ  |
| AYN   | ع  | ع  | ع  | ع  |
| GHAYN | غ  | غ  | غ  | غ  |
| FĀ    | ف  | ف  | ف  | ف  |
| QĀF   | ق  | ق  | ق  | ق  |
| KĀF   | ك  | ك  | ك  | ك  |
| LĀM   | ل  | ل  | ل  | ل  |
| MĪM   | م  | م  | م  | م  |
| NŪN   | ن  | ن  | ن  | ن  |
| HĀ    | ه  | ه  | ه  | ه  |
| WĀW   | و  |    |    | و  |
| YĀ    | ي  | ي  | ي  | ي  |

| Name | EF | MF | BF | IF |
|------|----|----|----|----|
| ALIF | ا  |    |    | أ  |
| BĀ   | ب  | ب  | ب  | ب  |
| TĀ   | ت  | ت  | ت  | ت  |
| THĀ  | ث  | ث  | ث  | ث  |
| JĪM  | ج  | ج  | ج  | ج  |
| ḤĀ   | ح  | ح  | ح  | ح  |
| KHĀ  | خ  | خ  | خ  | خ  |
| DĀL  | د  |    |    | د  |
| DHĀL | ذ  |    |    | ذ  |
| RĀ   | ر  |    |    | ر  |
| ZĀY  | ز  |    |    | ز  |
| SĪN  | س  | س  | س  | س  |
| SHĪN | ش  | ش  | ش  | ش  |
| SĀD  | ص  | ص  | ص  | ص  |

|               |              |              |             |             |             |
|---------------|--------------|--------------|-------------|-------------|-------------|
| FAT-HAH<br>اَ | DAMMAH<br>اِ | KASRAH<br>اِ | SUKOON<br>ْ | SHADDA<br>َ | MADDAH<br>َ |
|---------------|--------------|--------------|-------------|-------------|-------------|

baseline لا إله إلا الله محمد رسول الله  
overlap ligature ligature

82

Another class of problems is basically caused by the handwriting styles. Since Arabic writing differs radically from person to person. This is due to the fact that Arabic writing is typically more an art than a disciplined technique. Not to mention in some situations physiological and psychological state of the scripter, conditions and circumstances of the writing process itself. Therefore, this causes the introduction of various kinds of anomalies such as: broken letters within a word or a pseudo-word, touching letters between two adjacent letters, touching letters belonging to two adjacent lines (in particular ascenders and descenders of consecutive lines are frequently connected), words at different scales, words are not uniformly spaced, lines are not straight, etc. Such anomalies add more complexity to Arabic handwriting segmentation and recognition. Not to mention the quality of the image itself: noise and degradation, faded ink, page border marks, etc. These variability, anomalies and complexity make Arabic word segmentation (decomposition) in letters very delicate and not always ensured; this is what explains the unavailability of Arabic handwritten OCRs in the market.

## A NOVEL RECOGNITION METHODOLOGY

We are motivated by the following remark: “a human being can read the Arabic handwriting just fine. Why can’t the OCR system?” We stipulate a novel general unifying framework to tackle the problem by observing Arabic handwriting conjointly from both perceptual and linguistic points of view and hence analyze the underlying input signal from different granularity levels. Considering both human perception and cognition, such a unifying framework is cognitively more plausible to reflect and mimic the human reading process. This motivates our current granular approach having a human-like decision making capabilities for the analysis and the recognition of visual data. A recognition system should be constructed in a modular way with different levels of processing. Following this line, we emphasize developing a modular granular architecture for our general framework based on the following three abstraction layers corresponding to the three following granularity levels **(i)**, **(ii)** and **(iii)** :

### **(i) Low granularity level (or perceptual level) :**

Regardless of the classification technique that will be used in the recognition stage, we focus on developing a segmentation stage which is mandatory before and independent from the classification technique that will be used during the recognition stage. We adopted and developed a strategy similar to those advocated in (Marr 1982; Simon 1989; Amin 2000) that use perceptual features and segmentation methods and more specifically the notion of “*global visual indices*”, this constitutes a departure from the conventional ones, in that it consists to smoothly tackle the word as a whole and by consequent let the segmentation focus on only relevant indices at a higher degree of granulation rather than trying segmentation at the letter level. This stage is herein handled adequately by soft computing (SC) segmentation methods similar to those proposed in (Pal 2000).

### **(ii) Medium granularity level (associative or logical level):**

Which corresponds to the conventional recognition stage (classification technique), several methods have been used for resolving recognition from algorithmic, to various neural network models, HMMs, statistical, expert systems, SVMs, fuzzy logic, to name only a few. Usually this stage comprises two phases; a learning phase and a decision phase. Fuzzy and soft computing (Zadeh 1965, 1971; Siy and Chen 1974; Kickert and Koppelaar 1976; Pal et al. 2000; Pal and Ghosh 2000; Zadeh 1998) and especially hybrid fuzzy-neuro classification methods (Beldjehem 1993, 1994, 2002, 2004, 2006, 2008) that allows to handle adequately this stage by allowing premature decisions to be deferred and enabling to develop tolerant recognizers ensuring close resemblance with human like decision making. Furthermore hybridization should ensure that it provides application specific merits, besides the generic advantages. The Recognizer constitutes the core of any OCR system.

### **(iii) High granularity level (linguistic or cognitive level):**

Exploiting techniques from Arabic Natural Language Processing (NLP) technologies, and more specifically Arabic Computational Linguistics (CLs) will address these challenges effectively. In particular our focus is on the integration of the conventional post-processing stage very early within the recognition chain rather than to be handled manually as in the traditional OCR systems via human proofreading or automatic tool-based spell-checking. We stipulate to integrate it and perform it automatically within the recognition chain by the means of a fully integrated morpho-lexical analysis dedicated for generation/ segmentation/ recognition/ verification/ correction purposes. This will necessitate the construction of lexicon. This morpho-lexical analyzer constitutes a component within the OCR recognition engine itself rather than an independent component used in post-processing as in conventional OCR systems. This level carries well around the Arabic word’s morphology. The rational behind using morphology is obvious for the reader.

The original idea revolves around using Arabic word’s morphology in the segmentation/recognition of an Arabic word. We will investigate in the sequel in details this framework and in particular the two-folds **(i)**, **(iii)**. It is worth mentioning that the three levels may be either intermingled or executed sequentially. We use this framework to implement the GOAVMRec system that is an Acronym for **Granular Off-line Arabic Visual Morphological-based Recognizer**.

## ARABIC-WORD SEGMENTATION

This section deals with off-line Arabic segmentation, bearing in mind that Arabic handwriting is very hard to segment, because of the semi-cursive nature of the Arabic word (or pseudo-word), typical individuality of Arabic handwriting, anomalies we have already mentioned, not to

mention the writing conditions and the quality of acquisition. We distinguish between two different categories; the first category considers segmentation as mandatory, whereas the second category bypasses segmentation. It is worth mentioning that our framework can accommodate both categories of segmentation approaches as the three levels may be either intermingled or executed sequentially.

Most devised segmentation algorithms of the first category rather than trying to segment an Arabic word to letters, they try to segment it to graphemes. A grapheme may be a part of a letter (or a fragment), one or more letters. Basically, such algorithms differ only in the method used for detecting the segmentation points (SP) and the choice of the segmentation unit serving as input to the Recognizer. Such a unit influences significantly the features selection and primitives' extraction. Among them we developed and adapted a method that deals with *singularities and regularities* similar to those approaches advocated in (Marr 1982; Simon 1989; Miled and al. 1998; Amin 2000) among others.

The second category adopts segmentation-free approaches that bypasses the segmentation problem or in fact fuse segmentation with recognition in one stage (Allam 1995; Cheung et al. 2001). Bearing in mind that the shape of an Arabic letter is of major importance, it is worth stressing the importance of the visual aspect of cursive Arabic handwriting.

The segmentation stage is composed of the three conventional or standard phases as follows :

#### A. Low-level processing and low-level segmentation

#### B. Segmentation of a word to graphemes

#### C. Features selection and primitives extraction

Let us describe succinctly the involved tasks related to each phase:

#### A. Low-level processing and low-level segmentation

This Segmentation stage starts by performing various low-level processing steps followed by a low-level segmentation. This low-level processing comprises the following operations that enhance the quality of the image by reducing the noise, eliminating fluctuations and prepare it for the next phases of segmentation and for the recognition stage:

##### Pre-processing that allows

1. acquisition's noise filtering and removal
2. smoothing of the connected entities within the image
3. holes filling
4. global redressing of the image

This important step is necessary and useful in particular in handling degraded documents.

#### Low-level segmentation

1. segmentation of the image to lines
2. writing line detection
3. medium area detection
4. lines segmentation to connected entities
5. connected entities pre-classification

#### B. Segmentation of a word to graphemes

The second phase consists on the segmentation of the word into graphemes. Such a phase is ensured by two modules: a *kernel module* and an *analyzer module*. The kernel is script-independent and can be used either to Arabic or another language such as Latin, whereas the analyzer uses contextual information about the used script and more specifically knowledge about the Arabic script nature. The analyzer is complementary to the kernel and works at the grapheme level to compensate the deficiencies of the kernel. The following is a brief description of the tasks involved within the two modules:

##### The kernel ensures the following tasks

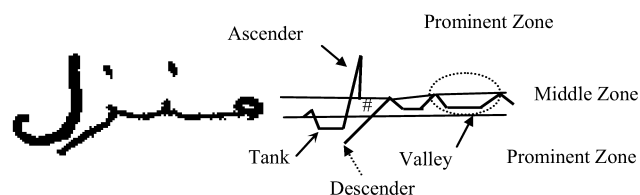
1. upper contour extraction using the *Freeman codes*
2. modeling and coding
3. primary segmentation points (PSPs) detection
4. writing line redressing and detection
5. decisive segmentation points (DSPs) selection
6. elimination of overlapping connections

##### The analyzer ensures the following tasks

1. graphemes detection, extraction and quantization
2. ligatures detection
3. connection lines and "kashida" detection
4. last character extraction

#### C. Features selection and primitives extraction

Taking into account the importance of the visual aspects of the Arabic handwriting, it is appealing to try to tackle the feature extraction phase by using a global view approach that works at the word level. In order to detect the information zones in the word, we try to describe the word (as a global entity) by a sequence of visual indices that capture perceptual features (As shown in Figure 3) and more specifically using the notion of "global visual indices" (Amin 2002) such an approach is well motivated and based on psycho-cognitive considerations of the *theory of first global vision* (Marr 1982).



**Figure 3 Different visual indices extracted from the tracing zone.**

A tracing in our framework is a conceptual concept of a connected graph (that is neither a diacritic nor an aleph) that is not necessarily a pseudo-word.



This phase starts by extracting the image external contours. The connected components of the word are first extracted and then pre-classified in three categories: tracings, diacritics, and “alifs”. Only the tracings are segmented by a method based on the extraction and analysis of the signal related to the general word shape using a perception based approach. A grapheme is represented by a characteristic vector of features, each feature belong to one of two different types of features, the first type includes structural (or topological) features that reflects the visual features related to human perceptions and observation such as: loops (either closed or open), ascenders, descenders, normalized size, and s.o, the second include the first 10 Fourier Descriptors.

As mentioned above diacritics are frequently used in Arabic script. The most pertinent ones are dots (single or multiple) that discriminate between letters having the same main body. Multiple dots come in two types: double and triple. In Arabic handwriting, the diacritics are so complex that we decompose each of these two types into their number of single dots. To refine the word description and to increase our information about the word, we separate detected diacritics (dots) into two distinct visual indices according to their relative position to the baseline.

Typically in Arabic, the diacritics (including dots) control the pronunciation of words and possibly their meanings, and even though due to their variable position, numbers, size and writing styles they are difficult to segment accurately and to relate to the appropriate corresponding character within a word (or a subword), they have a significant discriminating power and provide valuable information in recognizing Arabic letters when they are used appropriately.

## COOPERATIVE MORPHOLOGICAL-GUIDED RECOGNITION

Morphology analysis is the area of Computational Linguistics (CLs) that investigates word formation, including affixes, roots and patterns (or templates). Typically, Arabic words are divided into 3 groups: nouns, verbs, and particles. Particles are connection words such prepositions and pronouns. Nouns and verbs are derived from roots, which are linguistic units of meaning composed of 3, 4, and 5 letters. Arabic nouns and verbs are derived from roots by applying templates to the roots to generate stems. Addition of affixes to stems yields words. Applying templates often involves introducing infixes or deleting or replacing letters from the root (see Figure 4). Fortunately most Arabic words (except some proper nouns and newly included foreign words) are derived from a root. An interesting recent survey about Arabic morphology is provided in (Al-Sughayer and Al-Kharash 2004) and various solutions that attempt to implement Morphology analysis of Arabic in connections with information retrieval (IR), machine translation (MT) and natural language processing (NLP) have been proposed in the literature.

However, to the best of our knowledge no one exists in the context of Arabic handwriting Recognition, except work in (Amin 1984) but used only for post-processing purposes not for recognition. Within our framework the morpho-lexical analyzer is a basic building block that guides the recognizer in the course of the recognition and furthermore can enable a system to recognize a misspelled word, a misrecognized word or even a word that does not exist within the lexicon thanks to the word generation or the stemming capability. Such a framework subsumes lexicon-driven handwritten character string recognition approaches.

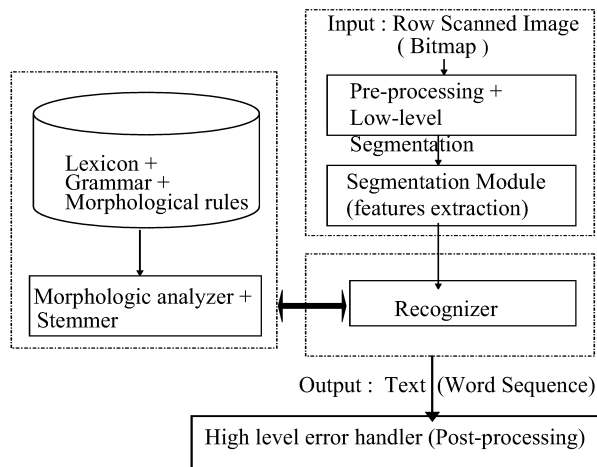
| Root<br>(جذر) | Template<br>(وزن) | Stem  | Word      |
|---------------|-------------------|-------|-----------|
|               |                   |       |           |
|               | فعليل             | عليل  | العليل    |
|               | فعال              | علام  | علامة     |
|               | مفاعل             | معالم | المعالم   |
|               | فاعل              | عالم  | عالمية    |
| علم           | فعل               | علم   | تعلمت     |
|               | مفعول             | معلوم | المعلومات |
|               | فعلول             | علوم  | العلوم    |
|               | مفعل              | معلم  | المعلمات  |
|               | تفعيل             | تعليم | التعليمات |
|               | إفعال             | إعلام | إستعلامات |

**Figure 4 Construction of Arabic words related to the root “علم” i.e., to know, either by derivation or by inflection or by both.**

The generation procedure of the morpho-lexical analyzer may be implemented efficiently using Arabic morphological rules; the automatic verification/correction of stems is adequately implemented using regular expressions and approximate string matching algorithms (that can be used in root detection and in verification for spell checking). The system starts by grouping graphemes (provided by the segmentation module) to attempt to form a candidate word (that may or may not be an actual word) that will be passed as input to both the morpho-lexical analyzer and the Recognizer for the segmentation/recognition purposes. This ensures that only valid meaningful words are recognized, the user interactively is demanded intervention for advice when needed; in particular when there are many possibilities and the system give several suggestions. The Arabic word’s morphology is by excellence the logical structure (even semantic) of a given Arabic word, whereas the visual data constitute the geometric (topological) structure of a given word. We need both of them for an effective recognition. This is the rational behind our approach and what motivates us to design a sort of guided morphological-based recognition system. The system granular architecture reflecting the three granularity levels of the framework and the independent high level error handler of the post-processing is illustrated in Figure 5, which is working outoff the recognition-time.

In the sequel we will focus our attention only on each component or subsystem that has not been already described :

Let us now describe the functioning of the recognition chain of our framework that is geared around the cooperative relationship between the segmentation module, the Recognizer and the morpho-lexical analyzer: Basically, the morpho-lexical analyser guides and gives clues to the Recognizer and the Recognizer in turn guides indirectly the segmentation module and vice-versa. From the computational point of view, they work concurrently and synergistically according to the *producer-consumer* model, in that the segmentation module attempt to produce gradually on a letter-by-letter basis (from graphemes) a candidate word and feeds the couple Recognizer/Morpho-lexical analyser which conjointly *cooperate* to consume (recognize/correct) the proposed word to obtain a meaningful (or valid) word and to inform the segmentation module. This cycle continues until the processing of the whole input data. A *morpho-lexical analyser* is a mechanism that attempts to identify the word basic morphemes: the prefix, the infix, the suffix and the root. This step is fully integrated within the classification stage. This ensures the effective recognition of the actual word. The lexicon may be easily implemented by a relational DB or taking into account that the Arabic has a rich lexicon, we may implement stem formation efficiently by keeping only the tri-consonant permutations (called radical) and using some morphological rules to generate all the other possible words when needed during the run-time. See Figure 4 for more details.



**Figure 5 The GOAVMRec three-levels granular cooperative recognition architecture**

For the convenience of approximate matching algorithms, the root will be represented by a pattern using the following regular expression: “\*ع\*ل\*م\*”, \* means zero or more letters. We are using fuzzy sets-based algorithms similar to those used for the recognition of imperfect

strings by Modechay (Mordechay et al. 1992) conjointly with algorithms for approximate string matching similar to those proposed by Ukkonen (Ukkonen 1985). The approximate string matching algorithm that we used starts progressively by attempting to detect the appropriate root, once the root is detected, using an edit distance a pairwise similarity index is computed for each possible template and the template with the maximum similarity is then determined; the corresponding stem is then selected, finally by applying morphological rules the word is then constructed (found) to guide cooperatively the Recognizer. This will necessitate building a component called a stemmer that strips common prefixes and suffixes from words. The stemmer is useful in case of failure of the morphological analyser, the stemmer takes control to complement and compensate such a failure and other deficiencies of the morphological analyser. For illustration purposes, assume that the string to be recognized is “المعلومات”. Assume that the Recognizer starts progressively by recognizing letter-by-letter until getting correctly the sequence “المعلوم” that enables the matching algorithm to detect (identify) the root “علم”, then maximum similarity gives the appropriate template “مفعول” that corresponds to the stem “معلوم” and hence a set of expecting words is generated (including the word “المعلومات”, etc. ) That guides the recognition of the rest of the string letter-by-letter. If for every remaining letter there is match between what the Recognizer recognizes and the morpho-lexical analyser suggests then it is a valid word. Otherwise there is an error and the algorithm fix it by replacing the erroneous letter (proposed by the Recognizer) by the correct one suggested by morpho-lexical analyser. In contrast if the string to be recognized is “المعلوماس”, when reaching the last letter, the algorithm detects that there is mismatch between what the recognizer proposes and what the morpho-lexical analyser suggests, and hence the letter “س” is simply replaced by the letter “ت”. Clearly it is a case of a misspelled word. This is because the lexicon is assumed to contain only correct words and the morpho-lexical analyzer too is assumed to generate only correct words (well-formed words). Thus allowing the premature detection or the prediction of the last character of an Arabic word. Whereas the Recognizer may either correctly recognizes a misspelled word that has been introduced erroneously initially by the writer himself or it is unable to recognize correctly a letter because of its inaccuracy. For such reasons we decided to give priority to the morpho-lexical analyzer.

