
# ADVCOMP 2012

The Sixth International Conference on Advanced Engineering Computing and Applications in Sciences

ISBN: 978-1-61208-237-0

September 23-28, 2012

Barcelona, Spain

**ADVCOMP 2012 Editors**

Sigeru Omatu, Osaka Institute of Technology, Japan

Toan Nguyen, INRIA, France

# ADVCOMP 2012

# Forward

The Sixth International Conference on Advanced Engineering Computing and Applications in Sciences (ADVCOMP 2012) held on September 23-28, 2012 in Barcelona, Spain, was a multi-track event covering a large spectrum of topics related to advanced engineering computing and applications in sciences.

With the advent of high performance computing environments, virtualization, distributed and parallel computing, as well as the increasing memory, storage and computational power, processing particularly complex scientific applications and voluminous data is more affordable. With the current computing software, hardware and distributed platforms effective use of advanced computing techniques is more achievable.

The goal of ADVCOMP 2012 was to bring together researchers from the academia and practitioners from the industry in order to address fundamentals of advanced scientific computing and specific mechanisms and algorithms for particular sciences. The conference provided a forum where researchers were able to present recent research results and new research problems and directions related to them. The conference sought contributions presenting novel research in all aspects of new scientific methods for computing and hybrid methods for computing optimization, as well as advanced algorithms and computational procedures, software and hardware solutions dealing with specific domains of science.

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

This event could also not have been a reality without the support of many individuals, organizations and sponsors. We also gratefully thank the members of the ADVCOMP 2012 organizing committee for their help in handling the logistics and for their work that is making this professional meeting a success. We gratefully appreciate to the technical program committee co-chairs that contributed to identify the appropriate groups to submit contributions.

We hope the ADVCOMP 2012 was a successful international forum for the exchange of ideas and results between academia and industry and to promote further progress in advanced scientific computing.

We hope Barcelona provided a pleasant environment during the conference and everyone saved some time for exploring this beautiful city.

**ADVCOMP 2012 Chairs:**

**ADVCOMP Advisory Chairs**
Chih-Cheng Hung, Southern Polytechnic State University, USA
Juha Röning, Oulu University, Finland
Sigeru Omatu, Osaka Institute of Technology, Japan
Erich Schweighofer, University of Vienna, Austria

**ADVCOMP Research/Industry Chair**
Jorge Ejarque Artigas, Barcelona Supercomputing Center (BSC-CNS), Spain
Helmut Reiser, Leibniz Supercomputing Centre (LRZ)-Garching, Germany

# ADVCOMP 2012

## Committee

### ADVCOMP Advisory Chairs

Chih-Cheng Hung, Southern Polytechnic State University, USA
Juha Röning, Oulu University, Finland
Sigeru Omatu, Osaka Institute of Technology, Japan
Erich Schweighofer, University of Vienna, Austria

### ADVCOMP 2012 Research/Industry Chair

Jorge Ejarque Artigas, Barcelona Supercomputing Center (BSC-CNS), Spain
Helmut Reiser, Leibniz Supercomputing Centre (LRZ)-Garching, Germany

### ADVCOMP 2012 Technical Program Committee

Witold Abramowicz, University of Economics - Poznań, Poland
Sónia Maria Almeida da Luz, Polytechnic Institute of Leiria, Portugal / University of Extremadura, Spain
Vincenzo Ambriola, Università di Pisa, Italy
Renato Amorim, University of London- Birkbeck, UK
Gabriel Amorós, Universitat de València, Spain
Sulieman Bani-Ahmad, Al-Balqa Applied University, Jordan
Roberto Beraldi, "La Sapienza" University of Rome, Italy
Simona Bernardi, Centro Universitario de la Defensa / Academia General Militar - Zaragoza, Spain
Ateet Bhalla, NRI Institute of Information Science and Technology - Bhopal, India
Muhammad Naufal bin Mansor, University Malaysia Perlis, Malaysia
Pierre Borne, Ecole Centrale de Lille - Villeneuve d'Ascq, France
Kenneth P. Camilleri, University of Malta - Msida, Malta
Juan-Vicente Capella-Hernández, Universitat Politècnica de València, Spain
Antonio Casimiro Costa, University of Lisbon, Portugal
Marisa da Silva Maximiano, Escola Superior de Tecnologia e Gestão - Instituto Politécnico de Leiria, Portugal
Vieri del Bianco, Università dell'Insubria, Italy
Javier Diaz, Indiana University, USA
Jorge Ejarque Artigas, Barcelona Supercomputing Center (BSC-CNS), Spain
Sameh Elnikety, Microsoft Research, USA
Simon G. Fabri, University of Malta - Msida, Malta
Umar Farooq, Amazon.com - Seattle, USA
Mehdi Farshbaf-Sahih-Sorkhabi, Azad University - Tehran / Fanavaran co., Tehran, Iran
Dmitry Fedosov, Forschungszentrum Juelich GmbH, Germany
Mohammad-Reza Feizi-Derakhshi, University of Tabriz, Iran
Dan Feldman, MIT, USA
Bin Fu, University of Texas - Pan American, USA

**Copyright Information**

For your reference, this is the text governing the copyright release for material published by IARIA.

The copyright release is a transfer of publication rights, which allows IARIA and its partners to drive the dissemination of the published material. This allows IARIA to give articles increased visibility via distribution, inclusion in libraries, and arrangements for submission to indexes.

I, the undersigned, declare that the article is original, and that I represent the authors of this article in the copyright release matters. If this work has been done as work-for-hire, I have obtained all necessary clearances to execute a copyright release. I hereby irrevocably transfer exclusive copyright for this material to IARIA. I give IARIA permission or reproduce the work in any media format such as, but not limited to, print, digital, or electronic. I give IARIA permission to distribute the materials without restriction to any institutions or individuals. I give IARIA permission to submit the work for inclusion in article repositories as IARIA sees fit.

I, the undersigned, declare that to the best of my knowledge, the article is does not contain libelous or otherwise unlawful contents or invading the right of privacy or infringing on a proprietary right.

Following the copyright release, any circulated version of the article must bear the copyright notice and any header and footer information that IARIA applies to the published article.

IARIA grants royalty-free permission to the authors to disseminate the work, under the above provisions, for any academic, commercial, or industrial use. IARIA grants royalty-free permission to any individuals or institutions to make the article available electronically, online, or in print.

IARIA acknowledges that rights to any algorithm, process, procedure, apparatus, or articles of manufacture remain with the authors and their employers.

I, the undersigned, understand that IARIA will not be liable, in contract, tort (including, without limitation, negligence), pre-contract or other representations (other than fraudulent misrepresentations) or otherwise in connection with the publication of my work.

Exception to the above is made for work-for-hire performed while employed by the government. In that case, copyright to the material remains with the said government. The rightful owners (authors and government entity) grant unlimited and unrestricted permission to IARIA, IARIA's contractors, and IARIA's partners to further distribute the work.

**Table of Contents**

# Power Quality Level Measurement and Estimation Issues

Vatau Doru
Electrical Power Engineering Department
"Politehnica" University of Timisoara
Timisoara, Romania
d_vatau@yahoo.fr

Surianu Flavius Dan, Olariu Adrian Flavius
Electrical Power Engineering Department
"Politehnica" University of Timisoara
Timisoara, Romania
flavius.surianu@et.upt.ro, adrian.olariu@et.upt.ro

*Abstract*—**The paper presents a few power quality monitoring aspects on the interface between the power transmission network (TNE) and the power distribution network (DNE), 110 kV voltage level, considering both the current situation and the perspective. Monitoring the power quality indicators, performed via dedicated portable or fixed analyzers, facilitates compliance to standard limits and create the database required for the completion and correction of the standards. Implementing the monitoring system within all the Transelectrica's substations, several advantages are achieved. The efficiency is increasing; the decisions are able to be taken more accurately at all the involved levels.**

***Keywords-power quality; monitoring system; real power; reactive power; quality indicators; power market.***

## I. INTRODUCTION

Power quality is a complex and controversial issue, the complexity resulting from the multitude of factors that condition it, and the controversy from how different researchers understand and present it differently. In 1985, the European Commission Directive 85/374 EC [1] established that electricity is a 'product', that requires clear definition of features. The voltage and frequency maintenance within the admissible limits and a pure sinusoidal voltage curve, without 'noise', are the key elements for an ideal electricity supply.

The paper presents a few power quality monitoring aspects on the interface between the power transmission network and the power distribution network, at 110 kV voltage level, considering both the current situation and the perspective. Permanently or temporarily power quality monitoring is performed in the common connection point, where the system operator / provider has an obligation to provide electricity within the quality parameters of the contract. The supplier / consumer is required to limit the system perturbations transmitted in the Romanian Power System below the quota. Knowing the situation within the power transmission network buses and the disturbance sources, it is required a complex measurement program, using acquisition and processing equipment dedicated to the private Power Transmission Network - Power Distribution Network interface.

The problems specific to the common connection point result in the under-load of the power transmission network and the power distribution network. Also, difficulties in maintaining the voltage within the admissible range in the power transmission network are recorded. The incidents that occur within the power transmission network cause large variations in voltage, gaps, short and long term power supply failures, leading to disturbances for the power supplied to consumers [2]. Power quality within the power distribution network is affected both by the voltage going out of the admissible range, and by the distortion of the voltage and power curves. In the power distribution network, power quality monitoring involves tracking it in the network buses and in the user common connection point, as well as establishing, for each connected user, the disturbance generated level.

The paper presents the structure of a power quality monitoring system, some experimental results and advantages obtained by using power quality monitoring system.

## II. POWER QUALITY WITHIN THE ROMANIAN POWER GRID COMPANY TRANSELECTRICA

Within the Romanian Power Grid Company Transelectrica the power transmission activity is based on the transmission network. It includes substations and overhead lines. The transmission network represents a national and strategic interest, having the rated phase-to-phase voltage greater than 110 kV. The power facilities managed and operated by Transelectrica (the Romanian Transmission System Operator) are the following ones:

79 substations:

- 1 x 750 kV substation;
- 36 x 400 kV substations;
- 42 x 220 kV substations;

8931.6 km OHLs:

- 154.6 km - 750 kV;
- 4703.7 km - 400 kV;
- 4035.2 km - 220 kV;
- 38 km - 110 kV (tine-lines with the power systems of the neighbouring countries);

and 218 transforming units totalizing 37565 MVA.

To assure the fulfilment of the power quality requirements, the Company is undergoing a rigorous maintenance program to maintain the transmission network power facilities technical status. Its main goal is the following one: to improve the transmission network safety operation to avoid situations that could lead to dangerous unwanted events both for the transmission network and for the population or environment.

The implementation of a substation remote control and monitoring system represents a priority for the Romanian

 ISBN: 978-1-61208-237-0

Power Grid Transelectrica. The improvement of the power transmission service efficiency and quality, the reduction of the dangerous events and of the operation and maintenance costs are direct consequences of such system.

To achieve this goal, remote control and monitoring centre have been established at each transmission branch. Currently the following ones are operating: The Remote Control and Monitoring Center of Timisoara and The Remote Control and Monitoring Center of Sibiu (Figure 1).

Applying the monitoring system at Timisoara and Sibiu Subsidiaries, based on the obtained results, its expansion to the entire country level can be achieved. The results provided by the monitoring system are able to sustain several decisions regarding the upgrading all the electrical installations at Timisoara and Sibiu Subsidiaries. Also, an entire imagine is provided concerning the electrical installations' upgrading for all the subsidiaries within Transelectrica.

Timisoara Transmission Subsidiary is operating across four counties: Timis, Arad, Caras-Severin and Hunedoara.

Timisoara Transmission Subsidiary contains:

- 400 kV substations: Arad, Mintia, Nadab;
- 220 kV substations: Arad, Baru Mare, Calea Aradului, Hasdat, Iaz, Mintia, Otelarie, Paroseni, Pestis, Resita, Sacalaz, Timisoara.

Sibiu Transmission Subsidiary is operating across the following six counties: Alba, Sibiu, Brasov, Mureş, Harghita and Covasna.

Sibiu Transmission Subsidiary contains:

- Brasov, Darste and Iernut 400 kV substation;
- Alba Iulia, Fantanele, Gheorgheni, Ungheni, Iernut 220 kV substations.

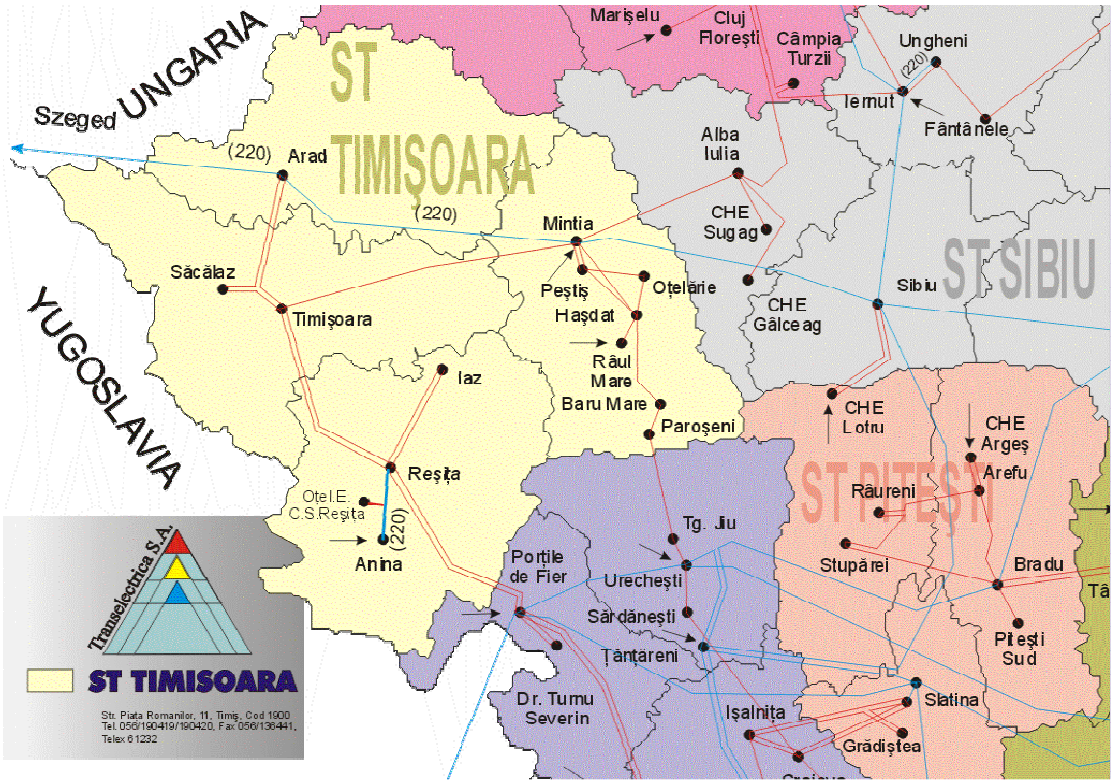



Figure 1. Timisoara and Sibiu's Remote Control and Monitoring Center

## III. POWER QUALITY MONITORING SYSTEM

The system is composed of multiple devices and software presented synthetically in Table I. The "Analyzer CEE" equipment is in according to [2] - [5] standards.

The PSTN modem is "U.S. Robotics Courier 56 K Business type for dial-up telephone line switchable. Selected communication speed is 19200 baud /s. This type of modem keeps its settings in case of accidental interruption of the supply voltage.

The data server is an HP personal computer Intel P4, 3 GHz. Due to large volume of data recorded for processing into various statistical forms, capacity is 1024 MB RAM, 120 GB SATA HDD. Auxiliary power is provided by UPS. Data Server has an LCD monitor 19", a multi color print A4. Fujitsu Siemens notebook communication with the server system is achieved by connecting the external modem described above, to the analogue telephone circuit.

Microsoft Windows NT operating system is used. On request, the data transmitted by the portable computer are automatically saved in a dedicated database. The system allows the external archiving of transmitted data on DVD RW and also their security.

 ISBN: 978-1-61208-237-0 

The entire database stored on the database server can be accessed, on demand, in order to generate one's own programs to process the primary data and to print the data, the graphs or the reports. Implementing the system requires carrying out works on two levels: in each measuring cell of the substations and in the central point.

TABLE I. EQUIPMENT USED FOR MEASUREMENTS

| No. | Equipment | Manufacturer | Type |
|---|---|---|---|
| 1 | Analyzer CEE | Power Measurement Canada | 7650 ION™ |
| 2 | Modem PSTN | US Robotics | Courier 56K Bussines |
| 3 | Server de date | Hewlet Packard | Procesor Pentium 4 |
| 4 | Software license | Power Measurement Canada | ION Enterprise 5.5 |

Figure 2 shows the system architecture in a simplified version that includes only one on-field location and the central point. As seen in the figure, the analyzers have been installed in fixed assembly, on the metering closet pertaining to the monitored cell. The database server and the dedicated application have been installed in the central point

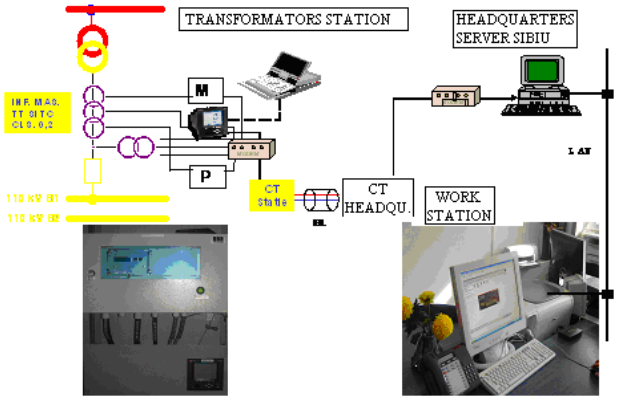



Figure 2. Implementing the system

## IV. EXPERIMENTAL RESULTS

Several experimental determinations have been performed within the following locations:

- Timisoara Transmission Subsidiary (400 kV substations: Arad, Mintia, Nadab; 220 kV substations: Arad, Baru Mare, Calea Aradului, Hasdat, Iaz, Mintia, Otelarie, Paroseni, Pestis, Resita, Sacalaz, Timisoara);
- Sibiu Transmission Subsidiary (400 kV substations: Brasov, Darste and Iernut; 220 kV substations: Alba Iulia, Fantanele, Gheorgheni, Ungheni, Iernut).

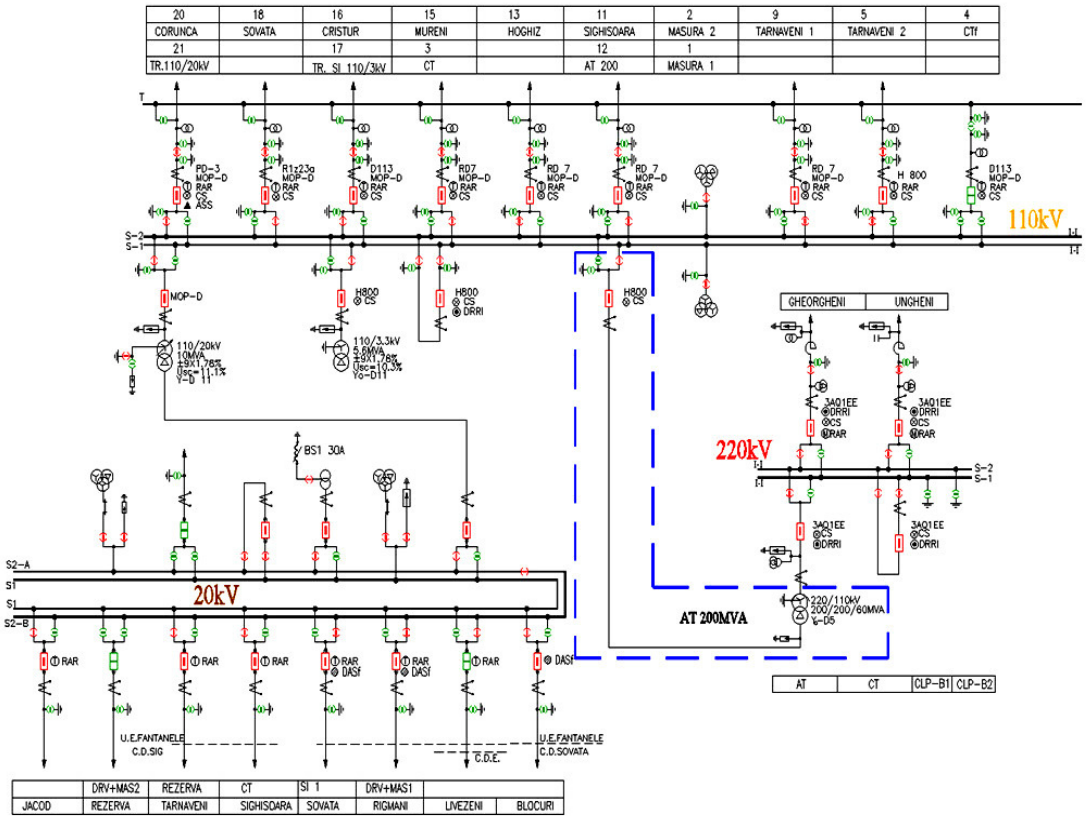


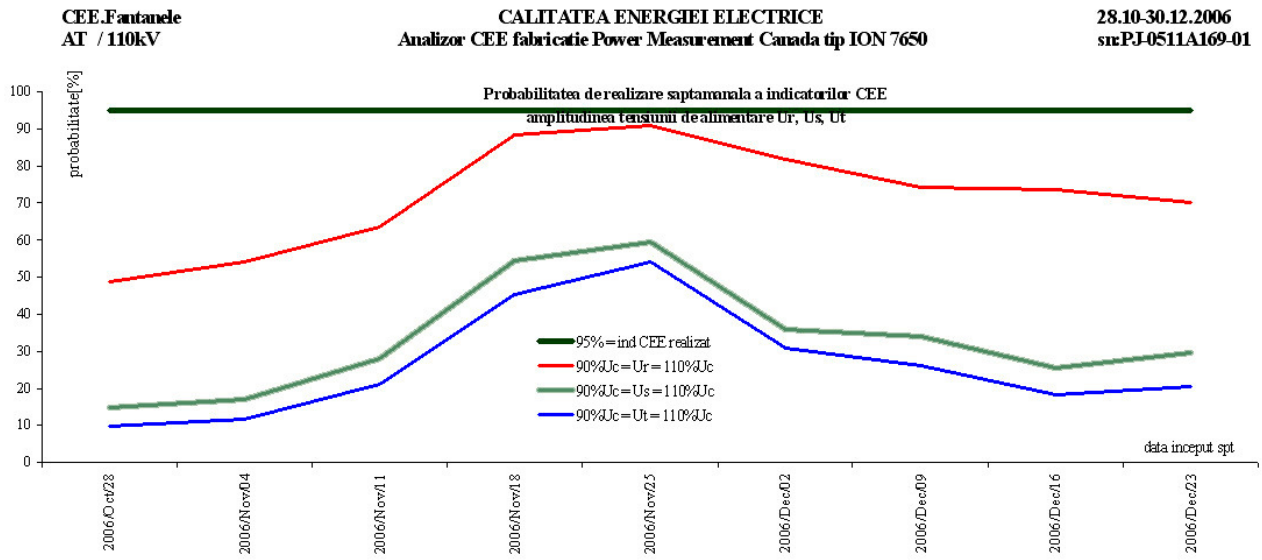

Figure 3. Fantanele substation one-line operating scheme

Part of the experimental determinations has been previously presented in other references such as [6]-[9].

In the following, only the experimental results corresponding to the 220 / 110 / 20 kV Fantanele Substation are presented (Figure 3).

Over 50 % of the 220 kV and 400 kV total overhead lines (OHL) length (987 km), within Sibiu Subsidiary, are located on mountain areas, with difficult access. These OHLs are operating in special environment conditions, high requests from the maintenance and operation point of view. The power produced within the major power plants is transmitted through these OHLs: Lotru, Mintia and Iernut. Within Fantanele substation, additionally, even it has been partially upgraded, power quality indices unsatisfied values have been recorded. The negative aspects refer to the supplied voltage magnitude: $U_R$, $U_S$, $U_T$.

Within Fantanele Substation the measurements have been performed between 28.10-30.12.2006. The very short measurement period is due to several difficulties in establishing the communication between the measurement point and the central one. The greatest real energy quantity received from 110 kV network is noted (according to Table II) being an important characteristic for 110 kV autotransformer.

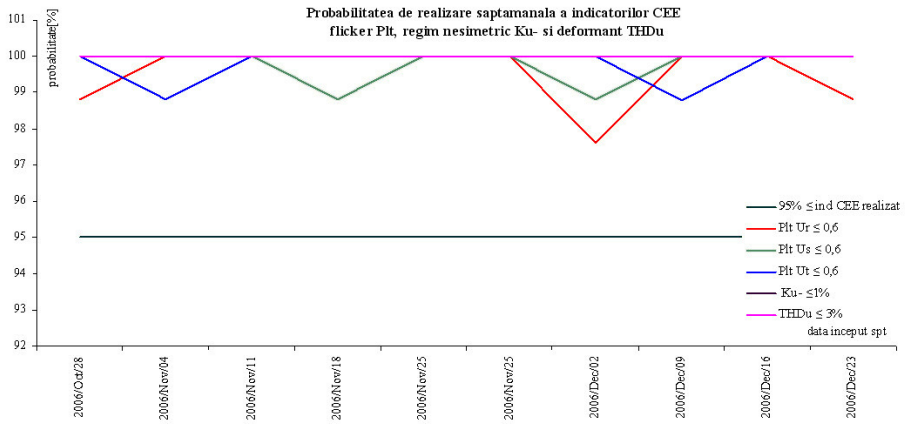



a)

| Nr crt | Amplitude of temporary overvoltages | Number of temporary overvoltages by amplitude and duration | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $U_R$ | | | $U_S$ | | | $U_T$ | | |
| | | $\Delta t < 1s$ | $1s \le \Delta t < 1min$ | $1min \le \Delta t$ | $\Delta t < 1s$ | $1s \le \Delta t < 1min$ | $1min \le \Delta t$ | $\Delta t < 1s$ | $1s \le \Delta t < 1 min$ | $1 min \le \Delta t$ |
| 1 | $110\%Uc < U < 120\%Uc$ | 12 | 3 | 175 | 8 | 7 | 162 | 14 | 11 | 106 |
| 2 | $120\%Uc \le U < 140\%Uc$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | $140\%Uc \le U < 160\%Uc$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Nr crt | Amplitude of voltage gaps | Number of temporary voltage gaps by amplitude and duration | | | | | | | | |
| | | $U_R$ | | | $U_S$ | | | $U_T$ | | |
| | | $10ms \le \Delta t < 100ms$ | $100ms \le \Delta t < 500ms$ | $500ms \le \Delta t < 1ms$ | $10ms \le \Delta t < 100ms$ | $100ms \le \Delta t < 500ms$ | $500ms \le \Delta t < 1ms$ | $10ms \le \Delta t < 100ms$ | $100ms \le \Delta t < 500ms$ | $500ms \le \Delta t < 1ms$ |
| 1 | $10\%Uc < \Delta U < 15\%Uc$ | 2 | 2 | 0 | 2 | 0 | 0 | 1 | 0 | 0 |
| 2 | $15\%Uc \le \Delta U < 30\%Uc$ | 1 | 2 | 0 | 6 | 2 | 0 | 0 | 1 | 0 |
| 3 | $30\%Uc \le \Delta U < 60\%Uc$ | 0 | 0 | 0 | 6 | 0 | 0 | 2 | 1 | 0 |
| 4 | $60\%Uc \le \Delta U < 99\%Uc$ | 1 | 0 | 0 | 12 | 14 | 4 | 8 | 3 | 0 |
| Nr crt | Measurement interval (on-of) week | Number of short and long voltage breaks by duration | | | | | | | | |
| | | $U_R$ | | | $U_S$ | | | $U_T$ | | |
| | | $\Delta t < 1s$ | $1s \le \Delta t < 3min$ | $3min \le \Delta t$ | $\Delta t < 1s$ | $1s \le \Delta t < 3min$ | $3min \le \Delta t$ | $\Delta t < 1s$ | $1s \le \Delta t < 3min$ | $3min \le \Delta t$ |
| 1 | TOTAL | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

b)

 ISBN: 978-1-61208-237-0 




c)

Figure 4. Synthesis of the PQ indicators of the Fantanele Substation (a, b and c)

TABLE II. AUTOTRANSFORMER 110 KV FANTANELE SUBSTATION OPERATING CONDITIONS

| Network element | Transformer tap ratio | Rated power | Tap | Current transformer tap ratio | Voltage transformer tap ratio |
|---|---|---|---|---|---|
| Autotransformer corresponding to 2006 year operating conditions | 220 / 110 kV | 200 MVA | 12 | 1200 / 5A | 110000 / 100V |
| | Out of service | Average loading level | Real energy transmitted to 110 kV network | Real energy received from 110 kV network | Power factor |
| | 02.06.06<br>11.07.06<br>16.08.06<br>16-20.10.06 | 9.33 % | 162.802 GWh | 11.7 GWh | 88.86 |

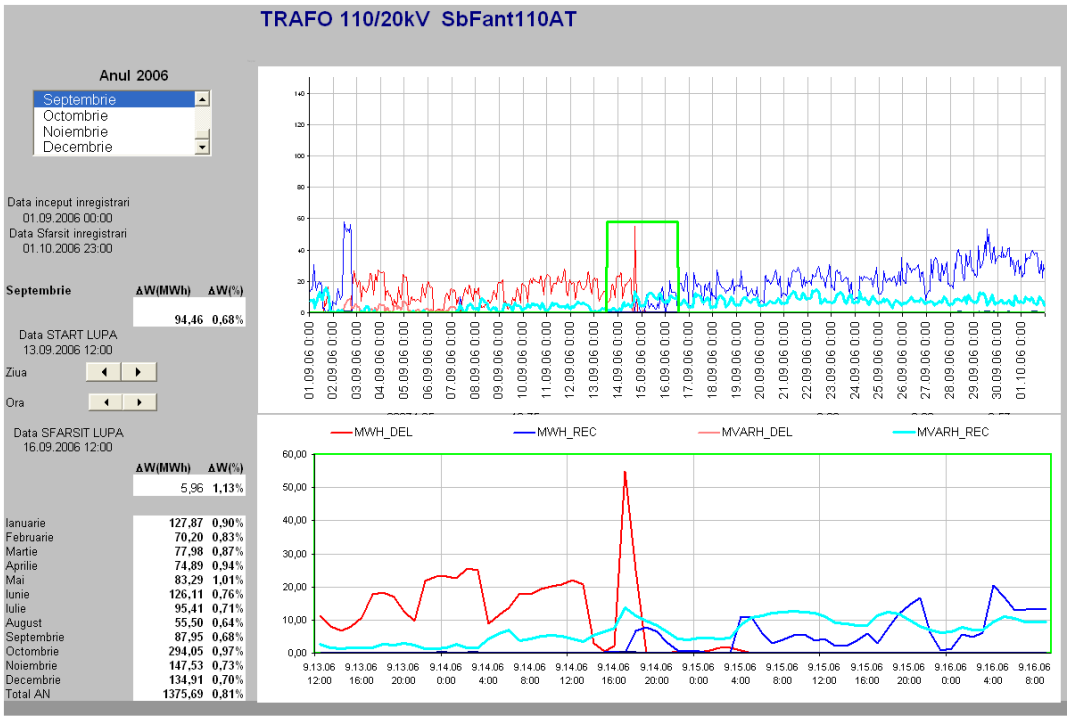



Figure 5. Autotransformer 110 kV Fantanele Substation load curves

The change of the real energy sense is followed by a significant reactive energy quantity. In specific cases the latter could overpass the real one (Figure 5). The 110 kV autotransformer power factor is 88.66 %, being the smallest in comparison to other analyzed network elements. The power quality indicator synthesis is presented in Figure 4. According to this figure the phase L1, L2, L3 supplied voltage magnitude and the long period flicker level Plt have not fitted between the admissible limits.

## V. CONCLUSIONS

The supplied voltage magnitude power quality indicator is the most frequent overpass of the imposed limits for all the measuring points, due to its increased value over 110 % $U_C$. The revision of the Electrical Transmission Network Technical Code has been proposed considering the technical equipment characteristics installed within 110 kV electrical installations and the experience achieved. The upper limit voltage magnitude variation from 121 kV to 123 kV has been proposed.

The feasibility study conducted for this system stresses that its implementation will ensure rapid access to information needed for all the responsible factors. It will therefore increase efficiency with regards to the establishment of concrete measures to reduce electromagnetic disturbances and to diminish their effects in order to:

- Reduce the additional losses in the power transmission network and for the consumers supplied directly from the power transmission network. It is achieved mainly by reducing the harmonics' level, voltage and power non-symmetries for this type of networks;
- Ensure the proper equipment operation with functions and performance affected by the harmonics and voltage and / or current non-symmetries presence;
- Reduce operation expenses for the preventive or corrective equipment maintenance that is affected by disturbances that damage power quality;
- Increase the life span of the power transmission network equipment and consumers supplied directly from the power transmission network, mainly by reducing the level of temporary over-voltages and the harmonic power and voltage on the network;
- Increase the generators, processing units, lines and electric motors efficiency;
- Reduce the costs of power generation / transmission and, in general, reduce investment in the Romanian Power System that would be necessary in order to cover the electromagnetic disturbances' effects caused by exceeding the admissible limits;
- Reduce the reactive power flow and reduce the reactive power exchanges between the power transmission network and the power distribution network;
- Reduce damage to consumers caused by voltage deviations from the rated value, by voltage gaps and by short term power supply failures.

The power quality monitoring system presented provides better results than the ones stipulated within the literature [10]-[13].

Implementing the monitoring system within all the Transelectrica's substations, several advantages are achieved. The efficiency is increasing; the decisions are able to be taken more accurately at all the involved levels.

Only limiting the advantages' evaluation at transmission network additional losses reduction, that affects the power quality, the following economic efficiency indices can be evaluated.

- Considering a 10 years period, the net updated revenue is NUR = 628,489 Euro. It presents the analyzed investment efficiency, for a considered study period and an update chosen rate. The limit condition the accept the investment is: $NUR > 0$.
- Profitability index Pi = 2,612. It represents the ratio between the sum of the yearly updated benefits and the sum of the yearly updated expenses, along the considered study period. To accept the investment this ratio has to be greater than 1.
- Internal return rate IRR= 45,14 %. It verifies if the investment is sensitive at greater updating rates, than the one chosen within the computing process.
- Updated recovery UR = approximately 2.5 years from the first operation of the investment. It highlights the capacity of the objective to refund the invested capital for its achievement. It is refunded based on the operation benefices, respectively from the number of years necessary to equal the investment value.

## ACKNOWLEDGMENT

This work was partially supported by the strategic grant POSDRU/88/1.5/S/50783, Project ID50783 (2009), co-financed by the European Social Fund – Investing in People, within the Sectoral Operational Programme Human Resources Development 2007-2013.

## REFERENCES

[1] 85/374/ EC: 1985, "The European Commission Directive".

[2] EN 50160: 1999, "Voltage characteristics of electricity supplied by public distribution systems".

[3] CEI 61000-4-7: 2000, "Electromagnetic compatibility (EMC) – Part 4 - 7: Testing and measurement technique - General guide on harmonics and interharmonics measurements and instrumentation, for power supply systems".

[4] CEI 61000-4-15: 2003, "Electromagnetic compatibility (EMC) - Part 4 - 15: Testing and measurement technique - Flicker meter - Functional and design specifications".

[5] IEEE 1159: 1995, "Recommended Practice on Monitoring Power Quality".

[6] D. Vatau and F.D. Surianu, "Monitoring of the Power Quality on the Wholesale Power Market in Romania", Proc. of the 9th WSEAS International Conference on Electric Power Systems, High Voltages, Electric Machines (POWER'09), Genova, Italy, October 17-19, 2009, pp.59-64.

 ISBN: 978-1-61208-237-0 

[7] D. Vatau, F.D. Surianu, A.E. Bianu and A.F. Olariu, "Considerations on the Electromagnetic Pollution Produced by High Voltage Power Plants", Proc. of European Computing Conference (ECC'11), Paris, France, April 28-30, 2011, pp. 164-170.

[8] P. Ehegartner, S. Jude, P. Andea, D. Vatau and F.M. Frigura Iliasa, "A Model Concerning the High Voltage Systems Impact on the Environment inside a Romanian Power Substation", Proc. of the 11th WSEAS International Conference on Automatic Control, Modelling & Simulation (ACMOS'09), Istanbul, Turkey, May 30 - June 1, 2009, pp. 413-418.

[9] D. Vatau, P. Andea, F.D. Surianu, F.M. Frigura Iliasa, S. Kilyeni and C. Barbulescu, "Overvoltage protection systems for low voltage and domestic electric consumers", Proc. of the 15th IEEE Mediterranean Electrotechnical Conference (MELECON 2010), Malta, Cyprus, April 25-28, 2010, pp. 1394-1397.

[10] T.L. Tan, S. Chen, and S.S. Choi, "An overview of power quality state estimation", Proc. of the 7th International IEEE Power Engineering Conference, (IPEC'05), Singapore, November 29 – December 2, 2005, pp. 271-276.

[11] R. Lima, D. Quiroga, C. Reineri, and F. Magnago, "Hardware and software architecture for power quality analysis", Computers & Electrical Engineering, vol. 34 (6), November 2008, pp. 520-530.

[12] M. Adam, A. Baraboi, C. Pancu, and A. Plesca, "Reliability centered maintenance of the circuit breakers", International Review of Electrical Engineering (IREE), vol. 5, no. 3, May/June 2010, pp. 1218-1224.

[13] C. Pancu, A. Baraboi, M. Adam and T. Plesca, "GSM Based Solution for Monitoring and Diagnostic of Electrical Equipment", Proc. of the 13th WSEAS International Conference on Circuits, Rodos, Greece, July 22-24, 2009, pp. 58-63.

 ISBN: 978-1-61208-237-0 

# Virtual Reality Technology Applied in Maintaining Interior Walls Painted Buildings

Alcínia Z. Sampaio, Daniel P. Rosário
Dep. Civil Engineering and Architecture
Technical University of Lisbon
Lisbon, Portugal
e-mail: zita@civil.ist.utl.pt, derosario@gmail.com

***Abstract*—In a building, the paint coating applied to interior walls conveys their aesthetic character and also performs an important function of protection. It is a construction component which is exposed to agents of deterioration related to its use, needing the regular evaluation of its state of repair. The proposed advanced computational model supports the performance of such periodic inspections and the monitoring of interior wall maintenance, using Virtual Reality (VR) Technology. Used during an inspection visit, the application allows users to consult a database of irregularities, normally associated with paint coating, classified by the most probable causes and by recommended repair methodologies. In addition, with this model, a chromatic scale related to the degree of deterioration of the coating, defined as a function of the time between the dates of the application of the paint and the scheduled repainting, can be attributed to each element of coating monitored. This use of VR technology allows inspections and the evaluation of the degree of wear and tear of materials to be carried out in a highly direct and intuitive manner. The developed computer application is an advanced computation tool with innovative visualization and interactive capacities, and so brings a positive contribution on the construction filed.**

***Keywords-Virtual Reality; maintenance; inspection; interaction***

## I. INTRODUCTION

The coating applied to building walls, naturally, performs an important aesthetic function: it is, however, essentially a protective element for the substrate on which it is applied as far as the action of environmental agents of wear and tear is concerned. The coating is fundamental to a proper overall performance of a building throughout its working life.

Materials frequently used in the coating of ordinary buildings are paint, varnish, stone and ceramics [1]. In Portugal, where interior walls are concerned, the most commonly used coating is paint. It is a multi-purpose material, used under a variety of decorative effects, based on a widely-ranging palette of colours, patterns and textures and is easily applied on any type of surface. In addition, paint, compared to other materials, is less costly; not only as far as the product itself is concerned, but also in its application, since relatively non-specialised labour is required. Nevertheless, as deterioration is a given, maintenance is needed.

Factors such as the constant exposure of the coating to the weather, pollutants and the normal actions of housing use, linked to its natural ageing and, in some cases to the unsuitable application of systems of painting give rise to its deterioration and to the appearance of irregularities, which can negatively affect its performance as both an aesthetic and a protective element. The weather significantly influence the state of use of peripheral walls of the building once the humidity through the wall thickness causing anomalies in the inner surface of the wall. According to Lopes [2], in normal conditions of exposure and when correctly applied, a paint coating can remain unaltered for about five years. Establishing suitable maintenance strategies for this type of coating is based on the knowledge of the most frequent irregularities, the analysis of the respective causes and the study of the most suitable repair methodologies.

Currently, the management of information related to the maintenance of buildings is based on the planning of action to be taken and on the log of completed work. The capacity to visualize the process can be added through the use of three-dimensional (3D) models which, facilitate the interpretation and understanding of target elements of maintenance and of 4D models (3D + time) through which the evolution of deterioration can be visually demonstrated and understood. Furthermore, the possibility of interaction with the geometric models can be provided through the use of Virtual Reality (VR) technology. The developed VR/4D model is an advanced computer tool in the maintenance field.

The work presented here is part of an on-going research project: *Virtual Reality technology applied as a support tool to the planning of construction maintenance*. PTDC/ ECM/67748/2006 [3] and as such is a component of the Project focussing on the support of the maintenance activity planning with particular reference to paint coating applied to interior walls of buildings for housing.

The completed computer virtual model identifies the elements of the building which make up the interior wall coating so that monitoring can take place. The application is supported by a database, created for the purpose, of irregularities, their probable causes and suitable repair processes, which facilitates the inspection process. The information is recorded and associated to each monitored element, allowing subsequently, the inspection and repair activity log to be consulted, thus providing a tool for the

 ISBN: 978-1-61208-237-0 

definition of a rehabilitation strategy. In addition, the model assigns a colour to each of the coating elements, the colours defined by the time variable, so that the evolution of the deterioration of the coating material is clearly shown through the alteration in colour. The prototype is, then, a 4D model.

The model integrates a virtual environment with an application developed in Visual Basic programming language. This allows interaction with the 3D model of buildings in such a way that it becomes possible to follow the process of monitoring the coating elements, specifically, painted interior walls, in terms of maintenance, throughout the life-cycle of the building.

Advantages of 4D virtual environments are found in improving communication, increasing insight, supporting collaboration, and supporting decision-making. Namely Fumarola [4] discusses different approaches on the generation of virtual environments of real world facilities supporting decision-making. The development of several applications, have been successfully implanted, in different branches of construction and education domain. The requisites of the real environment can be illustrated by the different VR models developed within the research project [5]: The implemented virtual prototypes concern the management of a lighting system in a building [6], the construction activity [7] and the maintenance of building façades [8]. The research project follows a previous work related to the development didactic VR models regarding construction processes: the construction of a roof [9] and bridge decks using two different methods (the cantilever process [10] and the incremental launching method [11]).

When comparing the present computer model with the previous work, the principal innovation concerns the incorporation of the capacity of changing the colour of the painted wall with the time parameter, when walking-through the virtual model. So the evolution of the deterioration of the coating material is visualized through the alteration in colour.

The paper presents the main aspects of the maintenance of buildings with a focus on maintenance of painted walls. Then the anomalies that most often originate in the painted finish, which are listed in an identical manner to that used in the virtual model database. The text describes how to interact with the virtual model on the conduct of inspections and monitoring of wall elements. Finally it makes a comparison with the traditional way of performing and points out the major benefits in the use of the interactive computational tool of maintenance support.

## II. MAINTENANCE

The General Regulations for Urban Buildings (RGEU) [12] stipulates the frequency of maintenance work, stating that existing buildings must be repaired and undergo maintenance at least once every eight years with the aim of eliminating defects arising from normal wear and tear and to maintain them in good usable condition in all aspects of housing use referred to in that document.

The time-limit indicated is applicable to all elements of the buildings generally. It is clear, however, that the regulatory period is too long for some specific components and that, frequently enough, the time-limits for action are not respected. There are, too, inefficient rent policies, leading to long periods without rehabilitation, and that the prevailing culture is one of reaction on the part of the various parties involved in the maintenance process. To these aspects should also be added the defects sometimes registered during the construction of property developments, exacerbating the poor state of repair of the buildings. This gives rise to numerous irregularities which, in turn, frequently leads to inadequate safety conditions.

According to Cóias [13], the purpose of maintenance is to prolong the useful life of the building and to encourage adherence to the demands of safety and functionality, keeping in mind the specific set of conditions of each case and its budgetary considerations. Satisfactory management of this activity is carried out by putting into practice a maintenance plan which must take into consideration technical, economic, and functional aspects arising with each case.

Collen [14] points out that investment in the maintenance and rehabilitation sector in Portugal is still weak compared to that in the same sector in the construction industry in the other countries of the European Community. She makes it clear, however, on a more positive note, that some measures have already begun to be implemented here: some urban regeneration programmes have been created, legislation, which focuses on the sustainability of buildings, has been laid down, and the revision of constructive solutions has been carried out, all with the objective of guaranteeing that the maintenance of built heritage be an integral part of the construction sector.

The maintenance of buildings, then, is an activity of considerable importance within the construction industry; its contributory aspects of conservation and rehabilitation work need to be supported by correct methodologies of action, underpinned by scientific criteria and by suitable processes for the diagnosis of irregularities and the evaluation of their causes. This paper aims to make a positive contribution to this field using the new computer technology tools of visualisation and interaction.

## III. PATHOLOGIES IN PAINT COATINGS

The technical document *Paints, Varnishes and Painted Coatings for Civil Construction* published by The National Laboratory for Civil Engineering (LNEC), defines paint as a mixture essentially made up of pigments, binder, vehicle and additives [1]. It has a pigmented, pasty composition, and when applied in a fine layer to a surface, presents, after the dispersion of volatile products, the appearance of a solid, coloured and opaque film [15].

The durability of the painted coating depends on the environment in which it is used, and on the surface it is applied to as well as the rate of deterioration of the binder in the paint. The influence of the environment is the result of the action, in conjunction or alone, of a variety of factors such as the degree of humidity, the levels of ultraviolet radiation, oxygen, ozone and alkalis, variations in temperature and of other physical or chemical agents whose effect depends considerably on the time taken to apply it

 ISBN: 978-1-61208-237-0 

[16]. When their influence is not counteracted or minimised, imperfections can arise in the coating film, such as, the appearance of defects in the layer or paint with the loss of functionality where the desired aim of the application is concerned. These irregularities manifest themselves in various ways and in different degrees of severity. Based on the study made of the causes of the defects, specific methodologies for their resolution were established. Figure 1 shows common defects in painted interior walls.

Figure 1. Swelling, efflorescence, cracking and blistering [17].