## ERROR HANDLING, CONTEXTUAL INFORMATION AND NATURAL LANGUAGE PROCESSING (NLP)

Handling errors depends on the nature of errors. On one hand even though a kind of built-in error handling word level has been incorporated early in the recognition chain by construction. Some residual errors pass the chain and need to be adequately handled at a higher level explicitly at the phrase level on the ASCII file within an editor once the recognition is completed. The *high level error handler* component for Arabic may be build to verify/correct

syntactic and semantic errors but by adding both syntax analysis and semantic analysis (or parsing) and even pragmatics. It works on the phrase level and is completely independent from the OCR engine as it might be thought of as a kind of a post-processing layer. Expressing Arabic grammar in terms of *Definite Clauses Grammars* (DCG) that extend the *context-free grammar* rules and exploiting the valuable information that “Tashkeel” brings, could implement such a layer effectively by writing them in Horn clauses using *Prolog* language. An exotic characteristic typical to Arabic is that it has three different forms: singular, dual, and plural used for number agreement handling and that can be used for correction. On the other hand to account for contextual information, some errors could be appropriately caught and handled by the *Pragmatic Analyser*, which has learning capabilities; the rational is to account for the vocabulary of the application domain where the OCR system will be deployed. This valuable knowledge could be exploited to fix some errors and hence to enhance the recognition overall performance. This knowledge is known as “contextual”. In-depth study will enable to identify most frequently encountered errors in connection with the Arabic script (letters sharing the same body, and so on.) and to use appropriate formalism to represent and learn error correction rules in connection with the domain application.

## CONCLUDING REMARKS

We have proposed a novel granular unifying framework that is cognitively plausible under which the GOAVMRec system is being implemented. A prototype for the proof of concept is available and is working for simulation purposes with a limited size lexicon. In particular, the morphological analyzer and the stemmer are being implemented using the *Prolog* language. Such a framework integrates conjointly both the perceptual and the cognitive aspects of the human reading process and ensures a cooperative granular processing of the underlying input signal from different granularity levels. The flexibility and expressivity of the framework allows coping with the complexity of the Arabic handwriting recognition problem, the associated granular architectures deals with different granules of information at different levels of abstraction, they are a spectrum of visual data and text constructs. This framework is language-independent. Intrinsic typical characteristics of an Arabic word make the morphological analysis sufficient to handle the Arabic handwritten text recognition at the word level. This framework subsumes the lexicon-driven approaches used in conventional OCR, in that it can recognize a word that does not exist within the lexicon. Indeed the lexicon itself is potentially or virtually of unlimited size. Automatic (or manual) building of public Arabic lexicon (from a very large corpus such as the holy Koran and other sources) to tackle real-world recognition problems in the field is of paramount importance for pursuing serious research in the fields of Arabic handwriting recognition, information retrieval (IR), machine translation (MT) and natural language processing (NLP). This is for training,

benchmarking purposes and for evaluating algorithms, systems and prototypes. It will ultimately constitute a brick toward building a *test bed* for Arabic language for various fields of research as it became clear that studies and applications for Arabic language can benefit both commercial applications and theoretical academic fields.

According to Zadeh (Zadeh 1994) “The exploitation of tolerance for imprecision and uncertainty underlies the remarkable human ability to understand distorted speech, decipher sloppy handwriting, comprehend nuances of natural language, summarize text, recognize and classify images and more generally, make rational decision in an environment of uncertainty and imprecision.” This ability is what granular soft computing (GrSC) systems in general and granular soft recognition architectures in particular try to capture by learning and emulate computationally.

Indeed Arabic handwritten recognition is a very complex and challenging problem ! It is still better known for its failures than for its successes. It is appealing to try to apply SC everywhere to cope with such a complexity; the later could be introduced and effectively contributes in any level of the three levels of granulation of our framework. Systems that combine CLs and granular GrSC techniques are approximate recognition models that have human-like behaviour and exhibits intelligence, learning, adaptability, evolvability, fault tolerance, flexibility, transparency, value approximation (Beldjehem 2006, 2008), they are tolerant to variation in writing styles so as to improve significantly accuracy and throughput. Hybridization can generalize over the large degree of variation between writing styles and ultimately recognition rules as well as error handling rules can be constructed automatically by learning using a data-driven approach. We hope that this paper will be a starting point for further incorporating and hybridization of CLs, NLP and GrSC in Recognition Technology.

## REFERENCES

- Abuhaiba I. S. I. and Ahmed P. 1993. Restoration of temporal information in off-line Arabic handwriting. *Pattern Recognition* 26, No. 7, 1009-1017.
- Abuhaiba I. S. I.; Sabri A. M.; and Green R. J. 1994. Recognition of Handwritten Cursive Arabic Characters. *IEEE Trans. Pattern Anal. Mach. Intell.* 16, No. 6, 664-672.
- Al-Badr B. and Mahmoud S. 1995. Survey and bibliography of Arabic Optical text recognition, *Signal Processing*, 41, 49-77.
- Allam M. 1995. Segmentation versus segmentation-free for recognizing Arabic text, *Proc. SPIE* 2422, 228-235.
- Allam A. 1995. Segmentation versus segmentation-free for recognizing Arabic text, *Proc. SPIE* 2422, 228-235.
- Almuallim H. and Yamaguchi S. 1987. A method for recognition of Arabic cursive handwriting, *IEEE Trans. Pattern Analysis Machine Intelligence* 9, No. 5, 715-722, (Sep).
- Al-Sughayer and I. A. Al-Kharash I. A. 2004. “Arabic Morphological Analysis Techniques.” *J. of the Amer. Soc. For Inf. Sc. And Tech.* 55, No. 3, 189-213.
- Al-Emami S. and Usher M. 1990. On-line recognition of handwritten Arabic characters 12, No. 7, 704-710, (Sep).

- Al-sheikh T.S. and El-Taweel S. G. 1990. Real-time Arabic handwritten character recognition, *Pattern Recognition* 23, No. 12, 1323-1332.
- Al-Ohali Y.; Cheriet M.; and Suen C. 2003. Databases for recognition of handwritten Arabic cheques, *Pattern Recognition* 36, 111-121.
- Amin A.; Kaced A.; Haton J.-P.; and Mohr R. 1980. Handwritten Arabic Character recognition by the IRAC system, in *Proc. Int. Conf. on Pattern Recognition*, Miami, Florida, USA, 729-731.
- Amin, A. 1984. Recognition of Handwritten Arabic Script and Sentences, in *Proc. IAPR 2*, Montreal, 1055-1057.
- Amin A. 2000. Recognition of printed Arabic text based on global features and decision tree learning techniques. *Pattern Recognition* 33, No. 8, 1309-1323.
- Beldjehem M. 1993. "Un apport à la conception des systèmes hybrides neuro-flous par algorithmes d'approximation d'équations de relations floues en MIN-MAX : le système Fennec." Ph.D. Thesis in Computer Science and Software Engineering (Artificial Intelligence), Université de la Méditerranée (Aix-Marseille II), Marseille (in French).
- Beldjehem M. 1993. "Fennec, un générateur de systèmes neuro-flous." in *Proc. les Actes des Applications des Ensembles Flous*, Nîmes, France, 209 -218 (in French).
- Beldjehem M. 1993. "Le système fennec." in *Electronic BUFEAL* 55, 95-104 (in French).
- Beldjehem M. 1994. "The fennec system." in *Proc. ACM Symposium on Applied Computing (SAC), Track on fuzzy logic in Applications*, 126-130, Phoenix, AZ (March)
- Beldjehem M. 2002. "Machine Learning based on the possibilistic-neuro hybrid approach: design and implementation." in *Electronic BUFEAL* 87, 95-104.
- Beldjehem M. 2002. "Learning IF-THEN Fuzzy Weighted Rules." in *Proc. International Conference of Computational intelligence, Nicosia*, North Cyprus.
- Beldjehem M. 2006. "Validation of Hybrid MinMax Fuzzy–Neuro Systems." in *Proc. International conference of NAFIPS*, Montreal.
- Beldjehem M. 2006. "Visual Processing of Arabic Handwriting: Challenges and Future directions". in *Proc. SACH 06*, Maryland.
- Beldjehem M. 2008. "Towards a Validation Theory of Hybrid MinMax Fuzzy–Neuro Systems." in *Proc. of the WSEAS International Conference*, Sofia.
- Beldjehem M. 2008. "Towards a Validation Theory of Hybrid MinMax Fuzzy–Neuro Systems." in *Proc. of the CIMSA International Conference*, Istanbul.
- Bozinovic R. M. and Srihari S. H. 1989. Off-line cursive script word recognition, *IEEE Trans. PAMI* 11, No. 1, 68-83.
- Cheung A.; Bennamoun M.; and Bergmann N. W. 2001. An Arabic optical character recognition system using recognition-based segmentation. *Pattern Recognition* 34, No. 2, 215-233.
- Hull J. J. 1994. A Database for handwritten text recognition research, *IEEE Trans. PAMI* 16, 550-554.
- Hyder. S. S. and Khoujah A. 1988. Character Recognition of Cursive Scripts. *IEA/AIE* 2, 1146-1150.
- Kickert W. J. M. and Koppelaar H. 1976. Application of fuzzy set theory to syntactic pattern recognition of handwritten capitals, *IEEE Trans. Syst. Man Cybernet.* 6, 148-151, (Feb.)
- Liu C. L.; S. Jaeger; and M. Nakagawa M. 2004. Online recognition of Chinese characters: the state of the art, *IEEE Trans. On PAMI* 26, No. 2, 198-213.
- Lorigo L. M. and V. Govindaraju V. 2006. Offline Arabic handwriting recognition: a survey, *IEEE Trans. PAMI* 28, No. 5, 712-724.
- Miled M. and al. 1998. "Multi-level Arabic Handwritten Words Recognition." in *Proc. SSPR/SPR*, Sydney, NSW, Australia, 944-951.
- Mordechay S.; Lim H.; and Shoaf W. 1992. The utilization of fuzzy sets in the recognition of imperfect strings, *Fuzzy Sets and Systems*, 49, 331-337.
- Marr D. 1982. *Vision, A Computational Investigation into Human Representation and Processing of Visual Information*, Freeman W. H. and Company, San Francisco.
- Pal S. K. and Ghosh A. (Ed.). 2000. *Soft Computing in Image Processing*, Physica-Verlag, Heidelberg.
- Pechwitz M.; Snoussi-Maddouri S.; Märgner V.; Ellouse N.; and Amiri H. 2002. IFN/ENIT database of handwritten Arabic words, *Proc. Colloque Francophone International sur l'Ecrit et le Document*, Hammamet, Tunisia, 129-136.
- Plamondon R. and Srihari S. N. 2000. On-line and off-line handwriting: A comprehensive survey, *IEEE Transaction on Pattern Analysis and Machine Intelligence* 22, No. 1, 63-84.
- Simon J. C. 1985. *La Reconnaissance des Formes par Algorithmes*, Edition Masson, 252.
- Simon J. C. 1989. *From Pixels to Features*, North Holland, Amsterdam.
- Simon J. C. 1992. Off-line cursive word recognition, *Proc. IEEE* 80, No. 7, 1150-1161.
- Siy P. and Chen C. S. 1974. Fuzzy logic for handwritten character recognition, *IEEE Trans. Syst. Man Cybernet.* 4, 570-575, (Nov. ).
- Souici-Meslati L. and Sellami M. 2006. A Hybrid Neuro-Symbolic Approach for Arabic Handwritten Word Recognition. *JACIII* 10, No. 1, 17-25.
- Suen C. Y.; Legault R.; Nadal C. P.; Cheriet M.; and Lam L. 1993. Building a new generation of handwriting recognition systems. *Pattern Recognition Letters* 14, No. 4, 303-315.
- Ukkonen E. 1985. Algorithms for approximate string matching, *Information and Control* 64, 100-118.
- Tappert C. C.; Suen C. Y.; and Wakahara T. 1990. The state of the art in on-line handwriting recognition, *IEEE Trans. On PAMI* 12, No. 8, 787-808.
- Zadeh L. A. 1965. "Fuzzy sets." *Info. Control* 89, 338-353
- Zadeh L. A. 1971. "Toward a theory of fuzzy systems." in : *R. E. Kalman, N. Declaris, Eds., Aspects of Network and System Theory (Holt, Rinehart and Winston, New York)*, 209-245.
- Zadeh L. A. 1994. "Fuzzy logic neural networks, and soft computing." *Communications of the ACM* 37, 77-84.
- Zadeh L. A. 1998. "Soft Computing, Fuzzy Logic and Recognition Technology." In *Proc. IEEE Int. Conf. Fuzzy Syst.*, Anchorage, AK, 1678-1679.

# BEHAVIOUR BASED PREDICTIVE MOTION CONTROLLER FOR A MOBILE ROBOT

Krzysztof Skrzypczyk

Krzysztof Fajarewicz

Adam Galuszka

Department of Automatic Control

Silesian University of Technology

Akademicka 16, 44-100 Gliwice

Poland

E-mail: {krzysztof.skrzypczyk,krzysztof.fajarewicz,adam.galuszka}@polsl.pl

## KEYWORDS

Predictive control, behaviour based control, identification.

## ABSTRACT

This paper presents an application of a predictive model of a range sensor to the behaviour-based motion controller of a mobile robot. The model predicts the sensor readings for a given time horizon using current sensor readings and velocities of the robot wheels assumed for this horizon. The novelty of the model presented in the paper comes from the fact that its structure takes into account physical phenomena and is not just a black box. From this point of view it may be regarded as a semi-phenomenological model. This model was applied to the control system designed as the behaviour-based one. The system is intended to navigate the Khepera<sup>TM</sup> mobile robot in a collision free way. The use of the predictive model of the sensor results in improvement of the quality of the motion of the robot what is proved by the results of simulations presented in the paper.

## INTRODUCTION

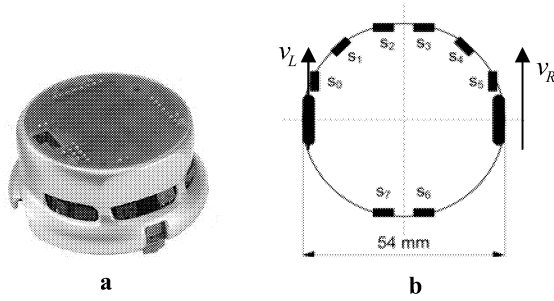
A mobile robot is a machine that is intended to operate in an environment a model of which is unknown apriori. The robot should also react dynamical changes of this environment. Inaccurate sensors, world unpredictability and imperfect control caused that the traditional, planner based approaches ended in failure. Therefore there was a need to develop more efficient and faster methods of collision free movement control. One of them is a purely reactive architecture introduced in (Braitenberg, 1984, Brooks 1991) which implements a control strategy as a collection of stimulus-reaction pairs. The system consists of a collection of purely reactive rules utilize minimal internal state. These systems use no internal models, perform no search. The only action they take is to assign appropriate control for a given input data (sensor reading). The next stage in evolution process of reactive systems was behaviour based architecture (Arkin, 1998, Brooks, 1991, Mataric, 1992, Michaud and Mataric, 1998). These systems embody some of the properties of reactive systems and may contain reactive components. However the primary feature of behaviour based systems is their distributed nature. They are built of a collection of schemas of reaction called behaviours devoid of centralized

reasoning module. The behaviour based systems are more efficient and flexible than purely reactive ones because they may use different forms of data representation. The aforementioned features of the behaviour based systems address them to work with inaccurate and uncertain sensors. The reason of that is there exists strong feedback in this system that can correct errors of sensor measurements. In the work by (Skrzypczyk, 2005), the behaviour based system was applied to control the Khepera<sup>TM</sup> robot. The sensors the robot is equipped with are characterized by very low accuracy and high uncertainty. Limited sensing range of these sensors implies some problems in the collision free navigation of the robot. Therefore some efforts were made to overcome this inconvenience. The common sense way to improve the work of such system is to apply a predictive algorithm. This is the reason that many predictive algorithms intended to control mobile robots have been reported in the literature. However many models of mobile robots usually takes into account only their kinematics and dynamics. Then variables describing the state of the mobile robot are: the location coordinates, the heading and sometimes the velocity of the robot. In such models usually readings from the distance sensors are not taken into account. Moreover, only present measurements are used without any anticipation of measurements. Such anticipation may improve the control quality. If, for example, during the operation of behaviour-based control the "avoid obstacles" rule is activated if the sensory reading exceed assumed threshold, then the control signals might be improved based on the anticipation of the sensory readings. In the literature several attempts of anticipation of sensory readings are reported but they are not serve during the control of the robot but for selection of landmarks (Fleischer et al. 2003). For example in the articles by (Duckett and Nehmzow 1999, Marsland et al. 2001) an artificial feedforward neural network was proposed as a predictor of further measurements of sensors. The proposed predictor was very simple - it contained only one layer and took as inputs only past readings from neighbouring sensors and did not use wheel velocities of the mobile robot. Such simplified model was good enough for the landmark selection based on the difference between the predicted and real readings but for the control purposes a more accurate model is needed. In the work by (Tani, 2001) the more precise model consisting the recurrent neural network was presented. In this paper predictive model of the robot sensor was applied to the

behaviour-based control system. Such approach (what was proved by experiments) gives very good results. The control system of the robot that utilizes this model can react faster to the approaching object.

## SENSOR MODEL

In this paper we utilize the sensor model for the Khepera<sup>TM</sup> robot. The derivation of the model was presented in (Fujarewicz, 2007) but to clearly address the problem the main principles of the model are presented in this section. Figure 1 presents the picture and the top view of the robot. The robot is driven by changing velocities of the left ( $v_L$ ) and the right ( $v_R$ ) wheel respectively. The robot is equipped with eight infra-red sensors with readings (signals) denoted as  $s_0$  up to  $s_7$ .



Figures 1: A picture of the Khepera<sup>TM</sup> robot used in experiments (a) and the layout of the sensors of this robot (b).