The information gained from the pathological analysis of this type of coating was used to draw up a database supporting the interactive application. These data support the creation of inspection files related to the elements which are monitored in each case studied.

In order to form a user-friendly database of relatable data, groups of pathologies, shown below in Table I, were considered. This classification provides the required automatism of access to the database and supports the presentation of synopses of the causes and repair methodology inherent in each pathology.

During the process of an on-site inspection, the user of the application can refer to the database in order to classify the abnormality being observed, consulting the list of defects, which includes, in addition to their identification, the most relevant characteristics and some of the causes that could be at the root of their development. Table II lists two of the irregularities from the classification: *Alteration in colour*.

The database was created with adequate relations between data, concerning each group of anomalies, in order to present the sequence of anomaly, provable cause and adequate repair work, to the engineer when it uses the virtual model in an inspection situation. The specialist must choose in each case the most appropriate sequence.

## IV. INTERACTIVE MODEL

The completed application supports on-site inspections and the on-going analysis of the evolution of the degree of deterioration of the coating [18]. The following computational systems were used in its development: *AutoCAD*, in the creation of the 3D model of the building; [19], for the programming of the interactivity capacities integrated with the geometric model; *Visual Basic 6* in the creation of all the windows of the application and in the establishment of links between components. All the systems were made available by the ISTAR/DECivil informatics laboratory of the Technical University of Lisbon.

TABLE I. CLASSIFICATION OF IRREGULARITIES

| *Classification* | *Irregularity* | *Repair methodology* |
|---|---|---|
| **Alteration in Colour** | Yellowing | - Cleaning the surface and repainting with a finish both compatible with the existing coat and resistant to the prevailing conditions of exposure in its environment |
| | Bronzing | |
| | Fading | |
| | Spotting | |
| | Loss of gloss | |
| | Loss of hiding power | |
| **Deposits** | Dirt pick-up and retention | - Cleaning the surface. |
| | Viscosity | |
| **Changes in Texture** | Efflorescence | - Removal by brushing scraping or washing;<br>- repainting the surface;<br>- When necessary apply sealer before repainting. |
| | Sweating | |
| | Cracking | |
| | Chalking | |
| | Saponification | |
| **Reduction in Adhesion** | Peeling | - Proceed by totally or partially removing the coat of paint;<br>- Check the condition of the base and proceed with its repair where necessary;<br>- Prepare the base of the paint work. |
| | Flaking | |
| | Swelling | |

TABLE II. IRREGULARITIES AND CAUSES

| *Classification* | *Irregularity* | *Characteristics and causes* |
|---|---|---|
| **Alteration in colour** | Yellowing | - A yellow colour caused by ageing of the film of the paint or varnish;<br>- Action of environmental agents (solar radiation, temperature oxygen and humidity) on the binder in the paint provoking changes in its molecular structure. |
| | Discolouration | - Partial loss of colour of the film of paint coating;<br>- Action of environmental agents (solar radiation, temperature, polluted atmosphere and chemically aggressive bases of application) on the binder and/or the pigments of the painted coating. |

The main interface gives access to the virtual model of the building and to the inspection and maintenance modules (highlighted in Figure 2). The first step is to make a detailed description of the building (location, year of construction, type of structure, etc.; see Figure 3) and representative modelled elements of the interior wall coating, so that they

 ISBN: 978-1-61208-237-0

can be monitored. The model is manipulated in the virtual environment by using the mouse buttons (movement through the interior of the model and orientation of the camera, Figure 4). The coordinates of the observer's position and the direction of his/her point of view are associated with the element during the process of identification.

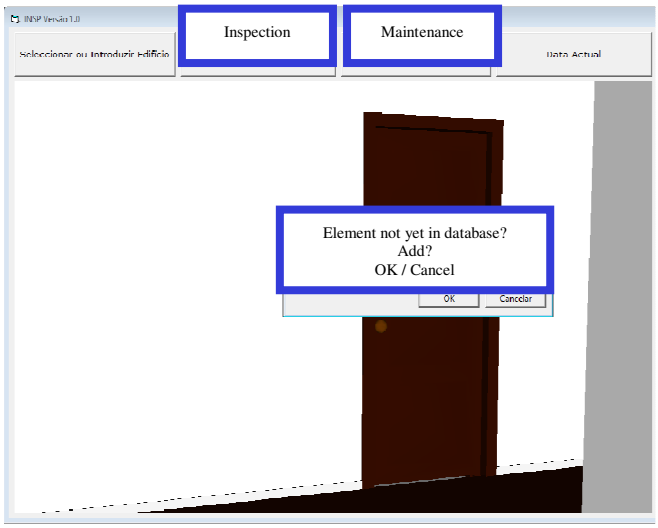



Figure 2. The main interface of the virtual application.

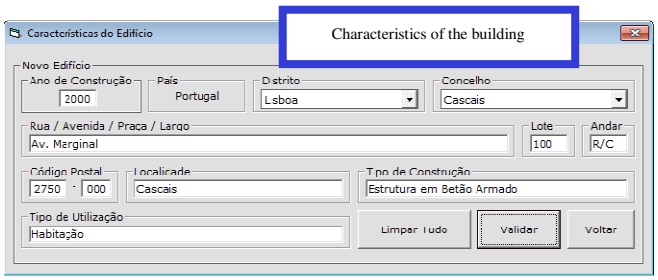



Figure 3. Interface for the detailed description of the building.

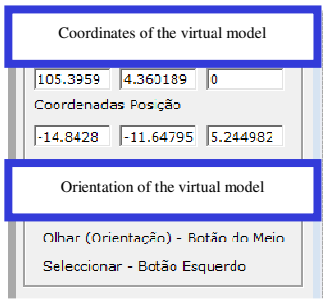



Figure 4. Coordinates and manipulation commands in the virtual model.

Thus, later, when an element in the database of the application is selected using the interface, the model is displayed in the visualisation window so that the target coating can be observed. Walking through the model with the aim of accessing all the elements of the building, the user needs to be able to go up and down stairs or open doors or windows.

The virtual model has been programmed, using the *EON system*, in such a way that these capacities are activated by positioning the cursor over the respective objects, in that way, the user is able to walk through the whole model. Each wall surface in each of the rooms of the house is a component which has to be monitored and, therefore, to be identified. Using the model, the user must click the mouse on an element, and the message *New Element* is shown (highlighted in Figure 2). Associated to this selected element is the information regarding location within the house (*hall*, *bedroom*), wall type (*simple internal masonry wall*) and coating (*paint*), as shown in Figure 5.

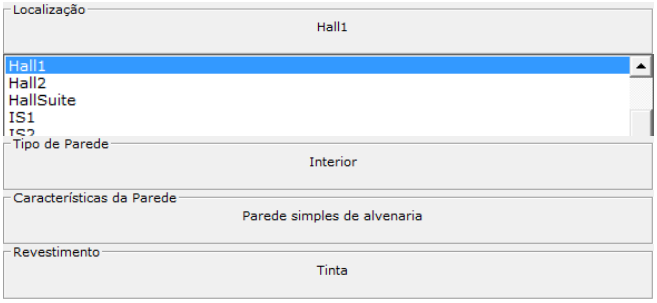



Figure 5. Identification of an element in the virtual model.

## A. *Making an Inspection*

Later, on an on-site inspection visit, the element to be analysed it selected interactively on the virtual model. The inspection sheet (Figure 6) is accessed by using the Inspection button which is found in the main interface (Figure 2). The data which identify the selected element are transferred to the initial data boxes on the displayed page (Figure 6).

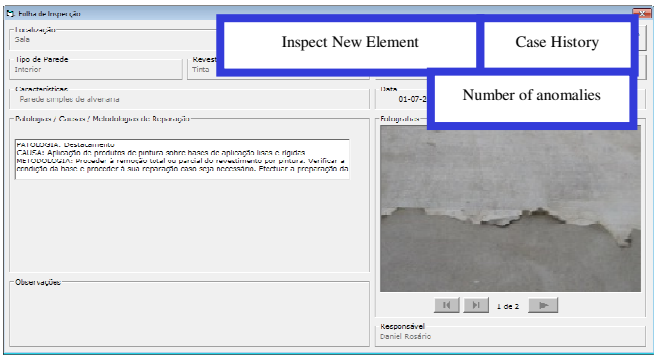



Figure 6. Presentation of the information introduced into the inspection sheet.

Next, using the database, the irregularity which corresponds to the observed defect, with its probable cause (ageing) and the prescribed repair methodology (removal and repainting) is selected (see highlighted area, Figure 7). The current size of the pathology should also be indicated since it reveals how serious it is (*area of pathology*, Figure 7). In the field *Observations*, the inspector can add any relevant comment (Figure 6), photographs obtained on site can also

be inserted into the inspection window and the date of the on-site visit and the *ID* of the inspector should also be added. Several different irregularities in the same coating can be analysed (field *Number of Pathologies*, Figure 6) and other elements can be analysed and recorded and defects observed.

Later, the files thus created, associated to each of the virtual model elements, can be consulted (*Case History* button in the *Interface* in Figure 6). This same window allows all the data referring to the building and to the completed inspection to be shown, in *pdf* format (Figure 8).

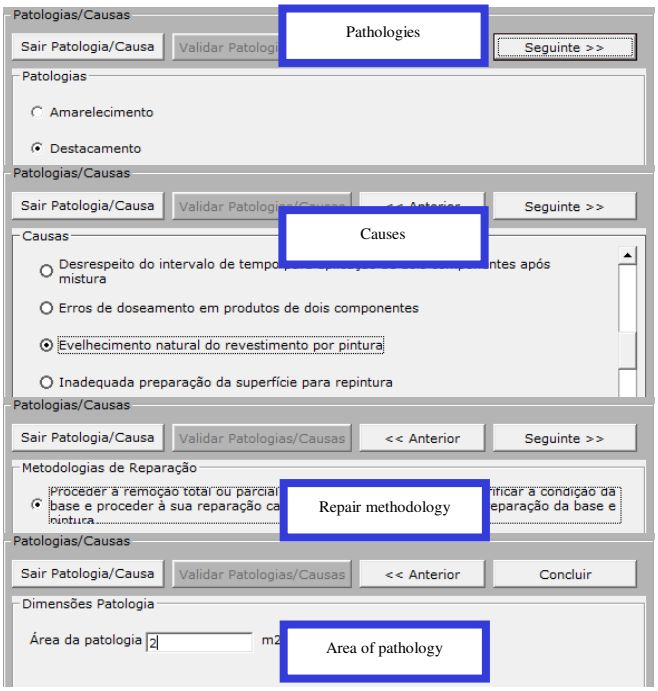



Figure 7. Interface for the selection of the irregularity, probable cause, area and repair methodology.

## B. *Maintenance Monitoring*

How long the working life of any construction component might be is an estimate and depends on a set of modifying factors related to their inherent characteristics of quality, to the environment in which the building is set and to its conditions of use [13]. In maintenance strategy planning the probable dates when adverse effects might occur in each of these elements must be foreseen, and the factors which contribute to defects must be reduced and their consequences minimized.

The completed model allows the user to monitor the evolution of wear and tear on the paint coating in a house. For this, technical information relative to the reference for the paint used, its durability and the date of its most recent application must be added (Figure 9) to each element through the *Maintenance Interface* (also accessed from the main interface, Figure 2)

Based on these data, it is possible to link in the date the virtual model is consulted and visualise, in the geometric model, the level of wear and tear as a function of time (see *state of repair*, Figure 9).


Figure 8. Inspection sheet.

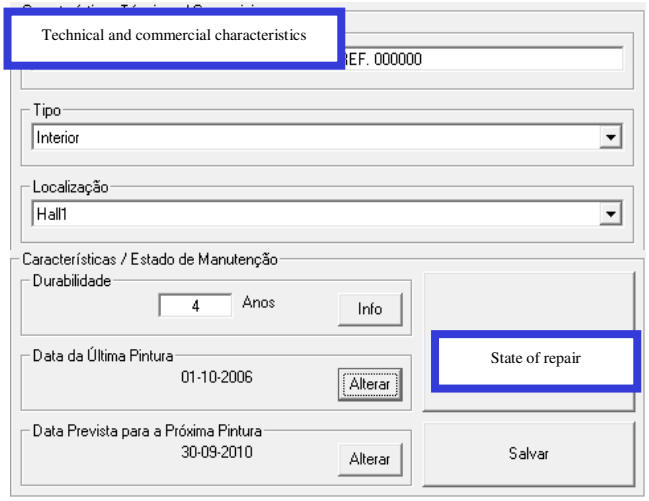



Figure 9. Technical characteristics and the durability of the paint-work.

The period of time between the date indicated and the date when the paint was applied is compared to the duration advised, in the technical literature, for repainting. The value given for this comparison is associated to the Red, Green, Blue (RGB) parameters which define the colour used for wall in the virtual model (Figure 10). In this way, the colour visualised on the monitored wall varies according to the period of time calculated, pale green being the colour referring to the date of painting and red indicating that the

 ISBN: 978-1-61208-237-0 

date the model was consulted coincides with that advised for repainting (Figure 10).

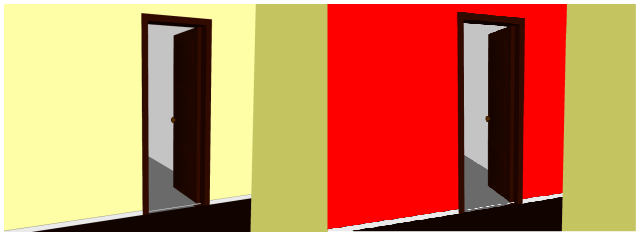


Figure 10. Chromatic alteration of the coating according to its state of deterioration.

The date for painting and repainting are saved to a list of coating elements to be monitored in the virtual model (Figure 11). When an element is selected from this list, the corresponding element is represented in the virtual model, through the preview window, in the colour that corresponds to the period of the consultation (Figure 10).

| Local | Initial date | Durability | Repainting date |
|---|---|---|---|
| Hall | 01-10-2006 | 4 | 30-09-2010 |

Figure 11. Information relevant to the monitored elements.

### C. *Principal innovations and benefits*

Normally, the inspection process of buildings is based on filling out paper files during the site survey. The inspector shall observe and analyze the anomalies observed, sort them and add comments that are timely, related to the degree of deterioration, so that, later, to establish a global repair work plan for the building.

With the use of computer system supported on the basis of data relating to the maintenance of the interior wall paint coating, the inspector shall select among the possibilities listed, the anomaly, possible cause and repair methodology that seems more appropriate. Moreover you can add photos. The definition of the repair plan is based on the analysis of the data entered and selected in the system. The designer must establish on the basis of the needs and economic availability of a proper maintenance plan, encompassing the various anomalies observed. The computer system allows registration, during the inspection, deficiencies and their seriousness, by supporting the planning of repair work.

The inspection process requires one or more visits to the site of the building [13]. There are made several observations which are annotated in sheets of paper. Based on these notes is elaborated a rehabilitation draft to the building. With the support of developed computer system the inspector can easily associate to the 3D model of the building the identification of the anomalies, know its extent and severity. The facility to incorporate photos from local helps assess the severity of repair work to be carried out later. The main innovation in the use of the model is to allow the designer to move within a virtual mode through the interior of the model, know the location, type and severity of each anomaly in relation to each wall area. In this way the designer can more easily analyze the global repair strategy. For an inspection based on paper there is no longer a global perspective of the house inspected and therefore the study of repair work to predict is more difficult. Furthermore, the virtual model has a history of documented inspections with timely remarks, comments and photos. You can add the repair operations after intervention and thus compare the observed problems and what the resolution applied.

## V. CONCLUSIONS

This application supports the maintenance of painted interior walls and promotes the use of IT tools with advanced graphic and interactive capabilities in order to facilitate and expedite the inspection process. The virtual model, moreover, allows users to see, in the virtual environment, the state of repair of the coating.

The information about pathologies, causes and repair methods, collected from a specialised bibliography, has been organised in such a way as to establish a database to be used as a base for the drawing up of a tool to support building maintenance. The main aim of the application is to facilitate maintenance enabling the rapid and easy identification of irregularities, as well as the possible prediction of their occurrence through the available inspection record. This analysis has been shown as playing an important role in conservation and in the reduction of costs related to the wear and tear of buildings and contributes to the better management of buildings where maintenance is concerned.

In addition to the inspection component, a maintenance component was developed which, being visualized in a VR environment, as well as being highly intuitive, facilitates the analysis of the state of repair of buildings. By means of a chromatic scale applied to the monitored elements, displayed in the walk-through of the geometrically modelled building, it is possible to identify the elements which, predictably, will need timely action. With the possibility of altering the time parameter freely, the user can carry out this analysis either for past instants or for future events, being able, in this way, to forecast future operations. This capacity of the model, therefore, contributes to the avoidance of costs associated to irregularities which, with the passage of time, become more serious and therefore more onerous.

## ACKNOWLEDGMENTS

The authors wish to thank the Foundation for Science and Technology for the financial support given for the development of the research project *Virtual Reality technology applied as a support tool to the planning of construction maintenance*. PTDC/ECM/67748/2006 [3].

## REFERENCES

[1] M. Eusébio and M. Rodrigues, “Paints, Varnishes and Painted Coatings for Civil Construction”, CS 14, National Laboratory for Civil Engineering, Lisbon, Portugal, 2009, ISBN: 9789724917627.

[2] C. Lopes, "Anomalies in painted exterior walls: technic of inspection and structural evaluation", Construlink Press, Monograph, nº22, Lisbon, Portugal, March/April 2004.

[3] A. Z. Sampaio and A. M. Gomes, "Virtual Reality technology applied as a support tool to the planning of construction maintenance", research project PTDC/ECM/ 67748/2006, FCT, Lisbon, Portugal, 2008-2011.

[4] M. Fumarola and R. Poelman, "Generating virtual environments of real world facilities: Discussing four different approaches", Automation in Construction , vol. 20 (2011), pp. 263–269, http://www.sciencedirect.com/science/article/pii/S0926580510002074#sec3. [retrieved: March, 2012]

[5] A. Z. Sampaio, C. O. Cruz, and O. P. Martins, "Didactic models in Civil Engineering education: Virtual simulation of construction works", Book: Virtual Reality, Ed. Jae-Jin Kim, ISBN: 978-953-307-518-1, chap. 28, 2011, pp. 579 – 598, http://www.intechopen.com/articles/show/title/didactic-models-in-civil-engineering-education-virtual-simulation-of-construction-works [Retrieved: January, 2012]

[6] A. Z. Sampaio, M. M. Ferreira, and D. P. Rosário, "Management system integration supported on Virtual Reality technology: The building lighting devices", Book: ENTERprise Information Systems, ISBN: 978-3-642-16401-9, Springer-Verlag Berlin Heidelberg, vol. 190, 2010, pp. 207-210. http://www.springerlink.com/content/978-3-642-16401-9#section=825347&page=1 [Retrieved: July, 2012]

[7] A. Z. Sampaio and J. P. Santos, "Interactive Models Supporting Construction Planning", Proc. CENTERIS 2011 - International Conference on ENTERprise Information Systems, Algarve, Portugal, Oct. 5-7, 2011, pp. 40-50.

[8] A. Z. Sampaio, A. R. Gomes, and J. P. Santos, "Virtual Environment in Civil Engineering: Construction and Maintenance of Buildings" Proc. ADVCOMP 2011, The 5th Int. Conf. on Advanced Engineering Computing and Applications in Sciences, Lisbon, Portugal, Nov. 20-25, 2011, pp. 13-20.

[9] A. Z. Sampaio, C. O. Cruz, and O. P. Martins, "Interactive models based on virtual reality technology used in Civil Engineering education", Book: Teaching through multi-user virtual environments: applying dynamic elements to the modern classroom, IGI Clobal, Ed. G. Vincenti, and J. Braman. ISBN: 978-1-61692-822-3, 2011, Chap. 21, pp. 387-413, http://resources.igiglobal.com/marketing/pdfs/vincenti/21.pdf

[10] P. F. Studer and A. Z. Sampaio, "Virtual reality technology applied to simulate construction processes", Book: Computational Science and Its Applications, Springer Berlin / Heidelberg, ISSN 0302-9743, vol. 3044/2004, pp. 817-826, http://www.springer link.com/content/1ww6af49ugj08d9d[Retrieved: July, 2012]

[11] O. P. Martins and A. Z. Sampaio "The incremental launching method for educational virtual model", Book: Cooperative Design, Visualization, and Engineering, Springer Berlin / Heidelberg, ISSN: 0302-9743, vol. 5738/2009, pp. 329-332, http://www.sprin gerlink.com /content/e63659p0746g4364/ [Retrieved: January, 2012]

[12] "RGEU - General Regulations for Urban Buildings", Decree-Law, nº 38 382, August 7, 1951, Lisbon, Portugal.

[13] V. Cóias, "Inspections and essays on rehabilitation of buildings". Lisbon, Portugal, IST Press, pgs 448, ISBN: 978-972-8469-53-5, (2ª Ed.) 2009.

[14] I. Collen, "Periodic inspections in buildings". Planet CAD studies, March, 2003. http://www.planetacad.com/presentationlayer/Estudo_01.aspx?id=13&canal_ordem=0403 [Retrieved: June, 2012]

[15] M. B. Farinha, "Construction of Buildings in practice: guide oriented to the development of processes and methodologies of construction", Vol. 2. Verlag Dashofer, ed. Psicosoma, Lisboa, 2010. http://www.psicosoma.pt/ [Retrieved: June, 2012]

[16] M. I. Marques, "Durability of plastic tint", ITMC 2, National Laboratory for Civil Engineering, Lisbon, Portugal, 1985.

[17] A. Moura, "Characteristics and conservation state of painted façades: study case in Coimbra", Master Dissertation in Construction, Coimbra, Portugal 2008.

[18] D. P. Rosário, "Virtual Reality technology applied on building maintenance: painted interior walls", Master Dissertation in Construction, Lisbon, Portugal, 2011.

[19] "EON Studio - Introduction to working in EON Studio", EON Reality, Inc. 2011. http://www.eonreality.com/ [Retrieved: January, 2012]

 ISBN: 978-1-61208-237-0 

# *Applying Neural Network Architecture in a Multi-Sensor Monitoring System for the Elderly*

Shadi Khawandi, Pierre Chauvet
University of Angers
Angers, France
chadi.khawandi@etud.univ-angers.fr,
Pierre.Chauvet@uco.fr

Bassam Daya
Lebanese University
Saida, Lebanon
b_daya@hotmail.com

*Abstract*— **One of three adults 65 years or older falls every year. As medical science advances, people can live with better health and alone up to a very advanced age. Therefore, to let elderly people live in their own homes leading their normal life and at the same time taking care of them requires new kinds of systems. In this paper, we propose a multi-sensor monitoring system for the fall detection in home environments. The system, which consists of a webcam and heart rate sensor, processes the data extracted from the two different sub-systems by applying neural network n order to classify the fall event in two classes: fall and not fall. Reliable recognition rate of experimental results underlines satisfactory performance of our system.**

***Keywords-Neural Network; fall detection; heart rate; webcam***

## I. INTRODUCTION

Falling and its resulting injuries are an important public-health problem for older adults. The National Safety Council estimates that persons over the age of 65 have the highest mortality rate (death rate) from injuries. Among older adults, injuries cause more deaths than either pneumonia or diabetes. The risk of falling increases with age. Demographic predictions of population aged 65 and over suggest the need for telemedicine applications in the eldercare domain. Many devices have been developed in the last few years for fall detection [1][2], such as a social alarm, which is a wrist watch with a button that is activated by the person in case he/she suffers a fall, and wearable fall detectors, which are based on combinations of accelerometers and tilt sensors. The main problem with social alarms is that the button is often unreachable after a fall, especially when the person is panicked, confused, or unconscious. For the wearable sensors, these autonomous sensors are usually attached under the armpit, around the wrist, behind the ear's lobe, or at the waist. However, the problem of such detectors is that older people often forget to wear them [3][4]; indeed, their efficiency relies on the person's ability and willingness to wear them.

The proposed system is composed of two different devices: webcam and heart rate sensor. The extracted data will be processed by a neural network for classifying the events in two classes: fall and not fall. Reliable recognition rate of experimental results underlines satisfactory performance of our system.

In this paper, we review some existing vision-based fall detection systems (Section II), and then we introduce our proposed system (Section III) with additional technical details. The experimental results are presented in Section IV, and finally, the conclusion is presented in Section V.

## II. RELATED APPROACHES

Information Technology combined with recent advances in networking, mobile communications, and wireless medical sensor technologies offers great potential to support healthcare professionals and to deliver remote healthcare services, hence, providing the opportunities to improve efficiency and quality and better access to care at the point of need. Existing fall detection approaches can be categorized into three different classes to build a hierarchy of fall detection methods. Fall detection methods can be divided roughly into three categories:

- Wearable Sensors (such as accelerometers or help buttons): These autonomous sensors are usually attached under the armpit, around the wrist, behind the ear's lobe, at the waist or even on the chest. Merryn [5] used an integrated approach of waist-mounted accelerometer. A fall is detected when the negative acceleration is suddenly increased due to the change in orientation from upright to lying position. A barometric pressure sensor was introduced by Bianchi [6], as a surrogate measure for altitude to improve upon existing accelerometer-based fall event detection techniques. The acceleration and air pressure data are recorded using a wearable device attached to the subject's waist and analyzed offline. A heuristically trained decision tree classifier is used to label suspected falls. Estudillo-Valderrama [7] analyzed results related to a fall detection system through data acquisition from multiple biomedical sensors then processed the data with a personal server. A wearable airbag was incorporated by Tamura [8] for fall detection by triggering airbag inflation when acceleration and angular velocity thresholds are exceeded. Chen [9] created a wireless, low-power sensor network by utilizing small, noninvasive, low power motes (sensor nodes). Wang [10] applied reference velocities and developed a system that uses an accelerometer placed on the head. However the problem of such detectors is that older people often forget to wear them, indeed their efficiency relies on the person's ability and willingness to wear them, moreover in the case of a

 ISBN: 978-1-61208-237-0 

help button, it can be useless if the person is unconscious or immobilized.

- Environmental Sensors: Environmental sensors based devices attempt to fuse audio data and event sensing through vibration data. Zhuang [11] proposed an approach the audio signal from a single far-field microphone. A Gaussian mixture model (GMM) super vector is created to model each fall as a noise segment. The pair wise difference between audio segments is measured using the Euclidean distance. A completely passive and unobtrusive system was introduced by Alwan [12] that developed the working principle and the design of a floor vibration-based fall detector. Detection of human falls is estimated by monitoring the floor vibration patterns. The principle is based on the vibration signature of the floor. The concept of floor vibrations with sound sensing is unique in its own way [13]. Pattern recognition is applied to differentiate between falls and other events. Toreyet [14] fused the multitude of sound, vibration and passive infrared (PIR) sensors inside an intelligent environment equipped with the above fusion elements. Wavelet based feature extraction is performed on data received from raw sensor outputs.
  Most ambient device based approaches use pressure sensors for subject detection and tracking. The pressure sensor is based on the principle of sensing high pressure of the subject due to the subject's weight for detection and tracking. It is a cost effective and less intrusive for the implementation of surveillance systems. However, it has a big disadvantage of sensing pressure of everything in and around the subject and generating false alarms in the case of fall detection, which leads to a low detection accuracy.
- Computer Vision Systems: Cameras are increasingly included, these days, in in-home assistive/care systems as they convey multiple advantages over other sensor based systems. Cameras can be used to detect multiple events simultaneously with less intrusion. Cucchiara [15] applied a multi-camera system for image stream processing. The processing includes recognition of hazardous events and behaviors, such as falls, through tracking and detection. The cameras are partially overlapped and exchange visual data during the camera handover through a novel idea of warping "people's silhouettes.
  From tracking data, McKenna [16] automatically obtained spatial context models by using the combination of Bayesian Gaussian mixture estimation and minimum description length model for the selection of Gaussian mixture components through semantic regions (zones) of interest. Tao [17] developed a detection system using background subtraction with an addition of foreground extraction, extracting the aspect ratio (height over width) as one of the features for analysis, and an event-inference module which uses data parsing on image sequences.
  Foroughi [18] applied an approximated ellipse around the human body for shape change. Projection histograms after segmentation are evaluated and any temporal changes of the head position are noted. Miaou [19] captured images using an Omni-camera called MapCam for fall detection. The personal information of each individual, such as weight, height and electronic health history, is also considered in the image processing task. Rougier [20] proposed a classification method for fall detection by analyzing human shape deformation. Segmentation is performed to extract the silhouette and additionally edge points inside the silhouette are extracted using a canny edge detector for matching two consecutive human shapes using shape context. With Visual fall detection, what appears to be a fall might not be a fall. Most of existing systems are unable to distinguish between a real fall incident and an event when the person is lying or sitting down abruptly.

## III. PROPOSED SYSTEM

This paper proposes a multi-sensor fall-detector system (Fig. 1) as a combination between two different commercial devices: a webcam and a heart rate sensor. Data extracted from the two sub-systems will be processed by the neural network [Multi-Layer Perceptron (MLP)] in order to detect the fall. Once the fall is detected, an emergency alert will be activated automatically and sent to care holders through an internet-based home gateway

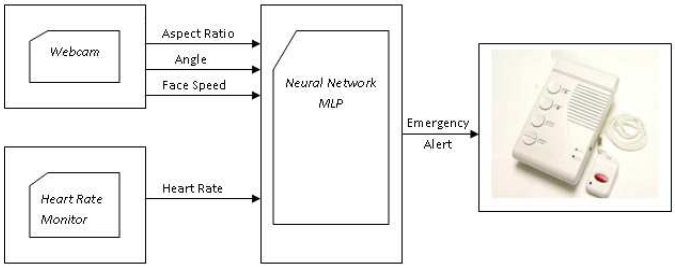



Figure 1. Overview of the proposed system.

### A. Webcam System

It is obvious that we need several webcams to cover the entire monitored zone and the switch between the webcams will be based on the face presence. In this paper, we present the webcam system as limited to one webcam, as it will be similar when having multiple webcams.

The webcam system is based on image processing in real time; this system detects the body and face of a person in a given area, collects data such as the aspect ratio, angle, and speed of movement of the person, then sends the extracted data to be processed by the MLP. The system starts by removing the background. After the silhouette is acquired, the next step is the skin color detection, which is an effective way often used to define a set of areas likely to contain a face or hands; then, the system detects the face. Then, features extraction is involved (speed of a person's movement, aspect ratio, and fall angle).

 ISBN: 978-1-61208-237-0 

*1) Background Subtraction*

Background subtraction (Fig. 2) is a particularly popular method to detect moving regions in an image by differentiating between the current image and a reference background image in a pixel-by pixel way

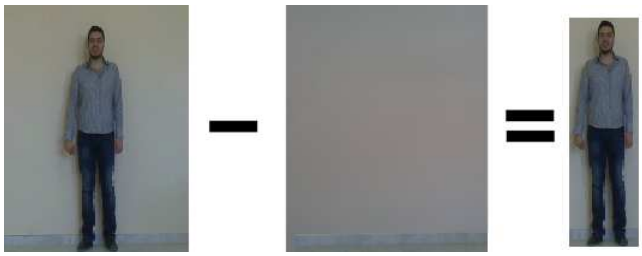


Figure 2. Background Subtraction.

*2) Skin-color and HSV detection*

The images captured by the webcams are then processed by the system to detect skin color. This is an effective technique for determining whether an image contains a face or hands. In this technique, the appropriate threshold values are defined for all the pixels in a color space (Fig. 3). Different color spaces are used to represent skin color pixels: RGB, RGB standard, HSV (or HSI), YCrCb, and HSV. After the detection of skin-color pixels, image filtering (erosion and dilation) is carried out

Figure 3. Image after skin color detection.

*3) Face Detection- Approximated Ellipse*

After identifying the skin areas, it is necessary to distinguish the face. For this, the shape of the detected object is compared with an ellipse. This correlation technique is very effective and efficient.

Based on the comparison with an ellipse, we may have more than one image, such as the hand. In order to solve this issue, each image will be converted into a binary image (black and white); then, the white contour will be replaced by black. In this state, the object representing the hand goes black but the object representing the face becomes black except the eyes and mouth. After this transformation, we compute the white surface in each picture, and the object having the greater white surface is the one of the face, and in this case, it is detected.

After calculating the white surface in each image, we found that the white surface in the face is greater than that in the hand; that is why this intelligent system detects the face (Fig. 4).

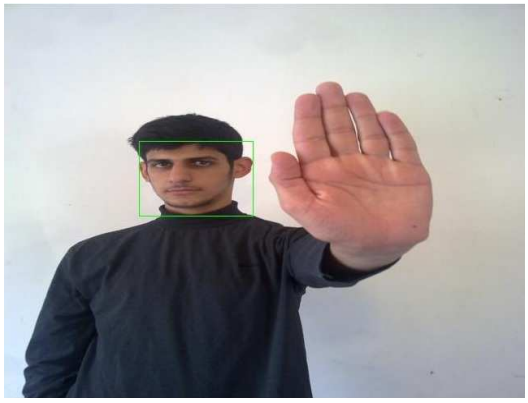


Figure 4. Image after skin color detection.

*4) Speed Extraction*

One major point in the recognition system is the feature extraction, i.e., the transition from the initial data space to a feature space that will make the recognition problem more tractable. So, we analyze the shape changes of the detected face in the video sequence. The planar speed of movement is calculated using the following formula:

Planar speed = distance/time (pixel/s);

- *Distance*: between the same face in consecutive frames (pixel);
- *Time*: processing time between two consecutive frames.

The range is from 90 to 700 pixels (it can vary depending on the quality of the pictures).

*5) Aspect Ratio and Angle Extraction*

- Aspect Ratio

  The aspect ratio of a person is a simple yet effective feature for differentiating a normal standing pose from other abnormal poses (Fig. 5). The aspect ratio of the human body changes during a fall. When a person falls, the height and width of his bounding box change drastically (height/width). The range is from 0.15 to 6 (it can vary depending on the dimensions of the subject or on the scaling camera to image coefficients).
- Angle

  Fall angle (θ) is the angle of a vertical line through the centroid of the object with respect to the horizontal axis of the bounding box (Fig. 5). The centroid (Cx, Cy) is the center of mass coordinates of an object. When a person is standing, we assume that he is in an upright position and the angle of a vertical line through the centroid with respect to the horizontal axis of the bounding box should be approximately 90 degrees. When a person is walking, the θ value varies from 45 degrees to 90 degrees. When a person is falling, the angle is always less than 45 degrees. For every frame, we calculate the fall angle (θ), and if θ value is

 ISBN: 978-1-61208-237-0 

less than 45 degrees, we confirm that the person is falling. The range is from 0 to 90 degrees.

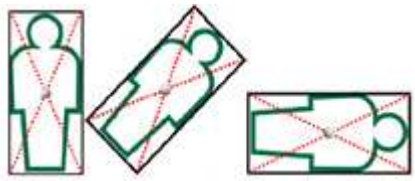


Figure 5. Bounding box and poses of human object.

### B. Heart Rate System

A heart rate monitor is a personal monitoring device, which allows a subject to measure his or her heart rate in real time or record his or her heart rate for later study. Early models consisted of a monitoring box with a set of electrode leads, which attached to the chest. This paper does not include the design of a heart rate monitor, but we will use an existing heart rate monitor. A Wi-Fi heart rate belt (HRM-2823) (Fig. 6) could be connected to a computer or Wi-Fi operator mobile; this monitor has professional software providing online data exchange model that allows the heart rate belt keeps connecting with the PC and transmits data to PC in real time. The idea is to have a "non-image"-related parameter involved in the fall detection in order to minimize the false alarms.

Figure 6. Heart rate monitor.

### C. Neural Network System

Throughout the years, the computational changes have brought growth to new technologies. Such is the case of artificial neural networks, that over the years, they have given various solutions to the industry. Designing and implementing intelligent systems has become a crucial factor for the innovation and development of better products for society. In our paper, we decided to design a neural network (Fig. 7) [21] that processes generated input data for classifying the events in two classes: fall and not fall. For the input data, the following sets of parameters are used for falling recognition:

- Speed: The planar speed is calculated using the formula: Planar speed = distance/time (pixel/s);
- Aspect ratio: The aspect ratio of a person is a simple yet effective feature for differentiating normal standing poses from other abnormal poses. The aspect ratio of the human body changes during a fall. When a person falls, the height and width of his bounding box changes drastically (height/width).
- Angle (degree): Fall angle is the angle of a vertical line through the centroid of an object with respect to the horizontal axis of the bounding box. The centroid (Cx, Cy) is the center of mass coordinates of an object. When a person is falling, the angle is always less than 45 degrees. For every frame, we calculate the fall angle (θ), and if θ value is less than 45 degrees, we confirm that the person is falling.
- Heart rate: Measures the number of heart beats per second (bpm).

We generated 5000 such sets of values, each having correspondence with real-life situations that can occur. We chose to have 2500 situations corresponding to non-fall situations and 2500 corresponding to fall situations. We decided that a fall situation occurs when we have measures of high speed, low aspect ratio, the angle under 45 degrees, and a heart rate close to normal. Having four types of data transmitted from the cameras and the sensor made our network have 4 inputs. With each frame, we receive a new set of data, which represent a new pattern that needs to be trained or tested by the network. For our network, we decided to implement a (MLP). The network is a feed-forward network with Back Propagation. The output consists of 1 element that can be 1 or 0. The value one has been assigned to the fall situation class and the value 0 correspond to the situation of non-fall. The training process allowed the neural network to automatically identify the regions in the input pattern space that contained the fall data points. For all of the simulations, we chose a sigmoid transfer function for the hidden layer and a linear transfer function for the output layer.

The network settings are presented below:

- Learning rule: BackPropagation
- Layers = 2
- Inputs = 4
- Hidden Neurons = 8
- Output Neurons = 1
- Transfer function hl = "logsig"
- Transfer function ol = "linear"
- Error function: MSE
- Goal = 0.01
- Max epochs = 10.000
- Momentum Coefficient = 0.01
- Learning method: gradient descent or Levenberg-Marquardt
- Learning rate = 0.01

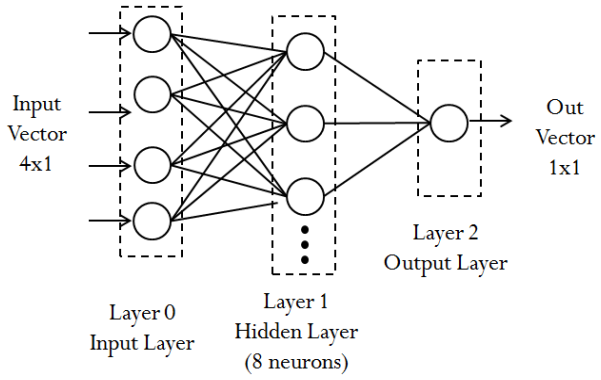



Figure 7. Neural network architecture.

## IV. RESUTLS AND DISCUSSION

### A. Preprocessing the training data

In principle, we can just use any raw input-output data to train our networks. However, in practice, it often helps the network to learn appropriately if we carry out some preprocessing of the training data before feeding it to the network. The ranges of our input are:

- Speed (pixel/s): from 90 to 700;
- Aspect ratio (height/width): from 0.15 to 6 ;
- Angle (degree): from 0 to 90;
- Heartbeat (bpm): from 70 to 200.

All inputs were normalized to [-3, 3]. The range of the output is [0, 1] (sigmoid function), so there is no need for scaling. The number of patterns that we used was 2500 for each class. Because, in our approach, we used batch training, there was no need for shuffling the order of the input patterns. We can observe that the Levenberg-Marquardt method [22] shows better results.

TABLE I. RESULTS FOR DIFFERENT LEARNING METHODS

| **Learning method** | **Num of Epochs** | **Goal** | **Performance** | | | |
|---|---|---|---|---|---|---|
| | | | Training (lastepoch) | Validation | | |
| | | | | Sensitivity | Specificity | Accuracy |
| traingdx | 56 | 0.1 | 0.0979 | 87.71 | 75.43 | 81.57 |
| trainlm | 3 | 0.1 | 0.0358 | 99.53 | 93.17 | 96.59 |
| traingdx | 10000 | 0.01 | 0.0156 | 92.37 | 96.12 | 98.07 |
| **trainlm** | **7** | **0.01** | **0.00307** | **100** | **98.29** | **99.15** |

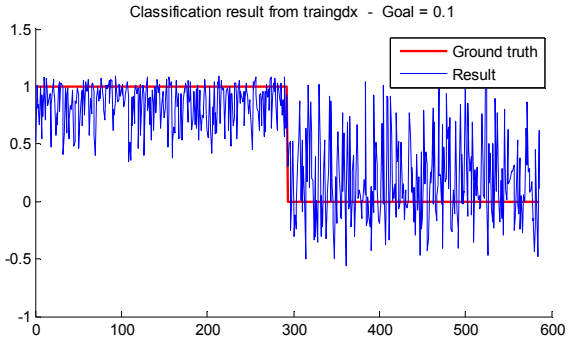



Figure 8. Classification result from traingdx (Goal = 0.1).

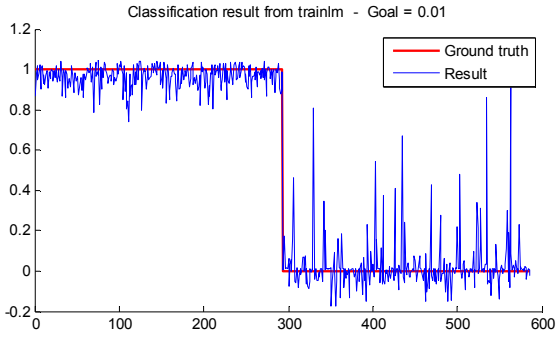



Figure 9. Classification result from trainlm (Goal = 0.01).

### B. Choosing the Initial Weights

The gradient descent learning algorithm treats all the weights in the same way, so if we start them all off with the same values, all the hidden units will end up doing the same thing, and the network will never learn properly. For that reason, we generally start off all the weights with small random values. Usually, we take them from a flat distribution around zero [-smwt, +smwt], or from a Gaussian distribution around zero with standard deviation smwt. Choosing a good value of smwt can be difficult. Generally, it is a good idea to make it as large as you can without saturating any of the sigmoid. We usually hope that the final network performance will be independent of the choice of initial weights, but we need to check this by training the network from a number of different random initial weight sets. The initial weights were generated using the functions:

- rands: return values between -1 and 1;
- midpoint: is a weight initialization function that sets weight (row) vectors to the center of the input ranges.
- initzero: initializing all the weight to zero.
- Nguyen-Widrow: the standard function in Matlab for initializing the weights

The best performance is obtained with the Nguyen-Widrow function [23].

TABLE II. DIFFERENT WAYS OF INITIALIZING THE WEIGHTS

| **Initial Weights** | **Num of Epochs** | **Performance** | | | |
|---|---|---|---|---|---|
| | | Training(lastepoch) | Validation | | |
| | | | Sensitivity | Specificity | Accuracy |
| rands | 4 | 0.00593 | 100 | 98.39 | 99.19 |
| Midpoint | 11 | 0.00849 | 100 | 98.93 | 99.46 |
| initzero | 11 | 0.00849 | 100 | 98.93 | 99.46 |
| Nguyen-Widrow | 7 | 0.00307 | 100 | 96.59 | 98.29 |

 ISBN: 978-1-61208-237-0 

### C. *Choosing the learning rate*

Choosing a good value for the learning rate η is constrained by two opposing facts:

- If η is too small, it will take too long to get anywhere near the minimum of the error function
- If η is too large, the weight updates will over-shoot the error minimum and the weights will oscillate, or even diverge.

The best performance was obtained for the learning rate of 0.001 (Table III). For the types of learning rates—fixed and adaptive learning rates—best performance was achieved with adaptive learning rate (Table IV). We tried to vary the number of hidden units from our network. We tried with 4, 8, and 12 hidden neurons and observed that the best performance was obtained for 4 neurons in the hidden neurons (Table V). The best performance was achieved with 4 neurons. The number of layers was also modified searching for the optimal network. It turned out to be that one hidden layer was enough for achieving good performances (Table VI). The best performance was obtained for 1 hidden layer.

TABLE III. RESULTS OBTAINED FOR DIFFERENT LEARNING RATES

| **Learning rate** | **Num of Epochs** | **Goal** | Training (lastepoch) | Validation | | |
|---|---|---|---|---|---|---|
| | | | | Sensitivity | Specificity | Accuracy |
| **0.001** | **14** | **0.01** | **0.00979** | **100** | **98.36** | **99.10** |
| 0.01 | 10 | 0.01 | 0.00985 | 100 | 97.75 | 98.55 |
| 0.1 | 11 | 0.01 | 0.0098 | 100 | 98.29 | 99.15 |
| 10 | 25 | 0.01 | 0.0099 | 98.61 | 98.95 | 98.78 |

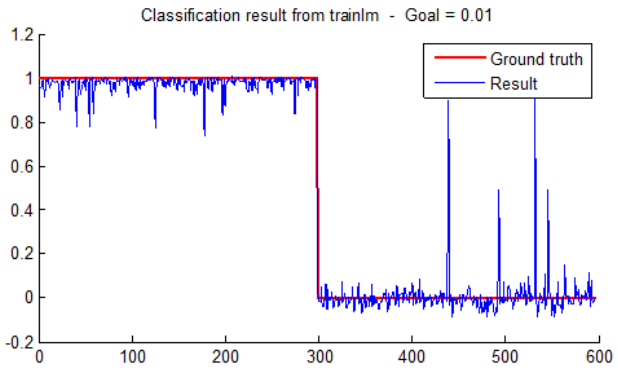



Figure 10. Classification result from trainlm (Lr = 0.001).

TABLE IV. RESULTS FOR DIFFERENT TYPES OF LEARNING RATES

| **Learning rate 0.1** | **Num of Epochs** | **Goal** | Training (lastepoch) | Validation | | |
|---|---|---|---|---|---|---|
| | | | | Sensitivity | Specificity | Accuracy |
| Fixed | 20 | 0.01 | 0.00911 | 100 | 97.56 | 98.12 |
| **Adaptive** | **14** | **0.01** | **0.00921** | **100** | **98.29** | **99.15** |

TABLE V. RESULTS FOR DIFFERENT TYPES OF HIDDEN UNITS

| **#of Neurons** | **Num of Epochs** | **Performance** | | | |
|---|---|---|---|---|---|
| | | Training (lastepoch) | Validation | | |
| | | | Sensitivity | Specificity | Accuracy |
| 4 | 14 | 0.00923 | 100 | 98.29 | 99.15 |
| 8 | 11 | 0.00959 | 100 | 98.29 | 99.15 |
| 12 | 12 | 009610. | 100 | 97.27 | 98.63 |

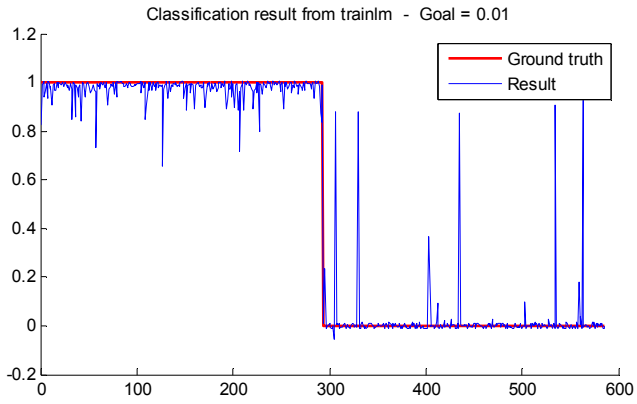



Figure 11. Classification result from trainlm (4 neurons 2 layers).