The predictive model of the sensors for the Khepera<sup>TM</sup> robot takes into account two velocities  $v_L$  and  $v_R$ . Moreover, we assume that the model for  $i$ th depends also on the neighbouring sensors:  $s_{i-1}$  and  $s_{i+1}$ . Hence, the general form of the model for  $i$ th sensor is as follows:

$$\frac{ds_i}{dt} = f_i(s_{i-1}, s_i, s_{i+1}, v_L, v_R) \quad (1)$$

The model (1) may be presented in a block-diagram form, see fig. 2. Several possible structures of the function  $f_i(\cdot)$  were examined. Finally the function quadratic with respect to  $s_i$  and linear with respect to  $s_{i-1}$  and  $s_{i+1}$  was chosen:

$$\frac{ds_i}{dt} = v_L \mathbf{p}'_L \mathbf{S} + v_R \mathbf{p}'_R \mathbf{S} \quad (2)$$

where

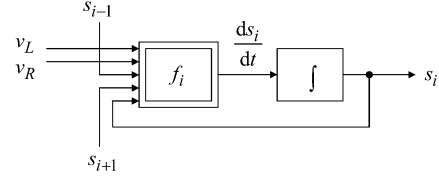
$$\mathbf{p}'_L = [p_{l1}, p_{l2}, \dots, p_{l5}], \quad \mathbf{p}'_R = [p_{r6}, p_{r7}, \dots, p_{r10}]$$

$$\mathbf{S} = [1, s_i, s_i^2, s_{i-1}, s_{i+1}]'$$

In the equation (2) velocities  $v_L$  and  $v_R$  are multipliers as in the model for the single sensor derived in (Fujarewicz, 2007). For  $v_L = v_R = 0$  the output of the model (2) is constant.

The model contains 80 parameters  $p_{ij}$ ;  $i = 1, 2, \dots, 8$ ,  $j = 1, 2, \dots, 10$ . To identify the model we

need to find their optimal values based on the data collected during identification experiment.



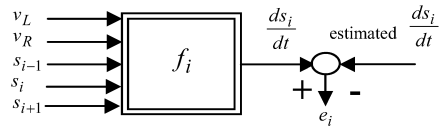
Figures 2: Block diagram of the model for one sensor.

## The model identification

The data contains sensor readings  $s_i$ ;  $i = 1, 2, \dots, 8$  and velocities  $v_L$  and  $v_R$  recorded at discrete moments of time  $\{0, h, 2h, 3h, \dots, Nh\}$  where  $h$  is the sampling time and  $N+1$  is the total number of the samples. There are several possible ways to identify the parameters. The problem of identification is not trivial because the model is the non-linear one. Fortunately, the task can be solved by using Least Squares (LS) method. Let us write the equation (1) where the left side is replaced by backward-difference estimator of the derivative:

$$\frac{s_i(t) - s_i(t-h)}{h} = f_i(s_{i-1}, s_i, s_{i+1}, v_L, v_R) \quad (3)$$

For one sensor we may write  $N$  such equations for  $t = h, 2h, 3h, \dots, Nh$ . The model (2) is linear with respect to parameters  $p_{ij}$ ;  $i = 1, 2, \dots, 8$ ,  $j = 1, 2, \dots, 10$  so it can be solved by Least Squares (number of equations is much greater than number of parameters). The approach is illustrated in fig. 3.



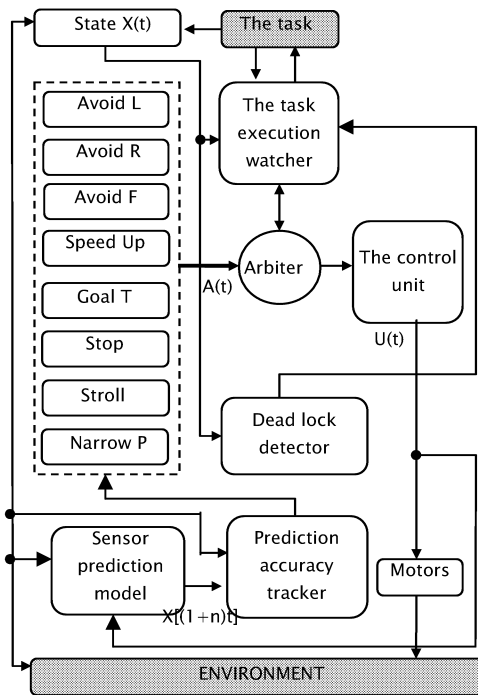
Figures 3: The idea of the identification: the minimized error is the difference between derivatives of the sensor reading taken from the model and estimated from the data.

## THE CONTROL SYSTEM

Architecture of the control system presented in this work is the behaviour-based one (Arkin, 1998, Brooks, 1991). The system is composed of the input-output reactive modules called behaviours. The behaviours process the state and the sensors readings into the proper set-points (linear and angular velocity) for the motion controller. The behaviours are coordinated by the arbiter with fixed priorities. A diagram of the controller is presented in fig.4. Its easy to distinguish five main modules of the controller:

- Behaviours definition module;
- Arbitration module;
- Control computation module;
- Task execution watcher module;
- Dead lock detector module;

Each of behaviour can be perceived as a pattern of a reaction of the system to a given stimulus that comes from an environment. The stimulus can be represented by the the reading of the sensors as well as by the internal state of the robot itself. The behaviour definition module consists of the eight behaviours. The four behaviours (*avoid left*, *avoid right*, *avoid front*, *speed up*) are designed to avoid collisions with the objects (obstacles) located correspondingly on the left, right, frontal and the back side of the robot body. Fifth behaviour (*goal tracking*) minimizes the distance between the robot and the target. The behaviour *stop* simply stops the robot in case the collision is detected or the target is reached by the robot. Sixth behaviour called *stroll* makes the robot goes straight in case no objects are detected.



Figures 4: The behaviour based control system architecture.

And the last behaviour – *narrow passage* stabilizes the robot movement. This behaviour prevents the robot oscillations while it goes through narrow passages. Each behaviour is designed as a function which maps the value of the selected element of  $X$  into the activation level of the given behaviour  $a_i \in (0,1)$ . In the discussed system each of behaviour is modeled by the sigmoid function.

The behaviours generate the control that is optimal from the perspective of their own "point of view". Therefore some coordination mechanism has to be used to obtain the final

control of the robot that is optimal from the perspective of the executed task. In the discussed system the method of priority arbitration was applied. The arbitration based on this method results in choosing  $k$ th behaviour which satisfies the following:

$$k = \max_{i=1,2,\dots,8} (a_i q_i)$$

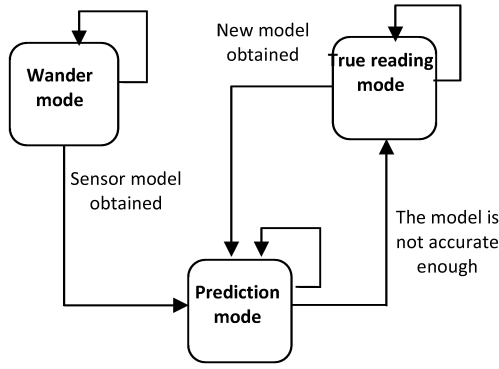
where  $q_i$  denotes the priority fixed to the  $i$ th behaviour. The activation level of the selected  $k$ th behaviour is used to compute the control of the robot. Both the angular and the linear velocity are defined by a heuristic function of the activation level of the selected behaviour.

The next module is called *the task execution watcher*. This module is designed as a finite state automaton. The function of this module is to supervise the process of the navigational task execution. The automaton is determined by four states.

The module starts its work in the state of waiting for the new task to do. If the new task is sent to the module it will switch itself to the state of execution of the task - the robot moves in collision free way toward the target. If the task is completed the module will send a message to the global coordinator and switch itself to the first state. If any exception happens during task execution (collision detection for instance) the robot will stop, send appropriate message to the global coordinator, and switch to the first state. The aim of this brief description is only to sketch the main principles and the ideas of the construction of the behaviour based motion controller module. Detailed description exceeds the scope of the work and is not its main subject. For detailed information about the system please refer to the work by (Skrzypczyk, 2005)

## PREDICTIVE CONTROL

The system utilizes the previously defined predictive model of the sensor to perform the navigational task more efficiently. Assuming that the robot moves with a given velocity  $(v_L, v_R)$  it is possible to forecast the presence of an obstacle and take a proper avoiding action earlier. The module that realizes this uses the model of the sensors of the robot. The model is identified on the basis of experimental training data collected during some run of the robot. There are three distinguishable states of the work of the controller. They are presented in fig. 5. In order to obtain the data, initially the system is switched into the state of wandering mode. This mode consists in exploring in a collision free manner the robot surroundings. In order to move the robot one of Braitenberg's algorithms was applied. During this operation the system collects the training data. If the robot records a given number of data it will start to identify the model (2). In this case the system switches into *the prediction mode*. Working in this mode the system navigates the robot to the target using prediction of the sensor readings obtained from the equation (3).



Figures 5: A diagram that illustrates the particular states of the work of the system.

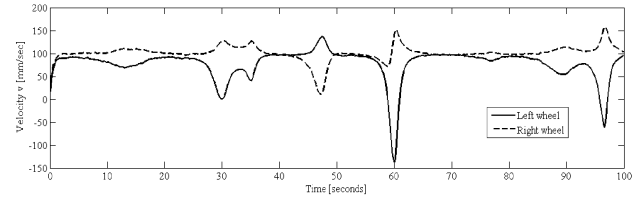
Of course there is the need to verify constantly the accuracy of the obtained model. In order to do it an additional module called *the accuracy tracker* is added to the system. The module monitors the difference between the predicted and the real value of the sensor reading. If the difference is greater than some threshold value the system is switched into the mode where it uses the true readings to perform the action. As the robot performs its mission the readings of the sensors as well as the velocities of the robot wheels are collected. The number of collected samples is determined by the time-window frame. Simultaneously as the robot works on the basis of the real sensor readings, the new model of sensor is computed. When the new model is prepared the system switches again to the prediction mode.

## SIMULATIONS

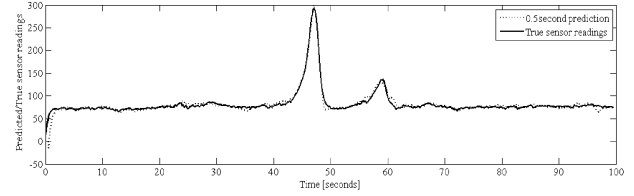
### Identification and validation of the sensor model

The model of the sensor (2) was identified using the method previously discussed. During the operation in the *wander mode* the system collected two data sets: identification and a validation data set. During both experiment the robot explored its workspace and was driven by Braitenberg's algorithm (Braitenberg, 1984). Velocities of both wheels and signals from eight sensors were recorded with sampling time  $h = 0.1$  [s]. The model (2) with parameters estimated based on the identification data set was tested on the validation data set. Below validation results are presented. Figure 6 presents velocities of wheels during the validation experiment. The next two figures contain sensory readings and their prediction for only one, 6th sensor.

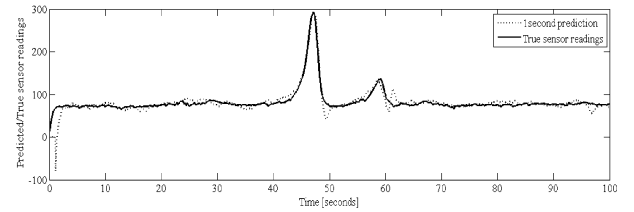
The prediction for time horizon  $\Delta t = 0.5$  [s] is presented in Fig. 7 by dotted line. Solid line represents the true (recorded) signal from the sensor. In fact the dotted line is a result of many simulations (for 0.5 [s]), each starting from true readings of all sensors. In other words: it tells us how the signal from the sensor has been predicted 0.5 [s] before. Fig. 8 presents predictions for longer time horizon 1[s]. One can observe that the differences are bigger than before. This is obvious: the longer is the prediction time horizon the bigger are differences.



Figures 6: The velocities of the robot wheels during the validation experiment.



Figures 7: Predicted vs. true sensor readings for one sensor during the validation experiment, prediction horizon = 0.5 second.

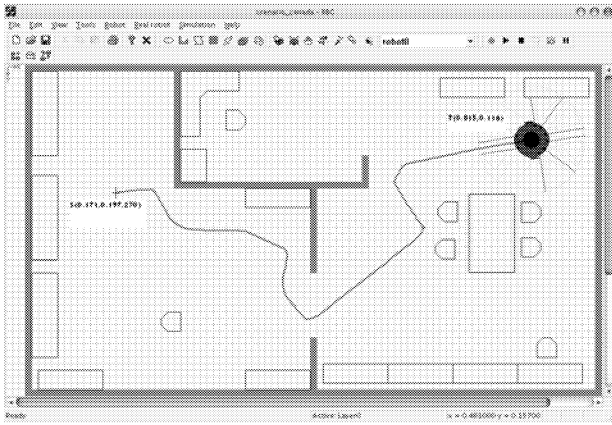


Figures 8: Predicted vs. true sensor readings for one sensor during the validation experiment, prediction horizon = 1 second.

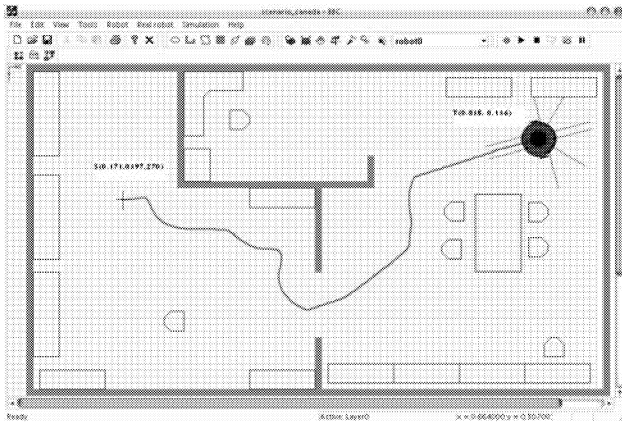
### Predictive work of the behaviour based controller

The system discussed in the paper was implemented in the M.A.S.S. simulation environment. In order to prove the efficiency of the approach discussed in this paper a number of simulations were carried out. Here in this section there are presented results of two exemplary experiments. Using the simulation software some scenario was modeled. The model of the environment corresponds to a typical office layout. Using this simulation tool, the control system presented in the paper as well as the prediction algorithm was implemented. In the fig. 9 results of the two experiments are presented. The experiments consist in navigation of the robot from the initial pose  $(0.171, 0.197, 270)$  [m, m, °] to the desired position  $(0.815, 0.116)$  [m, m]. Figure 9 presents the work of the system without the prediction mechanism. As one can see the robot reached the goal avoiding collisions. However, the path of the robot sometimes passes the obstacles by in a very close distance. Turning the prediction mechanism on improves the quality of the path of the robot. The prediction horizon was set to the value 0.5[s]. Thanks to the prediction of the sensory readings the system can react to the obstacle occurrence both faster and in a smoother way. The results of the simulation experiments were very rewarding. Therefore ongoing works on applying this methodology to a real Pioneer 3-DX™ robotic system have been carried on.





Figures 9: The collision free path obtained without using the prediction model.



Figures 10: The smoother collision free path of the robot – the result of applying the prediction model.

## CONCLUSION

In this paper the predictive model of sensory readings for the mobile robot and its application to the behaviour-based control system were proposed. The model of the sensor is a semi-phenomenological one that means the structure of the model takes into account the physical phenomena. The model contains parameters which are estimated based on experimental data. The model is non-linear but it is possible to estimate parameters using LS method. Numerical results of identification (based on the learning data) and validation (based on the test data) showed that the prediction of the model is satisfactory for time horizon of about 1 second which is enough long for control purposes. Thanks to this model there was possible to improve the efficiency of collision free motion control of the robot. Since the robot the work of which was simulated is equipped with a low accuracy and low range sensors, the prediction allowed speeding up the reactions of the system to the obstacle occurrence. The obtained paths of the collision free

movement of the robot are smoother and pass the obstacles by in a greater distance. The simulation studies of the discussed approach seem to be very promising. Therefore further researches have been carried on. These researches are focused on applying this approach to the real robotic system.

## Acknowledgements

This work has been supported by the BK grant no 209/RAu1/t.1 in the years 2008 (for the first and the second author).

For the third author the work has been supported by Polish Ministry of Science funds in the years 2007-2008.

## REFERENCES

- Althaus P., Christensen H.I., 2003. Behaviour coordination in structured environments. *Advanced Robotics*, 17(7).
- Arkin R. C., 1998. *Behaviour-Based Robotics*, MIT Press, Cambridge, MA.
- Bicho E., Schoner G., 1997. The dynamic approach to autonomous robotics demonstrated on a low-level vehicle platform. *Robotics and Autonomous Systems*, 21(1).
- Braitenberg V., 1984. *Vehicles: Experiments in synthetic psychology*, MIT Press.
- Brooks R.A., 1991. Intelligence without representation. *Artificial Intelligence*, (47).
- Duckett T., Nehmzow U., 1999. Learning to predict sonar readings for mobile robot landmark selection, Internal Report, University of Manchester, Manchester, UK.
- Fleischer, J., Marsland, S., Shapiro, J., 2003. Sensory anticipation for autonomous selection of robot landmarks, M. Butz et al. (Eds.): *Anticipatory Behaviour ...*, LNAI 2684, pp. 201–221, Springer-Verlag.
- Fujarewicz K., 2007. Predictive model of sensor readings for a mobile robot, *Proceedings of World Academy of Science, Engineering and Technology Volume 20 april 2007 issn 1307-6884*, pp. 162-166
- Marsland, S., Nehmzow, U., and Duckett, T., 2001. Learning to select distinctive landmarks for mobile robot navigation. *Robotics and Autonomous Systems*, 37:241-260.
- Mataric M.J., 1992. Integration of representation into goal-driven behaviour-based robots. *IEEE Transactions on Robotics and Automation*, 8(3).
- Skrzypczyk K., 2005. Hybrid control system of a mobile robot. PHD dissertation, Gliwice, Poland, 2005
- Tani, J., 1996. Model-based learning for mobile robot navigation from the dynamical systems perspective. *IEEE Trans. Syst. Man Cybern., Part B, Cybern.* 26(3), 421–436

## AUTHOR BIOGRAPHY

KRZYSZTOF SKRZYPCZYK was born in Tarnowskie Gory, Poland and went to the Silesian University of Technology, at Gliwice, where he studied automatic control and robotics. There he obtained his MSC degree in 2000 and the PHD in 2005. Since 2005 he has been working at The Silesian University of Technology in the Automatic Control Department. His research activities are focused on a mobile robot motion planning and control. Moreover he works on programming simulation tools for robotics.

# Cognitive Modeling of a Cooking Activity: Integration of the Contention Scheduling Theory in the Cognitive Architecture ACT-R

Pierre-Yves Groussard  
Hélène Pigot

Domus Laboratory  
University of Sherbrooke

email: pierre-yves.groussard@Usherbrooke.ca  
helene.pigot@Usherbrooke.ca

## KEYWORDS

Cognitive Modeling, ADL Simulation, Cognitive Science, ACT-R, Contention Scheduling

## ABSTRACT

According to the fact that human actions are goal directed, representation and management of goals are crucial to create models close to human behavior. In this paper, we present the Cooper and Shallice theory named Contention Scheduling (CS) for goals representation and management and we present our own goal model based on the CS theory and use ACT-R to implement it. ACT-R is a cognitive architecture well adapted to model human behavior and is often used to model cognitive theories. We finally present how we apply our goal model to an activity of daily living (ADL) and the results of a cooking activity simulation.