TABLE VI. RESULTS FOR DIFFERENT NUMBER OF HIDDEN LAYERS

| **# of Hidden Layers** | **Num of Epochs** | **Goal** | **Performance** | | | |
|---|---|---|---|---|---|---|
| | | | Training (lastepoch) | Validation | | |
| | | | | Sensitivity | Specificity | Accuracy |
| **1** | **11** | **0.01** | **0.00959** | **100** | **98.29** | **99.15** |
| 3 | 9 | 0.01 | 0.0097 | 100 | 96.55 | 98.01 |
| 5 | 12 | 0.01 | 0.09612 | 100 | 96.27 | 97.54 |

 ISBN: 978-1-61208-237-0 

### D. *Fall detection*

For fall incidents, the inputs (speed, ratio, angle, and heart rate) have to satisfy certain thresholds:

- 650 < Speed < 700
- 0.15 < Ratio < 1.5
- 0 < Angle < 45
- 70 < Heart rate < 110

TABLE VII. TESTING THE NETWORK

| Sensitivity | Specificity | Accuracy |
|---|---|---|
| **100** | 97.58 | **99.15** |

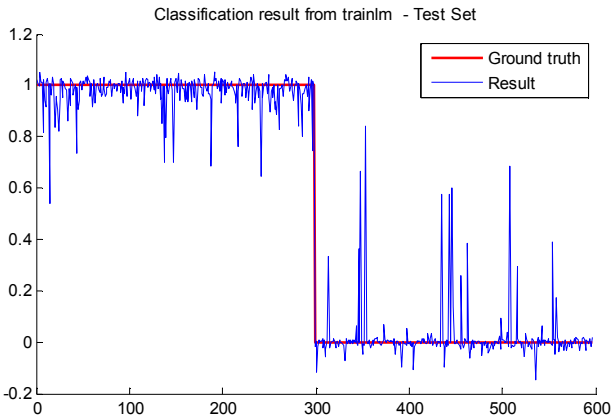



Figure 12. Classification result from trainlm (Test Set).

TABLE VIII. FALL DETECTION

| Speed | Ratio | Angle | Heart rate | Class |
|---|---|---|---|---|
| 666 | 0.46 | 15 | 104 | Fall |
| 689 | 0.99 | 5 | 70 | Fall |
| 698 | 0.19 | 26 | 91 | Fall |
| 650 | 0.56 | 16 | 86 | Fall |
| 674 | 0.42 | 17 | 99 | Fall |
| 583 | 4.4 | 60 | 110 | No Fall |
| 328 | 1.12 | 20 | 90 | No Fall |
| 147 | 0.32 | 45 | 86 | No Fall |
| 423 | 2.03 | 50 | 77 | No Fall |

## V. CONCLUSION AND FUTURE WORK

Fall-related injuries have been among the five most common causes of death amongst the elderly population. Falls represent 38% of all home accidents and cause 70% of deaths in the 75+ age group. Early detection of a fall is an important step in avoiding any serious injuries. An automatic fall detection system can help to address this problem by reducing the time between the fall and arrival of required assistance. In an eldercare context, false alarms can be expensive. Too many false alarms could result in a loss of trust, or worse, loss of use of the system. However, missing a single fall is the worst-case scenario. Identifying an acceptable false-alarm rate and understanding the conditions in which many false alarms occur is of vital use for the long-term success of an automated system. Healthcare video surveillance systems are a new and promising solution to improve the quality of life and care for the elderly, by preserving their autonomy and generating the safety and comfort needed in their daily lives. This corresponds to the hopes of the elderly themselves, their families, the caregivers, and the governments. The positive receptivity for video surveillance systems suggests that this technology has a bright future for healthcare and will advantageously complement other approaches (e.g., fixed or wearable sensors, safer home modifications, etc.) by overcoming many of their limitations. Better performances and results can be obtained by implementing neural network architecture when different methods of acquiring data are combined (wearable devices + webcam images). The presented work may be extended and enhanced, in a later phase, to include multiple webcams and other parameters that could help to address this problem by reducing the risk of false alarms and improving the time between the fall and the alarm.

## VI. REFERENCES

[1] N. Noury, T. Hervé, V. Rialle, G. Virone, and E. Mercier, “Monitoring behavior in home using a smart fall sensor and position sensors,” in IEEE-EMBS. Microtechnologies in Medicine & Biology, October, Lyon-France, 2000, pp. 607–610.

[2] N. Noury, A. Fleury, P. Rumeau, A. K. Bourke, G. O. Laighin, V. Rialle, and J. E. Lundy, “Fall Detection - Principles and Methods,” 29th Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society, Lion (France), August 2007, pp. 1663–1666.

[3] N. Noury, A. Fleury, P. Rumeau, A. K. Bourke, G. O. Laighin, V. Rialle, and J. E. Lundy, “Fall Detection - Principles and Methods,” 29th Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society, Lion (France), August 2007, pp. 1663–1666.

[4] A. Yamaguchi, “Monitoring behavior in home using positioning sensors” Int. Conf. IEEE-EMBS, Hong-Kong, 1998; 1977-79.

[5] M. J. Mathie, A. C. F. Coster, N. H. Lovell and B. G. Celler, Accelerometry: Providing an Integrated, Practical Method for Long-term, Ambulatory Monitoring of Human Movement, Journal for Physiological Measurement (IOPScience), Vol. 25, 2004.

[6] F. Bianchi, S. J. Redmond, M. R. Narayanan, S. Cerutti and N. H. Lovell, Barometric Pressure and Triaxial Accelerometry Based Falls Event Detection, IEEE Transactions on Neural Systems and Rehabilitation Engineering, Vol. 18, pp. 619-627, 2010.

[7] M.A. Estudillo-Valderrama, L.M. Roa, J. Reina-Tosina and D. Naranjo-Hernandez, Design and Implementation of a Distributed Fall Detection System - Personal Server, IEEE Transactions on Information Technology in Biomedicine, Vol. 13, pp. 874-881, 2009.

 ISBN: 978-1-61208-237-0 

[8] T. Tamura, T. Yoshimura, M. Sekine, M. Uchida and O. Tanaka, A Wearable Airbag to Prevent Fall Injuries, IEEE Transactions on Information Technology in Biomedicine, Vol.13, pp. 910-914, 2009.

[9] J. Chen, K. Kwong, D. Chang, J. Luk, and R. Bajcsy, Wearable Sensors for Reliable Fall Detection, 27th IEEE Annual Conference of Engineering in Medicine and Biology (EMBS), pp. 3551-3554, 2005.

[10] C. C. Wang, C. Y. Chiang, P. Y. Lin, Y. C. Chou, I. T. Kuo, C. N. Huang and C. T. Chan, Development of a Fall Detecting System for the Elderly Residents, 2nd IEEE International Conference on Bioinformatics and Biomedical Engineering, ICBBE, pp. 1359-1362, 2008.

[11] X. Zhuang, J. Huang, G. Potamianos and M. Hasegawa-Johnson, Acoustic Fall Detection Using Gaussian Mixture Models and GMM Super-Vectors, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.69-72, 2009.

[12] M. Alwan , P. J. Rajendran, S. Kell, D. Mack , S. Dalal, M. Wolfe, and R. Felder, A Smart and Passive Floor-Vibration Based Fall Detector for Elderly, IEEE International Conference on Information & Communication Technologies (ICITA), pp. 1003-1007, 2006.

[13] Y. Zigel, D. Litvak and I. Gannot; A Method for Automatic Fall Detection of Elderly People Using Floor Vibrations and Sound Proof of Concept on Human Mimicking Doll Falls, IEEE Transactions on Biomedical Engineering, Vol. 56, pp. 2858-2867, 2009.

[14] B. U. Toreyin, E. B. Soyer, I. Onaran, and A. E. Cetin, Falling Person Detection Using Multi-sensor Signal Processing, 15th IEEE Signal Processing and Communications Applications Conference, SIU & EURASIP Journal on Advances in Signal Processing, vol. 8, 2007 & 2008.

[15] R. Cucchiara, A. Prati, and R. Vezzani, “A multi-camera vision system for fall detection and alarm generation,” Expert Syst. J., vol. 24(5), 2007, pp. 334-345.

[16] C. H. Nait and S. J. McKenna, “Activity summarisation and fall detection in a supportive home environment,” Int. Conf. on Pattern Recognition (ICPR), 2004.

[17] J. Tao, M. Turjo, M. F. Wong, M. Wang and Y. P. Tan: Fall Incidents Detection for Intelligent Video Surveillance, Fifth IEEE International Conference on Information, Communications and Signal Processing, pp. 1590-1594, 2005.

[18] H. Foroughi, B. S. Aski, and H. Pourreza, Intelligent Video Surveillance for Monitoring Fall Detection of Elderly in Home Environments, 11th IEEE International Conference on Computer and Information Technology (ICCIT), pp. 219-224, 2008.

[19] S. G. Miaou, P. H. Sung and C. Y. Huang, A Customized Human Fall Detection System Using Omni-Camera Images and Personal Information, 1st Trans-Disciplinary Conference on Distributed Diagnosis and Home Healthcare (D2H2), pp. 39-42, 2006.

[20] C Rougier, J Meunier, A St-Arnaud, and J Rousseau, Robust Video Surveillance for Fall Detection Based on Human Shape Deformation, IEEE Transactions on Circuits and Systems for Video Technology (CSVT), Vol. 21, pp. 611-622, 2011.

[21] G. Miller, P. Todd, and S. Hedge, " Designing neur al networks using genetic algorithms," Proceedings of the International Confe rence on Genetic Algorithms 1989.

[22] H. Demuth, M. Beale, and M. Hagan. “Neural Network Toolbox 5, User’s Guide,” Version 5, The MathWorks, Inc., Natick, MA, revised for version 5.1.

[23] H. Demuth and M. Beale. “Neural Network Toolbox, User’s Guide,” Version 4, The MathWorks, Inc., Natick, MA, revised for version 4.0.4.

 ISBN: 978-1-61208-237-0 

# Application of the Stacking Regression in Context of the Soil-Water Modelling

Milan Cisty, Juraj Bezak, Jana Skalova
Department of Land and Water Resources Management
Slovak University of Technology
Bratislava, Slovakia
milan.cisty@stuba.sk, juraj.bezak@stuba.sk, jana.skalova@stuba.sk

***Abstract*— Modelling water transport in soil has become an important tool in simulating hydrological systems and agricultural productivity. Some of the data necessary for this modelling are usually easily available in competent institutions, but hydraulic soil properties (namely water retention curve) are only rarely easily available. The aim of this paper is to contribute to solving this deficit by evaluating so-called pedotransfer functions by data-driven modeling methods. Multi-linear regression, artificial neural networks, support vector machines and combination of these three methods in stacking model was evaluated. Work proves that stacking model yields more precise results than individual data-driven models and could be suggested for soil water modelling.**

***Keywords-soil-water modelling; pedotransfer function; data-driven model; stacking.***

## I. INTRODUCTION

Modelling water transport in soil has become an important tool in simulating hydrological systems and agricultural productivity. Models that deal with the transport of water and solutes range in scale from physically-based, fully distributed catchment models to the land parameterization scheme of general circulation models. Their practical application includes, e.g., systematic estimation of soil-water status to determine both the appropriate amounts and timing of irrigation. As is usual in any modelling, it depends on knowledge of the input data which are needed for the numerical simulations. Some of the data necessary for modelling water transport in soil (meteorological, climatic, hydrological or crop characteristics) are usually easily available in competent institutions, but hydraulic soil properties are only rarely easily available. These characteristics are therefore a key problem in the numerical simulation of a soil-water regime, and a modeller must deal with the problem of how to obtain them. The aim of this paper is to contribute to solving this task.

The water retention curve is one of the main soil hydraulic properties, which is used in simulating the water regime of soils. It represents the relationship between the water content and the soil's water potential (the potential energy of water per unit volume, which quantifies the tendency of water to move from one place to another). This curve is characteristic of different types of soil. It is used to predict a soil's water storage, the water supply to plants, and for other tasks in soil water modelling. A relatively large number of works have appeared in the past which were devoted to determining the water retention curve from more easily available soil properties such as particle size distribution, dry bulk density, organic C content, etc., e.g., [1][2][3]. In this context, Bouma [4] introduced the term "pedotransfer function" (PTF), which he described as "translating data that we have (soil survey data) into data that we need (soil hydraulic data)." In this paper, we will focus on point estimation methods of the PTFs, which follow the direct approach by estimating the water content at predetermined pressure heads.

Besides the application of the standard regression methods for solving this task, data-driven techniques appeared in the scientific literature in the second half of the previous decade as a tool for solving regression tasks in developing PTFs. However, there is no overall best data-driven technique which could be used in building hydrology models, because their suitability depends on the details of the problem, the data structure, the input data used, etc. For this reason various data-driven techniques are compared in this case study.

In the following part of the paper, the methods used in this study are briefly explained. Then the data acquisition and preparation is presented. In the "Results" part, the settings of the experimental computations are described in detail, and the "Conclusion" of the paper evaluates these experiments on the basis of the statistical indicators.

## II. METHODS USED TO FIT THE PEDOTRANSFER FUNCTIONS (PTFS)

The first approach for modelling the PTFs used in this paper is the application of *artificial neural networks* (ANNs). Briefly summarized, a neural network consists of input, hidden and output layers, all containing neurons. The number of nodes in the input layer (e.g., the soil's bulk density, the soil's particle size data, etc.) and output layer (various soil properties) correspond to the number of input and output variables of the model. So-called "learning" or "training" involves adjustment of the coefficients (i.e., the synaptic connections that exist between the neurons or weights), which are used for the transformation of the inputs to the outputs. For that reason, an important step in developing an ANN model is the training (computing) of its weight matrix. A type of ANN known as a multi-layer perceptron (MLP), which uses a back-propagation training algorithm, was used for generating the PTFs in our study. The training process was performed by the back propagation

 ISBN: 978-1-61208-237-0 

training algorithm. The basic information about the application of an ANN to regression problems is available in the literature and is well known, so we will not provide a more detailed explanation here.

The basic idea behind the second methodology applied – *support vector regression regression* - is to project the input data by means of kernel functions into a higher dimensional space called the feature space, where a linear regression can be performed for an originally nonlinear problem which is to be solved. The results of the regression are then mapped back to the input space. The kernel trick is a mathematical tool which can be applied to any algorithm which solely depends on the dot product between two vectors. Wherever a dot product is used, it is replaced by a kernel function. However, because kernels are used, the function never needs to be explicitly computed. This is highly desirable, because this higher-dimensional feature space could be unfeasible to compute.

The next important concept in SVM methodology is to fully ignore small errors (by introducing the variable ε, which defines what the "small" error is) to make the regression task dependent on a smaller number of inputs than were given in the original task, which makes the methodology much more computationally treatable. These crucial vectors of the inputs are called the support vectors.

In an ε-SVM regression [5], the goal is to find a function f(x) that at most has an ε deviation from the actually obtained targets $y_i$ (or f(x)) for the training data:

$$f(x) = w.\Phi(x) + b \quad w \in X, b \in R \tag{1}$$

where *f(x)* is the model's output, and input *x* is mapped into a feature space by a nonlinear function Φ(*x*) with the weight vector *w* and bias *b*.

The goal of a regression algorithm is to fit a flat function to the data points. "Flatness" means that one seeks a small w. One way to ensure this flatness is to minimize the norm, i.e. $\|w\|^2$. Thus, the regression problem can be written as a quadratic optimization problem:

$$\text{minimize} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \tag{2}$$

$$\text{subject to: } y_i - (w.\Phi(x) + b) \le \varepsilon + \xi_i$$
$$(w.\Phi(x) + b) - y_i \le \varepsilon + \xi_i^*$$
$$\xi_i, \xi_i^* \ge 0$$

where ξi, ξi* are slack variables that specify the upper and lower training errors, subject to an error tolerance ε (soft margin), and C is a positive constant that determines the degree of the penalized loss when a training error occurs. In Equation system (2), the first term of the objective function indicates the model's complexity, and the second term is the empirical risk. That is why this objective function simultaneously minimizes both the empirical risk and the model's complexity; the trade-off between these two goals is controlled by parameter C. An important characteristic of SVMs as a consequence of this fact is that a better ability to generalize could be expected, compared, e.g., with ANNs (the better results for the data which were not used for building the model), because unnecessarily complex models usually suffer from over-fitting.

The third approach applied is to build the ensemble of the data-driven models, is so-called stacking model based on the base learners described above. Approach evaluated in this study is to generate ensemble model by applying different learning algorithms contained in the ensemble (other ensemble schemes, e.g., bagging or additive regression usually consist of one type of model). This approach to ensemble modelling deals with the task of training a meta-level base model to combine the predictions of multiple base-level base models. In other words, stacking introduces the concept of 1) base models and 2) a meta model, which computes the final results and replaces the averaging procedure used, e.g., in bagging. In such a way, stacking tries to learn which base models are more reliable than others, using mentioned meta-model (it could be a different algorithm than the base models) to discover how best to combine the output of the base models to achieve the final results. The results of the base learners are de facto new data for another learning problem, and in the second step a meta learning algorithm is employed to solve this problem. Variant of this approach is described on Fig. 1, where SVM is abbreviation for support vector machines, ANN is artificial neural network and MLR is multiple-linear regression – which are models from which stacking ensemble model consist from in our study.

```
D = {(x1, y1), (x2, y2) ... (xm, ym)};   % input data set, xi is vector
Base learning algorithms: A = {SVM, ANN, MLR}
Meta learner: SVM
TRAINING:
For j=1:3 do        % Train a base learner hj by applying
    hj = Lj(D)      % corresponding algorithm Aj
 end
For i=1: m
    For j=1:3
    zit = hj(xi)    % Use hj to predict the training example xi
    end
D' = {((zi1, zi2, zi3),yi)}
                        end;
h'=L (D')           % Train the meta learner h' (SVM)
                    % applying it to data set D'
                    % cross-validation was used
                    % to optimize its parameters
TESTING:            H (x) = h' (h1(x), h2(x), h3 (x))
```

Figure 1. Stacking algorithm scheme

## III. STUDY AREA AND DATA COLLECTION

The data used in this study were obtained from a previous work [6]. An area of the Zahorska lowland was selected for testing the methods described. A total of 226 soil samples was taken from various localities in this area.

The soil samples were air-dried and sieved for a physical analysis. A particle size analysis according to four grain

categories was performed utilizing Cassagrande's methods. Category I means the percentages of the clay (diameter < 0.01 mm), category II - silt (0.01–0.05 mm), category III - fine sand (0.05–0.1 mm) and category IV - sand (0.1–2.0 mm). The dry bulk density, particle density, porosity and saturated hydraulic conductivity were also measured on the soil samples. The points of the drying branches of the PTFs for the pressure head values of -2.5, -56, -209, -558, -976 and -3060 cm were estimated using overpressure equipment (set for pF-determination with ceramic plates).

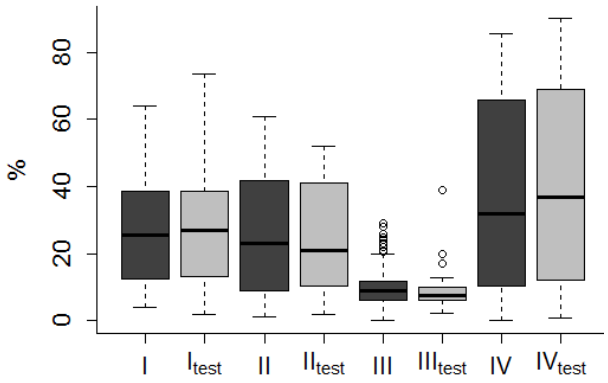



Figure 2. Comparison of grain categories (I, II, III, IV) in the training and testing data

A full database of the 226 samples and their properties were used for creating the input data for the modelling from which the training and testing subsets of the data were produced. The training data consist of 181 data samples and test data from 45 data samples. Statistically similar data should be in both data subsets; this condition is visualized by the boxplots on Fig. 2. In this figure, I, II, III and IV are grain categories in training set of data and the same identification with the subscript "test" is used for the test set. From this evaluation, it can be seen that category III will probably have the lowest impact on the pedotransfer function evaluation, but it will be included in the input data, anyway.

## IV. RESULTS

### A. Artificial neural networks

The first approach applied to determining the water retention curves in the presented work was the *artificial neural networks* methodology (ANN). In this work a multilayer perceptron with 4, 5, and 6 neurons in the hidden layer was tested; an ANN with 5 neurons in the hidden layer was finally chosen for the final neural network model used in the comparisons (it has the best results). A neuron with a hyperbolic tangent activation function was used in the hidden layer and a linear activation function in the output layer. The Levenberg-Maquardt method was used in the context of the back propagation method. The networks were trained to compute the water content at the pressure head value $h_w$ = -2.5, -56, -209, -558, -976, -3060 cm. The "hold-out" method was used for stopping the ANN to avoid overtraining, and this "hold-out" sample was 20% of the data from the training set.

Then the testing dataset was computed with the trained ANNs. The results with the regression coefficients are summarized in Table I. Three variants of the ANN with different hidden layer sizes and SVM are evaluated ($h_w$ - pressure head, H4 – H6 is the number of neurons in the hidden layer).

### B. Support vector machines

For a comparison with the ensemble approach, the given regression problem was also solved using *support vector machines (SVM)*. The estimation of the practical steps of the SVM regression are as follows: 1) selecting a suitable kernel and the appropriate kernel's parameter; 2) specifying the $\varepsilon$ parameter (2); and 3) specifying the capacity $C$ (2).

The radial basis function was chosen as the kernel function on a trial and error basis for which parameter $\gamma$ should be specified. The cross-validation methodology with 10 folds was used for finding the mentioned parameters of the SVM model.

In the training phase, SVM models for computing the water content for the pressure head values of $h_w$ = -2.5, -56, -209, -558, -976 and -3060 cm were created (on the basis of the particle size distribution as in the multi-linear regression case). Then the testing dataset was computed with the models obtained, and the final results were summarized with the help of the regression coefficients in Table I. The calculations of the SVM were performed using the LIBSVM library developed by Chang and Lin [8].

TABLE I. CORRELATION OF THE MODEL'S RESULTS WITH THE ACTUAL VALUES OF THE PTFS.

| $h_w$ [cm] | ANN – H4 | ANN – H5 | ANN – H6 | SVM | Stacking |
|---|---|---|---|---|---|
| -2.5 | 0.874 | 0.883 | 0.879 | 0.872 | 0.881 |
| -56 | 0.846 | 0.857 | 0.849 | 0.872 | 0.905 |
| -209 | 0.874 | 0.874 | 0.866 | 0.898 | 0.898 |
| -558 | 0.866 | 0.872 | 0.873 | 0.896 | 0.904 |
| -976 | 0.853 | 0.859 | 0.860 | 0.882 | 0.885 |
| -3060 | 0.833 | 0.846 | 0.852 | 0.880 | 0.890 |

### C. Stacking ensemble model

Stacked generalization (or stacking) is a way of combining the multiple models used in this work; it introduces the concept of a meta learner, the task of which is to combine the predictions of multiple base-level learners. In this work ANN, SVM and multi-linear regression were used as base learners. For stacking an ANN with four hidden units, the hyperbolic tangent activation function in the hidden layer and the linear function in the output layer were selected. The SVM as a base learner was not optimized (we did not include the parameter searching into the computational scheme), and the radial basis function kernel was chosen to maintain the nonlinearity. Parameter C was set as equal to the range of the output values [9], and parameter $\varepsilon$ in the $\varepsilon$-insensitive loss function was set to its default value 0.1 [7]. However, support vector regression was also used in the stacking model as a meta-learner. The SVM built a stacked model on top of the predictions of the base learners.

 ISBN: 978-1-61208-237-0 

In this case, its parameters γ, C and ε were optimized by tenfold cross-validation. The schema of this approach is in Fig. 1. The results with the regression coefficients are summarized in Table I.

From the results expressed by the correlation coefficient in Table I it can be seen that the stacking ensemble methodology evaluated in this case study give better results than when individual learners are used solo (ANN, SVM). Also, the application of the linear regression to the development of the pedotransfer function was evaluated in this work. Its main advantages are simplicity of implementation and interpretability; on the other hand, its shortcoming is that if the relationship between the input and output cannot be reasonably approximated by a linear function, the model will give poor predictions. This was also confirmed in this case study; the results in Table II, which were obtained by multi-linear regression, are generally worse than the results of the ANN and SVM alone (Table I) and significantly worse than the results obtained by stacking ensemble methodology evaluated in this work.

A more detailed evaluation of the various data-driven methods applied in this work is presented in Table II. For practical reasons (the limited extent of this paper) it is restricted only to an evaluation of the prediction of the water content for the pressure head value $h_w$ = -3060 cm. The results for the other pressure heads are similar from point of view of effectiveness of the algorithms used. In Table II the mean error (ME), mean absolute error (MAE), mean square error (MSE), root mean square error (RMSE), normalized root mean square error (NRMSE), percent bias (PBIAS), correlation coefficient (r), maximal difference between the simulated and actual values (maxD) and the minimal difference between the simulated and actual values (minD) are evaluated. The names of the models in the heading of Table II are clear from their abbreviations. From this analysis, it is evident that it is worthwhile to pay attention to the development and choice of the proper regression model when evaluating the pedotransfer function, because it can be seen that a relatively big difference is between the effectiveness of the worst performing model (MLR) and the best model. The models are ordered in columns according to their quality from worst to best.

TABLE II. EVALUATION OF THE VARIOUS MODELS FOR PREDICTION OF THE WATER CONTENT AT $h_w$ = 3060 CM BY DIFFERENT STATISTICS

| | **MLR** | **ANN** | **SVM** | **Stacking** |
|---|---|---|---|---|
| ME | -0.39 | -0.53 | -1.02 | -0.45 |
| MAE | 4.21 | 3.75 | 3.31 | 3.32 |
| MSE | 30.40 | 25.98 | 21.44 | 19.73 |
| RMSE | 5.51 | 5.10 | 4.63 | 4.44 |
| NRMSE | 57.50 | 53.10 | 48.30 | 46.30 |
| PBIAS | -1.80 | -2.40 | -4.70 | -2.10 |
| r | 0.82 | 0.85 | 0.88 | 0.89 |
| maxD | 9.83 | 7.12 | 6.24 | 6.40 |
| minD | -15.27 | -15.02 | -12.91 | -11.88 |

## V. CONCLUSION AND FUTURE WORK

This paper proposed and evaluated data-driven models for the development of pedotransfer functions for the point estimation of the soil-water content for six pressure head values $h_w$ from the basic soil properties (particle-size distribution, bulk density). The ensemble data-driven model (stacking) was compared to single data-driven models (artificial neural networks and support vector machines) and to a multiple linear regression methodology. The accuracy of the predictions was evaluated by the correlation coefficient between the measured and predicted parameter values and by other statistics. From the results obtained it was proved that nonlinear data-driven methods work significantly better than multi-linear regression and that even better results were obtained by using data-driven methods in an ensemble context.

However, several issues remain to be addressed by further research. Although in this work stacking performs well, it is not easy to give the reasons for selecting its particular components. This process is subjective in the present state of our knowledge (on the basis of trial and error), which should be improved in the future.

## ACKNOWLEDGMENT

This work was supported by the Slovak Research and Development Agency under Contract No. LPP-0319-09, and by the Scientific Grant Agency of the Ministry of Education of the Slovak Republic and the Slovak Academy of Sciences, Grant No. 1/1044/11 and 1/0243/11.

## REFERENCES

[1] S.C. Gupta and W.E. Larson, “Estimating soil water retention characteristics from particle size distribution, organic matter percentage, and bulk density”, Water Resour. Res. 15, pp. 1633-1635, 1979.

[2] W.J. Rawls, D.L. Brakensiek and K.E. Saxton, “Estimating soil water retention properties”, Trans. ASAE 25, pp. 1316-1320, 1982.

[3] B. Minasny, A.B. McBratney and K.L. Bristow, ”Comparison of different approaches to the development of pedotransfer functions for water retention curves”, Geoderma, 93, pp. 225–253, 1999.

[4] J. Bouma, “Using Soil Survey Data for Quantitative Land Evaluation”, Adv. Soil Sci., 9, pp. 177–213, 1989.

[5] V. Vapnik, ”The Nature of Statistical Learning Theory”, Springer, NY, 1995.

[6] J. Skalova, ”Pedotransfer functions of the Zahorska Lowland soils and their application to soil-water regime modelling”, Faculty of Civil Engineering STU Bratislava, (in Slovak), 2001.

[7] I.H. Witten, E. Frank and M.A. Hall, “Data mining”, Morgan Kaufmann Publishers, 2011.

[8] C.Ch. Chang and C.J. Lin, ”A library for support vector machines”, 2001. http://www.csie.ntu.edu.tw/~cjlin/papers/libsvm.pdf [retrieved: August, 2012]

[9] V. Cherkassky and Y. Ma, ”Practical selection of SVM parameters and noise estimation for SVM regression”, Neural Networks 17(1), pp. 113-126, 2004.

 ISBN: 978-1-61208-237-0 

# Expert System used for Power Quality and Environmental Impact Assessment

Doru Vatau
Electrical Power Engineering Department
"Politehnica" University of Timisoara, Romania
300223 Timisoara, Bd. V. Parvan, Nr. 2, Timis
doru.vatau@et.upt.ro

*Abstract*—**This paper presents an expert system used for power quality monitoring, transferred from power generation sector up to power delivery within the Romanian power market. First of all, the system allows the analysis of methods and means for making all maintenance works within substations, belonging to TRANSELECTRICA S.A., the National Company for Power Transmission, at high efficiency. Secondly, the system allows the high voltage facilities environmental impact assessment. Using a fuzzy logic based algorithm, it provides efficient measures in order to ensure the environmental parameters' quality, both local and regional. As a case study, a representative power substation within the Romanian Power System (Brasov Substation) is used.**

***Keywords-expert system; fuzzy logic; power quality; monitoring; environment; power market.***

## I. INTRODUCTION

Electricity is one of the greatest discoveries of mankind. It is now used in almost all areas of activity: agriculture, industry, medicine, scientific research etc. In electricity, there are, among others, highly topical issues such as:

- Sustainable use of energy resources;
- Quality of electricity supplied;
- Efficient use of generated electricity;
- Reducing the power facilities environmental impact of.

The implementation of a remote control and monitoring system represents a priority for the Romanian Power Grid Transelectrica. To achieve this goal, remote control and monitoring centres have been established at each transmission branch (Figure 1). Currently the following ones are operating: Remote Control and Monitoring Centres in Timisoara and Sibiu.

Timisoara Transmission Subsidiary is operating across four counties: Timis, Arad, Caras-Severin and Hunedoara. Timisoara Transmission Subsidiary contains:

- 400 kV substations: Arad, Mintia, Nada;
- 220 kV substations: Arad, Baru Mare, Calea Aradului, Hasdat, Iaz, Mintia, Otelarie, Paroseni, Pestis, Resita, Sacalaz, Timisoara.

Sibiu Transmission Subsidiary is operating across the following six counties: Alba, Sibiu, Brasov, Mureş, Harghita and Covasna. It contains:

- Brasov, Darste and Iernut 400 kV substation;
- Alba Iulia, Fantanele, Gheorgheni, Ungheni, Iernut 220 kV substations.

This paper presents the structure of an expert system used for power quality monitoring, some experimental results (for Brasov Substation) and advantages obtained by using the expert system. Also the 110 kV, 220 kV and 400 kV power facilities environment impact can be studied.

Using this expert system, optimal upgrading decisions have been able to be taken for all power substation. The additional transmission network losses, due to perturbations affecting the power quality, have been mitigated [1]-[2].

## II. EXPERT SYSTEM USED FOR POWER QUALITY AND ENVIRONMENTAL PARAMETERS MONITORING

The system is composed of multiple devices and software synthetically presented in Table 1.

Electric field is determined by measuring the potential gradient (electric field intensity) in kV / m, using the ICEMENERG gradient meter. It is part of the floating potential measuring type apparatus, the detector being included within the measuring probe. The measuring probe is a plane parallel dipole and is therefore made as a parallel plate probes isolated from them according to the IEC Standard 833 – "Measurement of the industrial frequency electric fields". The ICEMENERG gradient meter, according to the IEC 61786/1998 Standard is part of the single axe sensor measuring instruments, for measuring the human body electric field exposure.

Magnetic field is determined by measuring the maximum induction B in mT, for the points established using Tesla device monitor. The measuring device is part of the magnetic field measurement using a coil probe calibrated in a uniform magnetic field created by a solenoid with a suitable size to ensure the uniformity of the field. Measuring device complies with IEC 61786/1998 – "Measurement of electric and magnetic fields regarding the human exposure. Special requirements for measuring devices and rules".

The power quality analyzer is 7650TM type, considering the current regulations and standards [3]-[7].

The PSTN modem is "U.S. Robotics Courier 56 K Business type for dial-up telephone line switchable. Selected communication speed is 19200 baud / s. This type of modem keeps its settings in case of accidental interruption of the supply voltage.

The data server is an HP personal computer Intel P4, 3 GHz. Due to large volume of data recorded for processing into various statistical forms, capacity is 1024MB RAM, 120GB SATA HDD. Auxiliary power is provided by UPS. Data Server has an LCD monitor 19", a multi colour print A4. Fujitsu Siemens notebook communication with the server system is achieved by connecting the external modem described above, to the analogue telephone circuit.

 ISBN: 978-1-61208-237-0 

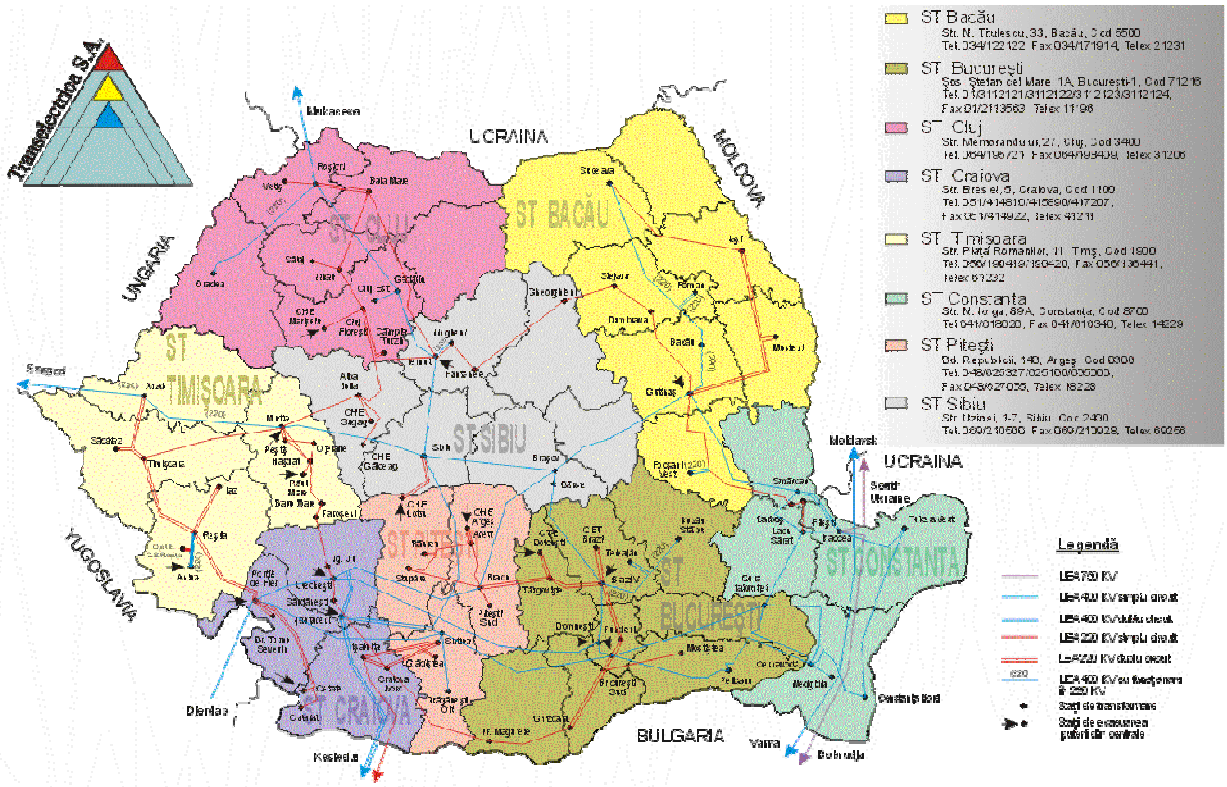



Figure 1. Transelectrica's Remote Control and Monitoring Centres

TABLE 1. EQUIPMENT USED FOR MEASUREMENTS

| No. | Equipment | Made by | Type |
|---|---|---|---|
| 1 | Electric field measurement equipment | ICEMENERG | Gradientmetru |
| 2 | Magnetic field measurement equipment | Conrad Electronic | Tesla Monitor |
| 3 | Portable computer | Fujitsu Siemens | Procesor Pentium 4 |
| 4 | Power quality analyzer | Power Measurement Canada | 7650 ION™ |
| 5 | PSTN modem | US Robotics | Courier 56K Bussines |
| 6 | Data server | Hewlet Packard | Procesor Pentium 4 |
| 7 | Software license | Power Measurement Canada | ION Enterprise 5.5 |

Microsoft Windows NT operating system is used. On request, the data transmitted by the portable computer are automatically saved in a dedicated database. The system allows the external archiving of transmitted data on DVD RW and also their security.

The entire database saved on the server can be accessed on demand to generate the own primary data processing programs, data listing, graphical plots, reports.

The implementation of the system requires the magnetic and electric field measurement within the substations. It is followed by the transmission and storage of the measurements reports at the central point.

The environment impact assessment fuzzy based expert monitoring system provides the substation human operator and the one from the central monitoring point with the following important information:

- Electric field intensity value and magnetic induction values in case of each working area, within the substation territory;
- If the measured values are below or exceed the imposed ones by the regulations;
- A solution set for eliminating or limiting the disturbing effects, having as a goal human beings' health protection that are working within the substation.

Currently, the environment impact assessment system is fully implemented within the Brasov substation. Due to the advantages obtained using this expert monitoring system, its implementation for all the Transelectrica's substations is going to be realized.

 ISBN: 978-1-61208-237-0 

The fuzzy logic based algorithm for environment impact parameters' evaluating within the substations is presented in Figure 2.

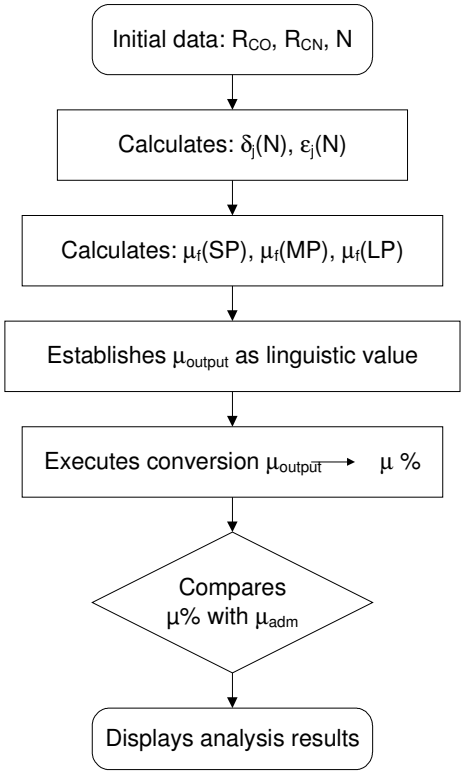



Figure 2. Block diagram of algorithm of establishing the environment impact parameters within the substations ($R_{CO}$ – electric field intensity value or magnetic induction value corresponding to the last maintenance work within the substation installations; $R_{CN}$ – electric field intensity value or magnetic induction value corresponding to measurement moment N; N – measurement number; $\delta_j(N)$ – rated error; $\varepsilon_j(N)$ – variation between two consecutive measurements; SP, MP, LP – fuzzy values; $\mu_f(SP)$, $\mu_f(MP)$, $\mu_f(LP)$ – fuzzy functions; $\mu_{output}$ – final fuzzy function having linguistic value; μ % – output function having numeric value; $\mu_{adm}$ – output function admissible value imposed by regulations)

The data obtained from the periodical measurements N and $R_{CN}$ are used to calculate rated error $\delta_j(N)$.

$$\delta_j(N) = 1 - \frac{R_{CO}}{R_{CN}} \tag{1}$$

as well as the variation of this value between two consecutive measurements $\varepsilon_j(N)$

$$\varepsilon_j(N) = \delta_j(N) - \delta_j(N-1) \tag{2}$$

Placing $\delta_j(N)$ and $\varepsilon_j(N)$ in fuzzy multitudes of Figure 3 and linking them to fuzzy values SP, MP, LP, we can pass to the calculation of the final functions of output to a fuzzy multitude

$$\mu_j(SP) = \underset{x_i}{MAX}\left[MIN\left(\mu(x_i)\right)\right] \tag{3}$$

$$\mu_j(MP) = \underset{x_i}{MAX}\left[MIN\left(\mu(x_i)\right)\right] \tag{4}$$

$$\mu_j(LP) = \underset{x_i}{MAX}\left[MIN\left(\mu(x_i)\right)\right] \tag{5}$$

The calculation of functions $\mu_f(SP)$, $\mu_f(MP)$ and $\mu_f(LP)$ is realized on the basis of Table 2 and Table 3. The following step consists of determining belonging function $\mu_{output}$ from the algorithm presented in Figure 2 in relations

$$\mu_{output} = MAX\left[\mu_f(SP), \mu_f(MP), \mu_f(LP)\right] \tag{6}$$

Through the conversion of the linguistic value of $\mu_{output}$ into a numerical value μ % and comparing the latter with $\mu_{adm}$.

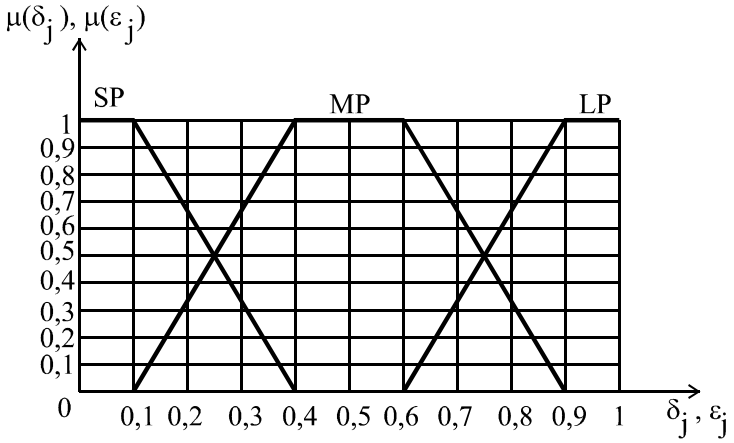



Figure 3. Fuzzy multitudes and belonging functions

 ISBN: 978-1-61208-237-0 

TABLE 2. FUZZY RULES

| $\delta_i$ \ $\varepsilon_j$ | SP | MP | LP |
|---|---|---|---|
| SP | SP | MP | MP |
| MP | MP | MP | LP |
| LP | LP | LP | LP |

TABLE 3. DECISION TABLE OF FUZZY MULTITUDES AND BELONGNING FUNCTIONS

| $x_i$ | input | output | | | $x_i$ | input | output | | |
|---|---|---|---|---|---|---|---|---|---|
| | $(\delta_j, \varepsilon_j)$ | $\mu(x_i, SP)$ | $\mu(x_i, MP)$ | $\mu(x_i, LP)$ | | $(\delta_j, \varepsilon_j)$ | $\mu(x_i, SP)$ | $\mu(x_i, MP)$ | $\mu(x_i, LP)$ |
| $x_1$ | (SP,SP) | 1 | 0.5 | 0 | $x_6$ | (MP,LP) | 0 | 0.5 | 1 |
| $x_2$ | (SP,MP) | 0.5 | 1 | 0.5 | $x_7$ | (LP,SP) | 0 | 0.5 | 1 |
| $x_3$ | (SP,LP) | 0.5 | 1 | 0.5 | $x_8$ | (LP,MP) | 0 | 0.5 | 1 |
| $x_4$ | (MP,SP) | 0.5 | 1 | 0.5 | $x_9$ | (LP,LP) | 0 | 0.5 | 1 |
| $x_5$ | (MP,MP) | 0.5 | 1 | 0.5 | | | | | |

The fuzzy logic based algorithm for evaluating the environment impact parameters within the substations is in patent phase. More details regarding this algorithm are going to be offered ion the future, once the patent phase is finished.

Experimental results from each point of measurement will be presented:

- Schedule normal operation of the substation for electricity conversion;
- Location point for measuring the delimitation of the area dotted element network;
- Continuity in the supply measurement point;
- Maintain the analyzer PQ mounted at the measurement point;
- Graphical representation of weekly analysis of the PQ indicators;
- Representation of numerical analysis for the annual PQ indicators;
- Nonframing indicators analysis within the limits allowed by their reflection in the real and reactive load curves;
- Reports measuring the electric field and magnetic, containing the following data: the test, test name, date of test, technical prescriptions, test results in table;
- Classification analysis / nonframings the values measured in the admissible limits.

The real and reactive energy is flowing in both directions through the measurement point being recorded by the PQ analyzer, but also by the metering system. This meter has been installed by the electric energy remote-metering within the wholesale market, providing a superior accuracy measurement PQ analyzer. The data recorded by the remote-metering of the electric energy within the wholesale market have been used in the following analysis. It has been developed a software dedicated for this analysis and it was represented the monthly real and reactive power evolution. Excel application is used to represent load curves (Figure 5), containing an area that allows the selection of alphanumeric interval analyzed. It also contains a graphical area plotting the time evolution of energy through the network. The user is requested to select the month from the list of options, which will be represented in the chart at the top, power flow evolution. A set of buttons is available for detailed analyses that allow the change of the interest area (green rectangle in the chart above), representing it within the chart at the bottom. This is the effect of the "magnifying glass" that allows the simultaneous observation of the development of both monthly and the interest area.

Starting from the in-depth analyses of the recorded events, the effective causes that are leading to power quality indicators mitigation have to be determined. Also, this analysis has to be correlated with the operation data from the substations and electric networks including the ones archived in SCADA systems. The power quality indicators' admissible limits' overpass analysis has been performed only within the Brasov substation measuring point. This case study has been chosen due to the fact that the supplied voltage magnitude had a spectacular evolution.

## 3. EXPERIMENTAL RESULTS

Several experimental determinations have been performed within the following locations:

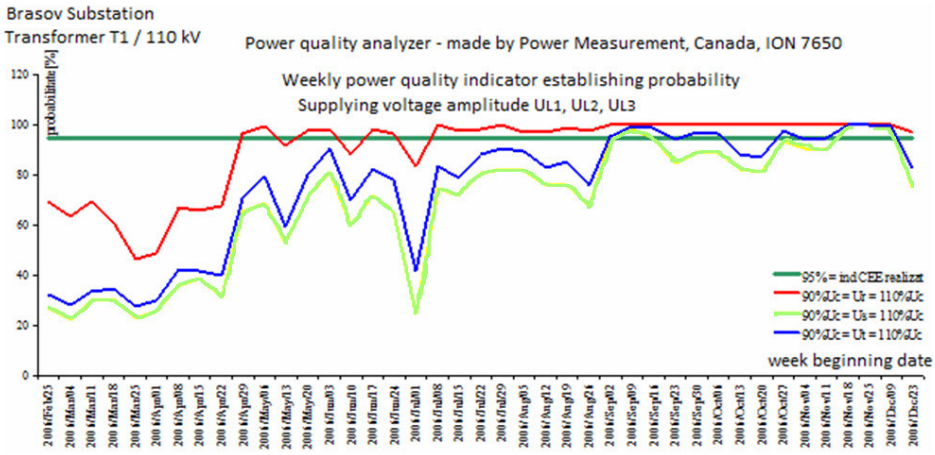
- Timisoara Transmission Subsidiary (400 kV substations: Arad, Mintia, Nadab; 220 kV substations: Arad, Baru Mare, Calea Aradului, Hasdat, Iaz, Mintia, Otelarie, Paroseni, Pestis, Resita, Sacalaz, Timisoara);
- Sibiu Transmission Subsidiary (400 kV substations: Brasov, Darste and Iernut; 220 kV substations: Alba Iulia, Fantanele, Gheorgheni, Ungheni, Iernut).

Part of the experimental determinations has been previously presented in other references such as [1]-[2] and [8]-[9].

In the following, only the experimental results corresponding to the 400 / 110 kV Brasov Substation are presented.

Within Brasov Substation the measurements have been performed between 25.02-30.12.2006. It contains the greatest number of 110 kV power supplies from all the analyzed substations: 21 OHLs (overhead lines) and 2 transformers. Concerning the other network elements, the greatest real energy quantity transported in 2006 is noted, according to Table 4. The load curve variation corresponding to the maintenance period between 03-14.07.06 is presented in Figure 5. The power quality indicator synthesis is presented in Figure 4.

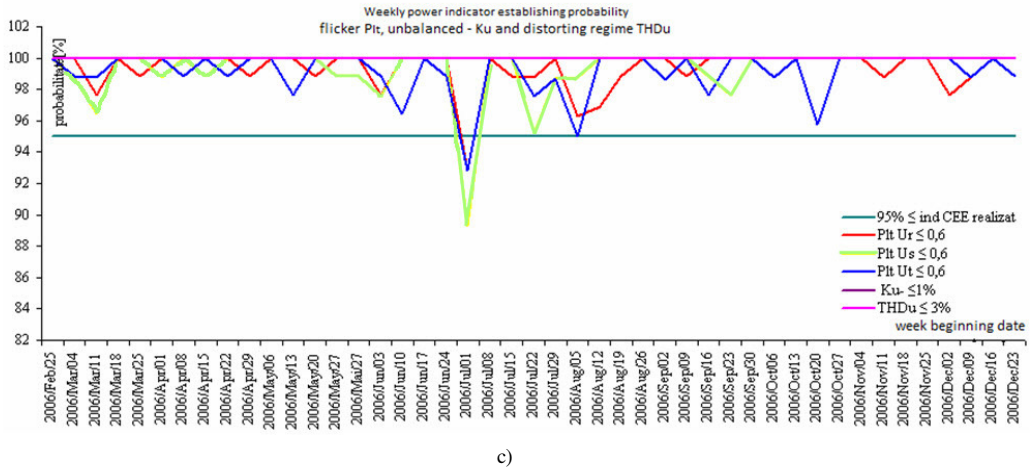
 ISBN: 978-1-61208-237-0 




a)

| No. | Temporary overvoltage magnitude | Temporary overvoltage number with respect to magnnitude and duration | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $U_{L1}$ | | | $U_{L2}$ | | | $U_{L3}$ | | |
| | | $\Delta t < 1$ s | 1 s ≤ Δt < 1 min | 1min ≤ Δt | $\Delta t < 1$ s | 1 s ≤ Δt < 1 min | 1min ≤ Δt | $\Delta t < 1$ s | 1 s ≤ Δt < 1 min | 1min ≤ Δt |
| 1 | 110 % Uc < U < 120 % Uc | 12 | 36 | 700 | 39 | 52 | 907 | 34 | 150 | 1199 |
| 2 | 120 % Uc ≤ U < 140 % Uc | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 1 | 0 |
| 3 | 140 % Uc ≤ U < 160 % Uc | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| No. | Voltage sag magnitude | Voltage sag number with respect to magnitude and duration | | | | | | | | |
| | | $U_{L1}$ | | | $U_{L2}$ | | | $U_{L3}$ | | |
| | | 100 ms ≤ Δt < 100 ms | 100 ms ≤ Δt < 500 ms | 500 ms ≤ Δt < 1 ms | 10 ms ≤ Δt < 100 ms | 100 ms ≤ Δt < 500 ms | 500 ms ≤ Δt < 1 ms | 10 ms ≤ Δt < 100 ms | 100 ms ≤ Δt < 500 ms | 500 ms ≤ Δt < 1 ms |
| 1 | 10 % Uc < ΔU< 15 % Uc | 3 | 1 | 0 | 1 | 2 | 0 | 6 | 3 | 0 |
| 2 | 15 % Uc ≤ ΔU < 30 % Uc | 6 | 3 | 0 | 8 | 2 | 0 | 4 | 5 | 0 |
| 3 | 30 % Uc ≤ ΔU < 60 % Uc | 1 | 6 | 0 | 1 | 2 | 0 | 0 | 1 | 0 |
| 4 | 60 % Uc ≤ ΔU < 99 % Uc | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 |
| No | Measurement period | Short and long period voltage sag number with respect to their duration | | | | | | | | |
| | | $U_{L1}$ | | | $U_{L2}$ | | | $U_{L3}$ | | |
| | | $\Delta t < 1$ s | 1 s ≤ Δt < 3 min | 3min ≤ Δt | $\Delta t < 1$ s | 1 s ≤ Δt < 3 min | 3min ≤ Δt | $\Delta t < 1$ s | 1 s ≤ Δt < 3 min | 3min ≤ Δt |
| 1 | 25.02-30.12.2007 | 12 | 2 | 15 | 9 | 2 | 16 | 0 | 0 | 0 |

b)




c)

Figure 4. T1 transformer 110kV Brasov Substation power quality indicator synthesis (a, b and c)

TABLE 4. T1 TRANSFORMER 110 KV BRASOV SUBSTATION OPERATING CONDITIONS

| Network element | Transformer tap ratio | Rated power | Tap | Current transformer tap ratio | Voltage transformer tap ratio |
|---|---|---|---|---|---|
| T1 Autotransformer corresponding to 2006 year operating conditions | 400 / 110kV | 250MVA | 8 and 7 | 1200 / 5A | 110000 / 100V |
| | Out of service | Average loading level | Real energy transmitted to 110 kV network | Real energy received from 110 kV network | Power factor |
| | 03-14.07.06<br>23-25.10.06<br>09.12.06 | 25.9 % | 570.3 GWh | 0 GWh | 99.23 |

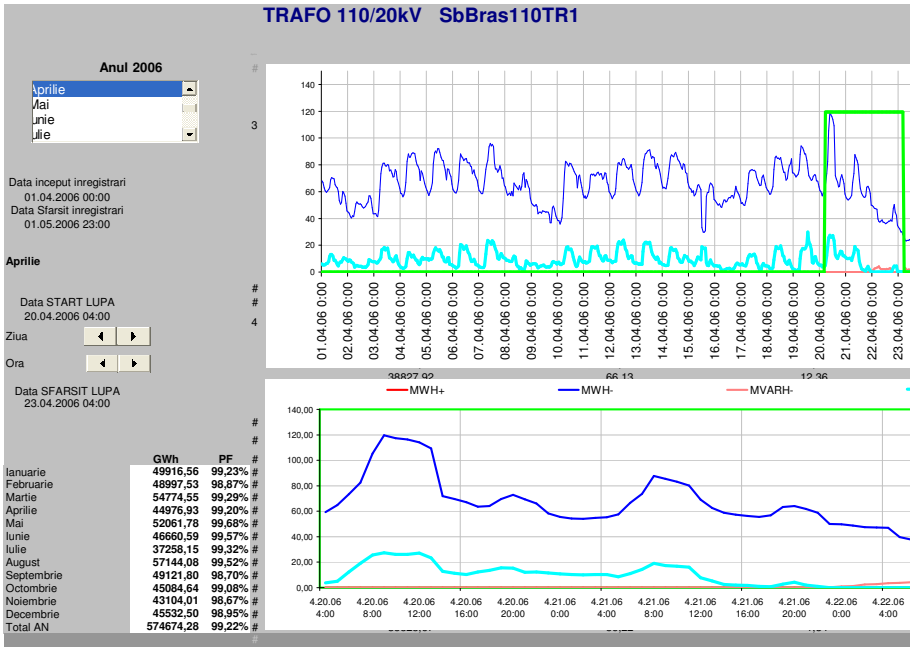



Figure 5. T1 transformer 110 kV Brasov Substation load curves

According to this figure the phase L1, L2, L3 supplied voltage magnitude and the long period flicker level Plt have not fitted between the admissible limits. The T1 transformer (type TTUS-FS 400 kV ± 8*12.5 % / 121 kV) is provided with the possibility of primary circuit voltage control by the under-load tap changer type T-III-1000, made by Reinheisen. It has been set on 8 position within the 25.02-21.04.2006 period and on 7 position between 21.04-30.12.2006. The tap changing is also highlighted by the supplied voltage magnitude rapid variation and also by the reactive energy before and after the commutation.

Starting with 2006 year, an upgrading process, for all the voltage levels, has been started within the substation. Obviously, the upgrading process is based on the recorded values. Several economical advantages have been obtained (the accidental events' number has been reduced, also the operation expenses).

CENELEC - Project ENV European standard 50166-1, Human exposure to low frequency electromagnetic fields (0 - 10 kHz) and the General rules of safety set by the Ministry of Labor and Social Protection and the Ministry of Health in Romania provide that the maximum permissible intensity of the electric field E = 10 kV / m for a time of 8 hours a day. In the conditions under which staff are exposed to E > 10 kV / m is recommended reducing the waiting time in the electric field using the formula t = 80 / E, where t is time in hours.

The E [kV / m] electric field measurement results, within the 400 kV Brasov substation, are synthesized within Table 5 (only few of the measured values are highlighted). Measurements in 148 points have been performed. In 92 points values greater the 10 kV / m have been found. In these areas with E > 10 kV / m necessary measures to protect staff in accordance with international and domestic rules. To protect human beings' health within the Brasov substation, the installation of several protection screens has been performed. Their role is to reduce the electric field intensity values in case of the main working areas. Also, have reduced work-time for the intense electric field areas.

CENELEC - Project ENV European standard 50166-1, Human exposure to low frequency electromagnetic fields (0 - 10 kHz) and the General rules of safety set by the Ministry of Labor and Social Protection and the Ministry of Health in

 ISBN: 978-1-61208-237-0 

Romania in Romania provide that the maximum allowable induction of the magnetic field B = 0.5 mT on exchange of work (for 8 hours daily). In the conditions under which staff are exposed to B = 5 mT duration of exposure will be less than 2 hours on the exchange work.

The B [mT] magnetic field measurement results, within the 400 kV Brasov substation, are synthesized within Table 6 (only few of the measured values are highlighted). All values measured induction B are much lower than the maximum allowable B = 0.5 mT and therefore are not necessary measures to protect personnel action against the magnetic field.

TABLE 5. 400 KV BRASOV SUBSTATION ELECTRIC FIELD INTENSITY VALUES

| No. | Measurement point | Electric field intensity E [kV / m] | | |
|---|---|---|---|---|
| | | $L_1$ | $L_2$ | $L_3$ |
| **Transformer 2 cell – 400 / 220 kV** | | | | |
| 1 | IO circuit breaker | 11 | 9 | 12.5 |
| 2 | Circuit breaker triggering mechanism | 16 | 12.5 | 16 |
| 3 | S $B_2$ $T_2$ insulating switch | 18 | 18 | 18 |
| **400 kV transfer bus-bar cell** | | | | |
| 13 | Bus-bar insulating switch 1 | 13.5 | 12.5 | 16 |
| 14 | Circuit breaker triggering mechanism | 16.5 | 11 | 12.5 |
| 15 | Bus-bar insulating switch 2 | 18 | 16 | 18 |

TABLE 6. 400 KV BRASOV SUBSTATION MAGNETIC FIELD VALUES

| No. | Measurement point | Magnetic induction B [mT] | | |
|---|---|---|---|---|
| | | $L_1$ | $L_2$ | $L_3$ |
| 1 | Transformer cuve $T_2$ | 0.069 | | |
| 2 | Circuit breaker triggering mechanism – $T_2$ | 0.018 | 0.022 | 0.016 |
| 3 | Bus-bar insulating switch 2 | 0.013 | - | 0.02 |
| 4 | Relays cabin | 0.05 | | |

## 4. CONCLUSIONS

For all the measuring points, the supplied voltage magnitude $U_R$, $U_S$, $U_T$, overpasses the imposed limits 99-121 kV, stipulated within the Electrical Transmission Network Technical Code [10]. The upper limit has been exceeded for a 95 % period of the week. The revision of the Electrical Transmission Network Technical Code has been proposed considering the technical equipment characteristics installed within 110 kV electrical installations and the experience achieved. The upper limit voltage magnitude variation from 121 kV to 123 kV has been proposed.

For a 95 % of the analyzed period, long term flicker has not been framed within the limits imposed by the IEC 61000-4-15:2003 [11].

The feasibility study conducted for this expert system stresses that its implementation will ensure rapid access to information needed FOR all the responsible factors. It is necessary to establish concrete measures designed for reducing electromagnetic disturbances and to diminish the following effects:

- Further losses' reduction within power transmission networks and consumers, mainly by reducing the level of harmonics, voltage unbalances and current;
- Proper equipment operation, for those cases when their functions and performance are affected by the harmonics' presence and voltage unbalances and / or current;
- Reducing the operation expenses for the equipment preventive or corrective maintenance, for those cases when they are affected by disturbances that damage the power quality;
- Increasing the efficiency of the generating units, processing units, lines and electric motors (including the ones for the substations' ancillary services) etc.;
- Reducing the costs of power generation / power transmission and, in general, reducing the investment within the National Power System. It would result from the need of over sizing the network elements to cover the effects of electromagnetic disturbance with offences against limits;
- Reducing the damages to consumers caused by voltage (the violation of the rated value, voltage gaps and short term interruptions);
- Establishment of concrete protecting measures to protect the operating personnel from 110 kV, 220 kV and 400 kV installations against the electric and magnetic fields, based on the literature study.

Using this expert system, optimal upgrading decisions have been able to be taken for Brasov substation. The

 ISBN: 978-1-61208-237-0 

additional transmission network losses, due to perturbations affecting the power quality, have been mitigated.

Currently, a profitability index Pi = 2.8 has been obtained. It represents the ratio between the sum of the yearly updated benefits and the sum of the yearly updated expenses, along the considered study period. Consequently, the system implementation had an important advantage compared to other systems describer within the literature [12]-[14].

## REFERENCES


[1] D. Vatau and F.D. Surianu, "Monitoring of the Power Quality on the Wholesale Power Market in Romania", Proc. of the 9th WSEAS International Conference on Electric Power Systems, High Voltages, Electric Machines (POWER'09), Genova, Italy, October 17-19, 2009, pp.59-64.

[2] P. Ehegartner, S. Jude, P. Andea, D. Vatau and F.M. Frigura Iliasa, "A Model Concerning the High Voltage Systems Impact on the Environment inside a Romanian Power Substation", Proc. of the 11th WSEAS International Conference on Automatic Control, Modelling & Simulation (ACMOS'09), Istanbul, Turkey, May 30 - June 1, 2009, pp. 413-418.

[3] CEI 61000-4-30: 2003, "Electromagnetic compatibility (EMC) – Part 4 - 30: Testing and measurement technique - Power quality measurement methods".

[4] CENELEC EN 50160: 1999, "Voltage characteristics of electricity supplied by public distribution systems".

[5] IEEE 1159: 1995, "Recommended Practice on Monitoring Power Quality".

[6] Power quality analyzer, Topas 1000, 3.3 version, User guide, LEM NORMA GmbH, Austria.

[7] IEEE 1459: 2000, "Standard definitions for the measurement of electric power quantities under sinusoidal, nonsinusoidal, balanced, or unbalanced conditions".

[8] D. Vatau, F.D. Surianu, A.E. Bianu and A.F. Olariu, "Considerations on the Electromagnetic Pollution Produced by High Voltage Power Plants", Proc. of European Computing Conference (ECC'11), Paris, France, April 28-30, 2011, pp. 164-170.

[9] D. Vatau, P. Andea, F.D. Surianu, F.M. Frigura Iliasa, S. Kilyeni and C. Barbulescu, "Overvoltage protection systems for low voltage and domestic electric consumers", Proc. of the 15th IEEE Mediterranean Electrotechnical Conference (MELECON 2010), Malta, Cyprus, April 25-28, 2010, pp. 1394-1397.

[10] Technical Code RET - Romanian internal standard.

[11] CEI 61000-4-15: 2003, "Electromagnetic compatibility (EMC) - Part 4-15: Testing and measurement technique - Flicker meter - Functional and design specifications".

[12] T.L. Tan, S. Chen, and S.S. Choi, "An overview of power quality state estimation", Proc. of the 7th International IEEE Power Engineering Conference, (IPEC'05), Singapore, November 29 – December 2, 2005, pp. 271-276.

[13] G. Putrus, J. Wijayakulasooriya, and P. Minns, "Power Quality: Overview and monitoring", Proc. of the IEEE International Conference on Industrial and Information Systems (ICIIS'07), Sri Lanka, August 9-11, 2007, pp. 551-558.

[14] R. Lima, D. Quiroga, C. Reineri, and F. Magnago, "Hardware and software architecture for power quality analysis", Computers & Electrical Engineering, vol. 34 (6), November 2008, pp. 520-530.


 ISBN: 978-1-61208-237-0 

# Intelligent Classification of Odor Data Using Neural Networks

Sigeru Omatu
Department of Electronics, Information, and Communication Engineering
Osaka Institute of Technology
Osaka, 535-8585, Japan
omatu@rsh.oit.ac.jp

Hideo Araki
Department of Computer Science
Osaka Institute of Technology
Hirakata, 573-0196, Japan
araki@is.oit.ac.jp

Toru Fujinaka
Graduate School of Education
Hiroshima University
Higashi-Hiroshima, 739-8524, Japan
fjnk@hiroshima-u.ac.jp

Mitsuaki Yano
Department of Electronics, Information, and Communication Engineering
Osaka Institute of Technology
Osaka, 535-8585, Japan
yano@elc.oit.ac.jp

***Abstract*—Metal oxide semiconductor gas (MOG) sensors and quartz crystal microbalance (QCM) sensors are used to measure several kinds of odors. Using neural networks to classify the measured data of odors, artificial electronic noses have been developed. This paper is to consider an array sensing system of odors and to adopt a layered neural networks for classification. Furthermore, we consider mixied effect of odors for classification accuracy. For simplicity, we will treat the case that two kinds of odors are mixed, since more than two becomes too complex to analyze the classification efficiency. In order to consider the mixed effect, we use as the test data two out of four kinds of odors. An acceptable result, although not perfect, has been achieved for the classification of mixed odors, by using aht lruered neural network.**

***Keywords*-odor classification; odor sensors; sensor array; mixed odors; neural networks.**

## I. INTRODUCTION

The problem of recognition and classification of odors are important to achieve the high quality of information like human being since the smell is one of five senses. We have used these five senses to enjoy comfortable human life with communication and mutual understanding. Artificial odor sensing and classification systems through electronic technology are called an electronic nose and they have been developed according to various odor sensing systems and several classification methods [1][2][3][4].

We have developed electronic nose systems to classify the various odors under different densities based on a layered neural network and a competitive neural network of the learning vector quantization method [5][6][7].

Based on the experience, we have developed a new measurement system such that precise evaluation of the odor can be done with many sensors and by controlling dry air flow rate, temperature, and humidity. Furthermore, we have attached a sensing system with both Metal oxide semiconductor gas (MOG) sensors array and quartz crystal microbalance (QCM) sensors array.

After brief survey of the electronic nose and its measurement and classification methods, we consider the electronic nose accuracy when two odors are mixed after each of the original odors has been classified precisely by using a neural network. We will consider the classification of mixed odors based on the sensing data by using QCM sensors. The QCM sensors are more sensitive than MOG sensors to some kinds of odor and the environmental condition for sensing odors. Using many QCM sensors, we will try to separate the mixed odors into the original odors based on the neural network classifier.

## II. HUMAN OLFACTORY PROCESSES

Although the human olfactory system is not fully understood by physicians, the main components about the anatomy of human olfactory system are the olfactory epithelium, the olfactory bulb, the olfactory cortex, and the higher brain or cerebral cortex, as shown in Fig. 1.

The first process of human olfactory system is to breathe or to sniff the smell into the nose, as shown in Step 1 of Fig. 1. The difference between the normal breath and the sniffing is the quantity of odorous molecules that flows into the upper part of the nose. In case of sniffing, most air is flown through the nose to the lung and about 20% of air is flown to the upper part of the nose and detected by the olfactory receptors.

In case of sniffing, the most air flow directly to the upper part of the nose interacts with the olfactory receptors. The odorous molecules are dissolved at a mucous layer before interacting with olfactory receptors in the olfactory epithelium, as shown in Step 2 of Fig. 1.

The concentration of odorous molecules must be over the recognition threshold. After that, the chemical reaction in each olfactory receptor produces an electrical stimulus. The electrical signals from all olfactory receptors are transported to olfactory bulb, as shown in Step 3 of Fig. 1.

The input data from olfactory bulbs are transformed to be the olfactory information to the olfactory cortex, as shown

 ISBN: 978-1-61208-237-0 

in Step 4 of Fig. 1. Then the olfactory cortex distributes the information to other parts to the brain and human can recognize odors precisely, as shown in Step 5 of Fig. 1. The other parts of the brain that link to the olfactory cortex will control the reaction of the other organ against the reaction of that smell. When human detects bad smells, human will suddenly expel those smells from the nose and try to avoid breathing them directly without any protection. This is a part of the reaction from the higher brain.

Finally, the cleaning process of the nose is to breathe fresh air in order to dilute the odorous molecules until those concentrations are lower than the detecting threshold, as shown in Step 6 of Fig. 1. The time to dilute the smell depends on the persistence qualification of the tested smell.

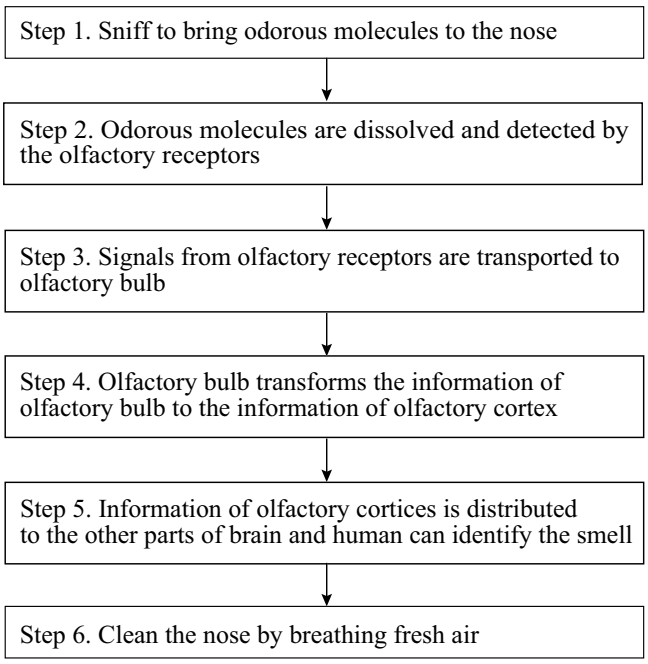



Fig. 1. Olfactory system.

## III. ELECTRONIC NOSE SYSTEM

The electronic nose system is an alternative method to analyze smell by imitating the human olfactory system. In this section, the concept of an electronic nose is explained. Then various sensors for odors applied as the olfactory receptors are explained. Finally, the mechanism of a simple electronic nose that will be developed in this paper is described in detail by comparing the function of each part with the human olfactory process.

The mechanism of electronic nose systems can be divided into main four parts as shown in Fig. 2.

### A. Odor delivery system

The first process of the human olfactory system is to sniff the odorous molecule into the nose. Thus, the first part of the electronic nose system is the mechanism to bring the odorous molecules into the electronic nose system. There are three

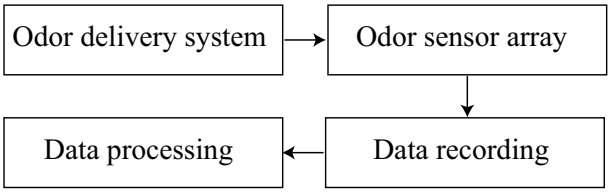



Fig. 2. Main parts of electronic nose systems.

main methods to deliver the odor to the electronic nose unit, sample flow, static system, and pre-concentration system.

The sample flow system is the most popular method to deliver odorous molecule to the electronic nose unit. Some carrier gas such as air, oxygen, nitrogen, and so on, is provided as a carrier gas at the inlet port to flow the vapor of the tested smell through the electronic nose unit via the outlet port. The mechanism to control the air flow of an electronic nose may contain various different parts such as a mass flow controller to control the pressure of the carrier gas, a solenoid valve to control the flow of inlet and outlet ports, a pump to suck the tested odor from the sampling bag in case that the tested odor is provided from outside, a mechanism to control humidity, and so on. Most commercial electronic noses contain complicated odor delivering systems and this makes the price of the electronic noses become expensive.

The static system is the easiest way to deliver odorous molecules to the electronic nose unit. The electronic nose unit is put into a closed loop container. Then an odor sample is injected directly to the container by a syringe. It is also possible to design an automatic injection system. However, the rate to inject the test odors must be controlled to obtain accurate results. Normally, this method is applied for the calibration process of the electronic nose. But, in this case, the quantity of the odor may not be enough to make the sensor reach the saturation stage, that is, the stage that sensor adsorbs the smell fully.

The pre-concentration system is used in case of the tested smell that has a low concentration and it is necessary to accumulate the vapor of the tested odor before being delivered to the electronic nose unit. The pre-concentrator must contain some adsorbent material such as silica and the tested odor is continuously accumulated into the pre-concentrator for specific time units. Then, the pre-concentrator is heated to desorb the odorous molecule from the adsorbent material. The carrier gas is flown through the pre-concentrator to bring the desorbed odorous molecules to the electronic nose unit. By using this method, some weak smells can be detected by the sensor array in the electronic nose unit.

### B. Odor sensor array

The second process of the human olfactory system is to measure various odors corresponding to various receptors in the human olfactory system. In order to realize many receptors artificially, we adopted two types of sensors. One is MOG type n [8] and the other is QCM type p [7]. The idea of the present

 ISBN: 978-1-61208-237-0 

paper is to use many sensors which are allocated in an array structure for each type of MOG and QCM types. This structure is adopted based on the human olfactory system. As we will explain in what follows, the odor sensors are not so small, which results in a large space to measure odors.

### C. Data recording

The data recording is corresponding to temporal memory for the human olfactory system. In the latter case, after learning odors we could identify an odor suddenly, we store sensing data of odors in a computer. To make reading and writing the data, we make an efficient structure of data base.

### D. Data processing

Using the data base of odors, we must apply an intelligent signal processing technique to recognize odors correctly. We make pre-processing the odor data such as noise reduction, normalization, feature extraction, etc. Then we use neural networks for classification of odors. Basically, we use a layered neural networks and competitive networks for odor classification since learning ability and robustness are important in odor classification. The most difficult and important process in the odor classification is to find excellent features which are robust for environment like temperature, humidity, and density levels of odors.

## IV. Principle of Odor Sensing

Nowadays, there are many kinds of sensors that can measure odorous molecules. However, only a few kinds of them have been successfully applied as artificial olfactory receptors in commercial electronic noses. We show two types of odor sensors which are major for odor sensing. One is a MOG sensor and the other is a QCM sensor, which will be explained in what follows.

### A. Principle of MOG sensors

MOG sensors are the most widely used sensors for making an array of artificial olfactory receptors in electronic nose systems. These sensors are commercially available as the chemical sensor for detecting some specific smells. Generally, an MOG sensor is applied in many kinds of electrical appliances such as a microwave oven to detect the food burning, an alcohol breathe checker to check the drunkenness, an air purifier to check the air quality, and so on.

The picture of some commercial MOG sensors are shown in Fig. 3. Various kinds of metal oxides, such as $SnO_2$, $ZnO_2$, $WO_3$, $TiO_2$ are coated on the surface of a semiconductor. But, the most widely applied metal oxide is $SnO_2$. These metal oxides have a chemical reaction with the oxygen in the air and the chemical reaction changes when the adsorbing gas is detected. The scheme of chemical reaction of an MOG sensor when adsorbing with the CO gas, is shown as follows:

$$\frac{1}{2}O_2 + (SnO_{2-x}) \rightarrow O^-\text{ad } (SnO_{2-x}) \quad (1)$$

$$CO + O^-\text{ad } (SnO_{2-x}) \rightarrow CO_2 + (SnO_{2-x}) \quad (2)$$

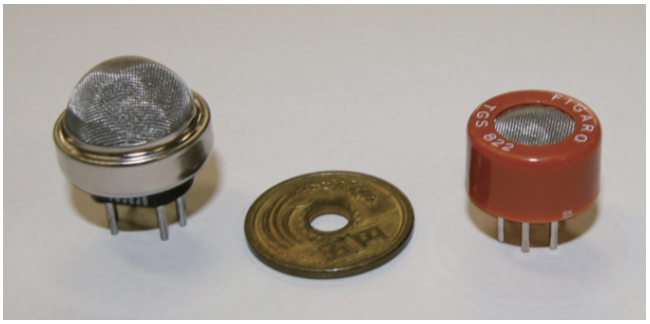



Fig. 3. MOG sensors.

When the metal oxide element on the surface of the sensor is heated at a certain high temperature, the oxygen is adsorbed on the crystal surface with the negative charge as shown in Fig. 4. In this stage, the grain boundary area of the metal oxide element forms a high barrier as shown in the left hand side of Fig. 4.

Then, the electrons cannot flow over the boundary and this makes the resistance of the sensor become higher. When the deoxidizing gas, e.g., CO gas, is presented to the sensor, there is a chemical reaction between negative charges of oxygen at the surface of the metal oxide element and the deoxidizing gas as shown in (1). The chemical reaction between adsorbing gas and the negative charge of the oxygen on the surface of MOG sensor reduces the grain boundary barrier of the metal oxide element as shown in the right hand side of Fig. 4. Thus, the electron can flow from one cell to another cell easier. This makes the resistance of MOG sensor lower by the change of oxygen pressure according to the rule of (2). Thus, (1) means that $CO_2$ is reduced and (2) means that CO is oxidized.

The relationship between sensor resistance and the concentration of deoxidizing gas can be expressed by the following equation over certain range of gas concentration:

$$R_s = A[C]^{-\alpha}$$

where $R_s$ =electrical resistance of the sensor, $A$ = constant, $C$ = gas concentration, and $\alpha$ =slope of $R_s$ curve. The electric

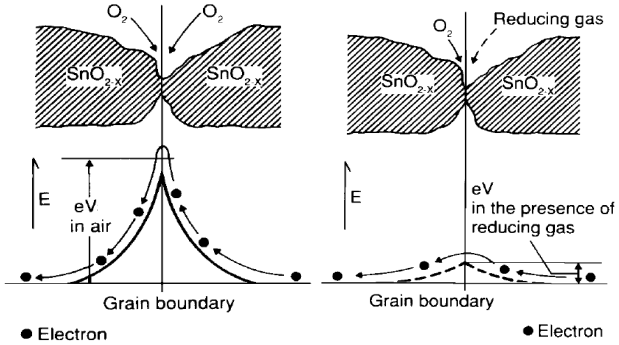



Fig. 4. Principle of MOG sensor [8].

circuit for the MOG sensor is shown in Fig. 5. Electrical voltages are provided to the circuit ($V_c$) and the heater of the sensor ($V_h$). When the MOG sensor is adsorbed with oxygen and the deoxidizing gas, the resistance of the sensor ($R_s$) is changed. Thus, we can measure the voltage changes ($V_{out}$) while the sensor is adsorbing the tested odor.

MOG sensors need to be operated at high temperature, so they consume a little higher power supply than the other kinds of sensors. The reliability and the sensitivity of MOG sensors are proved to be good to detect volatile organic compounds (VOCs), combustible gas, and so on [8]. However, the choices of MOG sensors are still not cover all odorous compounds and it is difficult to create an MOG sensor that responds to one odor precisely. Generally, most commercial MOG sensors respond to various odors in different ways. Therefore, we can expect if we use many MOG sensors to measure a smell, the vector data reflect the specific properties for the smell. Generally, it is designed to detect some specific

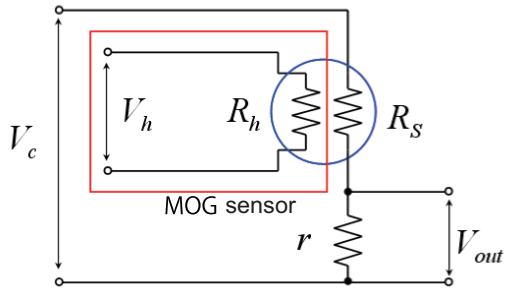



Fig. 5. Principle of MOG sensors.

smell in electrical appliances such as an air purifier, a breath alcohol checker, and so on. Each type of MOG sensors has its own characteristics in the response to different gases. When combining many MOG sensors together, the ability to detect a smell is increased. The main part of the MOG sensor is

TABLE I
LIST OF MOG SENSORS FROM THE FIS INC. USED IN THIS EXPERIMENT.

| Sensor Model | Main Detecting Gas |
|---|---|
| SP-53 | Ammonia, Ethanol |
| SP-MW0 | Alcohol, Hydrogen |
| SP-32 | Alcohol |
| SP-42A | Freon |
| SP-31 | Hydrocarbon |
| SP-19 | Hydrogen |
| SP-11 | Methane, Hydrocarbon |
| SP-MW1 | Cooking vapor |

the metal oxide element on the surface of the sensor. When this element is heated at certain high temperature, the oxygen is adsorbed on the crystal surface with the negative charges. The reaction between the negative charge of the metal oxide surface and deoxidizing gas makes the resistance of the sensor vary as the partial pressure of oxygen changes [8]. Based on this characteristic, we can measure the net voltage changes while the sensors adsorb the tested odor.

### B. Principle of QCM sensors

QCM sensors have been well-known to provide a very sensitive mass-measuring devices in Nano-gram levels, since the resonance frequency will change upon the deposition of the given mass on the electrodes. Synthetic polymer-coated QCM sensors have been studied as sensors for various gasses since a QCM sensor is coated with a sensing membrane works as a chemical sensor. The QCM sensors are made by covering the surface with several kinds of a very thin membrane with about 1 $\mu$m as shown in Fig. 6.

Since the QCM sensor oscillates with a specific frequency depending on the cross section corresponding to three axis of the crystal, the frequency will change according to the deviation of the weight due to the adsorbed odor molecular (odorant). The membrane coated on QCM sensor has selective adsorption rate for a molecular and the frequency deviation show the existence of odorants and their densities. Odorants and membranes are tight relation while it is not so clear whose materials could be adsorbed so much.

In this paper, we have used the following materials as shown in Table II. The reason why fluorine compounds are used here is that the compounds repel water such that pure odorant molecular could be adsorbed on the surface of the membrane. To increase the amount of odorants to be adsorbed it is important to iron the thickness of the membrane. In Table II, we have tried to control the density of the solute in the organic solvent. The basic approach used here is a sol-gel method. The sol-gel process is a wet-chemical technique used for the fabrication of both glassy and ceramic materials. In this process, the sol (or solution) evolves gradually towards the formation of a gel-like network containing both a liquid phase and a solid phase. Typical precursors are metal oxides and metal chlorides, which undergo hydrolysis and polycondensation reactions to form a colloid. The basic structure or morphology of the solid phase can range anywhere from discrete colloidal particles to continuous chain-like polymer networks.

TABLE II
CHEMICAL MATERIALS USED AS THE MEMBRANE WHERE E:ETANOL, W:WATER, DA:DILUTE NITRIC ACID, EA:ETHYL ACRYLATE, MTMS:TRIMETHOXY SILANE, PFOEA:PERFLUOROOCTYLETHYL ACRYLATE.

| Sensor number | Materials of membrane |
|---|---|
| Sensor 1 | E(4ml), DA(0.023ml), |
| Sensor 2 | W(3.13ml), E(4ml), EA(0.043ml), MTMS |
| Sensor 3 | W(3.13ml), E(4ml), EA(0.014ml), MTMS, PFOE |
| Sensor 4 | W(3.13ml), E(4ml), EA(0.015ml), MTMS, PFOE |
| Sensor 5 | W(0.30ml), E(4ml), EA(0.043ml), MTMS, PFOE |
| Sensor 6 | W(0.05ml), E(3.0ml), EA(0.043ml), MTMS, PFOE |
| Sensor 7 | W(0.30ml), E(3.2ml), EA(0.043ml), MTMS, PFOE |
| Sensor 8 | No membrane |

## V. ODOR SENSING SYSTEM

Generally, odor sensors are designed to detect some specific odor in electrical appliances such as an air purifier, a breath alcohol checker, and so on. Each of odor sensors such as MOG sensors or QCM sensors has itself characteristics in

 ISBN: 978-1-61208-237-0 

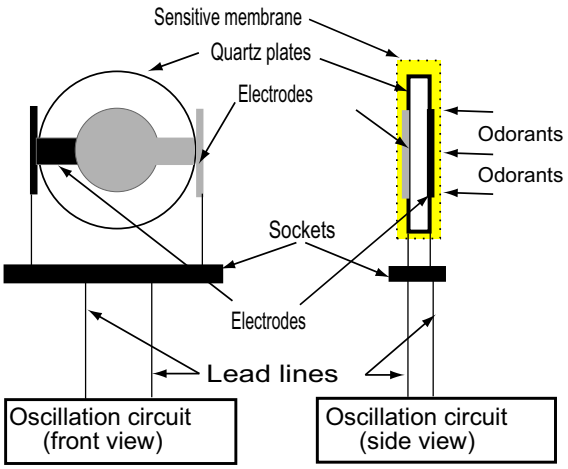



Fig. 6. Principle of QCM sensors. The odorants attached on a sensitive membrane will make the weight change of quartz plane. Thus, the original frequency of the crystal oscillation will become smaller according to the density of odorants.

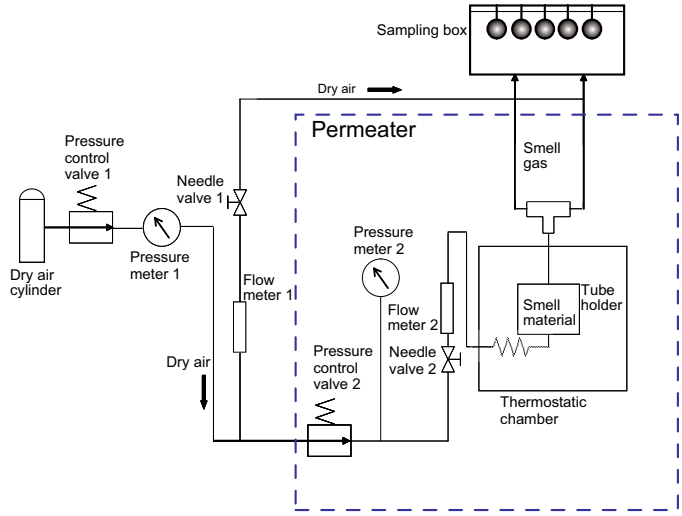



Fig. 7. Odor sensing systems. The air will be emitted from the dry air cylinder. Air flow is controlled by pressure control valves 1 and 2. By using the needle valve 2, more precise flow rate of the dry air can be achieved and the thermostatic chamber in the permeater can control the temperature of the dry air. Finally, the air is pull in the sampling box where the MOG sensors and/or QCM sensors are attached on the ceiling of the box.

the response to different odors. When combining many odor sensors together, the ability to detect an odor is increased. An electronic nose system shown in Fig. 7 has been developed, based on the concept of human olfactory system. The combination of odor sensors, listed in Tables I and II, are used as the olfactory receptors in the electronic nose.

## VI. Classification Method of Odor Data

In order to classify the odors we adopt a three-layered neural network based on the error back-propagation method as shown in Fig. 8.

The error back-propagation algorithm which is based on a gradient method is given by the following steps.

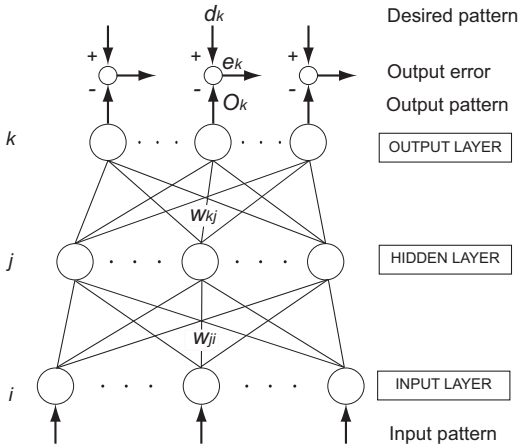



Fig. 8. Three layered neural network with the error back-propagation. The neural network consists of three layers, that is, an input layer $i$, a hidden layer $j$, and an output layer $k$. When the input data $x_i, i = 1, 2, \ldots, I$ are applied in the input layer, we can obtain the output $O_k$ in the output layer which is compared with the desired value $d_k$ which is assigned in advance. If the error $e_k = d_k - O_k$ occurs, Then, the weighting coefficients $w_{ji}, w_{kj}$ are corrected such that the error becomes smaller based on an error back-propagation algorithm.

*Step* 1. Set the initial values of $w_{ji}, w_{kj}, \theta_j, \theta_k$, and $\eta(> 0)$.
*Step* 2. Specify the desired values of the output $d_k, k = 1, 2, \ldots, K$ corresponding to the input data $x_i, i = 1, 2, \ldots, I$ in the input layer.
*Step* 3. Calculate the outputs of the neurons in the hidden layer and output layer by

$$\begin{aligned} \text{net}_j &= \sum_{i=1}^{I} w_{ji}x_i - \theta_j, O_j = f(\text{net}_j), f(x) = \frac{1}{1+e^{-x}} \\ \text{net}_k &= \sum_{j=1}^{J} w_{kj}O_j - \theta_k, O_k = f(\text{net}_k). \end{aligned}$$

*Step* 4. Calculate the error $e_n$ and generalized errors by

$$\begin{aligned} e_k &= d_k - O_k, \delta_k = e_k O_k(1 - O_k) \\ \delta_j &= \sum_{k=1}^{K} \delta_k w_{kj} O_j(1 - O_j). \end{aligned}$$

*Step* 5. If $e_k$ is sufficiently small for all $k$, END and otherwise

$$\begin{aligned} \Delta w_{kj} &= \eta O_j \delta_k, \quad w_{kj} \Leftarrow w_{kj} + \Delta w_{kj} \\ \Delta w_{ji} &= \eta O_i \delta_j, \quad w_{ji} \Leftarrow w_{ji} + \Delta w_{ji}. \end{aligned}$$

*Step* 6. Go to *Step* 3. Using the above recursive procedure, we can train the odor data. The measurement data is an eight-dimensional vector which are obtained with eight sensors stated in Table II.

## VII. Classification Results Using MOG Sensors

We have measured four types of tea using MOG sensors shown in Table I. The odors used here are shown in Table III. Note that the chemical properties of these odors are very

 ISBN: 978-1-61208-237-0

similar and it has been difficult to separate them based on the measurement data by using MOG sensors. We have examined

TABLE III
TEAS USED IN EXPERIMENT I.

| Label | Materials | Samples |
|---|---|---|
| A | English tea | 20 |
| B | Green tea | 20 |
| C | Barley tea | 20 |
| D | Oolong tea | 20 |

two examples, Experiments I and II. For Experiment I, we classify kinds of teas. The numbers of neurons for this experiment are eight in the input layer, four in the hidden layer, and four in the output layer, that is, 8-4-4 structure.

The number of training samples is fifteen and the number of test samples is five. We change the training data set for 100 times and check the classification accuracy for the test data samples. Thus, we have obtained 500 test samples as the total number of classification. The classification results are summarized in Table IV. Average of the classification is 96.2 %. For Experiment II, we consider to classify the five

TABLE IV
CLASSIFICATION RESULTS FOR EXPERIMENT I.

| | Classification results(96.2%) | | | | | |
|---|---|---|---|---|---|---|
| Odor data | A | B | C | D | Total | Correct |
| A | 500 | 0 | 0 | 0 | 500 | 100.0% |
| B | 0 | 494 | 6 | 0 | 500 | 98.8% |
| C | 0 | 71 | 429 | 0 | 500 | 85.8% |
| D | 0 | 0 | 0 | 500 | 500 | 100.0% |

kinds of coffees as shown in Table V where the smell data A, B, and C are the coffees of Mocha made from different companies. The numbers of neurons are eight in the input layer, five in the hidden layer, and five in the output layer, that is, 8-5-5 structure.