## INTRODUCTION

In order to adequately assist people with cognitive impairments, it is necessary to model human behaviors while performing activities. Such models are useful to validate cognitive orthoses and also to anticipate cognitive errors occurring during the activity realization. According to psychological literature, human activities are goal-directed (Piaget 1952). The goals are hierarchically structured where goals are subdivided into subgoals. Each goal may have one or more procedures to reach it. This procedure will be chosen according to the context. For instance, when one wants to drink coffee, he may pour instant coffee powder into boiled water or use an espresso machine.

One famous theory to represent and manage goals is the Contention Scheduling (CS) theory proposed by Norman and Shallice (Norman and Shallice 1986). In that theory a goal is linked to schemas that can be selected to accomplish it, depending on its activation level. The CS has been implemented to simulate tasks as coffee and breakfast preparation (Cooper and Shallice 2000). The goals sequencing is well represented but not the memory access. Cooper proposes

to improve its model by implementing it into the cognitive architecture ACT-R (Cooper 2002) but this proposition is not still realized. The cognitive architecture ACT-R is based on cognitive theories to simulate the human learning and reasoning processes (Anderson 1996). It has been extended with perceptual and motor modules (Anderson et al. 2004). However ACT-R lacks of facilities to manage goals.

The purpose of this paper is to present an implementation of the CS using ACT-R and to apply it on a human activity. This implementation is aimed to fulfill numerous situations where several goals are managed and to simulate the human behavior. After the CS and ACT-R presentations, we expose our implementation of the CS using ACT-R. It is then applied to model a cooking task, preparing a pudding. The results show how the model simulates activity realization and the various sequencing of goals selected.

## BACKGROUND

### Contention Scheduling Theory

Norman and Shallice have proposed a cognitive theory for goal representation and planning (Norman and Shallice 1986). Two main modules coexist to simulate how people plans and executes actions. The Supervisory Attentional System (SAS) is responsible of planning and controlling the actions while the contention scheduling (CS) is responsible of the action execution. The CS contains a set of goals required to perform a routine activity. For each goal several schemas should be used in order to reach that goal. After the SAS has planned the activity to be performed, the CS executes automatically a sequence of goals selecting the appropriate schemas. The CS model is constituted of three main networks: the resource network, the object network and the schema network. The schema network contains the goal and sub-goal decomposition of the routine activities, named tasks. The task is represented by a hierarchical tree where goal levels alternate with schema levels. For instance, two schemas nodes «making coffee with instant powder» and «making coffee with espresso machine» depends from the

goal node «making coffee». The selection of one or another schema depends on several influences coming from the schema itself and from the environment. The object network simulates the influences that determines the choice of an object, per se a schema. For instance, the sight of coffee powder and hot water will determine the selection of the «making powder with instant coffee» schema. The resource network represents the cognitive and physical resources needed to reach a goal.

## ACT-R

Numerous cognitive models use ACT-R to simulate learning task, reasoning and individual cognitive abilities (Anderson et al. 1997). ACT-R is a cognitive architecture, i.e. a computational model of human cognitive structures and processes based on psychological theories. It is composed of several modules corresponding to different brain areas. Each of these modules communicates with the model by the mean of buffers. In ACT-R the knowledge is represented by chunks. Chunks are composed of slots that contain information like a string or other chunks. A selection mechanism allows retrieving a chunk using activation function. The chunk selected, is the one matching the request and with the higher activation level. The procedural knowledge, which explains how to execute actions, is represented by rules. These rules lead to retrieve a chunk or to apply other rules. A rule describes what to do, under certain circumstances represented by the buffers content.

However just few models simulate human performing activities of daily living. One drawback is to express the goal management necessary when one performs activity. In general, cognitive models based on ACT-R use one goal which changes state during the model execution. Such implementation does not allow to make simulation of activities where goal sequences may change from one realization to another. For instance one would prefer once to pour instant powder before sugar or inversely sugar and then instant powder. Moreover at each step of an execution, the goal chosen for the next step can be different depending on the context.

## COGNITIVE BEHAVIOR MODEL

We propose a general model for goals representation and management based on CS theory goals and schema representation. According to this theory, an activity is defined by a goal (Figure 1). Each goal contains one or more schemas; each one describing a procedure to reach the goal. For instance, two schemas satisfy the goal "to make coffee" : "to make instant coffee" or "to make coffee with espresso machine". According to the CS theory, the schemas are defined by a set of goals. We introduce a new element the action. An action represents the concrete action, physical or cognitive, needed to reach a goal that cannot be split any more. The decision of splitting goals into sub goals or actions depends on

the level of granularity chosen to describe the activity. For instance, to describe the activity "making white coffee" an action should be "to take the milk" but if more granularity is suited the manual procedure like "to grasp the bottle" should be described. Figure 1 shows a goal that is reached using one of the two schemas. In the schema 1, three schemas are repertoried to reach the sub goal 1, whereas only one schema is known to reach the sub goal 2. For the schema two, the goal to reach couldn't be split and is linked to an action.

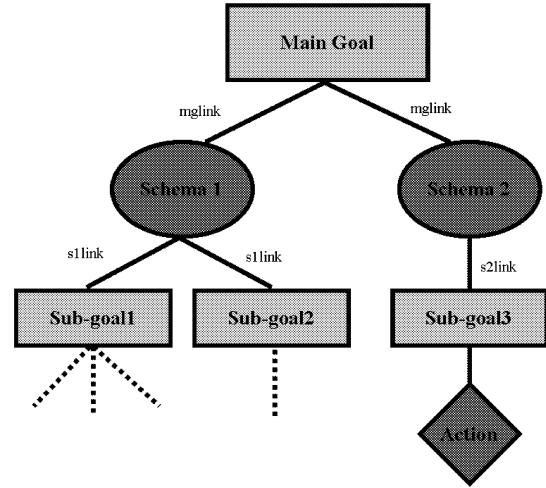


Figure 1: Simple representation of activation links between the chunks Goal, Schema and Action

To represent goals, schemas and actions, we define three types of chunks in ACT-R. The type of a chunk specifies the number of slots, i.e. information, which it contains. Specifying these types we insure that all goals and schemas have the same structure, i.e. the same number of slots. It induces the generalization of the retrieval process of goals and schemas. Figure 2 shows the definition of goal and schema chunks type. All schemas having the same value of glink slot that the current goal received activation. All goals having the same value of slink slot that the current selected schema received activation. Precond and postcond slots define the pre and post conditions needed to the chunk selection and the divisible slot define if the goal is split or not.

(chunk-type Goal name glink slink precond postcond divisible)  
(chunk-type Schema name glink slink precond postcond)

Figure 2: Goal and Schema chunk type definition

The selection of a particular goal activates the schemas associated to this goal. We use some of the slots to create links between a goal and the associated schemas to reach it and vice versa (Figure 3). These links are based on ACT-R activation mechanisms, so we call them "activation link". Two chunks A and B are linked by an "activation link" if they contain both a same chunk C, either in the slot glink or slink. Therefore when a Goal chunk is selected it transmits activation to all schemas that can accomplish it.

```

(MainGoal isa Goal
 name « MainGoal »
 glink mglink
 slink nil
 precond nil
 postcond « Task Done »
 divisible « true »)

(Schema1 isa Schema
 name « Schema1 »
 glink mglink
 slink s1link
 precond nil
 postcond « Task Done »)

(Schema2 isa Schema
 name « Schema2 »
 glink mglink
 slink s2link
 precond nil
 postcond « Task Done »)

(SubGoal1 isa Goal
 name « SubGoal1 »
 glink sg1link
 slink s1link
 precond nil
 postcond « sb1 Done »
 divisible « true »)

(SubGoal2 isa Goal
 name « SubGoal1 »
 glink sg2link
 slink s1link
 precond « sb1 Done »
 postcond « sb2 Done »
 divisible « true »)

(SubGoal3 isa Goal
 name « SubGoal3 »
 glink sg3link
 slink s2link
 precond nil
 postcond « Task Done »
 divisible « false »)

```

Figure 3: Goals and schemas representation of figure 1. The bold slots highlight the "activation link" between a goal and a schema

The model is divided in two parts (Figure 4). The first part, Memory Elements, is the mental representation of the goals, schemas and actions presented previously. The second part, "Cognitive Process", is the process which leads to the execution of the actions. A goal selection is made depending on the Schema Network activation and the Mental Environment Representation. The goal with the higher activation is selected and activations and inhibitions are sent to elements of the Schema Network. Two cases occur whether the goal could be split or not. If the goal can't be split an action is selected and executed. This execution modifies the activation level of the mental representation and schema network elements. If the goal can be split, a schema is selected to accomplish it. This schema selection directly influences a new goal selection and so forth until the selected goal can't be split.

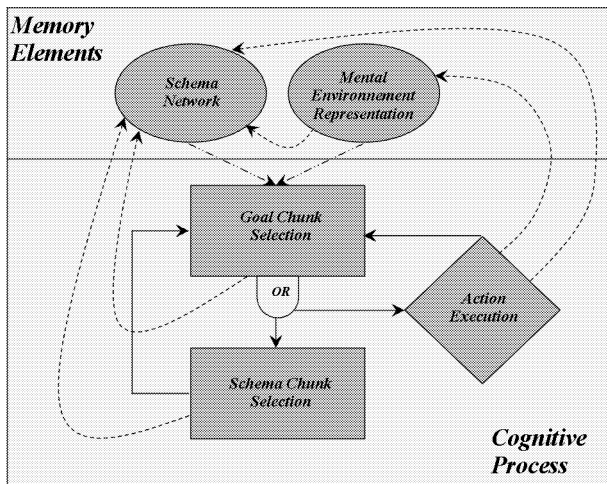


Figure 4: Schematic view of goal selection mechanism. The dotted line represents the activation/inhibition mechanism

## COOKING ACTIVITY SIMULATION

### Activity description

The cognitive behavior model is applied to simulate a cooking activity. It is used in a test, named Kitchen Task Assessment (KTA), to evaluate the autonomy and the level of assistance required by people with dementia of the Alzheimer type (Baum and Edwards 1993). It consists to realize a pudding using a commercial preparation. It is composed of six steps. The subject first measures ingredients then stirs the preparation, cooks it and pours the cooking mix in dishes. Finally, the subject cleans the utensils used while cooking. Each step can be broken down into sub-goal which permits to reach the goal of the step. The sequence of the goals could be rigid or it does not matter. For instance one must cook the preparation before pouring it while during the first step, the goal "measuring milk" is subdivided into two sub-goals "gathering milk" and "gathering the measuring cup". Either one picks up first the milk or the measuring cup.

A previous model has been built to simulate the pudding preparation (Serna et al. 2007). It reproduces the execution of the pudding preparation and simulates the decreasing performances due to cognitive diminishing occurring during the Alzheimer disease. However, this cognitive model is too rigid. Each sequence of goals is executed always in the same order. Based on the activity and objects representation of the pudding preparation proposed previously and the CS implementation of goals, we built a new model where the goal management is closer to the human behavior.

### Representation of the activity

Each step, of the cooking activity, is represented by a "Goal Chunk" and one or more "Schema Chunk" describing how the goal can be accomplished. Each goal that can be split has activation links with one or more schema. The others have activation links with an "Action Chunk" that represent the elementary action to reach it. The main goal of the activity has only one schema because we just simulate one recipe. Some of the steps, represented by goals, have several schemas to reach it because several procedures exist to reach it. For instance, it is possible to mix ingredients in the measuring cup and then to pour the mix in the saucepan or pour the ingredients in the saucepan and mix it after (Figure 5).

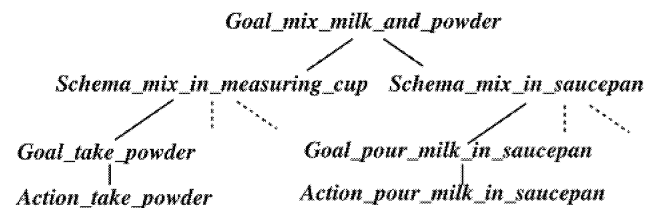


Figure 5: Extract of the cooking activity representation

All goals that couldn't be split anymore are linked to actions. We specify different types of action such as "take", "stir", who have particular slot depending on the action type. For example, the action "take" has a slot to specify the target and the type of the target to take. Therefore an activation link is created between the action "to take" and the objet to be taken. All the elements in the environment are also represented by chunks. Each element contains slots to characterize it. For instance the milk is an "ingredient", it is "liquid" and can be "measuring". A spoon is a "utensil", it can be "taken" and serves to "stir". Consequently we obtain a semantic network that allows having the correct object activated in function of the action. We use the ACT-R rules to implement the goal selection mechanism (Figure 4). We also implement a control mechanism to unsure the correct execution of the model. This mechanism is based on pre and post condition for goal selection. Each goal has a precondition required to its execution. If the goal selected doesn't match the precondition required, a case that normally not appears for not cognitive impaired people, a new request is made. Otherwise the goal is selected to pursue the activity.

## RESULTS AND DISCUSSION

The implementation of the CS in ACT-R has been tested on a cooking activity: make a pudding following the KTA protocol. The six steps of the recipe have been implemented. The trace simulation lists the actions across the time. These actions are performed to reach a goal, which could not be subdivided. The simulation exhibits the various ways to prepare the pudding. Like one watching the patient performing a task, the simulation trace shows actions but not goals selection. We then infer which goals and which schemas have been selected. The trace can shows different sequences of actions for a same step for two reasons. The first situation occurs when the same actions are reached but under different orders. This occurs when, in a schema, several goals have been reached in a different order. Figure 6 illustrates the various order of the actions presented with two different simulations. To pour the milk into the cup, one first takes the milk and then the measuring cup (figure 6.a) while the other prefers to first take the measuring cup (figure 6.b). The last goal of the schema « pouring milk in the measuring cup » remains always at the end, because the preconditions having the milk and having the measuring cup must be satisfied. In the model, the various sequences are allowed thanks to the goal activation computed during the simulation. Sometimes the chunk representing the goal « to take milk » got the higher activation while other times the chunk representing the goal « to take the measuring cup » receives a higher activation. At the beginning, each goal receives the same activation but at the selection time, ACT-R activation computation adds noise reinforcing one goal rather than the other.

The second situation appears when several schemas reach the same goal. Each of these schemas brings into play different goals. Doing one of these schemas leads to the selection of goals that not appear while selecting other schemas.

```

"Patient takes milk"
a. "Patient takes the measuring cup"
   "Patient pours milk in the measuring cup"

"Patient takes the measuring cup"
b. "Patient takes milk"
   "Patient pours milk in the measuring cup"

```

Figure 6: Extract of an execution trace of the measuring milk step

For instance when the patient wants to mix the powder with the milk, he could either pour the milk and the powder in the saucepan and mix them, or prefers to mix them into the measuring cup (figure 7). The model allows the two situations thanks to the sub symbolic mechanism of ACT-R. The two schemas are represented with chunks owning the same preconditions. When these preconditions are satisfied one schema or the other should be selected depending of their level of activation computed at that time of simulation.

```

"Patient pours measured milk in the saucepan"
"Patient takes powder"
a. "Patient pours powder in saucepan"
   "Patient takes the wooden spoon"
   "Patient stirs in:"
   SAUCEPAN
   "Patient stirs in:"
   SAUCEPAN

"Patient takes powder"
"Patient pours powder in the measuring cup"
b. "Patient takes the wooden spoon"
   "Patient stirs in:"
   MEASURING-CUP
   "Patient stirs in:"
   MEASURING-CUP
   "Patient pours the mix in the saucepan"

```

Figure 7: Extract of an execution trace for the stirring step

## CONCLUSION

ACT-R is a cognitive architecture often used to model cognitive phenomena, to simulate for instance memory process, learning task, or problem solving (Pavlik and Anderson 2005, Lebiere and Taatgen 1998, Danker and Anderson 2007). But few models are built in ACT-R to simulate executive functions like those involved when performing daily activities (Serna et al. Accepted, Salvucci et al. 2002). Until recently, ACT-R was based on a stack to manage goals needed to lead the model execution. The goals were put on or removed from the top of the stack. This goal management was criticized because it is not conform to the human goal management (Altmann and Trafton 2002). Especially, completion errors could not be explained by a goals stack. In ACT-R, goals could be processed like chunks using a state slot which indicates the progression of the goal (Taatgen 2007). However modeling daily activities lead to define various goals, like all goals used for describing the KTA activity. Then different goals at different times should

be chosen during the activity simulation. To solve this constraint, we propose to model CS theory in ACT-R. The CS theory is well adapted to simulate executive functions but not to represent memory elements or memory mechanisms. We propose a way to unify these two approaches to produce a more realistic model. Using the powerful of the CS theory with the ACT-R cognitive architecture we propose a mean to simulate, in a same model, executive functions and cognitive functions.

The model presented generalizes the use of goal in ACT-R using the CS theory. It discriminates the goal and the schema, defined as a procedure to reach a goal. According to ACT-R 6, goals, schemas and actions remain a standard unit of memory and is stored in the declarative memory. This model emphasizes the importance of a uniform representation of the goals. Each goal is described with a fixed number of slots, owning preconditions and postconditions. Fixing the number of slots leads to a uniform way to calculate activation levels of knowledge associated with the goals. The management of goal is facilitated by introducing schemas. It makes explicit the difference between a goal and the way to reach it. Then the mechanism of goal selection using the sub symbolic of ACT-R leads to execute goals in various orders. Furthermore, selecting one or another schema associated to a goal leads to simulate various ways to reach that goal.

This model is applied to the preparation of pudding. It illustrates how an activity should be simulated in various manner. The results show that the simulation of behavior is more conform to the reality where one chooses a way to do depending on the situations. Thanks to this model, the various ways to accomplish a task is simulated in so far they observe the preconditions of the goals. This model should be extended to allow learning process, as learning new schema, and cognitive errors. The cognitive errors, either procedural or declarative, should be easily introduced. A previous model has shown how to introduce errors in the pudding preparation (Serna et al. 2007). This should be apply to this model to reproduce omissions by modifying the activations parameters. In the previous model errors of sequence were induced by coding wrong procedures. In our model, wrong sequences could occur when a wrong schema or goal is selected but the preconditions are not checked enough. Introducing similarities between utensils and ingredients will also lead to substitutions errors.