The numbers of training samples are twenty and he number of test samples is fifteen. We change the training data set for hundred times and check the classification accuracy for the test data samples. Thus, we have obtained 1,500 test samples as the total number of classification. The classification results are shown in Table VI. From Table VI we can see that the

TABLE V
SMELL DATA OF TEAS USED IN EXPERIMENT II.

| Label | Materials | No. of samples |
|---|---|---|
| A | Mocha coffee1 | 35 |
| B | Mocha coffee2 | 35 |
| C | Mocha coffee3 | 35 |
| D | Kilimanjaro coffee | 35 |
| E | Char-grilled coffee | 35 |

total classification is 88.8%. Compared with Table IV, this case is worse by about 7%. In the latter case, the classification of Mocha coffees 1, 2, and 3 is not so good. The difference of the company may affect the classification results in a bad way although the smells are similar.

TABLE VI
CLASSIFICATION RESULTS FOR EXPERIMENT II.

| | Classification results(88.8%) | | | | | | |
|---|---|---|---|---|---|---|---|
| Odor data | A | B | C | D | E | Total | Correct |
| A | 1190 | 253 | 29 | 27 | 1 | 1500 | 79.3% |
| B | 225 | 1237 | 9 | 10 | 19 | 1500 | 82.5% |
| C | 142 | 7 | 1325 | 26 | 0 | 1500 | 88.3% |
| D | 9 | 14 | 3 | 1437 | 37 | 1500 | 95.8% |
| E | 0 | 18 | 0 | 11 | 1471 | 1500 | 98.1% |

TABLE VII
KINDS OF ODORS MEASURED OF EXPERIMENT III.

| Symbols | Kind of odors |
|---|---|
| A | Ethanol |
| B | Water |
| C | Methyl-salicylate |
| D | Triethyl-amine |

## VIII. CLASSIFICATION RESULTS USING QCM SENSORS FOR MIXED ODOR DATA

We have measured four types of odors, as shown in Table VII. The sampling frequency is 1 [Hz], the temperatures of odor gases are 24∼26 [°C], and the humidity of gas is 6∼8 [%]. To control the density of gases, we use diffusion tubes. Odor data are measured for 600 [s]. They may include impulsive noises due to the typical phenomena of QCM sensors. We call this experiment as Experiment III.

To remove these impulsive noises we adopt a median filter which replaces a value at a specific time by a median value among neighboring data around the specific time. In Fig. 9, we show the measurement data for the symbol A (ethanol) where the horizontal axis is the measurement time and the vertical axis is the frequency deviation from the standard value (9M [Hz]) after passing through a five-point median filter.

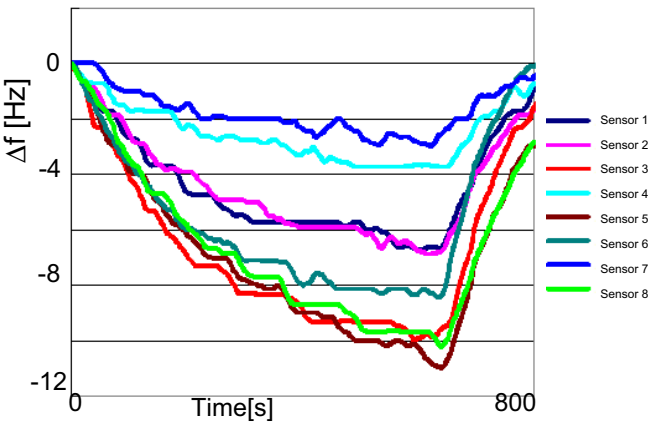



Fig. 9. Measurement data of Experiment III. Here, eight sensors are used and for 800 [s] the data were measured. The maximum value for each sensor among eight sensors is selected as a feature value for the sensor. Therefore, we have eight dimensional data for an odor and they will be used for classification.

 ISBN: 978-1-61208-237-0

TABLE VIII
TRAINING DATA SET FOR ETHANOL (A), WATER (B), METHYL-SALICYLATE (C), AND TRI-ETHYLAMINE (D) FOR EXPERIMENT III.

| Symbols | Output A | Output B | Output C | Output D |
|---|---|---|---|---|
| A | 1 | 0 | 0 | 0 |
| B | 0 | 1 | 0 | 0 |
| C | 0 | 0 | 1 | 0 |
| D | 0 | 0 | 0 | 1 |

TABLE IX
TESTING THE MIXED ODORS FOR EXPERIMET III. HERE, BOLD FACE DENOTES THE LARGEST VALUE.

| Symbols | Output A | Output B | Output C | Output D |
|---|---|---|---|---|
| A and B | **.673** | .322 | .002 | .001 |
| B and C | .083 | **.696** | .174 | .001 |
| C and D | .001 | .004 | .016 | **.992** |
| D and A | .003 | .003 | .002 | **.995** |
| A and C | **.992** | .006 | .002 | .000 |
| B and D | .003 | .003 | .003 | **.995** |

### A. Training for Classification of Odors

In order to classify the feature vector, we allocate the desired output for the input feature vector where it is nine-dimensional vector, as shown in Table VIII since we have added the coefficient of variation to the usual feature vector to reduce the variations for odors. The training has been performed until the total error becomes less than or equal to $0.5 \times 10^{-2}$ where $\eta$=0.3.

### B. Classification Results and Discussion for Mixed Odor Data

After training such that all the smells, A, B, C, and D, have been classified correctly, we have tested the mixed data sets such that two kinds of odors are mixed with the same rate where the data set of mixed smells are {A&B, B&C, C&D, D&A, A&C, B&D}. Then, the classification results are shown in Table IX where the values in boldface denote the top case where the maximum output values are achieved. The maximum values show one of the mixed odors. But, some of them do not show the correct classification for the remaining odor. Thus, we have modified the input features such that

$$z = x - 0.9y,$$

where $x$ is the feature, $y$ denotes the top value of each row in Table IX, and $z$ is a new feature. Using the new feature vector, we have obtained the classification results as shown in Table X. By changing the features according to the above relation, better classification results have been obtained. But, the coefficient 0.9 used in the above equation is not necessarily appropriate. The value might be replaced by the partial correlation coefficient in multivariate analysis.

## IX. CONCLUSIONS

In this paper, a new approach to odor classification has been presented and discussed by using MOG sensors and QCM sensors. After surveying the smell sensing and classification methods, we have examined two examples, Experiment 1 and Experiment II. Then, mixed effects of two different odors have been considered in Experiment III. From these results, we could know that the electronic nose systems might be applied to many applications in real world such that small detection of bad gas or uncomfortable smell, Odor classification of perfume of joss tics kinds, food freshness, etc.

TABLE X
TESTING THE MIXED ODORS WHERE EXCEPT FOR THE LARGEST VALUE THE TOP IS SELECTED AS THE SECOND ODOR AMONG THE MIXED ODORS FOR EXPERIMENT III. HERE, ASTERISK * DENOTES THE TOP EXCEPT FOR THE LARGEST VALUE.

| Symbols | Output A | Output B | Output C | Output D |
|---|---|---|---|---|
| A and B | .263 | **.290*** | .166 | .066 |
| B and C | .358 | .029 | **.631*** | .008 |
| C and D | .002 | .071 | **.644*** | .163 |
| D and A | **.214*** | .004 | .037 | .230 |
| A and C | .031 | .020 | **.527*** | .026 |
| B and D | .108 | .010 | **.039*** | .325 |

## ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant-in-Aid for Scientific Research(B)(23360175). The authors would like to thank JSPS to support this research work.

## REFERENCES

[1] Milke J. A. (1995), "Application of Neural Networks for discriminating Fire Detectors", *International Conference on Automatic Fire Detection*, AUBE'95, 10th, Duisburg, Germany, pp. 213–222
[2] Bicego M. (2005), "Odor classification Using similarity-Based Representation", *Sensors and Actuators B*, Vol. 110, pp. 225–230.
[3] Distante C., Ancona N., Siciliano P., and Burl M. (2003), "Support Vector Machines for Olfactory Signal Recognition", *Sensors and Actuators B*, Vol. 88, pp. 30–39.
[4] Gardner J. and Bartlet P. (1998), "Electronic Noses: Principles and Applications", *Oxford University Press*, Pennsylvania, USA, pp. 1–30.
[5] S. Omatu and M. Yano (2011), "Intelligent Electronic Nose System Independent on Odor Concentration", *International Symposium on Distributed Computing and Artificial Intelligence, Salamanca, Spain*, pp. 1–9.
[6] S. Omatu (2012), "Pattern Analysis for Odor Sensing System", *IGI Global*, pp. 20–34.
[7] T. Fujinak, M. Yoshioka, and T. Kosaka (2008), "Intelligent Electronic Nose Systems for Fire Detection Systems Based on Neural Networks", *The second International Conference on Advanced Engineering Computing and Applications in Sciences, Valencia, Spain*, pp. 73–76.
[8] General Information for TGS sensors, *Figaro Engineering*, available at www.figarosensor.com/products/general.pdf (2012).

 ISBN: 978-1-61208-237-0

# Hybridizing Direct and Indirect Optimal Control Approaches for Aircraft Conflict Avoidance

Loïc Cellier, Sonia Cafieri
*Lab. MAIAA*
*École Nationale de l'Aviation Civile*
*Toulouse, France*
*Email: {loic.cellier, sonia.cafieri}@enac.fr*

Frédéric Messine
*ÉNSÉÉIHT - IRIT*
*Université de Toulouse*
*Toulouse, France*
*Email: frederic.messine@n7.fr*

***Abstract*—Aircraft conflict avoidance is a crucial issue arising in air traffic management. The problem is to keep a given separation distance for aircraft along their trajectories. We focus on an optimal control model based on speed regulation to achieve aircraft separation. We propose a solution strategy based on the decomposition of the problem and on the hybridization of a direct and an indirect method applied on the obtained subproblems. Numerical results show that the proposed approach is promising in terms of reduction of computing time for conflict avoidance.**

***Keywords-air traffic management; conflict avoidance; speed regulation; optimal control; Pontryagin's maximum principle.***

## I. INTRODUCTION

In the context of Air Traffic Control (ATC), motivated by safety and efficiency reasons, tools for decision support are requested. To avoid the risk of collision, distances of separation must be respected. It is said that two aircraft are *in conflict* if the distances between them are less than $5$ NM horizontally and $1000$ ft vertically ($1$ NM (nautical mile) = $1852$ m and $1$ ft (feet) = $0.3048$ m).

Various methods of conflict detection and resolution have been proposed (see *e.g.*, [8]). They are based on different strategies that can be exploited to achieve aircraft separation, such as trajectory (heading), flight level or velocity changes.

Evolutionary computation based algorithms are widely studied in this context [5]. They are, in general, low time-consuming, but the global optimal solution and even a feasible solution (without conflicts) is not garanteed to be achieved in a given time. Several models of *optimal control* also appeared in this domain [1, 10, 11]. They are mainly based on changes of aircraft trajectories, putting the trajectory as a command on the system.

Recently, the European project ERASMUS (En-Route Air traffic Soft Management Ultimate System) [2] considered speed regulation and suggested a small velocity change range to enable a *subliminal control*, that is a speed control which is even not perceived and is performed without informing air traffic controllers. Velocity change approaches have also been recently studied in the context of mixed integer linear and nonlinear programming [4, 9, 11].

This work focuses on aircraft conflict avoidance problems solved through speed regulation. We propose an optimal control approach keeping the flight trajectories and focusing on velocity variations.

The paper is organized as follows. First, in Section II, we present an optimal control model for the addressed air traffic control problem which should be solved by small speed changes. In Section III, we propose stategies to deal with computational complexity. Particularly, we propose a *decomposition* of the problem by considering different time periods such that the separation constraints have to be imposed only in the time periods when the aircraft conflicts potentially occurr. In Section IV, we present an analytical resolution based on the *Pontryagin's maximum principle* (PMP) in the other time periods, *i.e.*, the ones where conflict have already been solved. In Section V, we discuss numerical results issued to the methods. In Section VI, conclusions are drawn.

## II. OPTIMAL CONTROL MODEL THROUGH VELOCITY REGULATION

We present an optimal control model to achieve separation based on speed changes only, keeping the aircraft trajectories unchanged. The *acceleration*, $u_i$ respective to each aircraft $i$, is then the command on the system.

In model ($\mathcal{P}$), $x_i$, $v_i$ and $u_i$ are respectively the position, the velocity and the acceleration (control) of aircraft $i$, with $I = \{1, ...,\text{n}\}$ and $n$ the number of aircraft involved; aircraft are expected to be at the same altitude (planar configuration, same flight level). For each aircraft $i$, velocity $v_i$ and acceleration $u_i$ are bounded (*i.e.*, belonging to $[\underline{v_i}, \overline{v_i}]$ and $[\underline{u_i}, \overline{u_i}]$ respectively).

We note by $t$, $t_0$ and $t_f$ the time, the initial time and final time respectively. Moreover, $D$ is the minimum required horizontal separation distance between two aircraft and $d_i$ is the direction (heading) of the $i^{th}$ aircraft. The final time $t_f$ of maneuvers is fixed and identical for all aircraft. The mathematical model is the following:

 ISBN: 978-1-61208-237-0

$$(\mathcal{P}) \begin{cases} \min\limits_{u} \quad \sum\limits_{i=1}^{n} \int_{t_0}^{t_f} u_i^2(t)dt & \\ \quad \dot{v}_i(t) = u_i(t) & \forall t \in [t_0, t_f], \forall i \in I \\ \quad \dot{x}_i(t) = v_i(t) d_i & \forall t \in [t_0, t_f], \forall i \in I \\ \quad \underline{u_i} \leq u_i(t) \leq \overline{u_i} & \forall t \in [t_0, t_f], \forall i \in I \\ \quad \underline{v_i} \leq v_i(t) \leq \overline{v_i} & \forall t \in [t_0, t_f], \forall i \in I \\ \quad x_i(t_0) = x_i^0 \qquad v_i(t_0) = v_i^0 & \forall i \in I \\ \quad x_i(t_f) = x_i^f \qquad v_i(t_f) = v_i^f & \forall i \in I \\ \quad D^2 - \parallel x_i(t) - x_j(t) \parallel^2 \leq 0 & \forall t \in [t_0, t_f], \forall i < j \end{cases}$$

We choose to minimize a quadratic energy-dependent cost function depending on speed variations. This criterion takes into account the contribution of each aircraft and also limits the penalization inequality between the aircraft.

Note that one of the main difficulties on this optimal control model is given by the constraints on the state variables $v$ and $x$. In the next section, we present resolution approaches tailored on the problem to achieve its efficient solution.

## III. Solution approach by decomposition of the problem

A typical solution approach for an optimal control problem like $(\mathcal{P})$ is based on the application of a *direct* method. It is based on a time discretization and leads to the solution of a nonlinear (continuous) optimization problem (NLP), which can be solved by standard NLP local solvers. For $(\mathcal{P})$, the corresponding NLP problem can be difficult to solve for large-scale problems, mainly due to the large number of variables and constraints. The *complexity* of the NLP corresponding to the *direct* method is $O(np)$ for the number of variables and $O(n^2p + p^2)$ for the number of constraints, where $n$ and $p$ are the number of aircraft and the number of time subdivisions respectively. For example, even on a simple conflict problem, with only 2 aircraft and a time window equals to $30'$ (with time subdivision equals to $15''$) the corresponding nonlinear problem has more than $240$ variables and $9000$ constraints.

We first recall that, in order to perform aircraft separation, a *detection* of potential conflict regions and a *resolution* step have to be carried out. The two steps can be performed at the same time by applying a direct method.

We propose to distinguish two discretization steps. The first one, for the detection, has to be tight enough to check if all constraints are respected. The second one, for the resolution, is used to decide the time frequency at which values the controls are computed and it can be larger than the previous one. For example, we used $15''$ for the detection and $1'$ or $5'$ for the resolution. As discussed in Section V, this strategy allows to reduce the number of variables and constraints of the nonlinear optimization problem to be solved.

Another possibility is to perform a pre-processing step to detect potential conflicts. Given aircraft predicted trajectories, one can check intersections of the trajectories and identify spatial regions where the separation constraints must be checked [7]. Once the different regions have been localized, one can exploit this information to devise a specific strategy of resolution aimed at reducing the computational complexity of the problem at hand. The main contribution of this paper is a strategy based on problem decomposition and related hybridization of optimal control solution methods.

Let *zone* be the region where for an aircraft pair separation constraints have to be verified and *postzone* be the following region where all the conflicts have been solved and when the aircraft are already separated.

For each aircraft $i$, let $x_i^{enter}$ be the *first* (by chronological order) $3D$ trajectory point for which there exists an aircraft $j$ ($j \neq i$) such that the Euclidean distance between $x_i^{enter}$ and the straight line corresponding to the $j^{th}$ aircraft predicted $3D$ trajectory is equal to the separation standard $D$. For each aircraft $i$, let $t_i^1$ be the time to reach $x_i^{enter}$ using the highest speed $\overline{v_i}$. Dually, for each aircraft $i$, let $x_i^{exit}$ be the *last* $3D$ trajectory point for which there exists an aircraft $j$ such that the Euclidean distance between $x_i^{exit}$ and the straight line corresponding to the $j^{th}$ aircraft predicted $3D$ trajectory is equal to the separation standard $D$. For each aircraft $i$, let $t_i^2$ the time to reach $x_i^{exit}$ using the lower speed $\underline{v_i}$.

For $n$ aircraft, setting the entry zone time equals to $t_1 := \min_{i \in \{1,\ldots,n\}} t_i^1$ and the exit zone time $t_2 := \max_{i \in \{1,\ldots,n\}} t_i^2$, we define conflict time phases for the whole problem. The *zone* and *postzone* correspond respectively to the time periods $[t_1, t_2]$ and $[t_2, t_f]$,

The *postzone* being characterized by the absence of separation contraints, it represents a subproblem easier to solve than the initial problem defined on the whole time horizon. We can apply the PMP [3] as discussed in the next section on the *postzone*. On the remaining time window, the direct method is applied. Numerical integrators of Euler-type are used to approximate the ordinary differential equations describing the system dynamic and different time discretization steps mentioned above are exploited.

## IV. Application of the Pontryagin's maximum principle

Without the separation constraint (difficult state constraints), we can easily apply the PMP, which gives us an analytical solution. In the *postzone* (time window $[t_2, t_f]$), as the aircraft conflicts have been solved, the necessity to check separation constraint does not exist anymore. The velocity and acceleration constraints are checked *a posteriori*. Hence, for each aircraft $i$ the following optimal control sub-problem $(\mathcal{P}_i)$ can be solved independently. We recall the assumption that aircraft are expected to be at the same

 ISBN: 978-1-61208-237-0

altitude (planar configuration, same flight level) so that two-components vectors appear in the formulation. The distinction between the two components of the direction (heading) vector $d_i = (d_i^X, d_i^Y)^T$ and the distinction between the position components $x_i = (x_i^X, x_i^Y)^T$ have been done to make easier the formalism.

$$(\mathcal{P}_i) \begin{cases} \min\limits_{u_i} \displaystyle\int_{t_2}^{t_f} u_i^2(t)dt \\ \dot{v}_i(t) = u_i(t) \qquad \forall t \in [t_2, t_f] \\ \dot{x}_i^X(t) = v_i(t)d_i^X \qquad \forall t \in [t_2, t_f] \\ \dot{x}_i^Y(t) = v_i(t)d_i^Y \qquad \forall t \in [t_2, t_f] \\ x_i^X(t_2) = x_i^{X_{t_2}} \qquad x_i^Y(t_2) = x_i^{Y_{t_2}} \qquad v_i(t_2) = v_i^{t_2} \\ x_i^X(t_f) \text{ free} \qquad x_i^Y(t_f) \text{ free} \qquad v_i(t_f) = v_i^{t_f} \end{cases}$$

We apply on $(\mathcal{P}_i)$ the *indirect* method. We introduce the co-state variables $z_0^i, z_1^i, z_2^i, z_3^i$, where $z_1^i, z_2^i, z_3^i$ are associated to $x_i^X, x_i^Y$ and respectively $v_i$. Writing the Hamiltonian

$$H_i \quad = \quad z_0^i u_i^2 + z_1^i v_i d_i^X + z_2^i v_i d_i^Y + z_3^i u_i \ ,$$

the co-state equations are:

$$\dot{z}_1^i = -\frac{\partial H_i}{\partial x_i^X} = 0 \ , \ \dot{z}_2^i = -\frac{\partial H_i}{\partial x_i^Y} = 0 \ ,$$
$$\dot{z}_3^i = -\frac{\partial H_i}{\partial v_i} = \ -(z_1^i d_1^X + z_2^i d_1^Y) \ .$$

By fixing $z_0^i = -1$, by using the PMP [3], we obtain:

$$u_i^* = \underset{u_i}{\operatorname{argmin}} \ H_i = \frac{z_3^i}{2}.$$

Solving the differential system composed by state and co-state equations and introducing six real constants $A_i$, $B_i$, $C_i$, $D_i$, $E_i$ and $F_i$, we obtain:

$$(\mathcal{S}_i) \begin{cases} z_1^i(t) = A_i \quad \text{and} \quad z_2^i(t) = B_i \ , \\ z_3^i(t) = -(A_i d_i^X + B_i d_i^Y)t + C_i \ , \\ \\ u_i(t) = -\dfrac{A_i d_i^X + B_i d_i^Y}{2}t + \dfrac{C_i}{2} \ , \\ v_i(t) = -\dfrac{A_i d_i^X + B_i d_i^Y}{4}t^2 + \dfrac{C_i}{2}t + D_i \ , \\ x_i^X(t) = -\dfrac{A_i (d_i^X)^2 + B_i d_i^X d_i^Y}{12}t^3 \\ \qquad + \dfrac{C_i}{4} d_i^X t^2 + D_i d_i^X t + E_i \ , \\ x_i^Y(t) = -\dfrac{A_i d_i^X d_i^Y + B_i (d_i^Y)^2}{12}t^3 \\ \qquad + \dfrac{C_i}{4} d_i^Y t^2 + D_i d_i^Y t + F_i \ . \end{cases}$$

From the terminal (position) conditions, $x_i^X(t_f)$ is free and $x_i^Y(t_f)$ is free, the PMP implies (*transversality conditions*: $z_1^i(t_f) = 0, z_2^i(t_f) = 0$, see [3]) that the real constants $A_i$ and $B_i$ are both equal to zero. This reveals that the optimal control corresponds to a constant acceleration.

This optimal acceleration depends only on the initial and final velocities ($v_i^{t_2}$ and $v_i^{t_f}$) and on the time window extremities ($t_2$ and $t_f$). More precisely, we obtain the following solution system for each instant $t$ belonging to $[t_2, t_f]$:

$$\begin{cases} u_i(t) = \dfrac{v_i^{t_f} - v_i^{t_2}}{t_f - t_2} \ , \\ v_i(t) = \dfrac{v_i^{t_f} - v_i^{t_2}}{t_f - t_2}(t - t_f) + v_i^{t_f} \ , \\ x_i^X(t) = \dfrac{v_i^{t_f} - v_i^{t_2}}{t_f - t_2} d_i^X \dfrac{t^2}{2} + (v_i^{t_f} - \dfrac{v_i^{t_f} - v_i^{t_2}}{t_f - t_2} t_f) d_i^X t \\ \qquad - (\dfrac{v_i^{t_f} - v_i^{t_2}}{t_f - t_2}(t_2 - t_f) + v_i^{t_f}) d_i^X t_2 + x_i^{X_{t_2}} \ , \\ x_i^Y(t) = \dfrac{v_i^{t_f} - v_i^{t_2}}{t_f - t_2} d_i^Y \dfrac{t^2}{2} + (v_i^{t_f} - \dfrac{v_i^{t_f} - v_i^{t_2}}{t_f - t_2} t_f) d_i^Y t \\ \qquad - (\dfrac{v_i^{t_f} - v_i^{t_2}}{t_f - t_2}(t_2 - t_f) + v_i^{t_f}) d_i^Y t_2 + x_i^{Y_{t_2}} \ . \end{cases}$$

Hence, starting from $t_2$, the problem can be analytically solved. Thus, just a discretization of the time window $[t_0, t_2]$ is needed.

## V. Numerical Results

In this section, we discuss numerical results obtained by applying the proposed strategies to solve the conflict avoidance problem. A computer 2.53 GHz / 4 Go RAM and the *MatLab v.* 7 environment are used. Data problems were randomly generated with the following characterics. The trajectory paths are straight. The horizontally separation norm is 5 NM. Most of the aircraft have a small *operating time* (*i.e.*, time before the first potential conflict), which is less than $15'$. Velocities are bounded, based on the ERASMUS project, by a small speed range, namelly: $[v_i^{t_0} - 6\% v_i^{t_0}, \ v_i^{t_0} + 3\% v_i^{t_0}]$ (where $v_i^{t_0}$ is the initial velocity of aircraft $i$). Acceleration are bounded, based on Eurocontrol's base of aircraft data [6], namelly $\overline{u_i} = -\underline{u_i} = 4000$ NM / h$^2$. Terminal conditions are returning to the initial velocities ($v_i^{t_0}$) at final time ($t_f = 30'$). The number of aircraft, the *collision proximity* (*i.e.*, the minimal distance between aircraft which could occur if no maneuvers are done), and the initial aircraft velocities are reported in Table I.

In Table II, we compare the results obtained by applying a direct method, with detection step and resolution step equal to $15''$ and $1'$ respectively, on the whole time window (without considering the *postzone*) and the results obtained by decomposing the problem and applying the direct and the indirect methods as described in the previous sections.

 ISBN: 978-1-61208-237-0 

Table I
TEST PROBLEMS CHARACTERISTICS: NUMBER OF AIRCRAFT, COLLISION PROXIMITY, AND INITIAL VELOCITY FOR PROBLEMS WITH 4 AND 6 AIRCRAFT.

| instances | number of aircraft | collision proximity | initial velocity |
|---|---|---|---|
| pb_n4a | 4 | 0 NM | 400 NM / h |
| pb_n4b | 4 | 2 NM | 400 NM / h |
| pb_n6b | 6 | 2 NM | 400 NM / h |
| pb_n6c | 6 | 3 NM | 400 NM / h |

Table II
COMPARISON OF NUMERICAL RESULTS OBTAINED WITHOUT AND WITH APPLICATION OF THE PMP ON THE POSTZONE : VALUE OF OBJECTIVE FUNCTION, NUMBER OF ITERATIONS, CPU TIME, FOR 4 AIRCRAFT PROBLEMS.

| | Application of the PMP on the *POSTZONE* | | | | | |
|---|---|---|---|---|---|---|
| | without | | | **with** | | |
| instances | objective | it. | time | objective | it. | time |
| pb_n4a | $1.8 \times 10^4$ | 258 | $22'$ | $1.9 \times 10^4$ | 148 | $\mathbf{1'54''}$ |
| pb_n4b | $8.0 \times 10^3$ | 228 | $16'30''$ | $1.0 \times 10^4$ | 188 | $\mathbf{3'20''}$ |

From Table II, we can see that with the application of the PMP on the *postzone*, the CPU times are significantly reduced with respect to the classical resolution based on the direct method applied on the whole time window, up to $90\%$ (see pb_n4a). The application of the PMP on the *postzone* allows us to tackle larger aircraft conflict avoidance problems because it reduces the time window where the direct method is applied, hence it reduces the number of variables and constraints of the NLP. We can then solve 6 aircraft conflict avoidance problems, as show in Table III, we compare results obtained by using different *resolution* time discretization steps. Like in Table II, the *detection* step is $15''$ and we applied the PMP on the *postzone*.

Table III
COMPARISON OF NUMERICAL RESULTS WITH DIFFERENT CONTROL TIME DISCRETIZATION STEPS: VALUE OF OBJECTIVE FUNCTION, NUMBER OF ITERATIONS, CPU TIME, FOR 6 AIRCRAFT PROBLEMS.

| | Control time DISCRETIZATION step | | | | | |
|---|---|---|---|---|---|---|
| | $1'$ | | | **$5'$** | | |
| instances | objective | it. | time | objective | it. | time |
| pb_n6b | $1.6 \times 10^4$ | 342 | $21'$ | $1.7 \times 10^4$ | **43** | $\mathbf{0'35''}$ |
| pb_n6c | $1.0 \times 10^4$ | 317 | $20'$ | $1.0 \times 10^4$ | **51** | $\mathbf{0'44''}$ |

From Table III, we emphasize the importance of the *resolution* time discretization step. We can see that with the *resolution* step equals to $5'$, the CPU times are significantly reduced with respect to the configuration with the *resolution* step equals to $1'$, up to $97\%$ (see pb_n6b). The two above comparisons (Tables II and III) show, on the one hand, the advantage of the application of the PMP on the *postzone*, and on the other hand, the benefit to hybridize the two methods to solve larger aircraft conflict avoidance problems.

## VI. CONCLUSION

We considered an optimal control model for aircraft conflict resolution based on speed changes. We proposed a strategy based on hybridization of the direct method applied to the conflict *zone* and the indirect method applied to *postzone* where conflicts have been solved. First numerical results validate our approach. They show that the proposed decomposition strategy is beneficial in the context of the considered control problem, significantly reducing the computational time for solving the aircraft conflict avoidance problem.

## REFERENCES


[1] A. Bicchi, A. Marigo, G. Pappas, M. Pardini, G. Parlangeli, C. Tomlin, and S.S. Sastry, *Decentralized air traffic management systems: performance and fault tolerance*, IFAC International Workshop on Motion Control, Grenoble, FRA, pp. $279 - 284$, 1998.

[2] D. Bonini, C. Dupré, and G. Granger, *How ERASMUS can support an increase in capacity in 2020*. In Proceedings of the $7^{th}$ International Conference on Computing, Communication and Control Technologies: CCCT 2009, Orlando, Florida, 2009.

[3] A.E. Bryson and Y.-C. Ho, *Applied optimal control – optimization, estimation and control*, Taylor & Francis Group, 1975.

[4] S. Cafieri, P. Brisset, and N. Durand, *A mixed integer optimization model for air traffic deconfliction*, In Toulouse Global Optimization workshop, Toulouse, FRA, pp. $27 - 30$, 2010.

[5] N. Durand, J-.M. Alliot, and J. Noailles, *Automatic aircraft conflict resolution using genetic algorithms*, In proceedings of the Symposium on Applied Computing, Philadelphia, ACM, 1996.

[6] Eurocontrol Experimental Center. *User manual for the base of aircraft data*, Technical report, Eurocontrol, 2004.

[7] H. Huang and C. Tomlin, *A network-based approach to enroute sector aircraft trajectory planning*, American Institute of Aeronautics and Astronautics, 2009.

[8] J.K. Kuchar and L.C. Yang, *A review of conflict detection and resolution modeling methods*, IEEE Transactions on Intelligent Transportations Systems, vol. 1, no. 4, pp. $179 - 189$, 2000.

[9] D. Rey, C. Rapine, R. Fondacci, and N.-E. El Faouzi, *Technical report on the minimization of potential air conflict using speed control*, Technical report, Laboratoire d'Ingénierie Circulation Transport, IFSTTAR, 2011.

[10] C. Tomlin, G.J. Pappas, and S.S. Sastry, *Conflict resolution for air traffic management : a case study in multi-agent hybrid systems*, IEEE Transactions on Automatic Control, vol. 43, no. 4, pp. $509 - 521$, 1998.

[11] A. Vela, S. Solak, W. Singhose, and J.-P. Clarke, *A mixed integer program for flight level assignement and speed control for conflict resolution*, In Joint $48^{th}$ IEEE Conference on Decision and Control and $28^{th}$ Chinese Control conference, Shanghai, China, pp. $5219 - 5226$, 2009.


 ISBN: 978-1-61208-237-0 

# COMBAS: A Semantic-Based Model Checking Framework

Eduardo González-López de Murillas, Javier Fabra, Pedro Álvarez, Joaquín Ezpeleta
*Aragón Institute of Engineering Research (I3A)*
*Department of Computer Science and Systems Engineering*
*University of Zaragoza, Spain*
*Email: {edugonza, jfabra, alvaper, ezpeleta}@unizar.es*

***Abstract*—The introduction of semantic aspects in scientific workflows is a powerful approach that allows the analysis of the workflow prior to its development and deployment. In this paper, the COMBAS framework for the semantic-based model checking processing is presented. COMBAS integrates the required languages and tools and implements its own algorithms in order to allow the verification of properties on a model specified with the U-RDF-PN formalism, a high-level Petri net-based formalism, which introduces parametric semantic annotations in the model. COMBAS facilitates the generation of temporal logic formulae to express the properties that are going to be verified in the model as well as it provides system designers with an RDF and CTL adapted environment to browse and review the results. The suitability of the proposed framework is demonstrated by means of its application to the analysis of the EBI InterProScan scientific workflow.**

***Keywords*-Semantic Annotated Processes; RDF; SMT; High-level Petri Nets.**

## I. INTRODUCTION

Scientific computing applications are being used in a broad spectrum of domains related to science and human life such as geography, biology, or the public sector, for instance. Scientific workflows are a special type of workflows, which often underlies many large-scale complex e-science applications, such as climate modelling, structural biology and chemistry, medical surgery or disaster recovery simulation, among others. Scientific workflows have been progressively improved by means of the introduction of new paradigms and technologies in order to achieve more complex challenges. Once the deployment and execution of such workflows has been carried out, the next challenge is focused on the incorporation of semantic Web techniques [1], [2] in order to analyse their behaviour.

With the development of semantic technologies, the incorporation of semantic aspects allows scientists to more efficiently browse, query, integrate and compose relevant cross discipline datasets and services [1]. Scientific workflow executions are expensive in the use of execution resources, as well as a time consuming activity. For this reason, it is of special interest to dispose of tools and techniques making possible the analysis of the workflow behaviour prior to its execution. The aim of such analysis would be to ensure a good behaviour (It is a waste of time and money to realize after 20 hours executing a task that the output has not the correct information to feed the next task!) as well as facilitating having a very efficient (from the budget and time points of view) resource utilization. The result of the analysis should allow predicting the quality of the results and also identifying those parameters suitable to get the expected outcome.

The introduction of semantic aspects in workflows requires new models and analysis techniques, able to deal with such semantic aspects, to be considered. With this respect, in this paper, the COMBAS framework is presented. COMBAS seeks at helping system designers (scientists and business process developers, among others) in the task of validation and property analysis of workflows that include semantic aspects in the task specification. The analysis process should be as follows. First, the designer models the workflow by means of a Petri net (the advantages of using Petri nets for this purpose has been widely discussed in the literature [3], [4]). The transitions of the Petri net model would correspond to system actions changing the workflow state (mainly, the execution of tasks the workflow requires). Semantic information is attached to the transitions, corresponding to the formal specification of the task and including the description of the input and output parameters. Since some of the outputs could correspond to data received or computed at runtime, formal symbolic parameters (as usual in procedure specification) will be used to represent such data. On the other hand, preconditions and postconditions for tasks should be considered and described as expressions involving the parameters in the task specification.

The U-RDF-PN formalism (Unary RDF Annotated Petri Net formalism) [5], is a subclass of Petri nets defined to consider this class of systems. Semantic annotations are specified using the RDF (Resource Description Framework) [6]. With respect to pre- and post-conditions, SMT (Satisfiability Modulo Theories) [7] solvers represent a very useful tool. They are able to work with logical predicates and solve the decision problem when using background theories. Therefore, they are able to determine the satisfiability of a collection of logical predicates, according to a certain combination of such theories. Finally, it is necessary then to have a tool, which allows us to verify the satisfiability of certain predicates, and also a language to express such predicates. The analysis of the workflow is carried out

 ISBN: 978-1-61208-237-0 

using model checking techniques, which define the way to verify the model through its reachability graph by means of queries expressed in terms of CTL (Computation Tree Logic) predicates.

The remainder of this paper is organized as follows. The design and implementation details of COMBAS are first depicted in Section II, and its application to a real problem in the scientific computing area is conducted in Section III. Section IV briefly introduces the related standards and tools for model checking analysis. Finally, Section V concludes the paper and addresses future directions of the work.

## II. COMBAS: A SEMANTIC-BASED MODEL CHECKER

The architectural view of the COMBAS framework is depicted in Figure 1. COMBAS integrates a set of tools and techniques to cover the full cycle for the model checking-based analysis of semantically annotated workflows: from the generation of U-RDF-PN models, the corresponding reachability graph and its Kripke-structure, the creation and edition of queries and CTL formulae, the execution of the model checking process and, finally, the results browsing and reviewing. Let us describe the comprehensive steps a system designer must perform in order to achieve the verification process using the framework:

1) First, the scientist must design the workflow or the process using a Petri net. For that purpose, the graphical tool Renew [8] can be used. The resulting model will be then exported to the PNML standard (Petri Net Markup Language, ISO/IEC 15909), an XML based description for Petri nets. The initial marking of the model as well as the model itself must be semantically annotatedby means of RDF, building a set of XML specification files (RDF graphs and patterns). A graphical assistant will help in this process.
2) Now some steps are taken in a transparent way for the user, who can review the results at each stage. The reachability graph generator uses the previous files as an input and generates two outputs: the reachability graph (RG), which is stored in the RDF Triple Store, and a set of XML files that contain the relation between states and their markings and a collection of RDF/XML files representing the RDFGs (RDF Graph) corresponding to the marking. The graphical representation of the reachability graph, which can be viewed with the tool Graphviz, is also generated at this step. Both the results from the previous processing and the reachability graph can be browsed by means of a Web viewer.
3) An annotated Kripke structure is constructed from this RG (RDF-KS). The reachability graph of a U-RDF-PN system can be easily transformed into an RDF Annotated Kripke Structure [5].
4) Now, it is necessary to create the CTL formula to be verified with the model checker. The designer can build this formula by means of a user-oriented interface provided by the Web editor, which will export it as an XML file compliant with the corresponding schema.
5) Finally, the model checker uses the CTL formula along with the RDF Kripke structure (RDF-KS) to compute and generate the output. This output consists of a collection of XML files that represent and relate states on the RG validated with the corresponding excerpts from the CTL formula. Both input files and results from the analysis at this stage can be viewed and browsed through the Web interface in a very convenient and intuitive manner.

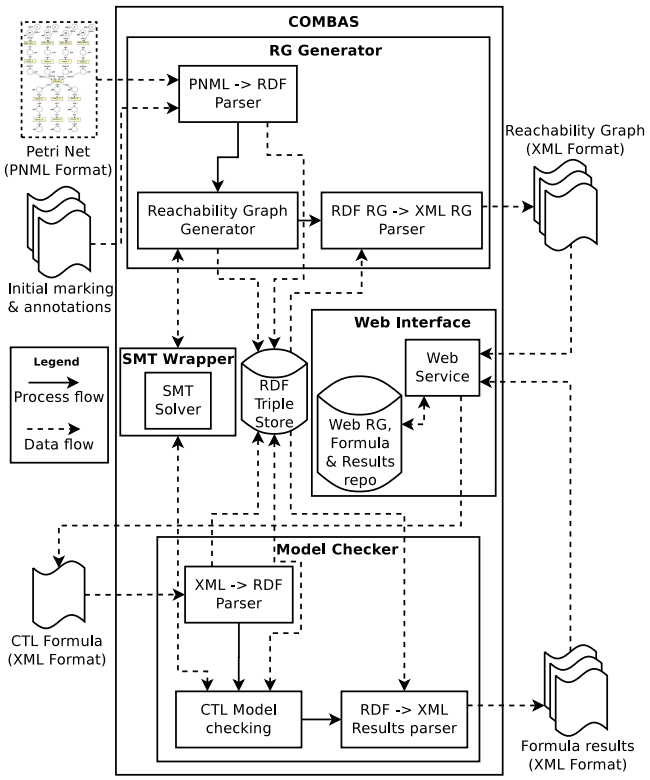



Figure 1. Architecture of the COMBAS model-checking framework.

All components in COMBAS expose an easy, flexible and usable interface, and the complexity of the graph generation, storage in the semantic triple store and verification processes are hidden from the user's perspective. During the model checking process, an RDF database (a triple store) is used. The COMBAS framework allows using several different RDF databases, such as the *AllegroGraph RDFStore* database or the *Virtuoso RDF Store*, for instance. The only requirement of a candidate RDF store is that it must allow to be accessed through a SPARQL interface. In this work the Virtuoso RDF Store has been used. As a result from the processing, the truth about the verification of the formula is obtained. Moreover, a graph depicting the

 ISBN: 978-1-61208-237-0

reachability graph states can be browsed using a graphical Web-based enabled interface. Doing so, it is possible to find the specific situations in which a predicate violates some wanted condition, having a better insight of the workflow behaviour, and thus facilitating the way to improve the workflow.

Internally, the generation of the reachability graph is based on the classical algorithm used for computation of the reachability graph in Petri nets and making the necessary modifications to adapt to the semantic nature of annotations [5]. Then, the implementation of our model checker is based on an adaptation of the labelling algorithm proposed in [9]. The inputs are an RDF-KS and an RDF CTL formula. Both inputs are stored into the RDF database following the corresponding RDF schemas. Basically, the algorithm computes the set of states of the input system $\mathcal{M}$ that satisfy the given CTL formula $\phi$. This process consists of three steps. Initially, the formula $\phi$ is translated into an equivalent formula in terms of the connectives $AF$, $EU$, $EX$, $\top$, $\wedge$ and $\neg$. The equivalence rules applied are defined as part of the labelling algorithm. Secondly, the states of $\mathcal{M}$ satisfying subformulas of $\phi$ ($\psi$) are labelled, starting with the smallest subformulas and finishing with the original formula. Finally, the algorithm returns the states labelled with $\phi$.

### A. Reachability graph generation

The reachability graph generation process needs a valid model to be used as an input: an U-RDF-PN class Petri net, with an initial marking corresponding to the system's initial state [5]. The semantic annotations can be linked to three different elements of the net: *arcs*, where it is possible to find RDF patterns; *transitions*, where guards can be defined; and *places*, where tokens are stored as RDF graphs. The generated reachability graph is stored also as RDF triples, according to the ontology depicted in Figure 2.

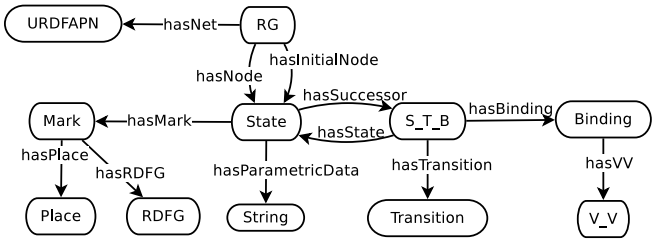



Figure 2. Ontology for the Parametric Reachability Graph.

In this paper, we are considering the parametric extension over U-RDF-PN as defined in [5]. The concept is very simple: to add parameters to the model, instead of static values. This way, representing general cases of a process or workflow consists of the addition of logical propositions to the information, which annotates each state in the reachability graph. Also the guards in transitions must be specified making use of such annotations.

Let us now briefly describe the main differences of the parametric approach with respect to the ordinary model checking approach presented in [5] regarding to the reachability graph generation. First of all, a parametric U-RDF-PN, which is an ordinary U-RDF-PN annotated with parametric statements in some of its transitions (as guards), places (as initial marking) and arcs (as part of the RDFG-Patterns), is used as an input. It has been necessary to implement a wrapper in order to make use of an SMT solver, crucial when working with logic and parameters. This way, the generation process is similar to the one of an ordinary U-RDF-PN, except for the following considerations:

1) The initial state is formed not only by the initial marking of every place in the net, but also by the parametric initial marking of each state. In case there is not such marking, the logical statement $true$ will be considered as initial parametric marking.
2) For every transition that has a parametric guard, the validity of such guard must be verified. This is performed making use of a SMT solver to check if the conjunction of the logical statements in the guard and the statements of the current state are satisfiable.
3) When generating a new state for the RG, its parametric marking is formed by the conjunction of the one of the parent state and the guard of the triggered transition.
4) Also when introducing a new state in the store, it is necessary to check if it is unique. To do so, the generator must compare the semantic part, and also the parametric part. When comparing the last one, an SMT solver is used to check the equivalence of both logical statements (P and Q), observing the satisfiability of the formula $P \rightarrow Q \wedge Q \rightarrow P$.

### B. Edition and visualization service

The basic features of this Web user interface are: 1) reachability graph visualization; 2) creation and edition of CTL formulae used by the Model Checker; and 3) review of Model checking results. The initial aim was to integrate all this functionalities in a single interface. Let us now detail the components of this interface.

**RG visualization.** As previously mentioned, the output of the RG generator consists of a main XML file, which describes the RG graph structure and a collection of XML files containing the RDF graphs that mark every state in the RG. In case we are dealing with a parametric model, also several *SMT-Lib* files will be generated. These files contain the parametric predicates corresponding to each state. The generated graph will contain a big amount of states, around hundreds in case it is a simple model, but it will raise exponentially as the complexity increases. Therefore, the review of the graph can be a tedious task. This was one of the principal aims to develop a viewer to represent the graph as a diagram, so it makes easier to check the marking of every state.

 ISBN: 978-1-61208-237-0 

The developed application makes possible to select a certain reachability graph, and visualize its structure in a graphical way, allowing to consult the marking in an easy and intuitive way.

**Formula edition.** Due to the purpose of this project, it is also important to provide an effective tool to edit CTL formulae. It represents an important benefit to the verification process, avoiding typing mistakes, and improving user experience because it reduces the learning curve from the user point of view and increases the formula creation and editing speed. That is why it is important to design a tool as intuitive as possible. It seemed right to develop a formulae constructor, which could provide a list of components and a mechanism to include them in the formula in an interactive way. Also, we thought it would be interesting to include a verification phase to make sure that all the input is well constructed, avoiding syntactic errors.

Among the features of this part, we find interesting the ability to create/edit formulae, visualize them in a graphical environment and the interactive creating using intuitive user interfaces.

**Model Checker results review.** A very important stage in the model checking process is the result review. In this stage, the user needs to verify the results and analyse them in order to identify the errors, if any, in the model. Therefore, it is crucial to provide the tools to manage the output of our Model Checker.

The developed interface makes possible to select a graph and inspect the execution result of a certain formula, using the reachability graph visualization, and checking the marking and satisfiability of the formula components in every state in the RG.

### C. Model checking process

The ontology represented in Figure 3 is used to specify the input CTL formula for the model cheker. It may contain any of the unary CTL operators $AF, EF, AX, EX, AG, EG, NOT$, or binary ones $AND, OR, AU, EU, IMP$, or any of the terminal nodes $TRUE, FALSE, RDFG, RDFGP, SMT-EXP$ being RDFG a RDF-Graph, RDFGP a RDF-Graph Pattern, and SMT-EXP a logical statement in SMT-Lib format.

The other input of our model checker is the reachability graph, which generation process has been previously explained. Because this RG may contain parametric data, we still need to make use of an SMT solver. So, the same SMT wrapper mentioned in II-A is required.

The verification process is similar to the one used with ordinary U-RDF-PN and CTL formulae as described in [5]. The main difference lies in the terminal node $SMT-EXP$, which is represented by a logical statement in SMT-Lib format. In order to know if a state satisfies such statement, we need to make use of the SMT wrapper. Therefore, the parametric marking of such state must be compared to the parametric statement in the formula. Being $P$ the parametric data stored in the state, and $Q$ the content of the $SMT-EXP$ node in the formula, the logic expression $P \vee Q$ is used to check the validity of the state parametric marking ($P$) and the formula parametric expression ($Q$). This way, we will know if both logic statements are compatible or contradictory.

## III. Analysis of the EBI InterProScan workflow

In order to show the suitability of the COMBAS framework, in this section the InterProScan workflow is going to be analysed. This workflow relies on the use of the EBIs WSInterProScan service [10], an its workflow can be checked out at the Myexperiment.org community [11]. The workflow receives as an input the protein sequence to be processed, a user email address for notification purposes and a few more parameters required for the analysis. Starting with this input, a protein sequence is searched inside a set of protein families and domain signature databases integrated in InterPro [12]. As a result, a set of matches are properly formatted and returned. These matches are also annotated with the corresponding InterPro and GO term assignments (for further explanations about these assignments, please refer to the experiment page). In order to execute the `runInterProScan` activity, two different Web services, `runInterProScan1` and `runInterProScan2`, are available in a repository. Both services are able to carry out the protein analysis, which represents the more expensive part of the experiment.

Figure 4 depicts the workflow modelled as a High-Level Petri net using the Renew tool [8] and semantically annotated according to the parametric U-RDF-PN formalism. All the data flowing through the workflow have been semantically annotated using instances based on the Protein Ontology [13]. Nodes corresponding to the original workflow are in dark grey, whereas some additional structures, which have been included in order to provide more generality in the generated reachability graph (representing a higher variety of cases and states to analyse) are in light grey.

Let us now briefly describe the parametric net. The seven annotated places at the top of the net are the main inputs of the workflow. Five of those inputs, which represent the job parameters and are also necessary to configure properly the service *InterProScan*, are grouped in the left side: goterms_default, async_default, crc_default, seqtype_default and email_address. These inputs are required for transition *job_params* to be enabled and then fired. In the right side, two places represent the protein sequence data: sequence_or_ID and input_datatype_default. They are also required to fire the transition *Input_data*. From the morphology of the net, we can observe that all mentioned parameters are necessary for the execution of the service *InterProScan*. When this service is executed, the process enters the bottom half of the net, in which the output is converted to the proper

 ISBN: 978-1-61208-237-0 

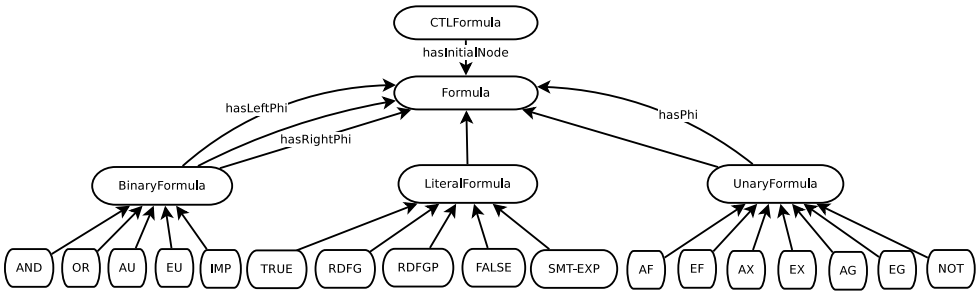



Figure 3. CTL formula ontology with parametric elements.

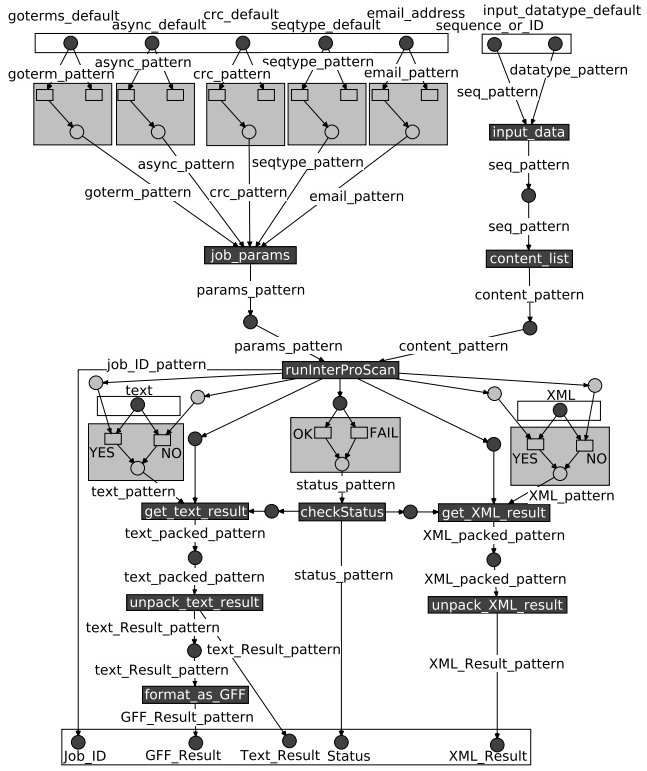



Figure 4. Parametric U-RDF-PN modeling the InterProScan workflow for the analysis of a protein sequence.

format. The are two places in charge of selecting the proper format, and they also represent an input of the workflow: 1) text and 2) XML. These places are semantically annotated, containing, each of them, the reference to a parameter. Below both places, a light grey section has been added. The purpose of this structure is to assign two different values to the mentioned parameters (true or false), so it is possible to analyse the behaviour of the workflow in any situation, using a single reachability graph. The same technique has been used in the top part of the net, with the five parameters described earlier. At the bottom, there are five output places: Job_ID, GFF_Result, Text_Result, XML_Result and Status. These are the places in which the results of the workflow will be stored. They represent the five different output data types we can obtain: the job identifier, the result in GFF, text and XML format, and the status value, respectively.

This model is then processed by the Reachability Graph Generator to compute the graph. The resulting reachability graph has been built in 1343 seconds, and it is composed of 798 states.

### A. Properties checking

Due to the size of the generated reachability graph, a manual checking would not be feasible, resulting in a tedious and complex task. This represents the perfect situation to apply the proposed solution to analyse the problem. Let us now design and formulate some different queries over the model in order to verify its correctness. These queries are expressed as formulae in CTL language and have been implemented using the corresponding XML format, which is ready to be processed by COMBAS. The Model Checker is then in charge of verifying the satisfiability of the formula.

The first query to be consider when verifying a workflow could be to satisfy if the process is able, in some situation, to end in a proper way. This will happen if the output places in the net are marked with a resulting graph. Five different output places are distinguished: Job_ID place; Text_Result place; GFF_Result place; Status place and, finally, XML_Result place. These places correspond to the ones located in the bottom of the net presented in Figure 4. Therefore, to verify if those places are reached, it is possible to check if there exists any state in which such places are marked by any graph (the empty graph represents the 'any graph' concept). This query corresponds to the following CTL predicate: $EF(RDFG_{Text\_Result} \vee RDFG_{GFF\_Result} \vee RDFG_{XML\_Result} \vee RDFG_{Status} \vee RDFG_{Job\_ID})$

This formula, when executed in the Model Checker, provides the next output:

```
INFO [main](CombasApp.java) Model satisfies the formula!
INFO [main](CombasApp.java) Output files generated.
INFO [main](CombasApp.java) Checked in 17159 millis
INFO [main](CombasApp.java) Formula: formula_1h9wbb3xwwj7c
INFO [main](CombasApp.java) Model: netId1337458105565_RG
```

This means that the model satisfies the formula and, therefore, it is possible to reach a state where any of the

specified states is marked. The next step would be to verify if the workflow is always able to finish. The corresponding formula is expressed as follows: $AF(RDFG_{Text\_Result} \vee RDFG_{GFF\_Result} \vee RDFG_{XML\_Result} \vee RDFG_{Status} \vee RDFG_{Job\_ID})$

The results from the model checking state that this formula is not satisfied, which means that the model will not always reach any of those places in the net. This can happen, for instance, when the net reaches a deadlock state with no transitions ready to fire.

Due to the morphology of the net, in order to trigger the transition named *runInterProScan* its source places must be marked. These places, which are referred to as *params* and *content_list* places, will be marked in some situation. Otherwise, the first query would not be satisfied (the one asking if it is possible to reach any of the output places). Then, it would be interesting to know if every time these two places are marked the workflow will reach any of the output places. This is the corresponding CTL query: $AG((RDFG_{params} \wedge RDFG_{content}) \rightarrow AF(RDFG_{GFF\_Result} \vee RDFG_{Text\_Result} \vee RDFG_{Status})$. When processed, the Model Checker states that the model satisfies the property.

Another useful query is related to verifying if every time a result is obtained (at least one of the output places is reached), the value of the status result is *true* (so the token in the place *status* has the value *true* for the *status* variable). The mentioned property in CTL is: $AG((RDFG_{Text\_Result} \vee RDFG_{GFF\_Result} \vee RDFG_{XML\_Result}) \rightarrow RDFGP_{Status(true)}$. The predicate satisfies the model, which means that every time we get a result, the status variable is set to *true*.

As shown, the model is very suitable to be analysed by means of CTL queries related to the properties we want to verify. Once the input model and its corresponding annotations have been defined, COMBAS allows to easily perform a model checking process, browse through the reachability graph checking in which states the conditions are not being verified, review the results, etc. No prior knowledge of the model checking algorithm is required for the scientist nor internal or technological details. Just using the interface with the framework allows a set of advanced features and tools for the analysis of complex systems.

## IV. Related work

There already exist numerous Model Checking tools, which can be classified according to the language they use to express models and properties. Some of the tools allow us to define properties in Computation Tree Logic and Linear Temporal Logic (CTL and LTL, respectively), although in this paper we will focus on the most widespread alternatives supporting CTL as the language to express properties. BANDERA uses code analysis techniques to verify properties on models defined in Java, while CADENCE SMV uses plain model checking to verify models expressed in Cadence SMV, SMV or Verilog languages [14]. Another example is NuSMV, which supports CTL, LTL and PSL to define properties, in order to do plain model checking on SMV models. APMC uses an approximate probabilistic technique on Reactive Modules, and properties are expressed in PCTL and PLTL languages [15]. PRISM supports CSL in addition to both previous languages, and uses a probabilistic model checking technique on models expressed in a big variety of languages, like PEPA, PRISM language or Plain MC [15]. Between the probabilistic model checkers, we find MCMR, which supports properties defined in CSL, CSRL, PCTL or PRCTL languages, and uses real time and probabilistic model checking on Plain MC models.

Other tools support the language $\mu$-calculus to describe properties, like TAPAs, which also supports CTL on CCSP models. ARC performs plain model checking of CTL* properties on AltaRica models. GEAR is an alternative tool, which accepts $\mu$-calculus properties too, among others like AFMC and CTL [16]. CWB-NC is a similar tool, and uses plain and temporized model checking on CCS, CSP, LOTOS and TCCS models, using AFMC, CTL and GCTL properties. Finally, MCMAS is a plain and epistemic model checker on ISPL models, which verifies CTL and CTLK properties [17].

Regarding the use of SMT solvers, there exists a wide variety solvers to choose from. The main characteristic aspects to choose among them are the collection of supported theories and languages, the programming language they are implemented in and also its portability and reusability features. However, there are other points to take into account, like the activity of the user community, how often new versions are released, and the quality of the documentation. Regarding to these concepts, the list of options is considerably reduced. CVC, OpenSMT and STP represent the most active projects. STP only supports formulas over the theory of bit-vectors and arrays, therefore it does not represent a valid solution for our problem, because the main theory we need support for is linear arithmetic, specifically over reals, integers and booleans. Although OpenSMT is a valid option, we choose CVC instead because it is a more stable and production ready alternative. More precisely, we choose CVC3, which is the stable version of CVC due to the instability of CVC4, which is in beta phase.

With respect to the description of predicates to perform the model checking process, the Satisfiability Modulo Theories Competition (SMT-COMP) is celebrated each year, since 2005, with the purpose of encouraging the development of SMT solvers, and also as an impulse to the adoption of the Satisfiability Modulo Theories Library standard (SMT-LIB) [18]. SMT-LIB is a format designed by the community that tries to unify the description of the background theories and the inputs/outputs for SMT solvers as well as to provide a collection of benchmarks to boost the development of this kind of tools. Therefore,

 ISBN: 978-1-61208-237-0 

compared to other formats like CVC language or DIMACS, SMT-LIB represents the most recommendable option to keep compatibility and portability for our predicates.

To sum up, there are a great variety of standards and tools that are involved in the model checking process. However, there are no frameworks that integrate them in a comprehensive way in order to allow scientists to analyse their scientific workflows prior to their deployment and execution, becoming COMBAS a suitable approach for this kind of scenarios.

## V. Conclusions and future work

In this work, the COMBAS framework to carry out model checking of semantically annotated processes and workflows has been presented. COMBAS represents a novel approach to add RDF semantic information to the source model as well as to achieve its processing and analysis. This will allow the scientists community to take advantages of the new semantic technologies and also to facilitate sharing workflows and tasks as well as reasoning about the results and behaviours. The framework allows verifying and viewing the intermediate structures and the results by means of a visual environment able to handle RDF and CTL aspects. The use of the Unary RDF Annotated Petri Net formalism (U-RDF-PN) for the modeling of processes has been extended with the addition of parametric aspects, allowing to consider a more flexible and powerful analysis for complex systems. Our proposal represents a novel approach to manage semantic-based computation problems by means of the integration of several model checking related standards and tools, and its suitability has been demonstrated in the analysis of the InterProScan workflow.

The COMBAS framework is currently being extended in order to integrate other standards and tools used in the scientific community. Also, it is being applied to solve some open challenges in the scientific workflow area, and the reachability graph generator is being adapted in order to be executed in grid environments, therefore improving the overall execution costs.

## Acknowledgments

This work has been supported by the research project TIN2010-17905, granted by the Spanish Ministry of Science and Innovation, and the regional project DGA-FSE, granted by the European Regional Development Fund (ERDF).

## References

[1] C. Berkley, S. Bowers, M. Jones, B. Ludäscher, M. Schildhauer, and J. Tao, "Incorporating Semantics in Scientific Workflow Authoring," in *SSDBM 2005*, 2005, pp. 75–78.

[2] C. Goble, J. Bhagat, S. Aleksejevs, D. Cruickshank, D. Michaelides, D. Newman, M. Borkum, S. Bechhofer, M. Roos, P. Li, and D. De Roure, "myExperiment: a repository and social network for the sharing of bioinformatics workflows," *Nucleic Acids Research*, 2010.

[3] W. M. P. van der Aalst, "The application of Petri nets to workflow management." *Journal of Circuits, Systems, and Computers*, vol. 8, no. 1, pp. 21–66, 1998.

[4] T. Gubala, D. Herezlak, M. Bubak, and M. Malawski, "Semantic composition of scientific workflows based on the Petri nets formalism," in *E-SCIENCE'06*.

[5] M. J. Ibáñez, J. Fabra, P. Álvarez, and J. Ezpeleta, "Model checking analysis of semantically annotated business processes," *Systems, Man, and Cybernetics – Part A: Systems and Humans*, 2012.

[6] P. Hayes, "RDF Semantics," W3C, Tech. Rep., February 2004, http://www.w3.org/TR/rdf-mt/.

[7] C. Barrett, R. Sebastiani, S. Seshia, and C. Tinelli, "Satisfiability modulo theories," *Handbook of Satisfiability*, vol. 4, 2009.

[8] O. Kummer and F. Wienberg, "Renew - the Reference Net Workshop," in *Tool Demonstrations, 21st International Conference on Application and Theory of Petri Nets*, 2000, pp. 87–89.

[9] E. A. Emerson, "Temporal and Modal Logic," *Handbook of Theoretical Computer Science*, vol. Formal Models and Semantics, pp. 995–1072, 1990.

[10] WSInterProScan, available at http://www.ebi.ac.uk /Tools/webservices/services/archive/pfa/wsinterproscan [retrieved: September, 2012].

[11] EBI's WSInterProScan workflow at MyExperiment community, available at http://www.myexperiment.org/workflows/814.html [retrieved: September, 2012].

[12] InterPro protein sequence analysis & classification, available at http://www.ebi.ac.uk/interpro/ [retrieved: September, 2012].

[13] Protein Ontology, available at http://bioportal.bioontology.org/ontologies/1052 [retrieved: September, 2012].

[14] D. Garlan, S. Khersonsky, and J. S. Kim, "Model checking publish-subscribe systems," in *SPIN'03*, 2003, pp. 166–180.

[15] M. Duflot, L. Fribourg, T. Hérault, R. Lassaigne, F. Magniette, S. Messika, S.Peyronnet, and C. Picaronny, "Probabilistic model checking of the csma/cd protocol using prism and apmc," *Electr. Notes Theor. Comput. Sci.*, vol. 128, no. 6, pp. 195–214, 2005.

[16] O. Grumberg and H. Veith, Eds., *25 Years of Model Checking - History, Achievements, Perspectives*, ser. Lecture Notes in Computer Science, vol. 5000. Springer, 2008.

[17] A. Lomuscio, H. Qu, and F. Raimondi, "Mcmas: A model checker for the verification of multi-agent systems," in *CAV*, 2009, pp. 682–688.

[18] SMT-Lib: The Satisfiability Modulo Theories Library, available at http://www.smtlib.org/ [retrieved: September, 2012].

 ISBN: 978-1-61208-237-0 

# Applications of Random Finite Element Method in Bearing Capacity Problems

Md Mizanur Rahman
School of Natural and Built Environments
University of South Australia, UniSA
Mawson Lakes, Adelaide, Australia
E-mail: Mizanur.Rahman@unisa.edu.au

Hoang Bao Khoi Nguyen
School of Natural and Built Environments
University of South Australia, UniSA
Mawson Lakes, Adelaide, Australia
E-mail: nguhy030@mymail.unisa.edu.au

***Abstract* – This paper aims to apply a new methodology in bearing capacity analysis, often referred to as random finite element method (RFEM). This method considers the variability of soil parameters within the finite element method (FEM) by generating a Gaussian random field for the parameters within finite elements. The local average subdivision method (LAS) was used in this study to generate the Gaussian random field. However, soil parameters are not, generally, randomly distributed within neighboring soil elements; they tend to be correlated over a distance. Thus, the correlation length, the distance over the soil parameters are correlated to each other, was considered in this study. The Monte Carlo simulation was done for bearing capacity problem and some statistical and probabilistic methods were applied for analyzing the results to get the failure probability of footing on clay. This study would help to understand the effect of variability of soil parameter by using RFEM; so that the safety issues of geotechnical design can be determined in terms of probability of failure.**

***Keywords – Random finite element method; Monte Carlo simulations; Gaussian random field generation; Statistical distribution; Probabilistic method; Correlation lengths; Risk assessment.***

## I. INTRODUCTION

In geotechnical engineering analysis, the soil parameters are often considered as constant within a soil layer. For instance, the bearing capacity of a strip footing is determined using constant value of $c$ and $\phi$ for each layer, where $c$ is the cohesion and $\phi$ is the angle of internal friction of soil particle. Thus, the equation for determining the bearing capacity on the surface of clay can be expressed as in Eq. (1) by Terzaghi [1]

$$q_u = cN_c . \quad (1)$$

where $q_u$ is bearing capacity, $c$ is cohesion of soil, $N_c$ is bearing capacity factor = 5.14 for $\phi$ = 0. Eq. (1) gives bearing capacity of clay based on cohesion, $c$ only. Thus, a design based on Eq. (1) can be conservative or optimistic with an associated risk based on how geotechnical engineer determine $c$ for a clay layer. In practice, a value of $c$ is approximated in such a way so that it gives a conservative estimation for design.

However, in reality, $c$ is not the same, i.e., it varies point to point in a soil layer because of complicated geological formation process. The deterministic approach as in Eq. (1) may be conservative, but do not consider realistic condition. Moreover, the implication of this simplification is not explored in details yet. Thus, it is necessary to consider the variation of $c$ in bearing capacity analysis and compare with deterministic approach. So, a new method, (RFEM) Random Finite Element Method was used in this study to calculate bearing capacity.

In RFEM, the material parameters are varied and distributed within finite elements. For example, the parameter $c$, in the above problem, is varied within finite elements for a clay layer. However, the variation of the parameters should be consistent with field condition. According to Fenton and Griffiths [2], the geotechnical parameters can possibly have several reasonable distributions, which include log-normal, normal and tanh bounded distribution. The parameter $c$ is generally assumed to be log-normally distributed with an advantage of avoiding negative $c$ that has no physical meaning [3, 4]. Then, based on the log-normal distribution, the Gaussian random field of $c$ is generated by Local Average Subdivision, LAS method for finite elements. However, the variation of $c$ within soil elements is not purely random; a smoother change of $c$ between two neighboring soil elements is expected than two elements at a distance apart.

A spatial correlation length is used within the random field to describe the distance over which random values tend to be correlated. When the correlation lengths in horizontal and vertical directions are same, the soil elements can be assumed as isotropic. Most of the previous studies focused on isotropic condition.

This paper focuses on the variability and the effect of anisotropic distribution of material parameters in geotechnical analysis. Some statistical and probabilistic methods, such as random field generator with log-normal distribution, correlation length and Monte Carlo simulations [2] are used within the finite element analysis. The probability of failure of footing on clay was obtained from cumulative distribution function [2].

## II. RANDOM FINITE ELEMENT METHOD

RFEM considers that the engineering parameters are distributed over a correlation length as a Gaussian random field [2] generated by local average subdivision [2] within

 ISBN: 978-1-61208-237-0 

finite elements. These techniques are discussed in following subsections.

### A. Correlation length

In reality, soil parameters within the neighbouring points are similar, i.e., correlated. The distance over that parameter is correlated is called correlation length or scale of fluctuation. According to Fenton and Griffiths [2], the correlation coefficient in isotropy can be determined as

$$\rho(|\tau|) = -2|\tau|/\theta . \quad (2)$$

where $\rho$ is correlation coefficient, $\tau$ is distance between two points. When the correlation lengths in horizontal and vertical direction are same, then it is called isotropic condition. Most of the previous studies are based on isotropic condition. However, in reality the correlation length in horizontal direction is higher than the vertical direction as geologically soil forms in horizontal layers [5, 6]. The condition, when horizontal and vertical correlation lengths are different, is called anisotropic condition. The anisotropy condition can be expressed as

$$\rho(|\tau_x|,|\tau_y|) = \exp\left\{-\sqrt{(2\tau_x/\theta_x)^2 + (2\tau_y/\theta_y)^2}\right\} . \quad (3)$$

where $\tau_x$ and $\tau_y$ are the distance in horizontal and vertical direction respectively, $\theta_x$ and $\theta_y$ are horizontal and vertical correlation lengths, respectively.

The correlation coefficient, $\rho$, takes an important role in the random field generation. They will take part in LAS process in order to correlate the parameters in finite element meshes. The functions of correlation coefficient are illustrated in Fig.1.

In isotropy, the coefficient is the same in both directions, thus it generates symmetric curve to the centre in both horizontal ($\tau_x$) and vertical directions ($\tau_y$) as shown in Fig. 1a. But, in anisotropy, the correlation coefficient is different in horizontal and vertical directions. In Fig. 1b, when the horizontal correlation length increases to a high value of 100, the correlation length becomes very close to 1.0 in horizontal direction. On the other hand, the vertical correlation coefficient is high at the middle and lower at the side. Hence, the parameters in horizontal direction will correlate better than the vertical direction.

### B. Local average subdivision

The LAS technique is one of the techniques that widely used to generate the Gaussian random field. It was introduced in Vanmarcke [7]. At first, the global average is generated with mean of zero and unit variance of 1 [2]. The global average is defined by local average theory as in Fenton and Griffiths [2]. The global average can be written in terms of expectation function as in Eq. (4)

$$E[X_T(t)] = E\left[\frac{1}{T}\int_{t-T/2}^{t+T/2} X(\xi)d\xi\right] . \quad (4)$$

where $T$ is the mesh size, $t$ is location at the centre of each mesh cell, and $\xi$ is location at moving average. The covariance of the local averages is defined by using a variance function as defined in Fenton and Griffiths [2].

$$\gamma(T) = \frac{2}{T^2}\int_0^T (|T| - \tau)\rho(\tau)d\tau . \quad (5)$$

where $\gamma$ = variance function. The variance function indicates the average correlation coefficient, $\rho$ between each pair of 2 separated points within the defined area, where $\rho(\tau)$ is defined in Eq. (3). Then, based on the covariance between pair of cells, the LAS process can be generated. In LAS process, one parent cell, $Q$ is subdivided into 4 equal cells, which are called child cells. The Fig. 2 illustrates the subdivision process, where $Q$ is the parent cell and $G$ is the child cell.

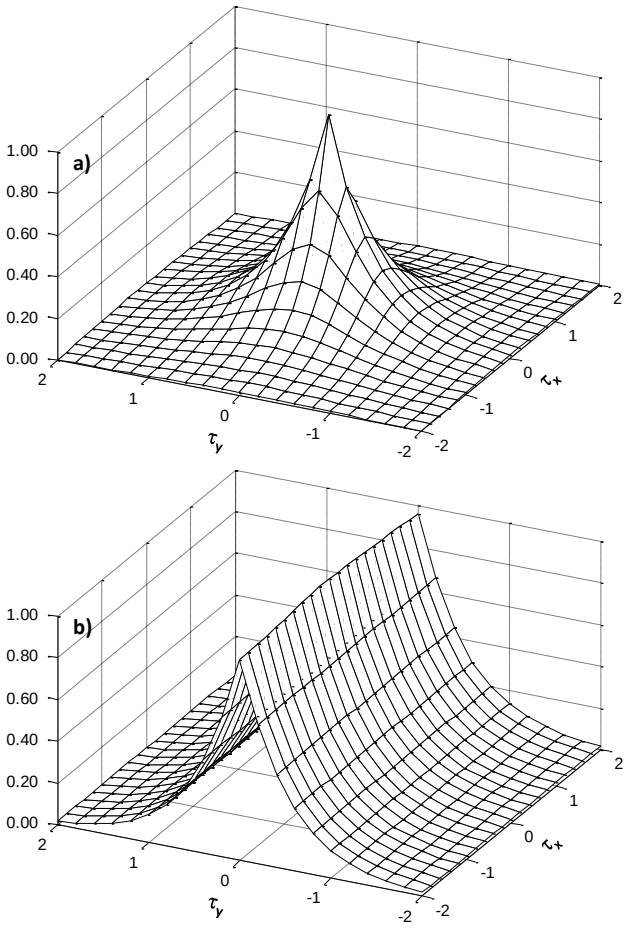



Figure 1. (a) Correlation coefficient in isotropy, $\theta_x = \theta_y = 1.0$ and (b) Correlation coefficient in anisotropy, $\theta_x = 100$ and $\theta_y = 1.0$

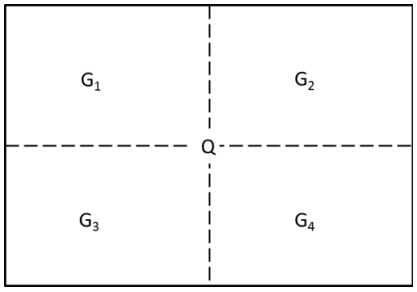



Figure 2. LAS process.

The function which is used for the LAS process can be expressed as

$$G = A^T Q + LU \,. \tag{6}$$

In this process, $U$ is indicated as the vector of independent standard normal random variables with mean zero and unit standard deviation.

The covariance described the relationship between the cells and can be written as the following equations.

Covariance between parent cells

$$R = E\left[QQ^T\right]. \tag{7}$$

Covariance between parent and child cells

$$S = E\left[QG^T\right]. \tag{8}$$

Covariance between child cells

$$B = E\left[GG^T\right]. \tag{9}$$

Then, the matrices $A$ and $L$ can be determined by

$$A = R^{-1}S \,. \tag{10}$$

$$LL^T = B - S^T A. \tag{11}$$

The method used to generate matrix $L$ is called matrix decomposition method, which is used to find the lower triangular matrix from the defined matrix. According to Fenton and Griffiths [2], the LAS technique is most reliable and efficient technique to generate the random field for RFEM and also give the best-fit results to the theory.

### C. Random field transformation

The cohesion of soil, $c$, is considered as log-normal distribution and thus the log-normal distribution transformation in Gaussian random field can be expressed as

$$X(i, j) = \exp[\mu + \sigma G(i, j)]. \tag{12}$$

where $X(i, j)$ is transformed random field, $\mu$ is mean, $\sigma$ is standard deviation, and $G(i, j)$ is random field generated by LAS process in Eq. (6). A distribution of $c$ is shown in Fig. 3.

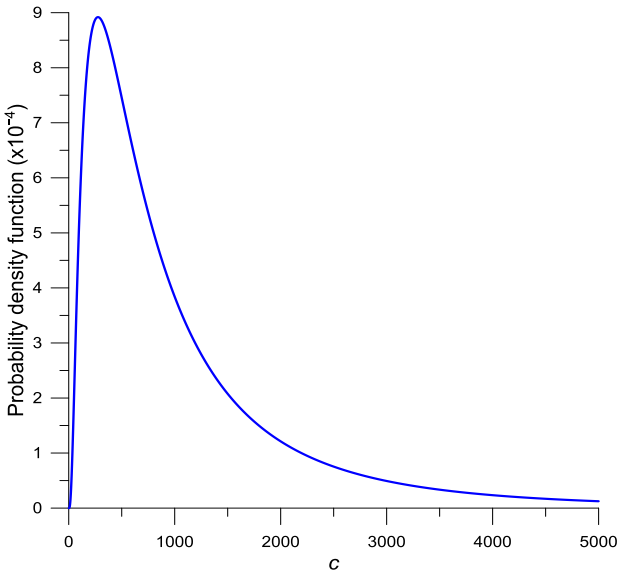



Figure 3. Log-normal distribution with μ=400 and σ=300

### D. RFEM and variability of soil parameters

An elastic-perfectly plastic stress-strain law with Tresca failure criterion is used in finite element formulation. The theoretical method is described in details in Chapter 6 of the text by Smith and Griffiths [8]. The software used in this study is called Mrbear2D and freely available online. The soil parameters including dilation angle, $\alpha$, elastic modulus, $E$ and Poisson's ratio, $\nu$ are assumed to be deterministic with specific constant values. However, the undrained shear strength parameter, $c$ was a variable in terms of coefficient of variation ($COV$) can be expressed as

$$COV = \sigma / \mu \,. \tag{13}$$

where $\mu$ is mean and $\sigma$ is standard deviation of $c$. The distribution of $c$ within finite element meshes for small correlation of $\theta_x = \theta_y = 0.1$ and $\theta_x = 10$ & $\theta_y = 4.0$ are shown in Fig. 4a & b. A smoother variation is apparent for higher correlation length.

## III. MONTE CARLO SIMULATIONS

The LAS technique will generate a random field in each Monte Carlo simulation [2]. This type of simulation is applied in order to consider the possible variability of

 ISBN: 978-1-61208-237-0 

parameter in geotechnical analysis. The Monte Carlo simulation process continues with LAS technique and distribution function until simulations obtain stable results. By using RFEM software, a reasonable number of 1000 FEM simulations were done to obtain stable result as shown in Fig. 5.

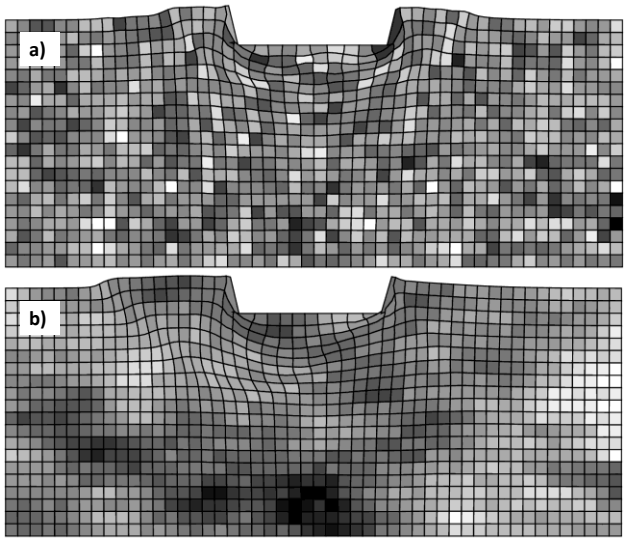



Figure 4. The mesh model with correlation coefficient; (a) $\theta_x = \theta_y = 0.1$ and (b) $\theta_x = 10$ and $\theta_y = 4$

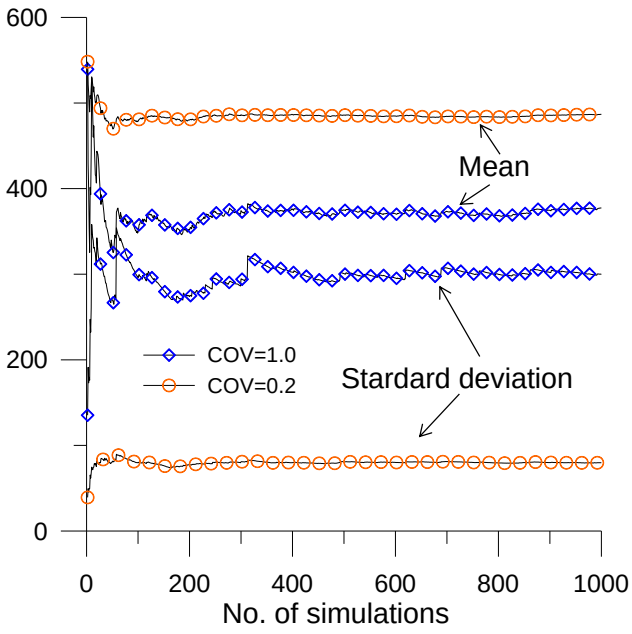



Figure 5. Mean value and standard deviation of the results through Monte Carlo simulations.

The y-axis and x-axis in Fig. 5 present bearing capacity and number of simulations, respectively. A number of 1000 FEM simulations, using Monte Carlo simulation, is adequate for this study. However, the effect of COV on bearing capacity is also apparent in the figure.

## IV. RESULTS

When doing the analysis with 1000 Monte Carlo simulations, there are 1000 bearing capacities for 1000 different $c$ fields. In RFEM, the cohesion, $c$ was the input with log-normal distribution. Then, the bearing capacity factor, $N_c$ will be determined by using cohesion and bearing capacity in Eq. (1).

$$N_c = q_u / \mu_c . \tag{14}$$

Because $c$ log-normally distributed, $N_c$ then can be considered as log-normally distributed. Thus, the mean and standard deviation of $N_c$ can be expressed as

$$\mu N_c = \sum_{i=1}^{1000} N_{ci} / 1000 . \tag{15}$$

$$\sigma N_c = \sqrt{\sum_{i=1}^{1000} (N_{ci} - \mu N_c) / 1000} . \tag{16}$$

where $N_{ci}$ is $N_c$ from each simulation, $\mu N_c$ is mean of 1000 $N_{ci}$, $\sigma N_c$ is standard deviation of 1000 $N_{ci}$. As $N_c$ log-normally distributed, the logarithm values of mean and standard deviation of $N_c$ were used in the probabilistic method for calculating the probability of failure. The conventional deterministic method in geotechnical engineering adopted Prandtl solution for the bearing capacity factor, $N_c$ of 5.14 [9]. Thus, the probability of failure is considered as the chance of the mean bearing capacity factor, $\mu N_c$ less than 5.14 [3, 10]. The probability can be expressed in terms cumulative function, $\Phi$

$$P[N_c < 5.14 / FS] = \Phi(\beta) . \tag{17}$$

where $\beta$ is reliability index and $FS$ is the factor of safety. The factor of safety is applied to minimise the probability of failure of the footing. The reliability index, $\beta$ of $N_c$, which is the expression of margin of safety, $M$ from its critical value ($M$=0) [5] can be defined as

$$\beta = \mu M / \sigma M . \tag{18}$$

where $\mu M$ is mean of margin of safety, $\sigma M$ is standard deviation of margin of safety.

In this case, the margin of safety, $M$ is the difference between the Prandtl solution from deterministic approach and the mean of bearing capacity factor [5]. Thus, mean and standard deviation of $M$ can be expressed as

 ISBN: 978-1-61208-237-0 

$$\mu M = 5.14/FS - \mu N_c . \quad (19)$$

$$\sigma M = \sigma N_c . \quad (20)$$

The probability of failure, hence, can be determined by the cumulative function of $\beta$. The cumulative function can be illustrated as Fig.6.

The correlation length in RFEM has impacts on the probability of failure for geotechnical problems [2] and this study considered different correlation lengths both for isotropic and anisotropic cases. The isotropic study with RFEM was mentioned by many authors includes Griffiths and Fenton [3], Vessia *et al.* [6] and Popescu *et al.* [11]. The probability of failure for isotropic condition for *FS*=3.0 and *COV* =1.0 is compared with other published data in Fig. 7. A good match is observed with Griffiths and Fenton [3], however significant discrepancies are observed with Kasama and Whittle [10]. It is worth nothing that this study used displacement based finite element method whereas Kasama and Whittle [10] used numerical limit analysis. However, the differences are not obvious and need further investigation. The probability of failure for anisotropic cases is also plotted in Fig. 7, $\theta_x$ = 1, 2 and 4 are plotted with varying $\theta_y/B$ ($B$ is width of footing). It shows that the probability of failure for anisotropic cases can be higher than the isotropic case irrespective of the method used in Griffiths and Fenton [3] and Kasama and Whittle [10]. However, the probability of failure in any case is not higher than 25%.

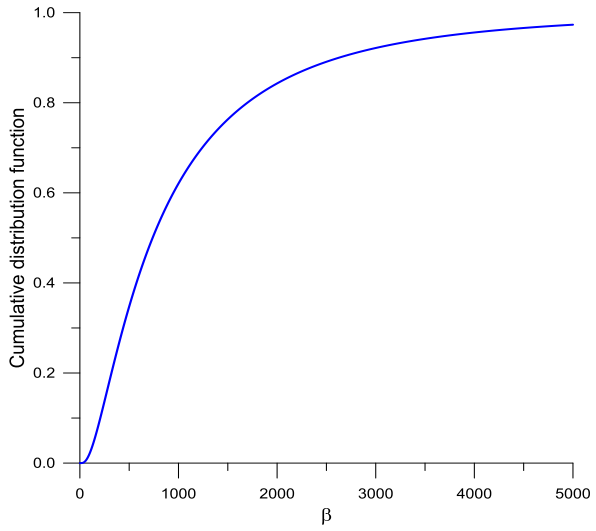



Figure 6. Cumulative density function of β

For better understanding the probability of failure in anisotropic condition, a contour map is presented for *COV*=0.1 in Fig. 8, where red is higher and blue is lower probability of failure. In Fig. 8, the probability of failure is very high at the small correlation length and the probability of failure decreases when correlation length increases. Generally, the probability of failure is affected by the correlation length and *COV* of *c* in RFEM. When the correlation length is high, the probability is low. In contrast, the *COV* is high; the probability of failure will be high. In anisotropy, the ratio of correlation length is increasing while the probability of failure is decreasing. It is interesting to note that the higher correlation length in horizontal direction ($\theta_x$) with $\theta_y$ in the range of 1 to 3 is better in terms of stability when comparing with higher correlation length in vertical direction ($\theta_y$). This is favorable as this is more realistic for general field condition. However, the opposite condition is not usual in field conditions though it is not impossible.

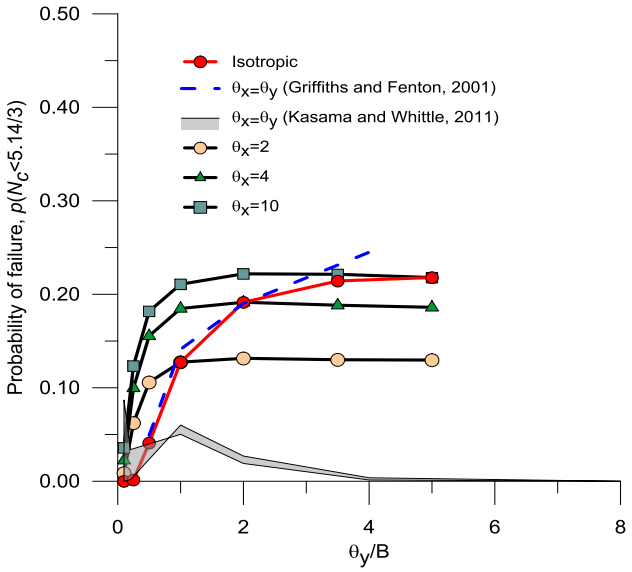



Figure 7. Probability of failure against vertical correlation length (*COV*=1.0)

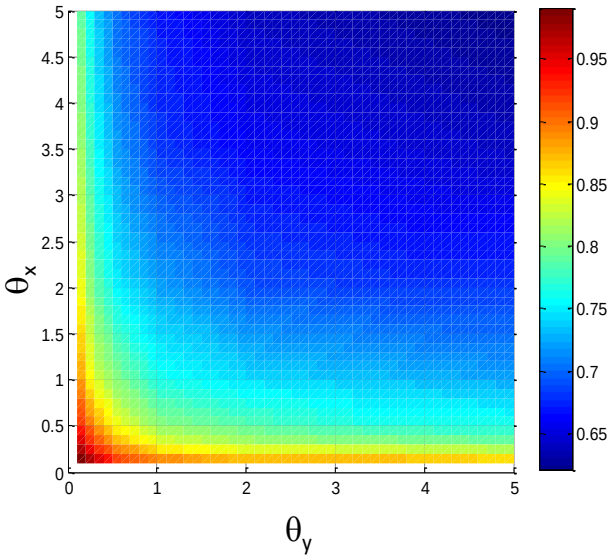



Figure 8. Probability of failure contour plot for COV=0.1 against horizontal and vertical correlation length.

 ISBN: 978-1-61208-237-0 

As mentioned previously, this study considered higher horizontal correlation length than the vertical direction and this is presented as the ratio of the horizontal and vertical correlation length, $\theta_x/\theta_y$. The effect of $\theta_x/\theta_y$ on the probability of failure is shown in Fig. 9.

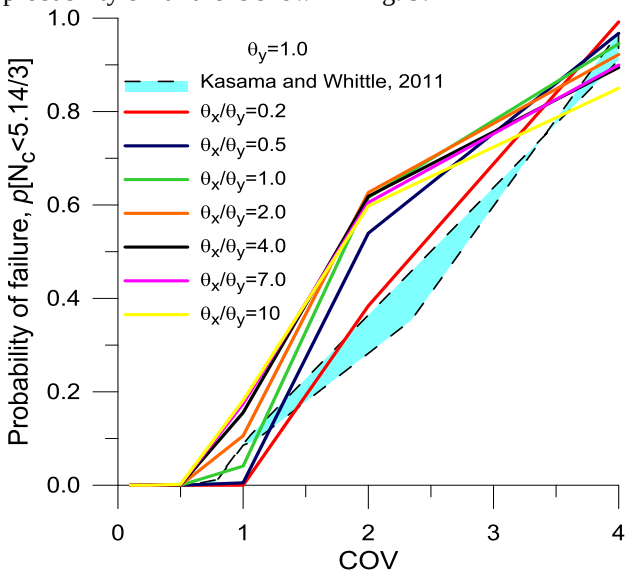



Figure 9. Probability of failure against COV

Fig. 9 shows that the probability of failure will be increasing when the *COV* of *c* increases. It is obvious that the results from anisotropy are significantly different than the isotropic results in Kasama and Whittle [10]. The results in both conditions do not match well when *COV*=2.0 to 3.0. For anisotropic conditions, at small *COV*, the greater the ratio of correlation length is, the higher the probability of failure is. In contrast, at high *COV*, the greater the ratio is, the smaller the probability of failure is.

## V. CONCLUSION

This paper discussed about the random finite element method (RFEM) and applied in bearing capacity problem. Then, some probabilistic and statistical methods used to evaluate the effect of the variability of soil parameter on the probability of footing failure. The correlation length of soil parameter within neighbouring soil elements and its effect on the probability of failure is explored. The major findings of this study are-

- The probability of failure for isotropic condition is different for different methods. This is not obvious, thus need further investigation.
- The probability of failure for anisotropic conditions is higher than the isotropic conditions for the cases presented in the study.
- A higher correlation length in horizontal direction is more favourable than a higher correlation length vertical direction. This is favourable to most of the general conditions.
- The factor of safety is also considered in the calculation of the probability of failure. The suitable factor of safety for footing design is 3.0, when the probability of failure is significantly small, particularly at the small *COV* and correlation length.

RFEM considers the variation of soil parameters within soil elements. It is more practical and realistic than deterministic approaches, which considers parameters are constant for all soil elements. By using RFEM, the chance of failure of footing can be determined more realistically. The probability of failure will vary against different correlation length and coefficient of variation (*COV*). In this case, RFEM can improve the accuracy and efficiency in geotechnical analysis when dealing with a distribution of parameters.

## ACKNOWLEDGEMENT

This study is partly supported by Early Career and New Appointee Researcher Development Award from Division of ITEE, University of South Australia. The software Mrbear2D used in this study, was coded and compiled by Griffiths and Fenton [3], is freely available online.

## REFERENCES

[1] Terzaghi, K., Theoretical soil mechanics. 1943: Wiley, New York.
[2] Fenton, G.A. and D.V. Griffiths, Risk assessment in geotechnical engineering. 2008, New Jersey: John Wiley and Sons, Inc.
[3] Griffiths, D.V. and G.A. Fenton, Bearing capacity of spatially random soil: the undrained clay Prandtl problem revistied. Geotechnique, 2001. 51(4): pp. 351-359.
[4] Rahman, M.M. and H.B.K. Nguyen, Spatial variability of material parameter and bearing capacity of clay, in In Proceedings of APMS International Conference on Applied Physics and Materials Science. 2012, Elsevier Ltd. : Dalian, China.
[5] Baecher, G.B. and J.T. Christian, Reliability and statistics in geotechnical engineering. 2003, New York: Wiley.
[6] Vessia, G., et al., Application of random finite element method to bearing capacity design of strip footing. Journal of GeoEngineering, 2009. 4(3): pp. 103-112.
[7] Vanmarcke, E.H., Random fields: Analysis and sysnthesis. 1984, Cambridge: MIT Press.
[8] Smith, I.M. and D.V. Griffiths, Programming the finite element method. 3rd ed. 1998, New York: John Wiley & Sons, Inc.
[9] Smith, I., Smith's elements of soil mechanics. 8th ed. 2006, Oxford, UK: Blackwell Publishing Ltd.
[10] Kasama, K. and A.J. Whittle, Bearing Capacity of Spatially Random Cohesive Soil Using Numerical Limit Analyses. Journal of Geotechnical and Geoenvironmental Engineering, 2011. 137(11): pp. 989-996.
[11] Popescu, R., G. Deodatis, and A. Nobahar, Effects of random heterogeneity of soil properties on bearing capacity. Probabilistic Engineering Mechanics, 2005. 20(4): pp. 324-341.