## REFERENCES

- Altmann E.M. and Trafton J.G., 2002. *Memory for goals: an activation-based model*. *Cognitive Science*, 26, 39–83.
- Anderson J.; Matessa M.; and Lebiere C., 1997. *ACT-R: A Theory of Higher Level Cognition and Its Relation to Visual Attention*. *Human-Computer Interaction*, 12, 439–462.
- Anderson J.R., 1996. *ACT: A simple theory of complex cognition*. *American Psychologist*, 51, 355–365.
- Anderson J.R.; Bothell D.; Byrne M.D.; Douglass S.; Lebiere C.; and Qin Y., 2004. *An Integrated Theory of the Mind*. *Psychological Review*, 111, 136–1060.
- Baum C. and Edwards D.F., 1993. *Cognitive performance in senile dementia of the Alzheimer's type: the Kitchen Task Assessment*. *American Journal of Occupational Therapy*, 47, no. 5, 431–436.
- Cooper R., 2002. *Order and disorder in everyday action: the roles of contention scheduling and supervisory attention*. *Neurocase*, 8, 61–79.
- Cooper R. and Shallice T., 2000. *Contention scheduling and the control of routine activities*. *Cognitive Neuropsychology*, 17, no. 4, 225–232.
- Danker J.F. and Anderson J.R., 2007. *The roles of prefrontal and posterior parietal cortex in algebra problem-solving: A case of using cognitive modeling to inform neuroimaging data*. *Neuroimage*, 35, 1365–1377.
- Lebiere C. W.D. and Taatgen N., 1998. *Implicit and explicit learning in ACT-R*, F. Ritter and R. Young, Nottingham: Nottingham University Press. *Cognitive Modeling II*. 183–193.
- Norman D.A. and Shallice T., 1986. *Attention to action: Willed and automatic control of behavior*. In *Consciousness and self-regulation: Advances in research and theory*. New York: Plenum., vol. 4, 1–18.
- Pavlik P.I. and Anderson J.R., 2005. *Practice and forgetting effects on vocabulary memory: An activation-based model of the spacing effect*. *Cognitive Science*, 29, 559–586.
- Piaget J., 1952. *The origins of intelligence in children*. New York: International Universities Press. (Original work published 1936).
- Salvucci D.D.; Kristen L.; and Macuga L., 2002. *Predicting the effects of cellular-phone dialing on driver performance*. *Cognitive Systems Research*, 3, 95–102.
- Serna A.; Pigot H.; and Rialle V., 2007. *Modeling the progression of Alzheimer's disease for cognitive assistance in smart homes*. *User Modeling and User-Adapted Interaction*, 17, 415–438.
- Serna A.; Pigot H.; and Rialle V., Accepted. *A computational model of activities performance decrease in Alzheimer's disease*. *International Journal of Medical Informatics*.
- Taatgen N.A., 2007. *The Minimal Control Principle*, Gray, W. D., New York: Oxford University Press. *Integrated Models of Cognitive Systems*. 368–379.

# Developing an Ontology Extraction Agent for a Biomedical Learning Social Network

S. Mohammed<sup>1</sup>, J. Fiaidhi<sup>1</sup> and O. Mohammed<sup>2</sup>

Department of Computer Science<sup>1</sup>

Department of Software Engineering<sup>2</sup>

Lakehead University, Thunder Bay, Ontario P7B 5E1, Canada

Emails: {mohammed, jfiaidhi, omohamme}@lakeheadu.ca

## INTRODUCTION

Classical search engines are successful in helping users find information based on keywords, but they are not successful in answering complex questions. Knowledge searches are needed to answer the complex questions, which are so predominant in the life sciences. Knowledge, however, is hardly ever used in traditional search engines. With the new initiatives of Semantic Web and Web 2.0, the new search engines are based on ontologies to provide knowledge and assist in answering complex questions. In practice, the name ontology covers a spectrum of useful artifacts, from formal upper-level ontologies expressed in first order logic (e.g., Basic Formal Ontology (BFO) and DOLCE) to the simple lists of user-defined keywords used, for example, to annotate resources on the Web. The latter are called "folksonomies" and play an important role in the Web 2.0. Folksonomy - also known as collaborative tagging- is a method of collaboratively creating and managing tags to annotate and categorize content. The difference of this knowledge system is that it is not only created and managed by experts but also by users in general. Some keywords are freely chosen to allow personalized annotation. Also linked to folksonomy, we find terms such as taxonomy, or practice and science of classification. The fields are organized in hierarchical structure and it is perfect for relationships such as network structures and folksonomies (e.g., MeSH).

However, considerable difficulty is involved in standardizing the ontologies, metadata formats, and knowledge bases that contain the wealth of information available to the biomedical community. For example, the Gene Ontology (GO) contains more than 120,000 terms and the Digital Imaging and Communication in Medicine (DICOM) contains more than 2,000 terms. Even when adopting ontological standards it does not ensure that everyone uses the same set of attributes. The vocabulary used for the values of attributes can be different making searches across multiple repositories difficult if not impossible. Vocabularies used can be different across users in hospitals, laboratories or healthcare facilities, different in multiple departments within an institution, and sometimes even within a department medical professionals might not use a standardized vocabulary. For example, consider the attribute Body Part in a DICOM image of the

lower half of the human face. One image might contain the word 'Jaws' as a value of the attribute Body Part. Another similar image might contain the word 'Mandible'. Yet another image might contain the SNOWMED id for that body part, such as T-D1217. The situation is further complicated by the fact that 'Mandible' and 'Jaws' are not equivalent – 'Mandible' (and 'Maxilla') are a sub-part of 'Jaws'.

This article proposes one possible solution to the above searching problem by developing an ontology agent extractor responsible for creating a virtual ontology for a given concept and rating the user ontologies by comparing it with the virtual ontology through an argumentation process. This agent allows biologists and medical learners to find the needle at the biomedical learning domain. Since the biomedical domain is quite huge, this article will focus only on the medical imaging subdomain based on the DICOM SR standard.

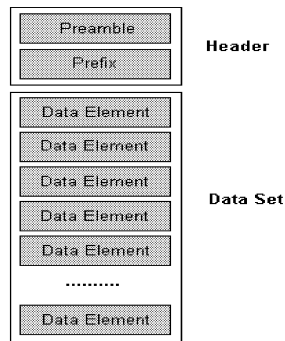
## The DICOM Standard and Biomedical Learning

Triggered by radiology and other medical informatics paradigms, standardized information models and workflows were world-wide defined based on DICOM. Actually, DICOM Structured Record (DICOM SR) is a medical image standard to represent medical images and the associated metadata for variety of biomedical reporting (<http://medical.nema.org/dicom>). This standard has been used for variety of biomedical applications including, patient-records systems, medical diagnoses, and e-learning systems (Fiaidhi, Orabi and Mohammed 2008). Each single DICOM image file contains both a header (which stores information about the patient's name, the type of scan, image dimensions, etc), as well as all of the image data or what is known as the data set. A data set represents an instance of a real world information/learning object. A data set is constructed of Data Elements. Data elements contain the encoded Values of Attributes of that object. A Data Element Tag uniquely identifies a Data Element. There are over 2000 tags used for the DICOM standard. For a full list of DICOM tags refer to 

|  |            |     |      |
|--|------------|-----|------|
| Queens   | University | Web | Site |
| <a href="http://www.sno.phy.queensu.ca/~phil/exiftool/TagNames/DICOM.html">www.sno.phy.queensu.ca/~phil/exiftool/TagNames/DICOM.html</a> |            |     |      |

 However, two types of Data Elements are defined at any DICOM SR report: (1) Standard Data Elements which have an even Group Number that is not (0000,eeee), (0002,eeee), (0004,eeee), or (0006,eeee) and (2) Private Data Elements have an odd Group Number that is not (0001,eeee), (0003,eeee), (0005,eeee), (0007,eeee), or (FFFF,eeee). Figure

1 illustrates the DICOM image file structure as well as an example of expressing it using the standard DICOM tags.



**Fig 1:** The DICOM File Structure.

Although there are many software tools that can be used to extract the metadata from the DICOM SR file (e.g. ExifTool (see [www.sno.phy.queensu.ca/~phil/exiftool/](http://www.sno.phy.queensu.ca/~phil/exiftool/))), there are no effective applications that can utilize such metadata for searching and identifying relevant DICOM objects. Actually, using only metadata is not sufficient to solve the problems of object accessibility (Zarraonandia et. al 2004). Consequently, we see a variety of silo implementations today making it almost impossible to achieve effective classifying, indexing, retrieval and process information about DICOM objects. Indeed, these problems will be solved only if the system and the learners use a common meaning for metadata values. For biomedical learning, solving these problems is of great importance. Making bonds and relations such as content, sequencing, and grouping between objects need to be included to make it possible, not only to carry out automatic tasks on these objects, but also to produce new knowledge from what already exists. The use of ontologies in the model of an e-learning system is an interesting solution. Ontology gathers the concepts, which represent the knowledge of a field in an explicit and formal specification (Studer, Benjamins and Fensel 1998). The use of ontologies and the semantic web for e-learning is well described in (Sampson et al.2004). This is further expanded by Aroya (Aroya and Dicheva 2004), who identify the need to capitalize on the use of (1) semantic conceptualization and ontologies, (2) common standardized communication syntax, and (3) large-scale service-based integration of educational content and functionality provision and usage. This article proposes a system that goes some way to meeting these aims. It also builds on the notion of social learning (Jung and Euzenat 2007) with the intention of "growing context" around the use of learning resources by the community.

## The DICOM Social Learning Network of Agents

Conceiving ontologies, according to Peter Mika (2007), as engineering artifacts allows us to objectify them. Problems arise with this simplistic view, however, as complications may arise. As the original community of agents evolves through members leaving and entering, or their commitments changing, a new consensus may shape up invalidating the knowledge codified in the ontology. To address the problem

of ontology drift, several authors have suggested emergent semantics as a solution (Peter Mika 2007). The expectation is that the individual interactions of a large number of rational agents would lead to communication, transaction, and interaction trends that could be observed as semantics. Ontologies would thus become an emergent effect of the system. While the idea quickly caught on due to the promise of a more scalable and easily maintainable semantic web, the agreement so far only extends to the basic conditions under which emergence would take place. Beyond the reasonable belief that individual actions in such a semantic-social network would lead to ontology emergence, there is a lack of an abstract model of such a system that could also explain the process of emergence. Thus, there appears to be a large conceptual gap in the literature between the vision and the details of implementations of various semantic architectures based on P2P, Grid, MAS and web technology. Hence, Ontology based social network models help in explicating relationships including social entities such as people and organizations, happenings such as events and finally locations. These relationships help in predicting the type of interactions that could occur between agents in a social network. The resulting predictions can help us identify the semantics that may emerge, hence allowing us to predict and better understand the process of emergence.

According to (Jung and Euzenat 2007), a semantic social network is composed of three superposed networks that are assumed to be strongly linked:

- **Social network** relating people on the basis of common interest;
- **Ontology network** relating ontologies on the basis of explicit import relationships or implicit similarity;
- **Concept network** relating concepts on the basis of explicit ontological relationships or implicit similarity.

In this article, we are introducing an ontology extraction agent (OEA) for the ontology network level for a DICOM-Based agent society. The agent society is created using the FIPA JADE middleware (Bellifemine et. al 2007). The JADE middleware provides completely distributed information, resources, controls, etc, on agent services. The JADE environment consists of containers, platforms, AMS, DF, etc. The Agent platform consists of the AMS (Agent Management System), the DF (Directory Facilitator), and the MTS (Message Transport System). In the JADE platform, agents that contain autonomous feature following the application program, user or require of environment and necessary, would appear in the system. The communication between peers through JADE platforms is done by the exchange of ACL messages between agents, in both wired and wireless network environments. The AMS agent provides a naming service. It contains the names and addresses of all existing agents. Agents can register their services with the DF, to enable an agent to search other agents providing a specific service in the DF. The MTS is a component used for supporting communications between agents.

## Designing An Ontology Extraction Agent



Societies need patterned behaviour to exist. Large-scale agent societies may contain a diversity of agents, each with differing abilities, ontology coverage and functionalities. When such an agent system is given a task, it must dynamically muster together a group of agents that collectively have the capability to accomplish the task. To do this, the agent society needs to be able to understand and group its agents. This section contains a proposed superstructure of an ontology extraction agent which is able to acquire the agents' ontology and group them according to a required concept. Hence, in a relatively closed agent society, agents have relative goals, similar actions, concepts and semantics. Therefore, an ontology agent extractor can have that role of classifying agents based on their basic concepts, predicates or actions. The ontology extractor uses a simple correlation method that can be implemented as follows:

(1) Start the Ontology Extraction Agent (OEA) with reference to standard domain ontology (e.g. DICOM Structured Report as fully described in W3C OWL format in [www.cs.man.ac.uk/~alanrw/dicom.owl](http://www.cs.man.ac.uk/~alanrw/dicom.owl)). Provide OEA with goal concept, action or predicate that you want OEA to group agents accordingly. Concepts represent expressions that indicate entities with complex structure that can be defined in terms of slots. Predicates are expressions that say something about the status of the world and can be true or false. Actions represent expressions that indicate actions that can be performed by some agent. Based on the required concept, predicate, action, the OEA generates a virtual ontology (e.g. Ontology for XRay Chest— Figure 2).

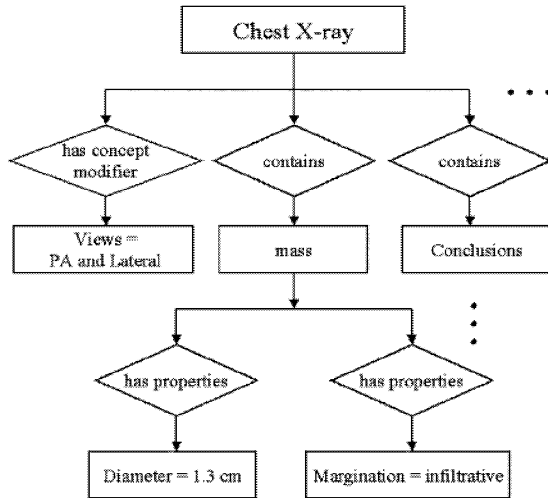


Fig. 2: An Example Schema for the Virtual Ontology for the Chest X-ray Concept as described in [12].

2) Navigate in the multi-agent system M to find first agent V1, which have the context relative to the generated virtual ontology in step 1. The user of the OEA may choose one of the following contexts to specify the required user perspective over the relation between the target agent ontology (V1) and the virtual ontology generated by the OEA:

- **Foundation Context** (connecting one ontology structure to another in terms of total or partial similarity)
  - *Is\_a*
  - *Part\_off*

- **Spatial Context** (connecting one entity to another in terms of relations between the spatial regions they occupy)
  - *Located\_in*
  - *Contained\_in*
  - *Adjacent\_to*
- **Temporal Context** (connecting entities at different times)
  - *transformation\_of*
  - *derives\_from*
  - *preceded\_by*
- **Participation Context** (connecting processes to their bearers)
  - *has\_participant*
  - *has\_agent*

First OEA communicates with the JADE Directory Facilitator (DF) the agent description (DFAgentDescription) which contains the registered users ontologies (refer to the following code). Later the OEA communicates with the target agent using an XML language codec (`jade.content.lang.xml.XMLCodec`) to verify the availability of the required context between the two ontologies (Target vs Virtual). In fact, the OEA communication with the target agent is driven by an argumentation engine built using tuProlog (Aroyo and Dicheva 2004). The argumentation engine utilizes the Prolog inference rules expressed by the tuProlog to activate series of context-matchers. The OEA argumentation engine accepts arguments from the target agent and tries to prove its availability at the virtual ontology according to the user required context. The OEA argumentation engine attempts to build an admissible set to support the user required relation or property via the use of one of the context-matchers. If the context-matcher finds a match, then the supplied arguments (i.e. concepts, predicates or actions) prove to be satisfactory and the name of the target agent is associated to matching context relation. Finding a match is based on the SELA Ontology Alignment through Negotiation Algorithm (Palmisano et. al 2006). A segment code of the *is\_a* context-matcher is given below:

```
import alice.tuprolog.lib
public class ArgumentationEngine {
    public boolean is-A(Ontology O1, Ontology O2) {
        /* if (# of concept schemas in O1 = # of concept schemas in O2 &&
        # of predicate schemas in O1 = # of predicate schemas in O2 && # of
        agent action schemas in O1 = # of agent action schemas in O2) */
        // Compare all schemas in O1 and O2 to see if they match
        boolean conceptSchemaMatch = true;
        boolean predicateSchemaMatch = true;
        boolean agentActionSchemaMatch = true;
        /* Try to prove the first argument on Concept Match */
        for (int i = 0; i < number of concept schemas; i++) {
            if(O1.getConceptSchema(i) != O2.getConceptSchema(i))
                conceptSchemaMatch = false;
        }
        /* Otherwise try to prove the Second argument on Predicate Match */
        for (int i = 0; i < number of predicate schemas; i++) {
            if(O1.getPredicateSchema(i) != O2.getPredicateSchema(i))
                predicateSchemaMatch = false;
        }
        /* Finally try to prove the Third argument on Action Match if the
        previous two fails*/
        for (int i = 0; i < number of concept schemas; i++) {
            if(O1.getSchema(i) != O2.getSchema(i))
                agentActionSchemaMatch = false;
        }
    }
}
```

```

    if (conceptSchemaMatch == true && predicateSchemaMatch == true
    && agentActionSchemaMatch == true) {
        return true;
    }
    else { /* The two agents does not share common ontology */
        return false;
    }
}
public boolean Part_Off(Ontology O1, Ontology O2) { ... }
public boolean Contained_In(Ontology O1, Ontology O2) { ... }
...
}
}

```

Repeat step 2 if other relations are requested..

3) Repeat step 2 on all other agents. If other agents found to have the same match as found in V1, then add them to a set G1 that match the target concept, predicate or action.

4) Repeat the steps 1-3 for any other target concept, predicate or action on agents outside the closure set of other agent groups. Hence, we can get G1, G2, G3...Gn. The relations between these agents can be visualized as an agent society network using the Java Universal Network Graph (JUNG) API (<http://jung.sourceforge.net/>).

Using the context-matchers, the multi-agent system is decomposed into series of agent groups. Figure 3 illustrates the overall architecture of our agent society system based on the use of OEA.

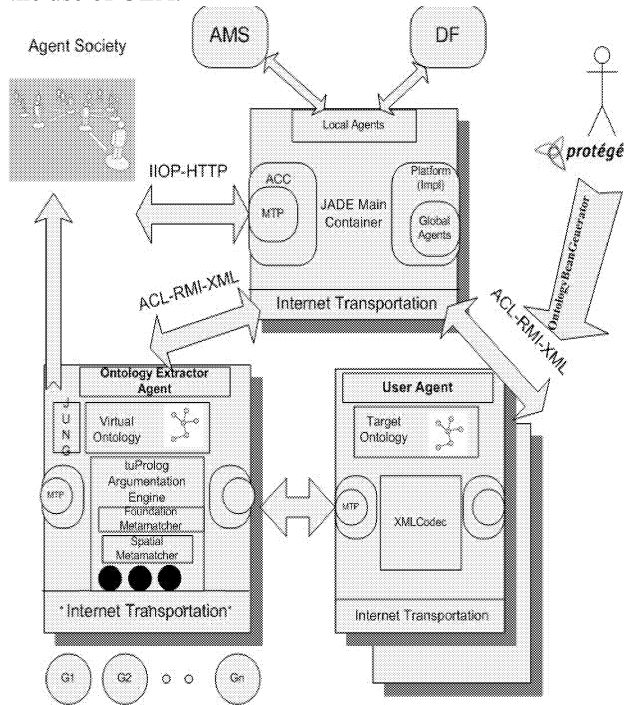


Fig. 3: The Overall Architecture of the JADE OEA System.