 ISBN: 978-1-61208-237-0 

# RADIC-based Message Passing Fault Tolerance System

Marcela Castro, Dolores Rexachs and Emilio Luque
*Computer Architecture and Operating Systems Department, Universitat Autònoma de Barcelona*
*marcela.castro@caos.uab.es; {dolores.rexachs; emilio.luque}@uab.es*

***Abstract*—We present an analysis design of how to incorporate a transparent fault tolerance system at socket level for message passing applications. The novel design changes the default socket model avoiding being unexpectedly closed due to a remote node failure. Moreover, a pessimistic log-based rollback recovery protocol added to this level makes possible restarting and re-executing a failed parallel process until the point of failure independently of the rest of the processes. This paper explains and analyzes the design time decisions. We tested and assessed them executing a master-worker (M/W) and Single Program Multiple Data (SPMD) applications which follow different communication patterns. Promising results of robustness in interprocess communication were obtained.**

***Keywords-Fault-tolerance; High-Availability; RADIC; message passing; socket.***

## I. INTRODUCTION

Fault tolerance (FT) solutions are regarded as a mandatory requirement for parallel applications since the probability of failure is higher in increasingly complex High Performance Computing (HPC) systems and with more components. There is a high risk of suffering an execution stop due to a node failure when the parallel application lasts more than the Mean Time Between Failures (MTBF) of the host system.

When a message passing application is executing in a cluster and suddenly one of the node fails, the communications established with the parallel processes in it also fall down. These communication errors would propagate causing fatal errors to the rest of the parallel processes.

The Figure 1 shows a typical communication level diagram of a message passing application. A failure at physical o networking levels usually spreads up errors to higher levels causing an undesirable execution stop of the application.

Socket [1] is a *de facto* standard application interface (API) of Portable Operating System Interface (POSIX) to use the transport level protocols like TCP or UDP [2]. This API is normally used for interchanging data packages between two executing processes in a cluster. The socket model is intended to do that, but a remote failure is treated by this API as a fatal error. However, controlling socket errors caused by a fall of remote peer would prevent the propagation of them to the upper levels of the message passing communication library and application.

The research work *Reliable Network Connections* [3] describes *rocks*, an approach which changes the normal behaviour of the diagram state of socket API by automatically detecting network connection failures, including those caused by link failures, extended periods of disconnection, and process migration, within seconds of their occurrence. When this kind of communication error happens, instead of closing unexpectedly the socket, the IP address is replaced by the new location of remote peer and the broken connection is recovered without loss of in-flight data as connectivity is restored.

Clearly, establishing reliable network connections instead of normal ones would contribute to provide a FT solution for message passing application avoiding unexpected fatal errors.

Message passing applications usually rely on rollback-recovery protocols to recover from failures. Most of these protocols were explained and classified by E.N. Elnozahazy [4]. RADIC *Redundant Array of Distributed Independent Controllers* [5] is a Fault Tolerance architecture for message passing applications that defines a proper model to apply a rollback recovery protocol using uncoordinated checkpoint and pessimistic log-based on receiver.

The approach of FT of this research work basically consist in modifying the socket model used by the upper levels indicated in Figure 1. The new model combines the use of reliable network connections with the models of RADIC architecture in order to provide a FT system for parallel application that would be used independently of what message passing library is in use. This independence let the FT be seen as an additional optional infrastructure service without requirements for the application.

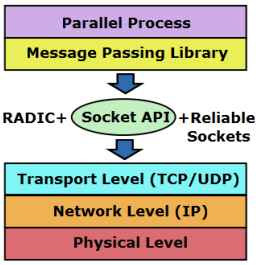



Figure 1. Socket Level

This paper is specially focused on explaining the requirements of the system, the problems that have fulfilled each of them and the corresponding solutions adopted in the design of a RADIC-based fault tolerance system at socket level.

A RADIC-based FT system inherits its properties of dis-

 ISBN: 978-1-61208-237-0

tributed and decentralized, which would facilitate to develop a scalable solution.

The content of this paper is organized as follows. In Section II we mention the related works. Section III defines the requirements to take into account in the design of this solution. The Section IV explains the problems and the solutions adopted to fulfill the requirements at socket level. The experimental evaluation is presented in Section V, and lastly, we state the conclusions and the future work in Section VI.

## II. Related Works

The approaches used to add FT in a message passing application can be classified into three groups according to [6]. First, the application can be changed adding the FT mechanisms. In relation to this approach, we can mention research works like [6] [7], which facilitate the programming tasks either defining FT programming patterns or adding new libraries to be called. Although this first group of FT solutions is likely to reach the best fit, it is expensive and not always applicable if the source code is not available. In the second group we can categorize the research works which locate FT algorithms in communication library. Most of the used solutions belong to this group, because the application does not need to be changed. We regard this kind of tools as an extension of the MPI communication library. MPICH-V Project [8] is an example of this case. Moreover, RADIC was previously implemented using this kind of strategy. [5]. Although the application is not changed it needs to be compiled again with the modified communication library. Furthermore, this could be a problem if we only have the executable programs and we still need to assure an error-free execution in spite of node-failures. Another drawback of this group is the need of adapting each MPI implementation to the specific FT strategy. Finally, available solutions at system level are also transparent for the application, but, most of them are very sensitive to changes in operative system versions and they are not easily portable to other architectures. An example using this last category is DMTCP [9], a checkpoint and restart tool for distributed applications which can be also used for message passing applications. However, scripts for checkpoint and restarting must be provided by the user. Our work fits in this last category as it works at system level.

On the other hand, there are three requirements to be covered by rollback recovery FT approaches. Firstly, **Protection** of information/state to continue computation. Secondly, **Detection** of the failure and lastly, **Restart** the computation reconfiguring the system to isolate the damage component and mask the errors. Fault tolerance solutions are not fully developed with all the requirements at system level.

For example, BLCR [10] is a well-known project of kernel-level process checkpoint. It can be used with multithreading programs but it does not support distributed or parallel process. This tool covers the protection and restarting requirements. The detection has to be added by the user.

DMTCP [9] (Distributed Multi-threaded Checkpointing) does not provide the detection requirement to be considered a FT solution.

DejaVu [11] is a transparent user-level FT system for migration and recovery of parallel and distributed applications. It provides the three mentioned requirements and implements a novel mechanism to capture the global state named online logging protocol. Although uncoordinated checkpoints are performed, it uses a coordinated mechanism to assure the global consistent state of them. This property can be a drawback to scale properly. In addition, DejaVu does not implement any log message protocol, so all parallel processes are forced to recover and restart in case of a node failure.

RADIC [5] meets the requirements of detection, protection and recovery. These tasks are carried out without any centralized element to keep the scalability of the running application. Thus, the behavior is completely distributed on the nodes of the clusters and the overhead added during the protection phase and in recovery is independent from the number of processes. We think this property is essential nowadays when the numbers of processors in the clusters are increasing so much. To protect it uses log message based on receiver rollback recovery protocol which facilitates the recovery tasks, but adding some overhead during the protection phase. The middleware we are presenting is based on RADIC and works at user level.

## III. Design Requirements

This section defines the two basic requirements to take into account during the design of the FT system. The first is that the design has to be located at socket level in order to achieve application and library independence. The second requirement is having properties of transparency, distribution, decentralization and scalability, which are going to be inherited from RADIC. This section begins with a brief explanation of RADIC architecture. Previous research papers can be consulted for detailed information [5] [12]. Finally, the concept of reliable sockets is defined, outlining how we can include them and what is required to do it.

### A. RADIC Architecture

RADIC architecture is based on uncoordinated checkpoints combined with pessimistic log-based on receiver. Critical data like checkpoints and message logs of each parallel process are stored on a different node from the one in which it is running. This selection assures application completion if a minimum of three nodes are left operational after *n* non-simultaneous faults. In short, RADIC defines the following two components also depicted in Figure 2

 ISBN: 978-1-61208-237-0 

- **Observer (Oi):** this entity is responsible for monitoring the application's communications and masks possible errors generated by communication failures. Therefore, the observer performs message logs in a pessimistic way as well as it saves periodically the parallel process state by checkpointing. Message logs and checkpoints are sent to protector **Ti-1**. There is an observer **Oi** attached to each paralell process **Pi**.
- **Protector: (Ti)** There is one running on each node which can protect more than one application process. In order to protect the application's critical data, protectors store that on a non-volatile media. In case of failure, the protector recovers the failed application process with its attached observer. Protector detects node failure by sending heartbeats to its neighbours.

Figure 2 shows the relationship between nodes running an application with RADIC fault tolerance architecture. Diagonal arrows represent critical data flow while horizontal ones represent heartbeats.

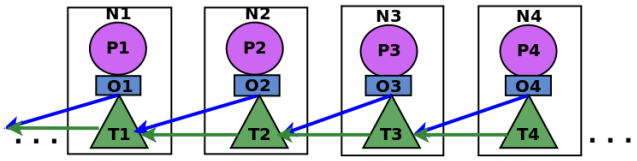



Figure 2. RADIC diagram shows each observer **Oi** sends the critical data to its protector **Ti-1**. Each protector **Ti** sends heartbeat signal to **Ti-1**

### B. Reliable sockets

TCP is a reliable transport protocol between two peers in a sense that every packet sent by one peer is assured to be delivered and received by the other peer respecting the sent order. Each peer uses send and receive buffers managed by flow-control to accomplish this reliability.

However, the TCP model does not provide a mechanism to recover the connection from a permanent failure of one of the peers, because this kind of situation is out of the scope of a transport protocol. Usually, the applications running on POSIX Operative System use socket API as I/O network interface to receive and send data through a TCP/IP connection. TCP connection failures occur when the kernel aborts a connection. This could be caused by several situations such data in the send buffer goes unacknowledged for a period of time that exceeds the limits on retransmission defined by TCP, or receiving a TCP reset packet as a consequence of the other peer reboots or closes the socket unexpectedly. Furthermore, when the kernel aborts the connection, the socket becomes invalid for the application.

If the application does not have a proper functionality to recover this invalid socket, usually the execution is aborted due to the unexpected situation.

*Rocks* architecture proposed by [3] defines the operation of a reliable socket by changing the default state socket diagram by a new one. This new operation does not allow the socket to be closed by such exceptional situation. Instead of that, the socket remains in a suspended state while a new address of the remote peer is got and the socket is reconfigured. This new socket behavior affects the process, not the internal TCP socket state maintained by the kernel.

This general idea is applicable to our design, but not the detailed behavior and implementation, because they considered only two peer applications and not parallel ones composed by several processes.

Furthermore, we need to incorporate to this socket level the functionality of the rollback recovery protocol defined by RADIC. To accomplish this task, we designed a new behavior of socket API considering reliable sockets and pessimistic based on receiver rollback recovery protocol.

Following the basic idea of reliable network connections, the FT logic needed for protection and for restarting a process is added by interposing socket functions as *socket, bind, listen, connect, send* and *recv*. The default diagram state of socket API is changed in order to not allow the socket to be closed when remote peer falls down. Next, the socket does not become invalid for the upper level process. RADIC recovery model determines that the failed processes are restarted in the protector node. As a result, the observer is able to re-configure the socket with this new address and then, the lost connection with the restarted process is re-established.

Taking into account that RADIC defines that the **observer** component is that one attached at each parallel process, the interposition library at socket level corresponds to this. Consequently, the library has to accomplish all the functionality of this component defined by RADIC. As we mentioned before, this paper is specially focused on defining this entity, because it is directly affected by the approach adopted. In contrast, the **protector** can be seen as an independent process that only interacts with other RADIC components like observers and other protectors. The functionality of protectors is completely defined, tested and explained in previous research works.

## IV. Fault tolerance using Secure Sockets

This section describes the three pieces of functionality needed to be incorporated at reliable socket level to get a RADIC **observer.** These three pieces are message log, checkpointing and restarting. Each of them presents different challenges to face, which are explained in the following subsections including the way they are overcame.

### A. Message Log

A pessimistic log-based on receiver rollback-recovery protocol has to be designed at socket level in order to assure that the state of each process is always recoverable. This kind of procedure can add some overhead during the normal execution (protection phase) but this way simplifies the

 ISBN: 978-1-61208-237-0 

recovery tasks because the effects of a failure are confined only to the processes that need to be restarting.

Log-based rollback-recovery assumes that all nondeterministic events can be identified and their corresponding determinants can be logged to stable storage. Receiving a packet is considered a nondeterministic event to log.

At first sight, it seems a simple challenge that can be solved interposing *recv* function and sending the received message to the protector afterwards.

But pessimistic logging protocols are designed under the assumption that a failure can occur after any nondeterministic event in the computation. This assumption is pessimistic since in reality, failures are rare. This property stipulates that if an event has not been logged on stable storage, then no process can depend on it. Because of that, a sender of a message needs to wait until the complete sent message is saved in stable storage before continuing its operation. Once a received message is completely saved on stable storage, an acknowledgment is sent to the sender.

To accomplish this requirement of acknowledgment of each received and saved package, we need to establish a communication between the two **observers** involved in each peer of a socket. We cannot use the application socket being interposed to send and receive acknowledge data because we can be interfering on the application protocol affecting the integrity of their messages.

Therefore, for each socket established by the upper level, the interposing library creates a new socket named **control-ft socket** used to interchange control data between two observers intercepting *send* and *recv* functions.

The Figure 3 shows how a message is treated since it is generated from the sender process. The *send* operation is interposed by sender observer **Os** which sends a numerated acknowledgement requirement to the receiver observer using the **control-ft socket** canal, represented by dotted lines. The message is sent to the receiver using the **real socket**, depicted as solid lines. The receiver observer **Or** interposes the *recv* operation and receives the acknowledgement requirement through the **control-ft socket** and the application message through the **real socket**. Observer **Or** sends the message to its protector. Once **Or** receives the ack of save operation, sends the ack to sender and finishes the *recv* interposed. Observer **Os** receives the ack indicating this message is correctly saved and it is not necessary to be resent anymore. Lastly, the *send* interposition is finished and the process resumes the processing. The gray block represents the tasks added by the logging message protocol during the failure-free execution.

### B. Checkpointing

Each parallel process has to be checkpointed periodically in order to save its state. In a log-based protocol, checkpointing is performed in order to limit the amount of work that has to be repeated in execution replay during recovery.

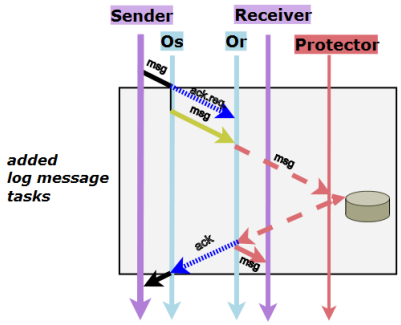



Figure 3. Message log: Real sockets: Solid lines - Control-Ft socket: dotted lines - RADIC sockets: dashed lines

This task is performed in an uncoordinated way, thus no centralized or blocking mechanisms are needed in the sake of scalability.

During the checkpoint, all the active communications of the observed parallel process need to be closed. The BLCR library being used to checkpoint processes recommends this procedure for two reasons. First, to avoid loosing in transit data, and second, because the socket and its corresponding connections have to be established again from scratch during the restarting in a new cluster node.

Therefore, all the opened sockets have to be closed before checkpointing and re-opened and re-established after it. To accomplish this task we need to keep the following data as it is not provided by the operative system:

- **Virtual socket**: It is the socket number id known by the parallel process achieved it during a *socket* or *accept* function.
- **Socket Type:** This type can be *connect, accept* or *listen*. It is used to identify which operation has to be performed to re-establish the socket after checkpoint or during restart.
- **Re-establish parameters:** The parameters used originally to execute the function *connect, accept or listen* interposed in order to re-execute the command after checkpointing or restarting.
- **Real socket**: Socket number id actually being in use to intercommunicate with remote process, getting during a re-open operation after checkpointing or restarting. The operative system delivers different id socket handler when a *socket* or *accept* function is re-executed. A one-by-one relation between virtual and real socket number is kept. During the interposition, the observer changes the virtual socket id referenced by the application by this real socket number. Therefore, the function is performed using the current real opened socket.

There is a problem to be solved when an *accept* type socket is re-established. During an *accept* re-execution, multiple client observers can be trying to re-connect at the same time. The server observer has to be able to recognize

 ISBN: 978-1-61208-237-0 

which is the other peer in order to continue the control message logging of the broken socket. Using operative system functions, the remote ip and port can be known but this data is not enough to identify uniquely the observer client process previously connected to this socket due to more than a parallel process can be executing in one node.

To accomplish this identity validation task, each observer sends a unique parallel process identification **(pid)** trough the **control-ft socket**. Consequently, during the re-establishment of an *accept* socket the former client connected can be uniquely identified to continue the log message associated to the virtual socket opened by the application.

For instance, a master accepts connections coming from two workers. They are connected using sockets 4 and 6 respectively. The socket 4 has socket 5 as **control-ft** and the 6 has the 7. The sockets are closed before checkpoint. After finishing checkpointing, the *accept* functions are executed to re-establish connections. But two observers associated to the two workers are trying to re-connect the unexpectedly closed socket at the same time. After connecting, the remote observer sends its identification using the control-ft. In this way, the server can determine which real socket corresponds with virtual socket 4 and which with the 6.

### C. Restarting

When a node fails, the protector recovers the processes which were being saved in it using the last checkpoint received. The processes are recovered in a spare node if there is one available or in the same protector node if there is not. Each process is restarted with its corresponding observer. This observer detects the restarting state and its behavior is different until the process arrives to the point of failure. This point is reached when the last received message in log is consumed by the restarting parallel processes.

Two important design decisions were made in order to get RADIC restart model at socket level. First of all, the sockets type *listen* which were active on failed host, are launched on this new host. These sockets are needed to be ready before re-connecting the sockets type *accept* being re-executed. In the second place, to re-execute the parallel process until the point of failure, the observer in restarting mode, intercepts the *recv* function and the contents is extracted from the log message previously saved by the protector.

## V. Experimental Results

We test the fault tolerance system for validating the functionality of RADIC and reliable connection at socket level. The principal aim is to assure that the mechanisms used to build the reliable tunnel connections and the log message protocol are working correctly. Our second aim is to know the overhead added in execution time by protection and recovery processes. We take some measures to have indicators for assessing how the system is working in terms of overhead and bandwidth consumed by the protection model.

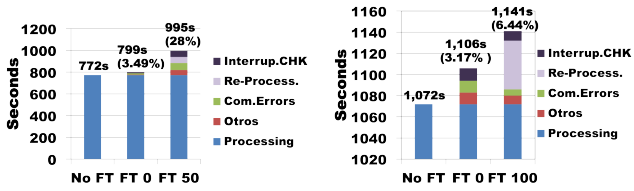
The experiments were executed on a cluster formed by 4 nodes Intel® Core™ i5-650 Processor 6GB RAM, Network Gigabit Ethernet. The OS used is Ubuntu 10.04 Kernel 2.6.32-33-server.

We use a sum of matrices Master/Worker and a heat-transfer SPMD applications based on TCP sockets, which follow different communication patterns in order to do a better test of the reliable socket model performed after checkpoints and in restart process.

We use three ways of execution. First, the normal without FT **No FT**, second using FT but without any node failure **FT 0** and lastly, we inject a fail in the process executing on the node **N3** some events after the first checkpoint, 50 in M/W **FT 50** and 100 in SPMD **FT 100**.

The M/W was executed with 4 workers, one per node. The first node executes the master and one worker. Master performs 2 checkpoint of 450Kb(avg) and workers 3 of 425kb(avg). SPMD is executed with four processes, also one per node. Each one executes 3 checkpoints of 1440kb(avg). The fall is injected in **N3** on both cases and the worker or spmd process are recovered in **N2.**

Two selected experiments are shown in Figure 4(a) and in Figure 4(b). The diagrams show the overhead time comparing the three ways of execution. These times are measured in the process executing in **N3.** The total execution is divided into: the seconds used by interruptions to perform checkpoints, time used for restarting and re-executing, seconds for recovering from communication errors due to node failures or remote checkpointing, time used by other reasons that are not measure by now, like detection of errors, and finally the base time that the application last without FT **(No FT).** The



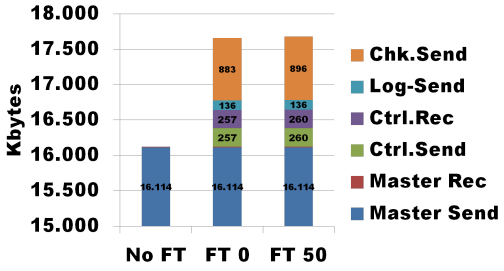

(a) MW Time overhead (b) SPMD Time overhead

Figure 4. Experimental results time execution results.

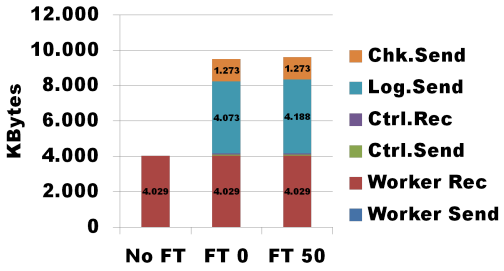
Figure 5(a) graphs the traffic sent and received by process master during the tree executions. and Figure 5(b) shows the same for the worker executed in **N3** that was directly affected by the failure. Master sent two checkpoint to protector. As the master receives very few data, the bytes sent to protector to log message is few. The data transfered by **control-ft** canal is proportionally low. On the other hand, the worker receives much more data, therefore, more kbytes of log are sent to protector.

Finally, in Figure 6, the traffic of the SPMD process executing in **N3** is shown. Similarly as it was observed

 ISBN: 978-1-61208-237-0




(a) MW Master process




(b) MW Failed Worker process

Figure 5. Master/Worker traffic overhead analysis

in previous executions, the overhead added by **control-ft** is low, and the log message is directly proportionated with the amount of received data.

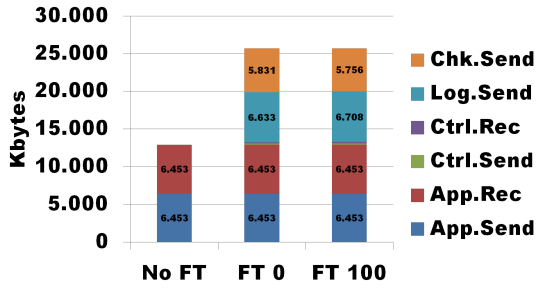



Figure 6. SPMD traffic overhead analysis

## VI. Conclusions and Future Work

The results show that the design made of a transparent and distributed fault tolerance system is appropriate. Message passing applications with different communication patterns are able to end successfully performing periodic checkpointing and restart and re-execute in case of node failure.

Using this approach it is possible to build a transparent fault tolerance middleware able to provide no fail stop to message passing application independently from the communication library in use. In that way, the user is not forced to choose a specific communication library to get the fault tolerance facilities. The library of preference can be chosen.

We are working on a set of experiments to prove we can use this system to fault tolerance applications using either MPICH or OPEN-MPI.

Future work will include as well an analysis of the scalability of the middleware, testing if the speedup of applications being protected can be kept in spite of using fault tolerance.

## Acknowledgments

This research has been supported by the MICINN Spain under contract TIN2007-64974, the MINECO (MICINN) Spain contract TIN2011-24384, the European ITEA2 project H4H, No 09011 and the Avanza Competitividad I+D+I contract TSI-020400-2010-120.

## References

[1] M. K. McKusick, K. Bostic, M. J. Karels, and J. S. Quarterman, *The design and implementation of the 4.4BSD operating system*. Redwood City, CA, USA: Addison Wesley Longman Publishing Co., Inc., 1996.

[2] R. Gordon, "The tcp/ip guide: A comprehensive, illustrated internet protocols reference." *Library Journal*, vol. 131, no. 1, pp. 146–146, JAN 2006.

[3] V. C. Zandy and B. P. Miller, "Reliable network connections," in *Proceedings of the 8th annual international conference on Mobile computing and networking*. New York, NY, USA: ACM, 2002, pp. 95–106.

[4] E. N. Elnozahy, L. Alvisi, Y.-M. Wang, and D. B. Johnson, "A survey of rollback-recovery protocols in message-passing systems," *ACM Comput.Surv.*, vol. 34, no. 3, pp. 375–408, September 2002.

[5] L. Fialho, G. Santos, A. Duarte, D. Rexachs, and E. Luque, "Challenges and issues of the integration of radic into open mpi," in *16th European PVM/MPI Users' Group Meeting on Recent Advances in PVM and MPI*, 2009, pp. 73–83.

[6] W. Gropp and E. Lusk, "Fault tolerance in message passing interface programs," *Int. J. High Perform. Comput. Appl.*, vol. 18, pp. 363–372, 2004.

[7] S. Rao, L. Alvisi, H. M. Viny, and D. C. Sciences, "Egida: An extensible toolkit for low-overhead fault-tolerance," in *In Symp. on Fault-Tolerant Comp.* Press, 1999, pp. 48–55.

[8] A. Bouteiller, T. Hrault, G. Krawezik, P. Lemarinier, and F. Cappello, "MPICH-V Project: A Multiprotocol Automatic Fault-Tolerant MPI," *IJHPCA*, vol. 20, pp. 319–333, 2006.

[9] J. Ansel, K. Arya, and G. Cooperman, "DMTCP: Transparent checkpointing for cluster computations and the desktop," in *IPDPS*, 2009, pp. 1–12.

[10] P. H. Hargrove and J. C. Duell, "Berkeley lab checkpoint/restart (blcr) for linux clusters," *Journal of Physics: Conference Series*, vol. 46, no. 1, p. 494, 2006.

[11] J. F. Ruscio, M. A. Heffner, and S. Varadarajan, "Dejavu: Transparent user-level checkpointing, migration, and recovery for distributed systems," *Parallel and Distributed Processing Symposium, International*, p. 119, 2007.

[12] G. Santos, A. Duarte, D. Rexachs, and E. Luque, "Providing non-stop service for message-passing based parallel applications with radic," ser. Lecture Notes in Computer Science, vol. 5168 LNCS, 2008, pp. 58–67.

 ISBN: 978-1-61208-237-0 

# A QoS Monitoring Framework for Composite Web Services in the Cloud

Rima Grati
University of Sfax
Multimedia, InfoRmation system & Advanced Computing Laboratory (Mir@cl)
Tunisia
rima.grati@gmail.com

Khouloud Boukadi
University of Sfax
Multimedia, InfoRmation system & Advanced Computing Laboratory (Mir@cl)
Tunisia
khouloud.boukadi@fsegs.rnu.tn

Hanene Ben-Abdallah
University of Sfax
Multimedia, InfoRmation system & Advanced Computing Laboratory (Mir@cl)
Tunisia
hanene.BenAbdallah@fsegs.rnu.tn

***Abstract*— Due to the dynamic nature of the Cloud, continuous monitoring of QoS requirements is necessary to manage the Cloud computing environment and enforce service level agreements. In this paper, we propose a QoS monitoring framework for composite Web services implemented using the BPEL process and deployed in the Cloud environment. The proposed framework is composed of three basic modules to: collect low and high level information, analyze the collected information, and take corrective actions when SLA violations are detected. This framework provides for a monitoring approach that modifies neither the server nor the client implementation. In addition, its monitoring approach is based on composition patterns to compute elementary QoS metrics for the composed Web service. In this paper, we illustrate our framework for the response time QoS requirement.**

***Keywords- Monitoring of Web service composition; Cloud environment; Service Level Agreement ; SLA violation***

## I. INTRODUCTION

Cloud computing has recently emerged as a new paradigm for hosting and delivering services over the Internet. It offers huge opportunities to the IT industry. In addition, it offers two advantages for business owners: it eliminates the requirement to plan ahead for provisioning, and it allows enterprises to start from the small and increase resources only when there is a rise in service demand. Besides these advantages, Cloud computing enables users to utilize services without the need to understand their complexity or acquire the knowledge and expertise to consume them [1]. It provides users with services to access hardware, software, and/or data.

Despite these advantages, business owners require that Cloud providers guarantee a pre-agreed upon set of Quality of Service (QoS) attributes, *e.g.*, response time, availability, security, and reliability. Face to these user requirements, and due to the dynamic nature of the Cloud, continuous monitoring of QoS attributes became mandatory to enforce Service Level Agreements (SLA) [2].

In fact, run-time monitoring has been in demand well before the Cloud. Several monitoring systems, *e.g.*, Ganglia [3], Nagios [4], MonaLisa [5], and GridICE [6] addressed monitoring of large distributed systems. However, these systems did not deal with problems induced by rapidly changing and dynamic infrastructures. This prompted the propositions of some monitoring approaches dealing with applications deployed on the cloud environment as a set of Cloud services [7][8]. Most of these approaches require modification of either the server or the client implementation code. However, to provide for independence of any Cloud provider/environment, monitoring should be performed without modifying the implementation of the deployed Cloud services. Furthermore, to the best of our knowledge, there is a lack of approaches dealing with monitoring of the service composition in a Software as a Service (SaaS) cloud environment.

In this paper, we propose a framework for QoS Monitoring and Detection of SLA Violations (QMoDeSV). This framework provides for the monitoring of composite services deployed on the Cloud. It is designed to handle the complete Web service composition management lifecycle in the Cloud environment, *i.e.*, composite Web service deployment, resource allocation, monitoring of QoS and SLA violation detection. In addition, QMoDeSV proposes a non-intervening modular approach for monitoring QoS attributes: QoS pertinent information is collected by "watching" locally each service component. Then, based on the composition pattern of the composite service, the overall QoS information is computed. This information is used by a separate module in the QMoDeSV framework to look for potential violations of SLA pre-agreed upon QoS attributes. The findings of this module can be very helpful for service providers, who can then take corrective actions to improve their services.

The remainder of this paper is organized as follows: Section 2, provides a background of Cloud computing and monitoring, then it overviews related works on monitoring. Section 3 presents our monitoring framework. Section 4 presents our running example. Section 5 summarizes the presented work and highlights some directions for future work.

 ISBN: 978-1-61208-237-0 

## II. RELATED WORK

### A. On SLA and Monitoring

Many definitions are proposed for cloud computing. The generally accepted definition is the one proposed by L. M. Vaquero [10]: "Clouds are a large pool of easily usable and accessible virtualized resources (such as hardware, platforms and/or services). These resources can be dynamically reconfigured to adjust to a variable load (scale), allowing also for an optimum resource utilization. This pool of resources is typically exploited by a pay-per-use model in which guarantees are offered by the Infrastructure Provider by means of customized SLAs."

This definition clearly emphasizes the role of Service Level Agreement in the context of the Cloud. It requires a Cloud provider to be able to propose and guarantee quality of services for the provided resources. That is, a Cloud provider must be able to both establish contracts, and *continuously monitor and verify the compliance of the offered QoS with the agreed-upon SLAs.*

Once services and business processes become operational, their progress needs to be managed and monitored both to gain a clear view of how services perform within their operational environment, and to take management decisions. *Monitoring* is the procedure of measuring, reporting, and improving the QoS of systems and applications delivered by the service. Monitoring consists also of verifying at run-time that the requirements, specified by the clients and the service providers, are met during execution. The contract signed between the clients and the Cloud provider is called SLA. It includes the non-functional requirements of the service specified as QoS, obligations, service pricing, and penalties in case of agreement violations.

Flexible and reliable management of SLA agreements is of paramount importance for both providers and consumers. On the one hand, prevention of SLA violations avoid penalties that providers have to pay and, on the other hand, based on flexible and timely reactions to possible SLA violations, user interactions with the system can be minimized.

### B. Works on monitoring

We classify works pertinent to monitoring into two categories: Web Service Monitoring, and Cloud Service Monitoring.

#### 1) Web Service Monitoring

Rosenberg et al. [11] propose a monitoring approach for Web services. Their approach relies on aspect oriented and object oriented programming techniques and does not require any access to the Java source code of the service implementation. The proposed approach requires information related to the implementation of the monitored Web service (*e.g.*, endpoint and reference to WSDL). It makes use of monitoring tools such as Jpcap to monitor only latency measurement.

Repp et al. [9] present an approach to monitor performance across network layers such as HTTP, TCP, and IP. Their approach aims at monitoring QoS (in terms of network measurements) and detecting SLA violation. In the case of an SLA violation, this approach proposes to reconfigure the system at real time to minimize the substitution cost. For this, it uses the *windump* tool which requires access to the hardware for monitoring. This work monitors only network measurements.

#### 2) Cloud Service Monitoring

Shao et al. [7] propose a Runtime Model for Cloud Monitoring (RMCM). RMCM uses interceptors (as filters in Apache Tomcat and handlers in Axis) for service monitoring. It collects all Cloud layer performance parameters. In the SaaS layer, RMCM monitors applications while taking into account their required constraints and design models. To do so, it converts the constraints to a corresponding instrumented code and deploys the resulting code at the appropriate location of the monitored applications. Thus, it modifies the source code of the applications.

Boniface et al. [8] propose a monitoring module that collects QoS parameters of Cloud Computing. They use a monitoring application component (AC) that must be first described and registered in the application repository. The AC collects QoS parameters at both the application and technical levels. This approach is complicated and hard to install due to the description and registration of AC. Furthermore, their approach remains unevaluated.

To the best of our knowledge, none of the discussed approaches deals with monitoring Web service *composition in the Cloud*. As we describe in the next section, our approach has two additional distinctive features: computing QoS metrics in a modular way based on the patterns used in the composite service deployed in a SaaS Cloud, and collecting information (low level and high level) then comparing these metrics to SLA.

## III. THE QMODESV MONITORING FRAMEWORK

The QMoDeSV framework aims at monitoring composite Web services deployed on the Cloud. Its run-time monitor is based on the workflow patterns used in the composition (BPEL process). It is designed to handle the complete Web service composition management lifecycle in the Cloud environment. The service composition lifecycle includes activities such as composite service deployment, resource allocation to the composite service, composite service monitoring, and SLA violation detection.

In our approach, we suppose that the composite Web service (*i.e.*, the BPEL process) is offered through a SaaS provider. the latter should propose the BPEL processes, the BPEL engine responsible for executing the processes instances, the database management system (DBMS) as well as the monitoring framework.

We consider that monitoring begins when the customer places a service composition request through a defined application interface to the Cloud provider. As depicted in Figure 1, the QMoDeSV framework is a two-level framework consisting of a design time module (the Extractor Module) and five run time modules (the RTP Extractor, the QoSCalculator, the Local Host Monitor, the Lo2Hi QoS Convertor, and the QoS Detector Violation). Once the

 ISBN: 978-1-61208-237-0 

composite Web service is invoked, the run time modules of QMoDeSV are executed. These modules run in parallel with the BPEL instance in order to detect possible SLA violations.

The remainder of this section describes the role of each module and how it interacts with the other modules.

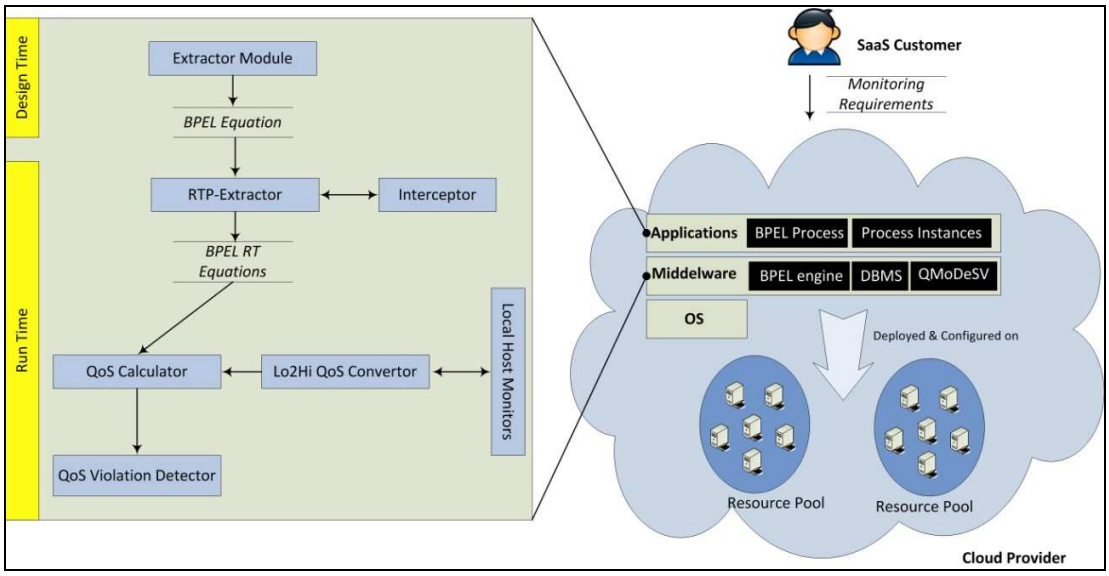



Figure 1. Overview of QMoDeSV architecture and module's interaction

### A. *The Extractor Module*

Web services can be composed using different patterns that are based on the usual workflow patterns. Usually, a complex Web service composition combines two or more of these patterns.

Our Extractor module can handle the following common workflow patterns:

- **Sequence pattern**: indicates that the components Web services are executed one after the other.
- **Parallel pattern:** indicates that two or more Web services can be executed in parallel.
- **Synchronization pattern:** indicates that the process will continue after the parallel pattern of the Web service is executed.
- **Exclusive choice pattern:** is a point in the process where a path is chosen from several available paths based on a decision or process data
- **Simple merge pattern:** defines a point in the flow of execution, where two or more alternative branches are merged.
- **Conditional pattern:** indicates that there are multiple services ($s_1$, $s_2$, …, $s_n$) among which only one service can be executed.
- **Synchronizing merge pattern:** marks a point in the process execution, where several branches merge into a single one.
- **Multi-merge pattern:** joins two or more different services without synchronization together.
- **Loop pattern:** indicates that a certain point in the composition block is executed repeatedly.
- **Deferred choice pattern:** describes a point in the composition where some information is used to choose one among several alternative branches. When one branch of the process is enabled, the others should be disabled

The Extractor Module is responsible for analyzing the composite Web service implemented as a BPEL process. It uses the pattern detection algorithm shown in Listing 1 to extract the used patterns from the BPEL process. The output of the Extractor Module is a design time equation containing the name of the components Web services as well as the patterns used for connecting the flows between these components.

 ISBN: 978-1-61208-237-0 

Listing 1. The pattern detection algorithm used by the Extractor Module

```
Input = ( BPEL process BP) // the composite Web service implemented using a BPEL process
Output = BPequation    //design time BPEL equation
Func
    GetBPTags (BPEL process BP) return ListOfTags  // this function parse the BPEL
    document and put each tag in a list box
    FilterUsedPatterns (ListOfTags) return ListOfUsedPatterns  // this function removes from
    the list box the set of used patterns
    Get WebServicesNames (ListOfTags) return  ListOfWebServices  // this function retrives
    the names of invoked Web services
    GetBPEquation (ListOfUsedPatterns, ListOfWebServices) return BPequation  // this
    function build the BPEL equation based on the identified pattern and the invoqued Web services


BEGIN
    String patterns[] = {"<sequence", "</sequence", "<receive", "<invoke", "<flow", "</
    flow", "<switch", "</switch", "<while", "</while", "<pick", "</pick", "<link", "</link"};

    for each BPEL process BP do
    {          List<String> ListOfTags  = GetBPTags (BP);
               List<String> ListOfUsedPatterns = FilterUsedPatterns (ListOfTags);
               List<String> ListOfWebServices = Get WebServicesNames (ListOfTags);
               String BPequation=GetBPEquation (ListOfUsedPatterns, ListOfWebServices);
               Print (BPequation);
    }

END
```

### B. *The Run Time Extractor Module*

The Run Time Extractor Module (see Fig. 2) "watches" the executed services and refines the equation obtained in the design time into a run time equation. The run time equation represents the execution path of the BPEL process instance. It is derived according to the patterns extracted at the design time. This module intercepts information about the executed service through the Monitoring thread: this latter is the extension of the API apache ODE (Orchestration Director Engine) [12]. The monitoring thread interacts with the BPEL engine to check the process states and informs the Run Time Extractor module to do a comparison between old and new process (see Fig. 2). After that, the Run Time Extractor Module extracts BPEL nodes to establish the execution graph of BPEL. Once the run time extraction path is done, the run time equation is established.

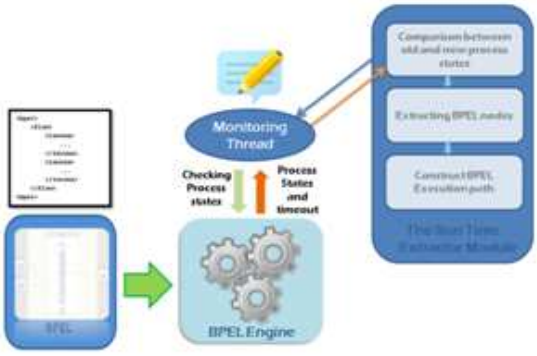



Figure 2. The Run Time Extractor Module

### C. *The QoS Calculator Module*

This module computes QoS metrics for the composite Web services. In its computations, it uses the values of the constituent services and the composition pattern.

We illustrate how the QoS calculator functions use the following QoS metrics:

- **Response Time (RT):** the time interval between when a service is invoked and when the service is finished.
- **Service Cost (C):** the price that a service requester has to pay for invoking the service.
- **Throughput (T):** represents the number of Web service requests at a given time period.
- **Reliability (R):** the probability that a request is correctly responded within the expected time.

The overall Web service QoS is derived based on the values collected locally for each constituent service and the composition pattern. For this, we adapt metrics proposed in [13] and [14]. The adapted metrics are instantiated by the QoS calculator based on the composition pattern detected by run time extractor.

Table 1 summarizes the QoS metrics we adapted to account for the composition patterns. To establish these metrics, we noted constituent Web services as $s_1$, $s_2$… $s_n$ and the Web service composition that includes these services as $S(s_1, s_2, \ldots, s_n)$. For the conditional pattern, we denote $p_i$ the probability that a service $s_i$ be selected. Finally we denote as $SO(s_i, p_i)$ the selection operation for the conditional patterns, which selects the service $s_i$ with an execution probability $p_i$.

 ISBN: 978-1-61208-237-0 

TABLE I. METRICS FOR COMPOSED WEB SERVICES

| Patterns | Response Time | Throughput | Reliability | Cost |
|---|---|---|---|---|
| Sequence | $\sum_{t=1}^{n} \mathrm{RT}(s_t)$ | $\frac{1}{\sum_{t=1}^{n} \frac{1}{T(s_t)}}$ | $\prod_{t=1}^{n} R(s_t)$ | $\sum_{t=1}^{n} C(s_t)$ |
| Parallel<br>Synchronization<br>Simple merge | $max\{RT(s_t)\}$ | $\min\{T(s_t)\}$ | $\prod_{t=1}^{n} R(s_t)$ | $\sum_{t=1}^{n} C(s_t)$ |
| Exclusive choice<br>Deferred choice | $RT(SO(s_t, p_t))$ | $T(SO(s_t, p_t))$ | $R(SO(s_t, p_t))$ | $C(SO(s_t, p_t))$ |
| Multi-choice/ conditional<br>Synchronizing merge | $\max\left(RT(SO(s_t, p_t))\right)$ | $\min\left(T(SO(s_t, p_t))\right)$ | $\prod_{t=1}^{n} R(SO(s_t, p_t))$ | $\sum_{t=1}^{n} C\left(SO(s_t, p_t)\right)$ |
| Loop | $n\ X\ RT$ | $\frac{1}{\sum_{t=1}^{n} \frac{1}{T}}$ | $R^n$ | $n\ X\ C$ |

For example, when considering the exclusive choice pattern, the response time is calculated by the selection operation, which selects one of the *n* possible Web services. In particular, it is defined as $RT(SO(s_t, p_t))$:. This pattern selects the service $s_t$ with a probability $p_t$ at design time. However, since at run time, the execution path is clear and this metric will be adapted by the QoS calculator into: $\sum_{t=1}^{n} \mathrm{RT}(s_t)$.

### D. The QoS Violation Detector

The QoS Violation Detector accesses the mapped metrics repository to get the mapped SLA parameters. These parameters are compared with the calculated values obtained from the QoS Calculator. In the case of a violation (none respect of SLA), it dispatches notification messages to the customer/provider to alert about the violation. An example of SLA violation threat can be an indication that the process consumed 5*ns* for a response time while the agreed response time is 3*ns*.

### E. The LHM and Lo2Hi QoS Convertor

The Local Host Monitor (LHM) process monitored values and is capable of measuring both hardware and network resources. It can be configured to access different virtual hosts at the same time to collect locally monitored values.

As shown in Figure 1, the Lo2Hi QoS Convertor interacts with two components: the LHM which monitors the resources, and the QoS Calculator which calculates the global obtained metric. Resources are monitored by the Local Host Monitor using arbitrary monitoring tools such as Gmond from Ganglia project [3]. Low level resource metrics include outbytes, inbytes, and packetsize. Based on the predefined mapping rules stored in a database, monitored metrics are periodically mapped to the SLA parameters. These mapping are obtained in a similar way to those in Grids where workflow processes are mapped to a Grid service in order to ensure their quality of service [15].

## IV. EXAMPLE

In this section, we illustrate the functioning of the Extractor Module and the Run Time Extractor Module. Our running example deals with the recruitment of an employee, which we modeled in BPMN (Fig 3). We consider a company named AdminCompany and a new employee called Joan. When Joan arrives to AdminCompany, his information should be collected and it is necessary to perform many activities in parallel such as, grant access to company information, sign some legal documents and set up her workstation. After that, the mode of remuneration should be selected either in cash or by check or by bank transfers.

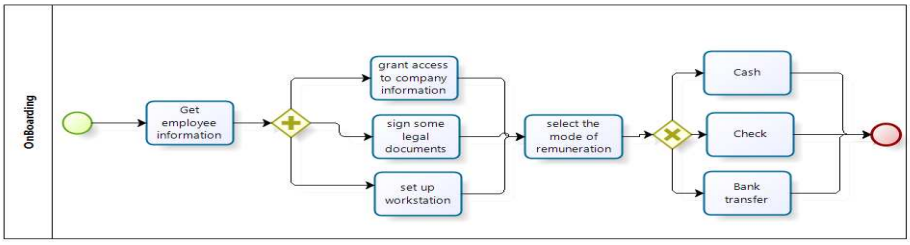



Figure 3. BPMN representation of the example

 ISBN: 978-1-61208-237-0 

For space limitation, we consider only the metric of response time (RT). First, the BPEL process corresponding to this example is implemented. Then the Extractor Module parses the BPEL process to extract the design time equation that represents the used patterns (Listing 2).