Figures 4 and 5 provides more details on the JADE OEA system by showing its major classes and a sequence diagram for interacting with it via the system GUI.

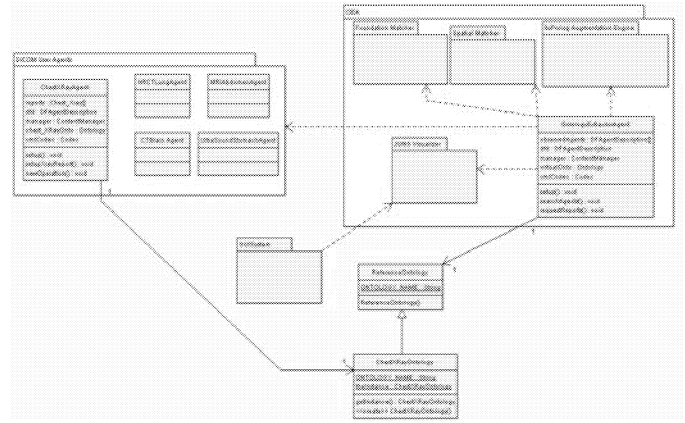


Fig. 4: The JADE OEA System Class Diagram.

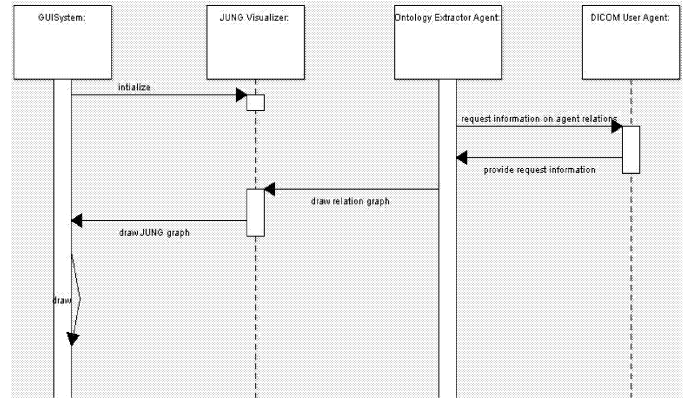


Fig. 5: A Sequence Diagram illustrating interaction with the JADE OEA System.

One more thing we need to mention in order to simplify generating a DICOM user ontology for each agent expressed in JADE format (i.e. Concept, Predicate, Action), the user may use Protégé 2000 editor along with the JADE ontology generator plugs-in (<http://protege.cim3.net/cgi-bin/wiki.pl?OntologyBeanGenerator>). Certainly, the user ontology needs also to conform to the current DICOM SR Version 3 specifications (<http://medical.nema.org/dicom/>).

## CONCLUSION

This article introduces an overview for the design of an ontology extractor agent capable of classifying DICOM user agents into groups of agents related according to certain context-matching relations. The designed OEA is part of an ongoing project for developing a P2P learning framework for sharing and searching of biomedical learning objects (Mohammed et. al. 2008). The architecture of the ontology extractor agent is different from the traditional research of mapping ontologies where the main goal is on finding ontology alignment. In such approaches heuristics are described for identifying corresponding concepts in different ontologies (e.g.comparing the names or the natural language definition of two concepts) and checking the closeness of two concepts in the concept hierarchy (e.g. PROMPT, RDFT). However, the problem is that the structures of all ontological data instances are heterogonous and the use of the ontological alignment approach for such structures proves to unsuccessful

(Chen, Tan and Lambrix, 2006),. Thus a new approach that uses an argumentation process to prove any two ontologies posses certain degree of contextual similarity is becoming more appropriate (Bryant and Krause 2006, Lee 2007) Our approach is based on argumentation employing several context-matchers for identifying complex and sensitive DICOM biomedical relationships between a required concept and target agents ontologies. The context-matchers represent tiny reasoners employing inferential rules written in tuProlog that can be integrated within the JADE multi-agents environment. The findings of these context-matchers are feed to JUNG API for visualizing the agents groups and their contextual relationships.. The contextual relations are carefully selected for DICOM SR classification. Figure 6 shows the JADE GUI running our OEA on a sample of 11-user agents employing variety of DICOM SR learning case studies.

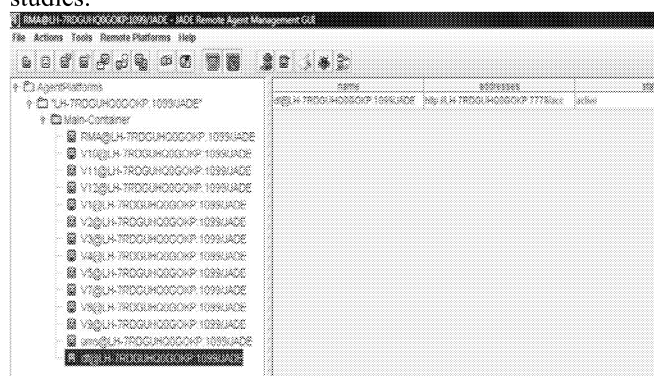


Fig. 6: The JADE OEA System Running a Pool of 11 Agents.

The result of the running our JADE OEA on a pool of 11 target user agents, given a Chest Xray concept, yields four groups (see figure 7): G1 (V1, V2, V3) for the Is\_a relationship, G2 (V4, V5, v7) for the Part\_Off relation and G3 (V10,V11, V12) for the Adjacent\_To relation and G4 ( V8, V9) for the Contained\_In relation.

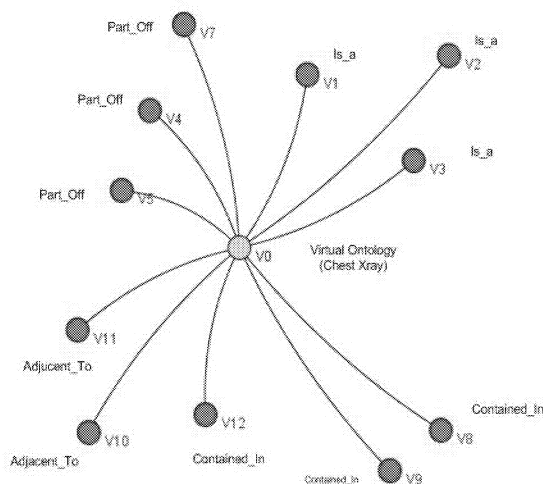


Fig. 7: The JADE OEA JUNG Visualization of the Society of Agents Groupings.

## REFERENCES

Aroyo,L. and Dicheva, D. 2004. The new challenges for e-learning: the

- educational semantic web. Educational Technology & Society, 7(4), 59-69.
- Bellifemine, F. Caire, G. and Greenwood, D. 2007, Developing Multi-Agent Systems with JADE, Wiley Series in Agent Technology, ISBN-13: 978-0-470-05747-6 - John Wiley & Sons.
- Chen, B., Tan, H. and Lambrix, P. 2006, Structure-Based Filtering for Ontology Alignment, 15th IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises, June 26-28 2006, pp364 - 369
- Fiaidhi, J, Orabi, M and Mohammed, S. 2008, Sharing DICOM Learning Objects within a mobile Peer-To-Peer podcasting environment, International Journal of Mobile Communications 2008 - Vol. 6, No.4 pp. 417 – 435
- Hussein,R. 2004, DICOM Structured Reporting: Part 1. Overview and Characteristics, RadioGraphics Journal, 2004;Volume 24:pp891-896.
- Jung,J. and Euzenat,J. 2007, Towards Semantic Social Networks, Lecture Notes in Computer Science , Springer Berlin / Heidelberg, ISSN 0302-9743, Volume 4519/2007, In Semantic Web: Research and Applications
- Lee, S. 2007,Ontology Based Context Alignment for Heterogeneous Context Aware Services, Springer Lecture Notes in Computer Science,volume 4761/2007, pp 40-46.
- Mika,P. 2007, Social Networks and the Semantic Web, Springer US , ISBN 978-0-387-71000-6.
- Mohammed, S. Orabi, A. Fiaidhi,J. and Passi, K. 2008, Developing a Web 2.0 RESTful Cocoon Web Services for Telemedical Education, IEEE International Workshop on Social and Personal Computing for Web-Supported Learning Communities (SPeL 2008), Turku-Finland (July 28- August 1, 2008).
- Palmisano, I. Iannone, L., Redavid, D. and Semeraro, G. 2006, Ontology Alignment Through Instance Negotiation: a Machine Learning Approach. Proceedings of the 3rd Italian Semantic Web Workshop (SWAP-2006), Pisa, Italy, December 18-20, 2006.
- Sampson, D., Lytras,M., Wagner, D. and Diaz, P 2004, Ontologies and the Semantic Web for E-learning. Educational Tech. & Society, 7(4), 26-28.
- Studer, R., Benjamins,V. and Fensel,D. 1998, Knowledge Engineering : Principles and Methods, Data and Knowledge Engineering (DKE), 25(1-2), 161-197.
- Zarraonandia, T., Doderio, J, Díaz, P. and Sarasa, A. 2004, Domain ontologies integration into the learning objects annotation process. Proceedings of the Workshop on Appl. of Semantic Web Technologies for e-Learning, 34-39.

## BIBLIOGRAPHIE

**SABAH MOHAMMED** is a Professor of Computer Science with Lakehead University since 2002. Prof. Mohammed graduated from Glasgow University (MSc in 1981) and Brunel University (PhD in 1986). His research interests include image processing, medical informatics and telemedical infrastructures.

**JINAN FIAIDHI** is a Professor of Computer Science with Lakehead University since 2002. Prof. Fiaidhi graduated from Essex University (PgD in 1983) and Brunel University (PhD in 1986). Her research interests include Multimedia Learning objects, P2P and M-Learning and telemedical education systems.

**OSAMA MOHAMMED** is a final year undergraduate student with the Department of Software Engineering,



# **INDUSTRIAL APPLICATIONS**



# MOVING CONTAINERS IN SMALL TERMINAL AS STRIPS PLANNING PROBLEM – PRELIMINARY RESULTS

Adam Galuszka and Krzysztof Skrzypczyk  
Institute of Automatic Control  
Silesian University of Technology  
Akademicka 16, 44-100 Gliwice,  
Poland  
E-mail: Adam.Galuszka@polsl.pl

## KEYWORDS

Container Terminal, Block World, STRIPS System, Linear Programming, Computational efficiency.

## ABSTRACT

In this paper Block World environment with STRIPS representation is proposed to solve problem of loading and unloading containers in small container terminal in Gliwice, Poland. It is assumed that containers are represented by blocks and reachstackers are represented by robot arms. STRIPS planning with complete information is usually hard problem (very often at least at least NP-complete (e.g. optimal planning in Block World environment is NP-complete). Planning in the presence of incompleteness is much more harder. It causes difficulties in solving real application problems. To increase efficiency of solving such problems a transformation to Linear Programming Problem is proposed. Also some preliminary results are shown.

## INTRODUCTION

Operating in container terminal is a source of many problems: how to load and unload containers, how to organize storage yards, how to plan re-handling operations in dynamic and high uncertain environment. Our intention is to propose efficient in time method for support decision-making processes in small container terminal in Gliwice, Poland ([www.ptkholding.pl/index.php?page=strona&id=56&lang=en](http://www.ptkholding.pl/index.php?page=strona&id=56&lang=en)). In table 1 general technical data of this terminal are presented.

Table 1. General data of the terminal

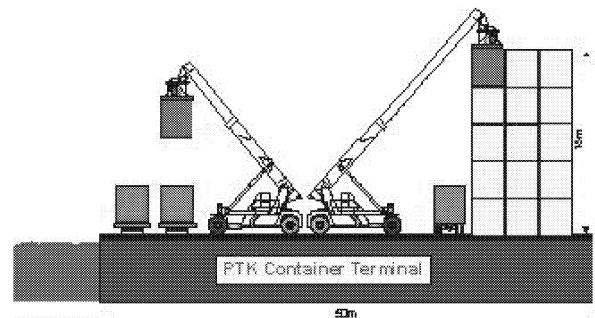
| Data             | Currently             | Project               |
|------------------|-----------------------|-----------------------|
| Area             | 30.000 m <sup>2</sup> | 74.000 m <sup>2</sup> |
| Capacity         | 1.700 TEU             | 3.000 TEU             |
| Track length     | 620m                  | 620m                  |
| Number of tracks | 2                     | 6                     |

The loading-unloading situation is schematically shown in the fig.1 and the reachstacker technical data are presented in table 2. It should be noted that all operations are performed only by reachstackers. There is an operational difference between cranes that are usually considered in literature (see e.g. Kim

and Kim 2002), and reachstackers. In case of cranes, when a driver of an outside truck requests an inbound container that has other containers on top of it, a crane must remove the containers on top of the target container (it is called ‘re-handling’ operation). In case of reachstackers, also all containers between reachstacker and target container must be removed (compare fig.1). It implies that the number of re-handling operations increases. It is important to minimize re-handlings in order to increase performance of the terminal. We propose to model this problem using STRIPS representation.

Table 2. Reachstacker technical data

|                  |  |
|------------------|--|
| Quantity         | currently: 2<br>project: 4                                     |
| Lifting capacity | 1st rank - 45 tons<br>2st rank - 31 tons<br>3st rank - 15 tons |
| Stocking height  | 5 levels   |



Figures 1. Loading-unloading situation

The goal of the project is to efficient (in time) plan the reachstacker moves in the terminal, assuming that the target containers are reached and the number of re-handlings is minimized.

## THEORETICAL BACKGROUND

In the paper problem environment was modelled as Block World with STRIPS representation. This domain is often used to model planning problems (Nilson 1980, Boutilier and Brafman 2001, Kraus et al. 1998, Smith and Weld 1998, Slaney and Thiebaux 2001) because of complex operator interactions and simple physical interpretation. Starting from

1970s STRIPS formalism introduced by Nilson (1980) is popular for planning problems (Weld 1999). Planning problems are PSPACE-complete in general case (see e.g.: Bylander 1994, Baral et al. 2000), even in Block World environment are not easy (here the problem of optimal planning is NP-complete – Gupta and Nau 1993).

The case of Block World problem where the table has a limited capacity corresponds to a container-loading problem (Slavin 1996, Slaney and Thiebaux 2001). In real situation decision problems at container terminals are more complex and divided into several groups: arrival of a ship, unloading and loading of a ship, transport of containers from and on a ship, stacking of containers (see e.g. [www.ikj.nl/container/decisions.html](http://www.ikj.nl/container/decisions.html)). Since arrival of a ship and containers transport are usually treated as scheduling and allocation problems (e.g. Imai 2001; Bish 2001), problems of loading and unloading and container stacking can be treated as planning problems (e.g. Avriel 2001). In natural way containers can be treated as blocks and transfer cranes (in our case reachstackers) as robots that are stacking and unstacking blocks.

## STRIPS REPRESENTATION

In general, STRIPS language is represented by four lists (C; O; I; G) (Bylander 1994, Nilson 1980):

- a finite set of ground atomic formulas (C), called conditions;
- a finite set of operators (O);
- a finite set of predicates that denotes initial state (I);
- a finite set of predicates that denotes goal state (G).

Initial state describes physical configuration of the blocks. Description should be complete i.e. should deal with every true predicate corresponding to the state. Goal state is a conjunction of predicates. Predicates in I and G are ground and function free. In multi-agent environment each agent defines own goal. This description does not need to be complete. The algorithm results in an ordered set of operators (i.e. action sequence) that transforms an initial state into a goal situation. Operators  $O$  in STRIPS representation consist of three sublists: a precondition list (pre), an add list (add) and a delete list (del). Formally an operator  $o \in O$  takes the form  $pre(o) \rightarrow add(o), del(o)$ . The precondition list is a set of predicates that must be satisfied in world-state to perform this operator. The delete list is a set of predicates that become false after executing the operator and the add list is a set that become true. Two last lists show effects of the operator execution in the problem state.

As an example consider the classical Block World operator that removes block  $x$  from block  $y$  (Nilson 1980):  $unstack(x,y)$ . Definition of this operator takes the form:

*precondition list & delete list:*  $handempty, clear(x), on(x,y)$   
*add list:*  $holding(x), clear(y)$ .

To model the operation of moving containers by reachstacker

we assumed following set of conditions:

$C = \{$  *clear*( $x$ ) – there is no container on container  $x$ ,  
*clear\_side*( $x$ ) – there is no container between reachstacker and container  $x$ ,  
*on*( $x,y$ ) – container  $x$  is placed on container  $y$ ,  
*before*( $x,y$ ) – container  $x$  is before container  $y$ ,  
*reachstacker\_empty*( $index$ ) – reachstacker arm of ‘ $index$ ’ reachstacker is empty  
*holding*( $x, index$ ) – reachstacker ‘ $index$ ’ is holding container  $x$  },

where  $x,y,z$  denotes containers and ‘ $index$ ’ is an index of reachstacker.

The operator that removes container  $x$  that is before  $z$  from  $y$  by ‘ $index$ ’ reachstacker can be defined using STRIPS representation as follow:

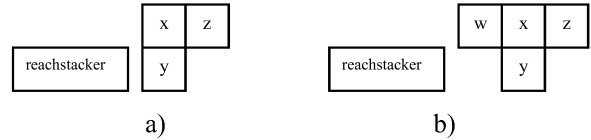
*Ustack*( $x,y,z,index$ ):

*precondition list & delete list:*

*clear*( $x$ ), *clear\_side*( $x$ ), *on*( $x,y$ ), *before*( $x,z$ ),  
*reachstacker\_empty*( $index$ )

*add list:* *clear*( $y$ ), *clear\_side*( $z$ ), *holding*( $x,index$ )

The illustration of this operator is shown in fig.2. In case a) the operator can be applied, in case b) the operator can not be applied, because *clear\_side*( $x$ ) condition is not true.



Figures 2. Illustration of *unstack*( $x,y,z,index$ ) operator

As it was mentioned above planning in Block World environment is a difficult computational problem. Because of complexity, to plan in the terminal efficiently a heuristic of polynomial transformation of STRIPS planning problem to linear programming problem (LP) is used (Galuszka and Swierniak 2004). This is done because LP problems are known to be computational easy (Chaczijan 1979). The cost of this approach is that LP result may be non-interpretable for some initial states of STRIPS planning problem (what is followed by assumption  $P \neq NP$ ).

## STRIPS SYSTEM AS LINEAR PROGRAMMING PROBLEM - PRELIMINARY SIMULATIONS RESULTS

Linear programming problem is defined as:

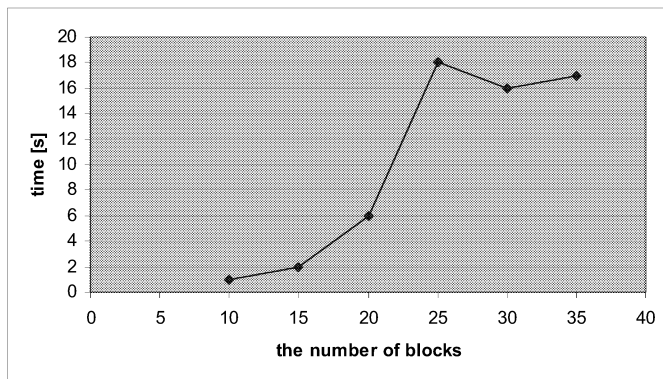
$$\begin{aligned} \text{Max } f^*x \quad \text{subject to: } A^*x &\leq b \\ x \quad \quad \quad Aeq^*x &= beq. \end{aligned}$$

Following (Bylander 1997) the transformation from planning to Linear Programming is based on mapping of conditions and operators in each plan step to variables. Truth values of conditions are mapped to 0 and 1 for the planning without



incompleteness, and to any values between 0 and 1 for planning with incomplete information. The objective function reaches the maximum if the goal situation is true in last step of planning.

Simulations in Block World environment with Strips representation transformed to Linear Program has been earlier (Galuszka and Skrzypczyk 2006, Galuszka 2007). Below (fig.3) we present computational efficiency of planning as Linear Program for problem sizes from 10 to 35 blocks (it corresponds to Linear Program variables number from 1620 to 64015). Test cases have been implemented in MATLAB.



Figures 3. Efficiency of planning

The single train capacity is about 60 containers of 20 TEU (is limited by track length) and it is the maximal size of Block World problem to be solved in the final version of designed system.

## CONCLUSION

In the paper the problem of moving containers by reachstackers is modeled as block world environment with STRIPS representation. Method for solving problem by its translation to Linear Programming problem is proposed. Translation to Linear Programming allows to efficient search for the solution. That is because planning in the presence of incompleteness is usual at least NP-complete problem, Linear Programming is polynomial-time complete problem and translation from STRIPS to Linear Programming is also polynomial (Bylander 1997).

## Acknowledgement

This work is a result of educational and research co-operation between PTK Holding SA Container Terminal, Poland and Institute of Automatic Control of Silesian University of Technology, Poland. This work has been supported by BK funds for the first author and Ministry of Science and Higher Education funds in the years 2008-2010 for the second author.

## REFERENCES

- Avriel, M., Penn, M., Shpirer, N., Witteboon, S. 1998. Stowage planning for container ships to reduce the number of shifts, *Annals of Operations Research* 76, 55--71
- Baral Ch., Kreinovich V., Trejo R. 2000. Computational complexity of planning and approximate planning in the presence of incompleteness. *Artificial Intelligence* 122, 241--267
- Bish, E.K., Leong, T.Y., Li, C.L., Ng, J.W.C., Simchi-Levi, D. 2001. Analysis of a new vehicle scheduling and location problem, *Naval Research Logistics* 48, 363--385
- Bylander T. 1994. The Computational Complexity of Propositional STRIPS Planning. *Artificial Intelligence* 69, 165--204
- Bylander T. 1997. A Linear Programming Heuristic for Optimal Planning. *Conf. American Association for Artificial Intelligence* (1997)
- Chaczijan L.G.: A polynomial algorithm for linear programming. *Dokl. Akad. Nauk SSSR*, 244, 1979, 1093-1096
- Galuszka A., Swierniak A. 2004. Translation STRIPS Planning in Multi-Robot Environment to Linear Programming. *Lecture Notes in Computer Science*, 3070 / 2004, 768--773
- Galuszka A., K. Skrzypczyk. 2006. Block World Planning with Uncertainty and Sensing Actions as Linear Programming Problem. 4th Int. Ind. Simulation Conf., Palermo, Italy, June 5-7, pp. 234-236.
- Galuszka A. 2007. Linear And Integer Programming Large Scale Heuristic For Strips Planning. *European Simulation Multiconference, ESM'2007*, October 22-24, 2007, Malta
- Gupta N., Nau D.S. 1992. On the complexity of Blocks-World Planning. *Artificial Intelligence* 56(2-3), 223--254
- Howe A.E., Dahlgren E. 2002. A Critical Assessment of Benchmark Comparison in Planning. *Journal of Artificial Intelligence Research* 17, 1--33
- Imai, A., Nishimura, E., Papadimitriou, S. 2001. The dynamic berth allocation problem for a container port, *Transportation Research B* 35, 401--417
- Kim K.H and H.B. Kim. 2002. The optimal sizing of the storage space and handling facilities for import containers, *Transportation Research B* 36, 821--835
- Klir G.J., T.A. Folger. 1998. Fuzzy sets, uncertainty, and information, Prentice-Hall International, Inc. USA
- Koehler J., Hoffmann J. 2000. On Reasonable and Forced Goal Orderings and their Use in an Agenda-Driven Planning Algorithm. *Journal of Artificial Intelligence Research* 12, 339--386
- Kraus S., Sycara K., Evenchik A. 1998. Reaching agreements through argumentation: a logical model and implementation. *Artificial Intelligence* 104, 1--69
- Nilson N.J. 1980. Principles of Artificial Intelligence. Toga Publishing Company, Palo Alto
- Slaney J., Thiebaux S. 2001. Block World revisited. *Artificial Intelligence* 125, 119--153
- Slavin T. 1996. Virtual port of call. *New Scientist* (June 1996) 40--43
- Swierniak A., A. Galuszka. 1999. Betweenness and indistinguishability in modelling and control of uncertain systems, *Proc. of 18th IASTED Conference Modelling, Identification and Control*, (Innsbruck 1999), 56--58
- Weld D.S., Anderson C.R., Smith D.E. 1998. Extending Graphplan to Handle Uncertainty and Sensing Actions. *Proc. 15th National Conf. on AI*, 897--904
- Weld D.S. 1999. Recent Advantages in AI Planning. *Technical Report UW-CSE-98-10-01*, AI Magazine

# INTRODUCTION TO COMPARISON OF TRADITIONAL AND VIRTUAL PATTERNS DESIGN IN 3D

Agnieszka Cichocka \*\*\*

Pascal Bruniaux\*

\*Laboratoire GEnie et Matériaux TEXtiles

(GEMTEX), UPRES EA2161

École Nationale des Arts et Industries Textiles (ENSAIT)

9 rue de l'Ermitage, BP 30329, 59056 ROUBAIX Cedex 01, France

Faculty of Textile Engineering and Marketing

Department of Clothing Science and Technology

Ul. Żwirki 36, 90-942 Łódź, Poland

[agnieszka.cichocka@ensait.fr](mailto:agnieszka.cichocka@ensait.fr), [pascal.bruniaux@ensait.fr](mailto:pascal.bruniaux@ensait.fr),

## KEYWORDS

Garment pattern process generation, modelling of virtual garment, adaptive model

## ABSTRACT

This paper presents a garment pattern generation process and modelling of virtual garment design method in 3D. Additionally this work characterizes our global project on virtual clothing design which contains the conception of virtual adaptive mannequin, and also of the creation and modelling of garment in 3D. According to the ideas of mass customization and e-commerce, as well as the need of numerical innovations in garment industry we employ a model described by [Cichocka et al 2007] of virtual garment and methodology enabling to conceive the virtual clothing directly on an adaptive mannequin morphotype in 3D. Moreover this work is a methodology conception study for development of garment in 3D and 2D suited for the industry. In our project we want to create the method which will give us possibility to create a perfect garment, and which will take into account all peculiarities of human body. The opportunities to ponder upon technical parameters such as ease handling make it a perfect study example. Patterns can be formed by either a 2D or 3D process. Often a combination of methods is used to create the patterns. Furthermore, problem of asymmetry of human body is taken into account. In 2D method we made two patterns: left and right side separately to obtain best result and to get possibility to compare two cases: when symmetric pattern is used and when we use pattern consisted of two parts specially made for each side of the body. In the present context we use the example of a basic women's shirt: bodice and sleeve. We used the method of making pattern employed in Russia. There are several differences between Russian and French method what is interesting as well. In 2D technology we obtained measurements by scanning human body. Finally the superposition of virtual and real pattern was done in order to visualise the right results.

## INTRODUCTION

It is very important to know information about human body outer shape description, size and measurements for the design clothes and manufacturers. There are many problems

in the garments production that are related to the features for a customer to get a garment maximally conformable. Currently the most important challenge of the clothing industry is improvement of the basic stage of designing process.

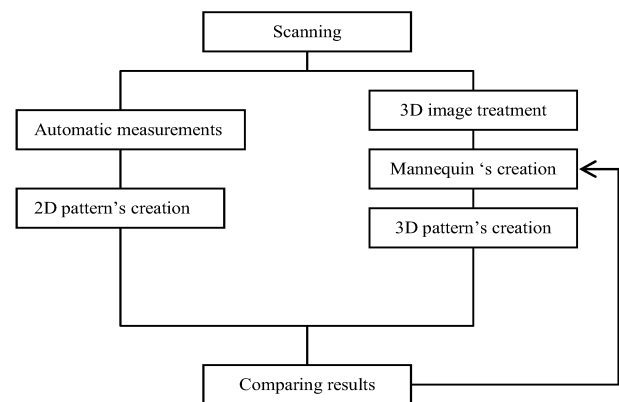


Figure 1. Scheme of proposed approach

Figure 1 presents the scheme of our approach of pattern development and garment simulation which lies in the domain of mass customization and personalization.

In this stage the model appearance and its patterns elaboration is achieved. It's very important to create good quality pattern and finally good fit garment in the end in order to be able to compete successfully on the world market.

Nowadays, textile manufacturers use 2D design of patterns. By this traditional method the pattern can be created approximately. Nevertheless a new technology, technology of future - 3D modelling of pattern of garments is coming. Creating the style of garments for a concrete person - customer in a 3D environment following the layout of the details of the style to 2D, would allow combining the advantages of mass production and tailoring. Moreover it can be possible creation of individual garment with good fitting, excluding the process of trying on. This technology also allows us to solve one problem more, like problem of asymmetric body. To implement such technology we need digital description of the human body shape. Anthropometrical points are used for human body description. The anthropometrical coordinates are detected and mathematical simulation of the bearing surface is

acquired. In 2D technology we obtained measurements by scanning human body. Today the world is looking for the new methods of garments designing which could promise the creation the most appropriate garments for each customer using the computing technique.

Until now some approximate surface layout methods are being used in the cloth designing praxis, which are known as pattern systems. The most progressive of those are based on the knowledge about the human body. Nevertheless, since a plane draft of a non-existent dimensional model is being created, the reflexive transformation does not deliver the expected results without additional corrections, fitting and correction of patterns.

## GOOD FIT IN MASS CUSTOMIZATION

Just about ten years ago approximately half of all consumers are unable to find ready-made garments that fit them properly. 70-80% of garments did not correspond to the reported size, forcing consumers to contend with a wide variation in fit and mail order houses to experience a 30% return rate due to poor fit.

Inappropriate garment sizing has always caused problems for manufacturers, retailers and customers. Often it results in financial loss. The result is costly inefficient in production and inventory planning, markdowns, lost market shares and and high volumes returned.

All individuals are different, having their own proper morphology explained in a form of measurements and proportions. If they all followed a normal distribution, with fixed mean and standard deviation, statistical models used for processing anthropometric data and converting it into size measurement charts would be right on target. Many current sizing systems are based on statistical models derived from outdated or inappropriate anthropometrical data. The result is far from perfect. Nowadays the scanner technology gives us possibility to obtain good result and in consequence to develop the garment production.

In recent years, the industry has been moving toward a highly demanding environment, based on quick response. This requires decreasing the time gap between determining body measurements and delivering a garment that fits them properly. The process of mass customization should reduce cost in terms of manufacturing labor and overhead, eliminating the need to stock infrequently requested sizes. Customer satisfaction should increase, thanks to better fitting garment, particularly for people outside the normal size range. In today's apparel market, consumers desire to personalize the style, fit, colour of the clothes.

Companies are being forced to react to the growing demand individualization. At the same time, cost cutting remains of principal importance due to the competitive pressure in global markets. Thus, making enterprises more customers centric efficiently is a top management priority in most industries. Mass customization and personalization are key strategies to meet this challenge. Companies like Procter Gamble, Lego, Nike, Adidas, Land's End, Brooks Brothers, or Levi Strauss, among others, have started large-scale mass customization programs. Nowadays we can use of the new technologies, including the body scanner showed in Figure 2.

In the apparel industry, several new technologies have helped mass customization operations which we also use in our work:

1. Body scanner (for collecting body measurements);
2. Computer-aided design (CAD) systems for pattern making.

Custom patterns can be generated in four basic ways:

- Traditional pattern making and grading processes can be expanded by creating multiple sets of patterns that will fit a variety of different body proportions. For example, patterns can be generated for every possible combination of waist and hip measurement.
- Traditionally graded patterns can be used in another way by selecting the closest-fitting pattern and applying automated alterations to custom-fit the pattern.
- Traditional pattern drafting techniques can be used to automatically generate a pattern directly from a set of body measurements.
- New software programs are being developed that actually "un wrap" a 3D representation of a garment to make a 2D pattern shape.

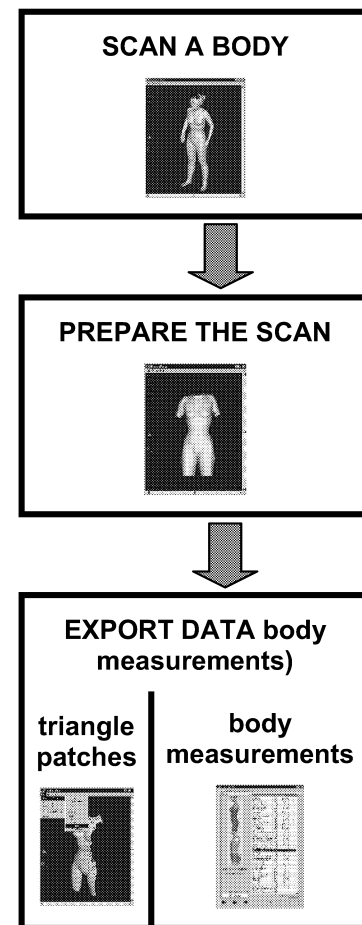


Figure 2. 3D body scanning technology for fashion and apparel industry

## 2D PATTERN'S MAKING METHOD

Nowadays it is one of the most popular methods in the world and widely used in industry. The most common 2D pattern making methods are flat, drafting and reverse engineering. In the flat method, a pattern is generated from an existing foundation pattern called a slope or block. A slope is a pattern that has no seam allowances or style lines. From a slope a numerous garment styles can be generated. The patternmaker creates a new style by adding design details such as a collar, pocket and pleats. The flat pattern making method is widely used in the ready-to-wear market because it is fast and accurate.

In the drafting method, patterns are made directly from measurements taken from a pre-existing garment, an individual or a body form. Using the collected measurements, the pattern is drawn directly using software. There are a lot of methodologies of drawing patterns. It is also interesting to test different methods, to compare them. We carry out small research of comparing French and Russian techniques.

### Comparing methods

Basically these two methods are quite similar. However there are some interesting differences between them. Common structure is identical, but methodology of geometrical construction is different. We can see on the Figure 3 that we have several main peculiarities.

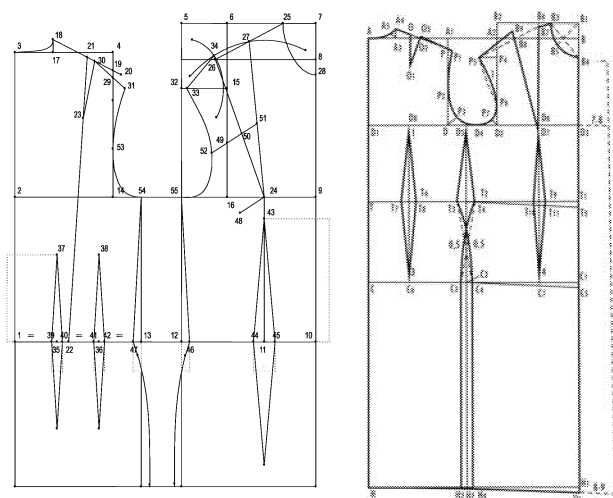


Figure 3. French (left) and Russian (right) methods

In French method the first stage of 2D pattern design is to define the framework describing the front and back of the garment. In the first phase we use the measures of maximum height (front bodice length) and width (chest girth) of the human body. Secondary, it is necessary to place the lines of construction related to the measurements reported on the human body in the following order: contour chest related to the waist line, the side lines and the width back line correlated with the middle front and back, spacing breasts line. The analysis of this method presents that the shoulder slopes positioning is defined by lines or points positioned at 25, 26 for the front and 18, 20 for the back. These values are defined only for the basic pattern. Another flaw is related to

positioning the width back points located at half of distance between the points 15 and 16 for the front, and a third of the distance between the points 4 and 14 on the back. Another criticism is the definition of neckline points set proportionally in relation to total dimension of neckline (1/6). However the creation of the front dart is a properly defined by the point 34 representing the intersection of the large circle giving a proper closure of the dart and small sized circle representing the half-shoulder value. It noted that for the back, this twice condition is not met because, the dart does not close properly. Moreover, for the front dart, the point 33 positioning the acromion is valid only for the basic pattern. During the construction of the width back point (52), the method imagine shutting dart as the 50 point is superimposed on 51 and 52 out of 49, this method could have been imagined with a technique similar to the front dart. Finally, positioning of waist darts seems correct.

Russian method presents the dress pattern construction which we modified in order to make a shirt. In this technique we also draw the framework with the dimensions describing all parts of garment (fronts and backs). During the positioning of construction lines we take into account the measurement of human body and the value of ease attributed to the considered lines of garment (for example the point of top centre back in relationship with waist line TT1 and hip line CC1). These imposed ease values are constant and independent of variation human body measurement. In point of view mass customization process and evolution of human body dimensions this represents a weak point of introduced methods. Moreover this method doesn't use position of chest line, only chest point is determined (D7). Comparing to the French method, there is not the same location for armhole depth line and the chest line. Waist line straight in French method, it is sloping on the front in Russian method (T5T4) which take into account the shape of abdominal. Furthermore the methodology of the top back dart location in Russian mode seems correctly putted in point of view of sew operation than in French method. Finally we decided to use the Russian method to make 2D pattern.

### Asymmetry of human body

In common, the construction of garment pattern is executed for only one half of human body. It is mean that we consider that our body is perfectly symmetric and that we reflect pattern just making in order to have 2 parts covering whole body. Unfortunately a perfect body doesn't exist in real word. Every human body is asymmetric. One hand is longer than another, right shoulder is lower than left one. Thus, it's make the process of creating garment good fitting more complicated. We have two possibilities: to make symmetric pattern for both different half of human body or adapting 2D pattern for each side. In our case of customisation garment project we are chosen the last one. Simultaneously the symmetric pattern making method is used with appropriate modification imposed. In this case we use the measurements of the most advanced right side of the body. The human body represents the entire physical structure of a human organism. The human body consists of a head, neck, torso, two arms and two legs. One of the most important things at the beginning of pattern making process is to take measurements correctly because the quality of the future

garment strongly depends on it. Our studied body also shows that several measurements are not symmetrical for the left and right side of the body Figure 4 concerning left and right arms, shoulders and chests. There are not huge variations but to customize the pattern it could be important. This is one of the reasons why we make patterns for the right and left side of the body and sleeves separately.

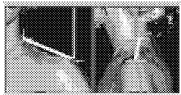

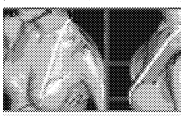

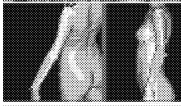

|   |                                  |                    |
|---|----------------------------------|--------------------|
|  | 3030<br>Shoulder width left      | 13.1 cm (5.17 in)  |
|  | 3031<br>Shoulder width right     | 12.3 cm (4.84 in)  |
|  | 4080<br>Bust point to neck left  | 25.5 cm (10.04 in) |
|  | 4081<br>Bust point to neck right | 25.0 cm (9.85 in)  |
|  | 8030<br>Arm length left          | 56.4 cm (22.20 in) |
|  | 8031<br>Arm length right         | 58.8 cm (23.14 in) |

Figure 4. The examples of differences in parts of studied human body

The differences between right and left pattern are shown on Figure 5. The black one presents the right pattern and the red one presents the left pattern. The width of the right shoulder equal 12,3cm differs significantly from the left one when they are superposed. Also the distance from chest point to neck left is 25,5cm and the right equal 25 cm influence on the variation between two patterns.

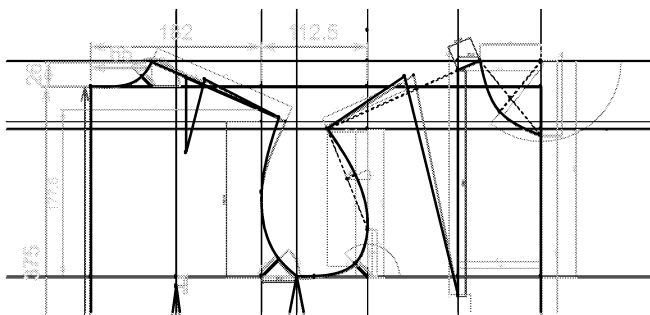


Figure 5. Asymmetry between right and left bodice

Sleeves construction shows also the dissimilarity between the right and left part when the length of right arm is 58,8 cm and the length of left arm is 56,4 cm which we can see in the Figure 6.

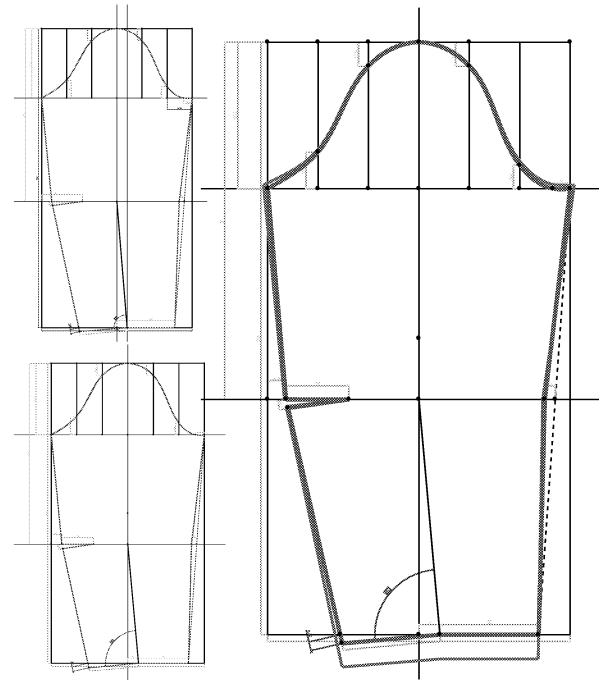


Figure 6. Asymmetry between right and left sleeve

### 3D PATTERN'S MAKING METHOD

One of the main advantages of 3D pattern making process is the ability to generate patterns that provide excellent fit. No doubt that it is important to use CAD systems in the garment designing process. These systems give the opportunity to suddenly react on model, size, assortment, order and other changes faster and more precise and the preparation of garment production is faster too. The current computer systems have wide designing possibilities and automatic garment preparation functions. Nevertheless the currently used CAD/CAM systems have some weak points.

In the industrial computer aided design systems the garment design is implemented in a 2D environment. The plane-like designed garment details are matched dimensionally to check the compatibility of contours and the volume of the garment comparing to the body.

Designing systems is to create 3D solutions for garment creation, 2D pattern pieces creation implementing 3D systems and forces. These tendencies are important for all CAD systems. As we can see Figure 7 describes one conclusion for all of them - to elaborate 3D systems able to make the garment directly on the human body.

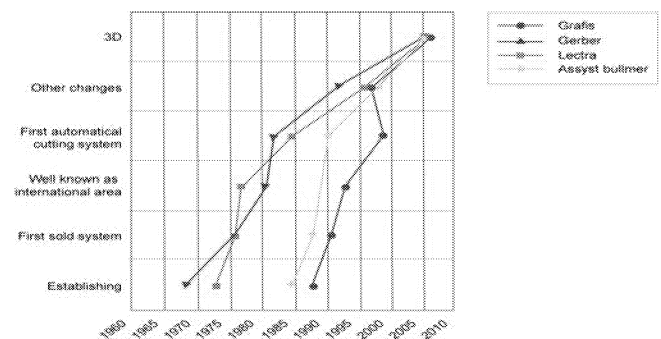


Figure 7. Evolution of different CAD systems

## Designing the garment in 3D

To start design the garment in 3D the model of human body is required. The first step consists of acquisition and image treatment of the data directly issued from a body scan presented in the points cloud. In order to adapt the human data to 3D, a cloud of points must be interpreted to obtain interpolation of human data and understood by CAD environment at once [Boudjemaï 2006]. In our case the RAPIDFORM software was used to transform the cloud of points by triangulations process to final superficial form of virtual mannequin. The used mannequin model is controlled by a variety of parameters which are directly related to the curves controlling the mannequin surface was presented in our previous work [Cichocka and all 2006]. More detailed study about the methodology of conception an adaptive mannequin morphotype presents [Cichocka 2008].

After the conception of virtual adaptive human body succeeded we start the next stage. The previous phase existing under the mannequin's contours form will make the support for proposed control point's arrangement (ease model). We must highlight that the ease model describes the essential of the virtual garment design process. Furthermore it depends and remains in a close relationship with the garment model. We establish our process in different parts:

1. Defining the contours on the body in order to place the garment control points on them Figure 8.

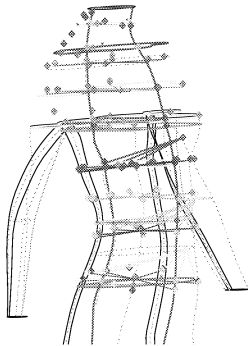


Figure 8. Control point's arrangement on the shirt contours

2. Defining the ease points that are projected in space by translating the points on the contours defined previously.
3. Joining the ease points to form new contours in space Figure 9.

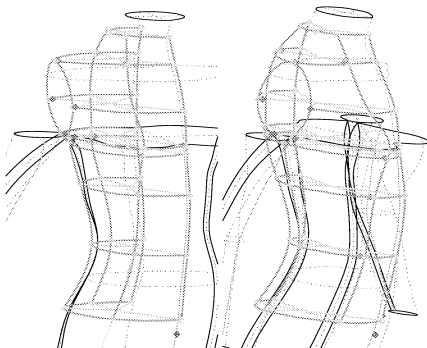


Figure 9. Conception of contours shirt

4. To connect the contours with another one to form outlines of the surface to be created.
5. To form a surface between the connected contours.
6. Flatten the surface in order to obtain a new flat surface - the patterns.

Management of this control point's arrangement known as ease model was completed with the significant numbers of ease parameters. These parameters are placed on those human body contours, selected to fit to the shirt model. The repartition of controlling points was also important in point of view the opportunity of future administration of ease model. We have also a possibility to change the value of the points influencing on the shape, we can change the distance between body and garment. The human body characterizes very complex set of shapes. Proposed model of ease consists of different number of points placed on the contours, so it was essential to locate more points on the curvatures places on the human body than on the places which are "flats" and straights.

Accordingly to classification of garment (garment destined for the lower limbs, upper limbs and bodice), the appropriate contours of human body must be chosen. It means that the established segments determine the different lines issued from the human morphology and also they have to match up to the analysed kind of garment. Moreover the garment conception imposes very strong constraints in choice of contours because in the fact the garments present a combination of different contours which contain at least two proposed segments.

In order to have a correct course of virtual garment design, it was necessary to throw additional contours. Besides the addition of these curves and the controlling points, it was necessary to put them in the places which are essential to model of basic shirt. In our case the supplementary contours are situated between the waist line and bottom line and chest line. The proposed model of garment for the basic lady shirt model is presented on the Figure 10.

Respecting the procedure of pattern making, firstly one front and one back were created only. After that the procedure was completed to cover second part of human body.

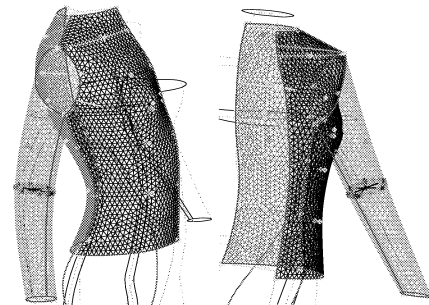


Figure 10. Model of virtual garment

## SIMULATION RESULTS

Generally, in garment industry the beginning of garment production means patterns making process which is done in two stages: firstly in 2D usually using the CAD software, and the second, by another person called pattern maker. This

process is very laborious and arduous and needs the designer's competence and experience which influence strongly at the prototyping and pattern design process. There is no need to mention that in fact it's an intuitive process. For that reason, it is very difficult to simulate the designer's work. The differences appearing during the sketch reading process can be seen often between designer and pattern maker. Proposed methodology 3D could aid in solving this misunderstanding an attitude for direct pattern making process during garment design course. We obtained the 2D pattern converted from 3D. We can notice some visual differences between the 2D and 3D pattern, especially distinguished on the pattern sides which are non straight Figure 11. The difference appears because our 3D pattern followed the curves of the body, so the darts are getting automatically. This means that the drafts take into account the seams side. In traditional method the drafts are imposed inside the pattern.

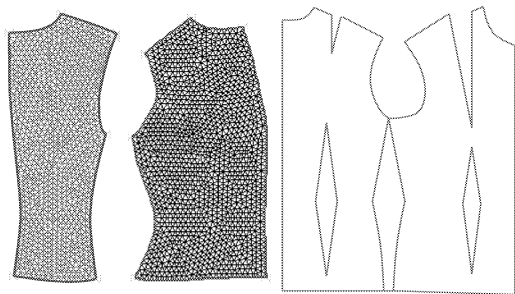


Figure 11. Comparison virtual and traditional pattern of shirt for the right bodice

The all pieces of shirt will be available at the end of ease identification process. Next, using interface enabling the conversion from 3D to 2D, the post-treatment is executed. This process in general consists of the triangulation of resultant surfaces. We must to define triangulation parameters appropriate to the model so the further analysis can be continued. During this action, an analysis of surface deformation in warp and weft directions is carrying out, as showed in bottom on Figure 12.

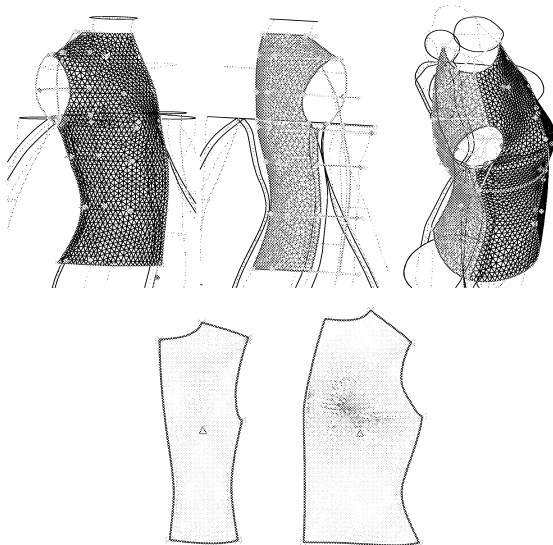


Figure 12. Post-treatment of shirt surface - front and back of shirts

The different colors illustrate the stretched and compressed zones. Basing on this analysis, we are able to predict if the draft is required and where is the best position to put it. On the chest zone we can see that the colour varied from blue to red which means the necessity of dart.

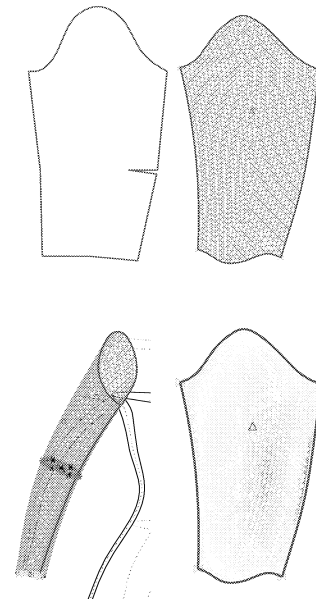


Figure 13. Comparison virtual and traditional pattern of right sleeve and post-treatment of sleeve

Figure 13 compare two studied pattern making methods with an example of sleeve. Method 2D standardizes the form of armpit, against the 3D method which personalize the look of this curve. Moreover adjusted sleeve follows the morphology of arms. Additional benefits of 3D sleeve construction are simplicity and speedy process of sleeve creating. Bottom line present a line double curvature which allows separate the interior and exterior sleeve seams automatically. The post-treatment justifies the presence of the dart at the elbow level that can be found in the pattern created with 2D method.

Also it was very important to take into account that in general, during the pattern making process the seam values are contained in overall pattern surface, in order to enabling the assembling of garment pieces. While our virtual 3D shirt supported on mannequin, represents a pattern surface without the seam values. Our pieces of shirt are assembled by putting their edges together, that is indicated by the outside and inside seam lines. It means that we do not consider the influence of seam presence in the garment and it is except in this simulation.

## CONCLUSION

This article presents a new approach of garment design process modelling. Proposed 3D method of garment making provides the effect of the relative ease centred (barycentre), - distance related to ease centred points arrangement. Whereas 2D methods of pattern making process offer ease with constant values that take into account the fabric and the size of person.

The advantage of this approach is that it is preformed with the real data of person that garment is customized for. This

process can be called also virtual tailoring. Another benefit is that 3D technology gives automatically the basic pattern. Moreover this method offers the designing power i.e. possibility to changing the positions of curves for different garment models and styles. Furthermore pattern control is carried out and integrated directly into the method, for example the shoulder line is unique in 3D. Whilst in 2D method we have 2 control lines (front, back). The interest of our proposed methodology of garment design is that it could be applicable in almost any software working in 3D. Moreover during the garment design process we are able to visualise the results in the form of 2D patterns immediately. However, in garment conception, the sleeve of the basic shirt follows the morphology of arm by adding the supplementary contours at elbow and wrist level giving an adjusted sleeve. To accomplish the process of garment design and give more realistic image transfer the stage of fabric simulation is required in order to verify the correct drape of shirt and finally validate the overall process [Cichocka 2008]. The main objective for the future will be to have the opportunity the exchange models of human body while maintaining an association with the garment. So the basic pattern must be dependent on different morphotypes. Following this idea the necessity to have a different basic pattern for the same model of garment which stays in relationship with each morphotype appears. In addition a fabric simulator integrated in the same software to accomplish the virtual try-on comfortably is necessary.

## REFERENCES

- Boudjemaï F. "Reconstruction de surfaces d'objets 3D a partir de nuages de points par réseaux de neurones 3D-SOM", thèse de USTL Lille, 2006
- Cichocka A., Bruniaux P. & Koncar V. "Modelling of Virtual Garment Design in 3D", Research Journal of Textile and Apparel, Vol. 11 No. 4, pp. 55-63, 2007
- Cichocka A., Bruniaux P. "Adaptive model of the human body – methodology of design of the mannequin morphotype" Industrial Simulation Conference 2006, Italy
- Cichocka A. PhD Thesis "Contribution à la modélisation et à la simulation de vêtements sur mannequin adaptatif", July 2008, France

**AGNIESZKA CICHOCKA** is preparing her PhD (in July 2008) at ENSAIT Roubaix in France and at Technical University of Lodz in Poland since 2004. She is doing her research in the field of modeling and simulation of garment and mannequins for a 3D design process. She participates to industrial projects with Lectra CAD company specialized in garment industry.

**PASCAL BRUNIAUX** was born in Denain, France in 1959 and went to the University of Lille, where he studied Automatic control and obtained his PhD in 1988. He worked at HEI engineering institute and in 1990 he came at ENSAIT, the textile engineering institute as Associate Professor. He is teaching in the field of mass customization and textile design process using virtual reality tools. He is doing his research in the field of fabric drape modeling and new approach to garment design using 3D CAD tools.



# **AUTHOR LISTING**



## AUTHOR LISTING

|                      |             |                     |          |
|----------------------|-------------|---------------------|----------|
| Acharya A.....       | 37          | Labelle S. ....     | 47       |
| Banerjee A. ....     | 37          | Liu Y. ....         | 7        |
| Beldjehem M. ....    | 23/37/63/81 | Mélange T. ....     | 15       |
| Belkacem C. ....     | 75          | Mohammed O.....     | 99       |
| Bruniaux P.....      | 110         | Mohammed S. ....    | 99       |
| Camilleri L. ....    | 52          | Mousavi S.M.....    | 31       |
| Cichocka A. ....     | 110         | Nachtegaele M. .... | 15       |
| Fiaidhi J. ....      | 99          | Pigot H.....        | 94       |
| Fujarewicz K.....    | 89          | Robillard P.N. .... | 47/57    |
| Galuszka A. ....     | 89/107      | Schumann J. ....    | 7        |
| Grogono P. ....      | 9           | Skrzypczyk K.....   | 89/107   |
| Groussard P.-Y. .... | 94          | Soleymani A. ....   | 31       |
| Gupta P. ....        | 7           | Sussner P.....      | 15       |
| Hardy S. ....        | 57          | Yao J. ....         | 5        |
| Hélène P.....        | 75          | Yousefi H.....      | 31       |
| Jafari A. ....       | 31          | Zadeh L.A. ....     | 13/46/73 |
| Kerre E.E.....       | 15          |                     |          |
| Konar A. ....        | 37          |                     |          |