Listing 2. Equation obtained in Design Time

**Sequence** (get employee information, **Flow** (grant access to company information, sign some legal documents, set up workstation),
**Sequence** (select the mode of remuneration), **Switch** (cash, check, bank transfer))

This equation will be refined through the Extractor Run Time Module to obtain a run time equation, corresponding to the set of invoked services as well as the patterns used for connecting the flows between them (Listing 3).

Listing 3. Equation obtained at Run Time

**Sequence** (get employee information, **Flow** (grant access to company information, sign some legal documents, set up workstation),
**Sequence** (select the mode of remuneration), **Sequence** (check))

For example, for the sequential pattern, the response time is defined as the sum of the response times of the constituent Web services. For the flow pattern (which includes parallel, synchronization and simple merge pattern), the response time is defined as the maximum response time of the constituent Web services (grant access to company information, sign some legal documents, set up workstation).

The values calculated and obtained for the composite Web services will be compared to the agreed SLA

## V. CONCLUSION AND FUTURE WORK

Monitoring Web services composition published in Cloud based on the patterns used in BPEL process remains an open research issue in Cloud computing. In this paper, we presented QMoDeSV, a novel architecture for monitoring and detecting SLA violations in Cloud computing environment.

Our framework is designed to handle the complete Web service composition management lifecycle in the Cloud environment and SLA violation detection. In addition, QMoDeSV proposes a non-intervening modular approach for monitoring QoS attributes: QoS pertinent information is collected by "watching" locally each service component. Then, based on the composition pattern of the composite service, the overall QoS information is computed. This information is used by our framework to detect potential SLA violations. Our framework can be very helpful for service providers, who can then take corrective actions to improve their services and to avoid penalties.

In our future endeavor, we will focus on the LHM and Lo2Hi modules responsible for managing the mapping of resource metrics gathered from Cloud environment to obtain SLA parameters.

## REFERENCES

[1] R. Buyya, C. S. Yeo, and S. Venugopa, "Marketoriented cloud computing: Vision, hype, and reality for delivering it services as computing utilities" In Proceedings of the 10th IEEE International Conference on High Performance Computing and Communications (HPCC-08), pp102-110

[2] A. Al-Flasi and M.A. Serhani, "A Framework for SLA-Based Cloud Services Verification and Composition", International Conference on Innovations in Information Technology, Abu Dhabi, UAE, April 2011. pp. 363-370

[3] M. L. Massie, B.N. Chun, and D. E. Culler, "The ganglia distributed monitoring system: Design, implementation and experience," Parallel computing, vol. 30, pp. 200, 206

[4] "Nagios." http://www.nagios.org/. [retreived: 7,2012]

[5] H. Newman, I. Legrand, P. Galvez, R. Voicu, and C. Cirstoiu, "MonALISA : A distributed monitoring service architecture," in Proceedings of CHEP03, La Jolla, California, 2003. pp. 214-220

[6] S. Andreozzi, N. De Bortoli, S. Fantinel, A. Ghiselli, G. L. Rubini, G. Tortone, and M. C. Vistoli, "GridICE: A monitoring service for grid systems," Future Gener. Comput. Syst., vol. 21, no. 4, pp. 559–571, 2005

[7] J. Shao, H. Wei, Q. Wang, and H. Mei, "A Runtime Model Based Monitoring Approach for Cloud," in Proceedings of 2010 IEEE 3rd International Conference on Cloud Computing (CLOUD 2010), I. C. Society, Ed. Miami, Florida: IEEE Computer Society, 2010, pp. 313–320.

[8] M. Boniface, B. Nasser, J. Papay, S. C. Phillips, A. Servin, X. Yang, Z. Zlatev, S. V.Gogouvitis, G. Katsaros, K. Konstanteli, G. Kousiouris, A. Menychtas, and D. Kyriazis, "Platformas- a-Service Architecture for Real-Time Quality of Service Management in Clouds," in Proceedings of the 2010 Fifth International Conference on Internet and Web Applications and Services (ICIW '10). Washington, DC, USA: IEEE Computer Society, 2010, pp. 155–160

[9] N. Repp, R. Berbner, O. Heckmann, and R. Steinmetz, "A Cross-Layer Approach to Performance Monitoring of Web Services," in Proceedings of the Workshop on Emerging Web Services Technology. CEUR-WS, Dec 2006. pp. 140-148

[10] Vaquero L M, Rodero-Merino L, Caceres J , and Lindner M, "A Break in the Clouds, Towards a Cloud Definition", Computer Communications Review, 2009, Vol. 39, No. 1, pp. 50-55.

[11] F. Rosenberg, C. Platzer, and S. Dustdar, "Bootstrapping Performance and Dependability Attributes of Web Services," in Proceedings of the IEEE International Conference on Web Services (ICWS'06). IEEE Computer Society, 2006, pp. 205– 212.

[12] Apache ODE http://ode.apache.org/ [retreived: 7, 2012]

[13] H. San-Yih, Wang H, S.Jaideep , and P. Raymond. "A probabilistic QoS Model and computation Framework for Web Services based workflow" In Proc of ER2004, pages 596-609, Sanghai, November 2004. pp. 254-260

[14] Michael C. Jaeger, Gregor Rojec-Goldmann, and Gero Muhl. "QoS Aggregation for Web Service Composition using Workflow Patterns" EDOC '04 Proceedings of the Enterprise Distributed Object Computing Conference, Eighth IEEE International. pp. 52-59

[15] D. Kyriazis, K. Tserpes, A. Menychtas, A. Litke, and T. Varvarigou, An innovative workflow mapping mechanism for grids in the frame of quality of service, Future Generation Computer Systems 24 (6) (2008) pp. 498–511.

 ISBN: 978-1-61208-237-0 

# Cost Optimization and Quality of Service Assurance in WAN-based Grid System

Marcin Markowski
Department of Systems and Computer Networks
Wrocław University of Technology
Wrocław, Poland
marcin.markowski@pwr.wroc.pl

***Abstract*—Grid systems become the common solutions supporting scientific researches, business analysis, entering even the world of entertainment. Modern grids should satisfy following requirements to be considered as useful and convenient tools: cost-effectiveness, performance, quality of service, reliability, security. The critical part of grid system is a communication network, responsible for proper and reliable data transfer between computing nodes. In case of grids where data transfer delays (QoS) and reliability are not crucial, public networks (usually in conjunction with VPN technology) are used. For better performance and stable communication parameters private wide area networks (WANs) are build and utilized in grid environments. WAN-based grid systems require specific designing methods in order to ensure the cost and QoS optimality. Some problems typical for wide area networks must be considered: topology of private WAN, capacity of channels, routing (flow assignment), allocation of computing nodes. In the paper the model of WAN-based grid system is presented and the cost minimizing and QoS ensuring algorithm is proposed. Both network optimization issues and distributed computing optimization issues are considered simultaneously. Algorithm may be useful as well for designing as optimizing of WAN-based grids.**

***Keywords-grid networks; capacity and flow assignment problem; distributed computing; wide area networks; algorithm***

## I. INTRODUCTION

Computational power demanded for latest applications is incessantly increasing. Huge amounts of data are collected and analyzed for different purposes: scientific (i.e., biomedical simulation), business (statistics, trends analyzing), security (pattern, i.e., malware recognition) and other. Since computational requirements often exceed the possibilities of single host, then distributed solutions like grid computing and cloud computing become more and more useful and popular. With distributed technologies, usually unused computing resources may be better utilized. Many hosts in public networks use only few percents of processing power (examples may be DNS servers or even production servers after usual work hours). In data centers, it is common that servers use only 20%-30% of processing power on average [1]. Other hosts and clusters, dedicated for computing purposes, are not loaded all the time, some services are less utilized during nights and the computing power is being wasted. Grid technology allows utilizing spare computing resources distributed in remote locations, and computations may be done in parallel on many physical systems.

Modern grids should satisfy following requirements to be considered as useful and convenient tools: latency, bandwidth, reliability, fault-tolerance, jitter control and security [7]. Then, the critical part of grid system is a communication network, responsible for proper and reliable data transfer between computing nodes. Communication between nodes may be based on public network. In this case it is rather difficult to ensure minimal data transfer delays (QoS) and reliability, because public lines are shared with other users and are susceptible to overloading and security attacks. For better performance and stable communication parameters private wide area networks (WANs) are build and utilized for grid environments.

In grid optimization issues, that may be found in the literature, different task scheduling problems are considered [2, 3, 4], denoting that the structure, capacity and flow routes in the communication network are given. Such solutions are not able to ensure QoS and reliability in the network layer. There is a lack of solutions for simultaneously optimizing of task scheduling and network parameters (i.e., topology, capacity and flow). The problem of WAN-based grid optimization, considered in the paper, represents more complex approach, taking into account the structure of the communication network. The problem consist in task assignment to grid nodes simultaneously with network capacity and flow routes assignment. The combined optimization criterion includes the computing cost and the network cost. The optimization parameters are: task assignment to nodes of grid, location of grid management centre, capacities of channels and flow routes.

The paper is organized as follows. In Section III the optimization problem has been formulated and explained in details. The mathematical model of WAN-based grid, with decision variables and constraints has been presented in Section IV. Approximate algorithm for considered problem is proposed in Section V. Section VI concludes the paper.

## II. RELATED WORK

Issues in designing and optimizing of grid systems are well known in literature [2-6]. Itami *et al* [3] relates to real-time distributed systems, where the rapid and reliable communication between nodes is critical. The authors proposed event-triggered distributed object models and developed a distributed computing environment. Sterritt *et al.*

 ISBN: 978-1-61208-237-0 

[4] consider the problem of autonomic computing. Autonomic computing refers to self-managing and self-configuring distributed and grid systems that are able to grow and increase they complexity without or with only the little involvement of administrator.

Grid optimizing solutions, that may be found in literature do not take into consideration simultaneous optimization of tasks assignment and network capacity. Usually overlay network, connecting grid nodes is being considered and bandwidth to each computing node is denoted. Such simplification allows to construct simpler algorithms, but the performance of solutions is strongly dependent on topology and capacity of communication network. Building grid networks on WAN base implies additional optimization issues. Designing of WANs involves such problem as: topology assignment, channels' capacities assignment and routing assignment (known also as flow assignment). In the classical capacity and flow assignment (CFA) problem, there are two design variables: channel capacities and flow routes (routing) [8, 9, 10]. The goal is to select those variables in order to minimize the criterion function, for example the total average delay per packet in wide area network or the capacity leasing cost [10]. Algorithm proposed in the paper allows to obtain better results, related to grid algorithms known in literature, because network capacity and network routing are considered and optimized. The advantages of such approach are better fixing of communication network to grid demands, less QoS and reliability inconveniences, finally lower costs of supporting the grid network.

Optimization problem in distributed computing systems, with different criterions and constraints set was the subject of our previous considerations [11]. The model of distributed computing system was proposed and optimization problem with the criterion combined of quality of service (indicated by average packet delay) and computing cost was solved. The WAN-based grid model, proposed in [11] is used in the paper in order to formulate an optimization problem and design an approximate algorithm.

## III. WAN-BASED GRID OPTIMIZATION PROBLEM

Consider grid system build of certain number of computing nodes connected to nodes of the WAN network. In each time period (in example each second) the portion of data is generated and requires to be processed. It has to be divided into separable parts (usually called blocks or subtasks) an sent to proper computing nodes. We denote than computational outlay needed in each time period is similar and each generated computational task may be easily divided into subtasks. We denote that during processing of particular block computing nodes do not have to exchange any information with nodes processing other subtasks. After processing, result data must be sent to the main node and the final result is compiled from all subtasks. The main node, managing realization of computational tasks we call the computing management centre (CMC). Management centre divides tasks into blocks, transmits blocks to grid nodes, receives and collects result data for each block, and compiles final result. The grid node, in which CMC is located, also takes part in blocks processing. Blocks and results transferred between this node and CMC are not sent through the network. Each grid node communicates only with computational centre node.

Processing block of data is connected with specified processing cost. Cost may represent money (for buying computational power), resource utilization or other virtual costs. Costs are specified for each of the blocks. Blocks may differ in computational effort needed to process them, then processing costs may also differ. Management centre is located in one of the grid nodes, since it is source or destination of all data transmitted in the network, then the proper allocation of CMC has a critical impact on the quality of service in the network. Maintaining of CMC in the node generates also some maintaining cost in each time period. Minimizing of total processing cost and maintaining cost is one of the objectives of grid optimization.

The computational power resources (also called resource capacity [6]) of grid nodes are limited and determined for each node. It is denoted that the total resources capacity of grid is enough to presses all generated blocks.

For each block two parameters must be specified: size of the block and size of the results. Size of the block is the amount of data that must be sent from CMC to the computing node chosen to process the block. Size of results generated during block processing is the amount of data that size must be sent from processing node to CMC when processing is finished.

It is assumed that communication network for grid system is the packed switched network. Channel allocation (network topology) is given, capacity of each channel must be assigned in optimization process. Possibility of assignment of channels capacity is very important especially for CMC node, which generates and receives all data sent between nodes. We may ensure enough bandwidth to computing management centre node in order to operate all transmissions.

On the basis of above assumptions, considered problem we formulate as follows:

Given:

- number of grid nodes, number of channels of wide area network,
- for each node: computational capacity of node,
- for each channel the set of possible capacities and costs (i.e., cost-capacity function),
- list of candidate nodes for Computing Management Centre, for each candidate node the value of maintaining cost,
- number of blocks which must be processed in each time period,
- for each block: size of block, size of result data, computational outlay needed to process block, costs of processing at each computing node,
- maximal acceptable average delay in the network.

Minimize:

- linear combination of supporting cost of the network (capacity leasing cost) and the total computing cost (total cost of data processing and maintaining cost of management centre).

 ISBN: 978-1-61208-237-0 

Over:
- CMC allocation,
- block allocation at computing nodes,
- channel capacities,
- flow routes (routing).

Subject to:
- channel capacity constraints
- multicommodity flow constraints
- computing capacity of nodes constraints
- QoS (delay) constraints

We assume that channels' capacities can be chosen from discrete sequence defined by ITU-T (International Telecommunication Union – Telecommunication Sector) recommendations. The formulated above problem is NP-complete, as more general that capacity and flow assignment problem [10, 12].

## IV. MODEL OF WAN-BASED GRID SYSTEM

In this section, the mathematical model of the WAN-based grid systems is presented. Then the optimization criterion is proposed and optimization problem is formulated. Developing of the system model is necessary in order to implement optimization algorithms. The model is based on graph theory. Selection of channel capacity, CMC allocation and block-to-node allocation are modeled with binary decision variables. Important constraint, connected to decision variables are also introduced. Model of WAN-based distributed system was developed and presented in our previous work [11]. Model presented in this section is an adaptation of the previous one.

### A. Variables and Parameters

Let $n$ be the number of computation nodes of the considered grid system. Let $b$ be the number of channels of the wide area network connecting grid nodes. For each channel $i$ there is the set $C^i = \{c_1^i,...,c_{s(i)}^i\}$ of alternative values of capacities. Let $D^i = \{d_1^i,...,d_{s(i)}^i\}$ be the set of leasing costs corresponding to channel capacities from the set $C^i$. Let $x_j^i$ be the discrete decision variable, connected with capacity choice for channel $i$:

$$x_j^i = \begin{cases} 1, \text{ if the capacity } c_j^i \text{ is assigned to channel } i \\ 0, \text{ otherwise} \end{cases}$$

Exactly one capacity from the set $C^i$ must be chosen for channel $i$, then the following condition must be satisfied [11]:

$$\sum_{j=1}^{s(i)} x_j^i = 1 \quad \text{for} \quad i = 1,..,b \tag{1}$$

Let $X_r$ be the set of all variables $x_j^i$ which are equal to one. $r$ is the number of iteration, since in successive chapters we propose approximate iterative algorithm.

Let $y_h$ be the discrete decision variable, connected with allocation of management node:

$$y_h = \begin{cases} 1, \text{ if the management centre is located in node } h \\ 0, \text{ otherwise} \end{cases}$$

Let $Y_r$ be the set of all variables $y_h$ which are equal to one. Since the management centre is located in one node only, the following condition must be satisfied [13]:

$$\sum_{h=1}^{n} y_h = 1 \tag{2}$$

Let $u_h$ be the cost of maintaining of the management centre in the node $h$.

Let $t$ denotes the number of blocks, which must be proceeded in the considered time period. The following given data are connected with each block:
- $v_l$ is the size of block $l$. In each time period the data of size $v_l$ must be sent from management node to the proper computing node;
- $w_l$ is the size of data generated as result of proceeding block $l$. Data of size $w_l$ must be sent from computing node to the management node at each time period;
- $p_l$ is the computational outlay needed to proceed block $l$ – measured in [instructions];
- $q_m^l$ is the cost of proceeding block $l$ in the computing node $m$, $Q^l = \{q_1^l,...,q_n^l\}$ is the set of proceeding costs of block $l$ in all computing nodes.

Let $z_m^l$ be the discrete decision variable, connected with grid node choice for proceeding block $l$:

$$z_m^l = \begin{cases} 1, \text{ if block } l \text{ is proceeded in grid node } m \\ 0, \text{ otherwise} \end{cases}$$

Let $Z_r$ be the set of all variables $z_m^l$ which are equal to one. In [11], we have proposed the following condition, that must be satisfy in order to ensure that each of the blocks is proceeded exactly in one node:

$$\sum_{l=1}^{t} \sum_{m=1}^{n} z_m^l = t \tag{3}$$

Each of the grid nodes has a limited computing power. Let $e_m$ be the computing power (also called computational capacity) of node $m$, given in [instructions per second] (IPS). In order to guarantee that the optimization problem has a solution, we propose the following condition which must be satisfied for given data:

$$\sum_{m=1}^{n} e_m > \sum_{l=1}^{t} p_l \tag{4}$$

Let $r_{mk}$ be the average traffic rate sent from node $m$ to node $k$ in each time period. In packet switched networks the flow between nodes is realized as a multicommodity flow.

 ISBN: 978-1-61208-237-0 

Since, we denote that in the considered networks only proceeding block and results of proceedings block are sent (there is no other network traffic) then $r_{mk}$ consists of packed exchanging between computing nodes and management node only. $r_{mk}$ we calculate as follows:

$$r_{mk} = \sum_{l=1}^{t} \left( z_m^l y_k w_l + z_k^l y_m v_l \right)$$

The triple of sets $(X_r, Y_r, Z_r)$ is called a selection. Let $\Re$ be the family of all selections. The selection $(X_r, Y_r, Z_r)$ defines the unique wide area network and distributed computer system, because:

- $X_r$ determines the values of capacities for channels of the WAN,
- $Y_r$ determines the allocation of CMC at the node of WAN.
- $Z_r$ determines the blocks' allocation at the grid nodes of distributed grid.

### B. Criteria and Constraints

Let $T(X_r, Y_r, Z_r)$ be the minimal average delay per packet in the wide area network in which values of channel capacities are given by set $X_r$ and traffic requirements are given by sets $Y_r$ and $Z_r$ (depend on management node allocation and assigning of block to computing nodes). $T(X_r, Y_r, Z_r)$ can be obtained solving a multicommodity flow problem in the network [10, 12]:

$$T(X_r, Y_r, Z_r) = \min_{\underline{f}} \frac{1}{\gamma} \sum_{x_j^i \in X_r} \frac{f_i}{x_j^i c_j^i - f_i}$$

subject to:

- $\underline{f}$ is a multicommodity flow satisfying the traffic requirements $r_{mk}$ given by $Y_r$ and $Z_r$ ,
- $f_i \le x_j^i c_j^i$ for every $x_j^i \in X_r$
- $f = [f_1, ..., f_b]$ is the vector of multicommodity flow, $f_i$ is the total average bit rate on channel $i$ , and $\gamma$ is the total packet rate generated and sent through the network by computational nodes and management node.

Let $A(Y_r, Z_r)$ is the computing cost, composed of proceeding costs of block at computational nodes and cost of maintaining of management node. We propose to calculate it as follows:

$$A(Y_r, Z_r) = \sum_{h=1}^{n} y_h u_h + \sum_{l=1}^{t} \sum_{m=1}^{n} q_m^l$$

Let $B(X_r)$ be the regular leasing cost of channel capacities, given with the formula:

$$B(X_r) = \sum_{x_j^i \in X} x_j^i d_j^i$$

Then, the we propose following objective function for the channel capacities, routing assignment and location of management centre and task scheduling problem:

$$OBJ(X_r, Y_r, Z_r) = \sigma A(Y_r, Z_r) + B(X_r)$$

Computing cost may not be the money but also computational power, CPU utilization or other abstract cost. Network maintaining cost is usually the money.

Let $T^{\max}$ be the maximal acceptable average packet delay in the network. $T^{\max}$ defines the level of quality of service (QoS) in the grid network. Quality of service in the network become degraded when average packet delay exceeds $T^{\max}$ .

### C. Problem Formulation

Above definitions allow us to formulate the WAN-based grid optimization problem:

$$\min_{(X_r, Y_r, Z_r)} OBJ(X_r, Y_r, Z_r) \quad (5)$$

Subject to:

$$(X_r, Y_r) \in \Re \quad (6)$$

$$T(X_r, Y_r, Z_r) \le T^{\max} \quad (7)$$

$$\sum_{l=1}^{t} z_m^l p_l < e_m \text{ for each } m = 1,.., n \quad (8)$$

## V. APPROXIMATE ALGORITHM

The problem (5)-(8) is NP-complete, as more general than classical CFA problem, which is NP-complete [12, 13]. NP-completeness is defined as the set of decision problems that can be solved in polynomial time on a nondeterministic Turing machine (Nondeterministic-Polynomial time). It means that complexity of the problem increases very quickly as the size of the problem (for example number of possible values of capacity for each channel) grows.

In the paper, an approximate algorithm for problem (5)-(8) is proposed. Unlike exact algorithms, approximate ones usually are able to find suboptimal solutions, not far from optimal. An advantage of approximate algorithms is the computational time – finding optimal solution for NP-complete problems takes very long time. Some exact algorithms for different WAN optimization problems were proposed by Markowski and Kasprzak [13, 14, 15].

Proposed algorithm starts with assignment of an acceptable solution of the problem. Acceptable solution is the selection $(X_r, Y_r, Z_r)$ and flow vector $\underline{f}$ , satisfying conditions (6) - (8). For the considered problem, finding acceptable solution consist in allocating of computing management centre, assignment tasks to computing nodes in order to satisfy requirement (8) and find solution for CFA problem satisfying (7). In case that the problem has no solution (i.e., it is impossible to build the network satisfying flow demands with given budget restriction), it is discovered during this stage and algorithm finishes. The second phase of an algorithm is optimization phase, while sub-optimal solution is being found.

 ISBN: 978-1-61208-237-0 

*A. Initial Solution*

Three main tasks appear during initial phase: choosing allocation for computing management centre, tasks allocation over the computing nodes, capacities of channels and flow assignment.

Few strategies for choosing the node for CMC allocation in distributed computing systems were proposed in [11]:

- Maximal computing power of candidate node
  The grid node maintaining CMC also takes part in blocks processing. Moreover, block processed in CMC node and results of them are not sent through the network. Locating CMC in the node of maximal computing power allows to minimize the data transfer in the network. To valuate quality of node, according to this criterion, we use the value of computational capacity $e_m$.
- Maximal capacity of node
  Since all blocks (except those processed by CMC node) must be send to grid nodes and results must be send back, links adjacent to CMC node must ensure enough capacity.
  Let $P_m$ be the sum of capacities of all channels adjacent to node $m$:

$$P_m = \sum_{x_j^i \in X_r} x_j^i c_j^i p_m(i) \tag{9}$$

  where

$$p_m(i) = \begin{cases} 1, & \text{if } i\text{ - th channel is adjacent to } m\text{ - th node} \\ 0, & \text{otherwise} \end{cases}$$

  To valuate quality of node, according to this criterion, we use the value of $P_m$.
- Location of node
  In order to minimize the traffic in the whole network, it is beneficial to locate the CMC in the centre of the grid network.
  Let $v_{gh}$ be the distance, in hops, between node $g$ and node $h$. It means that $v_{gh}$ is the minimal number of channels between nodes $g$ and $h$. Let $V_g$ be the distance between node $g$ and all other nodes of distributed grid, defined as follows:

$$V_g = \sum_{h=1}^{n} v_{gh} \tag{10}$$

  Choosing the node for allocating CMC we should choose such nodes $g$, for which the value of $V_g$ is minimal.

Another aim of an initial phase of an algorithm is task allocation to the computing nodes. We propose two strategies for initial task allocation:

- Regular tasks distribution approach.
  In this strategy, tasks are being allocated to the grid nodes in proportion to the computational power of each grid node.
- Regular traffic distribution approach.
  Task are being allocated to grid nodes in proportion to the capacity of node (the sum of capacities of all channels adjacent to the node). The constraint of the node computational power (4) must be satisfied.

Finally, capacity and flow (CFA) assignment problem for the initial phase may be solved using in order to satisfy requirement (6). We propose simple method, since optimal solution of CFA problem is not necessary on this phase. We start from maximal capacity for each channel of the network. Then, we minimize the capacities as long as (6) is satisfied.

*B. Suboptimal solution*

We start from the initial selection obtained in first phase. Then, in consecutive iterations we try to improve the solution by changing CMC allocation node, tasks allocation at grid nodes, routing and channel capacities. Algorithm finishes when there is no possibility of improving present solution.

To get the best choice we have to test all possible pairs of variables $y_h \in Y_r, y_m$ or $x_j^i \in X_r$, $x_s^i$ or $z_m^l \in Z_r, z_h^l$ using a local optimization criterion. Because of the different nature of the variables denoting channel capacity choice, block assignment and the computing centre allocation, we have to formulate three different criteria – one for each of decision variables $x_j^i$, $y_h$ and $z_m^l$.

*Proposition* 1. If the selection $(X_t, Y_t)$ is obtained from the selection $(X_r, Y_r)$ by complementing the variable $x_j^i \in X_r$ where $j < s(i)$ by another variable $x_s^i \in X_t$ then only the channel capacity change is being considered. We propose the following local optimization criterion on variables $x_j^i$ where $j < s(i)$:

$$\Delta_{js}^i = \begin{cases} Q(X_r, Y_r, Z_r) - d_j^i + d_s^i & \text{for } c_s^i > f_i \\ \infty & \text{for } c_s^i \le f_i \end{cases}$$

Complementing of variables $x_j^i$ means that value of total average delay in the network changes, that may affect restriction (6). Then we propose criterion for estimating of the total average delay after complementing:

$$\Delta T_{js}^i = T(X_r, Y_r, Z_r) + \frac{\sigma}{\gamma}\left(\frac{f_i}{c_s^i - f_i} - \frac{f_i}{c_j^i - f_i}\right) \text{ for } c_s^i > f$$

*Proposition* 2. If the selection $(X_t, Y_t)$ is obtained from the selection $(X_r, Y_r)$ by complementing variable $y_h \in Y_r$ by another variable $y_m \in Y_t$ then only allocation of CMC is being changed. Then, the traffic requirements between nodes change, channels' capacities do not change and blocks' allocation at computing nodes do not change. Then, to evaluate the pair $(y_h, y_m)$ we propose following criterion:

 ISBN: 978-1-61208-237-0 

$$\delta_{hm}^{CMC} = \begin{cases} Q(X_r, Y_r, Z_r) - u_h + u_m \\ \qquad \text{if } \tilde{f}_i < x_j^i c_j^i \text{ for } x_j^i \in X_r \\ \qquad \text{and } P_m \geq \sum_{l=1}^{t} (w_l + v_l) \\ \infty \quad \text{otherwise} \end{cases}$$

Value of total average delay, obtained in result of complementing is estimated as follows [11]:

$$\Delta T_{hm}^{CMC} = \frac{\sigma}{\gamma} \sum_{x_j^i \in X_r} \frac{\tilde{f}_i}{x_j^i c_j^i - \tilde{f}_i} \quad \text{if } \tilde{f}_i < x_j^i c_j^i \text{ for } x_j^i \in X_r$$

where

$$\tilde{f}_i = \sum_{h=1}^{n} \sum_{\pi_{hm}^a \in \Pi_{hm}} V_{hm}^a(i) \bar{f}_{hm}^a$$

$$\bar{f}_{hm}^a = \frac{m_{hm}^a}{m_{hm}} r_{mk}, \quad m_{ej} = \sum_{\pi_{ej}^a \in \Pi_{ej}} m_{ej}^a$$

$$m_{hm}^a = \min_{i \in \pi_{hm}^a} \left( x_j^i c_j^i \right) \text{ for } x_j^i \in X_r$$

$\Pi_{hm}^a$ denotes the $a$-th path from node $h$ to node $m$,

$$V_{hm}^a(i) = \begin{cases} 1 \text{ if } i\text{-th channel belong to path } \Pi_{hm}^a \\ 0 \text{ otherwise} \end{cases}$$

$\Pi_{hm}$ denotes the set of all paths from node $h$ to node $m$.

$\tilde{f}$ is the 'new' traffic flow in the network, after reallocating the CMC. After reallocation all routes in the network must be redirected from paths [*old CMC node; other nodes*] to paths [*new CMC node; other nodes*]. It is simply calculated as follows. For each node, we find all possible routes (paths) from that node to new CMC node. They are denoted by the set $\Pi_{hm}$. Then we allocate the traffic along all found routes, proportionally to they residual capacities.

*Proposition* 3. If the selection $(X_t, Y_t)$ is obtained from the selection $(X_r, Y_r)$ by complementing variable $z_m^l \in Z_r$ by another variable $z_h^l \in Z_t$ then the allocation of $l$-th block is being changed. So, the traffic requirements between nodes change, channels' capacities and CMC location do not change. We propose following criterion for evaluating the pair $\left(z_m^l, z_h^l\right)$:

$$\delta_{hm}^{l} = \begin{cases} Q(X_r, Y_r, Z_r) - q_m^l + q_h^l \\ \qquad \text{if } \tilde{f}_i < x_j^i c_j^i \text{ for } x_j^i \in X_r \\ \qquad \text{and } \sum_{l=1}^{t} z_m^l p_l < e_m \\ \infty \quad \text{otherwise} \end{cases}$$

Value of total average delay, obtained in result of complementing [12]:

$$\Delta T_{hm}^{l} = \frac{\sigma}{\gamma} \sum_{x_j^i \in X_r} \frac{\tilde{f}_i}{x_j^i c_j^i - \tilde{f}_i} \quad \text{if } \tilde{f}_i < x_j^i c_j^i \text{ for } x_j^i \in X_r,$$

where

$$\tilde{f}_i = f_i - f_{il}'' + \sum_{e=1}^{n} \sum_{\pi_{hm}^a \in \Pi_{hm}} V_{hm}^a(i) \bar{f}_{hm}^a$$

$$\bar{f}_{hm}^a = \frac{m_{hm}^a}{m_{hm}} r_{mk}, \quad m_{ej} = \sum_{\pi_{ej}^a \in \Pi_{ej}} m_{ej}^a$$

$$m_{hm}^a = \min_{i \in \pi_{hm}^a} \left( x_j^i c_j^i - f_i + f_{il}'' \right) \text{ for } x_j^i \in X_r$$

$f_{il}''$ is the part of the flow at $i$-th channel. It corresponds only to the packets connected with $l$-th block. $\Pi_{hm}^a$, $\Pi_{hm}$ and $V_{hm}^a(i)$ are defined like previously.

Replacements of decision variables are made in order to obtain the distributed computing network with the possible least value of criterion function *OBJ*. We should choose such pairs $y_h \in Y_r, y_m$ or $x_j^i \in X_r$, $x_s^i$ or $z_m^l \in Z_r, z_h^l$, for which the value of the criterion $\delta_{hm}^{CMC}$, $\Delta_{js}^i$ or $\delta_{hm}^l$ is minimal and increase of value of average total delay $T(X_r, Y_r, Z_r)$ is minimal.

### C. Calculation Scheme

Initial Phase

- Step 1.1. Choose node for computing management centre. Evaluate nodes using one of criteria defined in subsection A. Choose node with maximal computing power $e_m$, maximal capacity of node $P_m$ or the node with the best location according to criterion (10). Also combined criteria may be used, in example $e_m / V_m$ or $(e_m P_m) / V_m$.
- Step 1.2. Allocate tasks to the computing nodes, according to regular tasks distribution strategy or regular traffic distribution strategy.
- Step 1.3. Assign maximal possible capacity for each channel. Solve FA problem [8]. Calculate total average delay in the network. If $T(X_r, Y_r, Z_r) \leq T^{\max}$ then decide that problem (5)-(8) has no solution - algorithm finishes.
- Step 1.4. In consecutive steps, find the less utilized channel and decrease its capacity. Where there in no possibility for capacity reduction without violating requirement (6), then calculate value of objective function and remember it as $OBJ_{\min}$. Remember actual sets of decision variables as $(X_{\min}, Y_{\min}, Z_{\min})$.

  Go to optimization phase .

Optimization Phase

- Step 2.1. Perform $r = 0$. $(X_r, Y_r, Z_r) = (X_{\min}, Y_{\min}, Z_{\min})$.
- Step 2.2. Perform $r = r + 1$. Choose pair $y_h \in Y_{r-1}, y_m$ or $x_j^i \in X_{r-1}$, $x_s^i$ or $z_m^l \in Z_{r-1}, z_h^l$, for which the value of the criterion $\delta_{hm}^{CMC}$, $\Delta_{js}^i$ or $\delta_{hm}^l$ is minimal and, respectively, $\Delta T_{hm}^{CMC} \le T^{\max}$, $\Delta T_{js}^i \le T^{\max}$ or $\Delta T_{hm}^l \le T^{\max}$. If there is no such pair for which $\delta_{hm}^{CMC}$, $\Delta_{js}^i$ or $\delta_{hm}^l$ is less than $OBJ_{\min}$ then stop, $(X_{\min}, Y_{\min}, Z_{\min})$ is sub-optimal solution of problem (5)-(8). Otherwise swap values of variables of chosen pair.
- Step 2.3. Solve the flow assignment problem in WAN, where traffic requirements are given by CMC allocation and blocks allocation at nodes and channels' capacities are given by set $X_r$. Calculate $OBJ_r(X_r, Y_r, Z_r)$. If $OBJ_r < OBJ_{\min}$ (better solution is found), then assign $(X_r, Y_r, Z_r) = (X_{\min}, Y_{\min}, Z_{\min})$, and $OBJ_{\min} = OBJ(X_{\min}, Y_{\min}, Z_{\min})$. Go to step 2.2.

### D. Experiments and Analisis

Experiments conducted with proposed algorithm will validate the quality of approximate solutions and computational properties of an algorithm. Since the presented problem is NP-complete, there are no effective algorithms for finding optimal solutions - a brute force method may be used for small size problems but it takes exponential time. The proposed algorithm allows to find suboptimal solution in linear time.

Important parameter is the quality of solutions calculated by an algorithm. It may be measured as the distance between optimal and suboptimal solution. Precise distance may be calculated for small size problem, when optimal solution is find with the brute force method.

## VI. CONCLUSION

In the paper, the WAN-based grid optimization problem with combined cost function was formulated and an approximate algorithm was proposed. The considered problem is far more general than the similar problems presented in the literature, because network optimization (capacity and flow routes) is carried out simultaneously with block assignment and management centre allocation. Algorithms proposed so far for grid networks in the literature do not take into consideration network optimization problems, then performance of grid may be affected by network overloads.

Considering two different kinds of cost in criterion function is very important from practical point of view, since computing cost and supporting cost of the network are significant and carried regularly. Computing cost may not be the money but also computational power, CPU utilization or other. In some application computing cost may be more important and less important in other. With dual-cost criterion algorithm may be better fitted to different user demands – we may define the importance of each cost using proper parameter in criterion function. Since considered problem is NP-complete, the big advantage of proposed approximate solution in short computing time needed to find solution, even for grid composed of hundreds of computing nodes.

## REFERENCES


[1] D. Meisner, B.T. Gold, and T.F. Wenisch, "PowerNap: Eliminating Server Idle Power", Proc. The 14th international Conference on Architectural Support for Programming Languages and Operating Systems, ACM, New York, 2009, pp. 205-216

[2] H. Attiya and J. Welch, Distributed Computing: Fundamentals, Simulations, and Advanced Topics, 2 ed., Wiley-Interscience, 2004

[3] Y. Itami, T. Ishigooka, and T. Yokoyama, "A Distributed Computing Environment for Embedded Control Systems with Time-Triggered and Event-Triggered Processing", Proc. The 14th IEEE International Conference on Embedded and Real-Time Computing Systems and Applications, 2008, pp. 45–54.

[4] R. Sterritt and D. Bustard, "Towards an Autonomic Computing Environment", Proceedings of the 14th International Workshop on Database and Expert Systems Applications, 2003, pp. 699 – 703.

[5] F. Travostino, Grid Networks, J. Wiley & Sons, 2006.

[6] S. Demeyer, M. De Leenheer, J. Baert, M. Pickavet, and P. Demeester, "Ant colony optimization for the routing of jobs in optical grid networks", Journal Of Optical Networking, vol. 7, no. 2, pp. 160-172, 2008.

[7] M. Baker, R. Buyya, and D. Laforenza, Grids and Grid Technologies for Wide-Area Distributed Computing, Software — Practice and Experience, Hoboken, NJ: Wiley, 2002.

[8] L. Fratta, M. Gerla, and L. Kleinrock, "The Flow Deviation Method: An Approach to Store-and-Forward Communication Network Design", Networks, Vol. 3, 1973, pp. 97–133.

[9] M. P. Clark, Data networks, IP and the Internet: protocols, design and operation, John Wiley & Sons, 2003.

[10] M. Pioro and D. Medhi, "Routing, Flow, and Capacity Design in Communication and Computer Networks", Elsevier, Morgan Kaufmann Publishers, San Francisco, 2004.

[11] M. Markowski, "Resource and Task Allocation Algorithm for WAN-based Distributed Computing Environment", International Journal of Electronics and Telecommunications, Vol. 56, No 2, 2010, pp. 197-202

[12] A Kasprzak, Topological Design of the Wide Area Networks, Wroclaw University of Technology Press, Wroclaw, 2001.

[13] M. Markowski and A. Kasprzak, "The web replica allocation and topology assignment problem in wide area networks: algorithms and computational results", Lecture Notes in Computer Science, 3483, pp. 772-781, 2005.

[14] M. Markowski and A. Kasprzak, "The Three-Criteria Servers Replication and Topology Assignment Problem in Wide Area Networks", Lecture Notes in Computer Science, 3982, pp. 1119-1128, 2006.

[15] M. Markowski and A. Kasprzak, "An approximate algorithm for replica allocation problem in wide area networks", Proc. 3rd Polish-German Teletraffic Symp. PGTS 2004, VDE Verlag, Berlin, pp. 161-166, 2004.


 ISBN: 978-1-61208-237-0 

# A Fuzzy Test Cases Prioritization Technique for Regression Testing Programs with Assertions

Ali M. Alakeel
Faculty of Computing and Information Technology
University of Tabuk
Tabuk, Saudi Arabia
alakeel@ut.edu.sa

***Abstract*—Program assertions have been recognized as a supporting tool during software development, testing, and maintenance. Therefore, software developers place assertions within their code in positions considered to be error prone or have the potential to lead to software crash or failure. Like any other software, programs with assertions have to be maintained. Depending on the type of modification applied to the modified program, assertions also may have to go through some modifications. New assertions may also be introduced in the new version of the program while some assertions may be kept the same. This paper presents a novel approach for test cases prioritization using fuzzy logic for the purpose of regression testing programs with assertions. The proposed approach builds upon previous research in the fields of assertions-based software testing and assertions revalidation. In a first step, our method utilizes fuzzy logic concepts to measure the effectiveness of a given test case in violating a program assertion. The result of the first step is then used in prioritization test cases during the regression testing of programs with assertions. The main objective of this research is to show that fuzzy logic concepts may be employed to measure the *effectiveness* of a given test case in violating programs assertions during the regression testing of a modified program.**

***Keywords--Regression Testing; Fuzzy Logic; Program Assertions; Software Testing; Software Test Data Generation.***

## I. INTRODUCTION

Program assertions have been recognized as a supporting tool during software development, testing, and maintenance, e.g., [1-5]. Therefore, software developers place assertions within their code in positions considered to be error prone or have the potential to lead to software crash or failure, e.g., [4]. An assertion specifies a constraint that applies to some state of computation. When an assertion evaluates to a *false* during program execution (this is called assertion violation), there exists an incorrect state in the program. Many programming languages support assertions by default, e.g., Java and Perl. For languages without built-in support, assertions can be added in the form of annotated statements. For example, Korel and Al-Yami [2], presents assertions as commented statements that are pre-processed and converted into Pascal code before compilation. Many types of assertions can be easily generated automatically such as boundary checks, division by zero, null pointers, variable overflow/underflow, etc. For this reason and to enhance their confidence in their software, programmers may be encouraged to write more programs with assertions.

Recognizing the importance of program assertions, some recent research efforts have been devoted for the development of algorithms and methods specifically designed for programs with assertions. For example, Korel et al. reported in [6] an algorithm for assertions revalidations during software maintenance. In [3], an algorithm is presented for the efficient processing and analysis of a large number of assertions present in the program. Also, a regression testing method for program with assertions was proposed in [7].

Like any other software, programs with assertions have to be maintained. Software maintenance usually involves activities during which the software is modified for different reasons. Some of the reasons for which software may be modified are fixing faults, introducing a new functionality, improving the performance of some parts of the software through the introduction of new algorithms, etc. A study in [8] shows that there is a probability of 50-80% of introducing faults to the modified software during software maintenance. For this reason regression testing is performed during software maintenance for the purpose of testing the modified software. There exists many regression testing methods which may be classified as specification-based or code-based. Specification-based regression testing strategies, e.g., [9-11] generate test cases based on the specification of the software, while code-base regression testing, e.g., [7], [12-15] strategies depends on the software structural elements to generate test cases.

Regression testing is a very labor intensive and may be responsible for approximately 50% of software maintenance's cost [16]. In a systematic software development environment, all types of regression testing methods usually involve the usage of an original test suite which is used for the purpose of testing the original program before it has been modified. Sensible regression testing methods have to utilize existing test suite in some form. For example, a simple regression testing strategy would rerun existing testing suite, as it is, on the modified program while introducing new test cases to test new features. Although this method is simple, it is not practical for commercial software because existing test suite is usually very large and may take weeks to rerun on the new modified software. Therefore, regression test selection techniques, test suite minimization technique and test case prioritization techniques are proposed in the literature.

To mitigate the cost associated with running the whole existing test suite, the main objective of regression test selection techniques, e.g., [17-18], and test suite minimization techniques, e.g., [19-20], is to select a representative subset of the original test suite using information about the original program, its modified version and the original test suite. It should be noted that both of the regression test selection and test suite minimization techniques eliminate some elements of the original test suite which may undermine the performance of

 ISBN: 978-1-61208-237-0 

these techniques. Test case prioritization techniques, e.g., [21-24] order elements of the original test suite based on a given criterion. Furthermore, test case prioritization techniques do not involve the selection of a subset of the original test suite. In this presentation, we will concentrate on test case prioritization techniques, therefore regression test selection and test suit minimization will not be discussed any further.

Depending on the type of modification applied to the modified program which includes assertions, assertions also may go through some modifications. New assertions may also be introduced in the new version of the program while some assertions may be kept the same as in the original program. This paper presents a novel approach for test cases prioritization using fuzzy logic for the purpose of regression testing programs with assertions. The main objective of this research is to show that fuzzy logic concepts may be employed to measure the *effectiveness* of a given test case in violating programs assertions during the regression testing of a modified program. The proposed method builds upon previous research in the fields of assertions-based software testing and assertions revalidation reported in [6-7]. In a first step, our method utilizes fuzzy logic concepts [25-27] to measure the effectiveness of a given test case in violating a program assertion. The result of the first step is then used in prioritization test cases during the regression testing of programs with assertions.

The rest of this paper is organized as follows. Related work is discussed in Section II. We present our proposed fuzzy test cases prioritization model in Section III. Our conclusions and future work is discussed in Section IV.

## II. RELATED WORK

Previous research in using fuzzy logic for the purpose of test case prioritization is scant. In [28], a fuzzy expert system is reported where this system is used for a telecommunication application. To build the required knowledge base for the expert system reported in this research, the researchers had to acquire knowledge from different sources such as customer profile, past test results, system failure rate, and the history of system architecture changes. Although this expert system has shown promising results with respect to the *specific* application it was designed for, it is necessary to acquire a new knowledge base for new applications. Also, the proposed method in [28] treats the software under test as a black box; therefore, it cannot be used for the purpose of regressing testing programs with assertions.

### A. Regression Tesing for Programs with Assertions

This section briefly introduces the concept of regression testing for programs with assertions. For more detail, the reader is referred to [7]. Given an original program $P_o$ and a modified version of this program $P_m$, let $A_o= \{a_{o1}, a_{o2}, a_{o3}, \dots a_{on}\}$ be a set of assertions found in $P_o$ and $A_m= \{a_{m1}, a_{m2}, a_{m3}, \dots a_{mz}\}$ be a set of assertions found in $P_m$. Let $V \subseteq A_m$ be a set of assertions that are nominated for revalidation [6], using previous test suits, during the process of regression testing of $P_m$. Depending on the type of modification applied to the modified version, $P_m$, some assertions may have been kept the same; some assertions may have been modified, and new assertions may have been introduced. The main objective of regression testing for programs with assertions reported in [7] is to reduce the cost of regression testing of programs with assertions through the utilization of previous test suits that are used during the initial development process. Furthermore, this method concentrates on assertions that are kept the same and those which are modified; new assertions are not covered because new test cases must be generated to explore these assertions. The main elements of this method are described in the next paragraph.

Let $a_{mi} \in A_m$ be an assertion found in $P_m$. Assume that $a_{mi}$ was not changed from its original form in $P_o$ nor was it affected by the modifications [6] introduced to produce $P_m$. Therefore, $a_{mi}$, will be nominated, by the proposed approach, to belong to the set V, i.e., $a_{mi} \in V$. Suppose that assertions-oriented testing as reported in [2], has been performed on the original version $P_o$ and a set of test cases were generated during this process and were kept for later usage during regression testing. Specifically, let $a_{ok} \in A_o$ be an assertion found in $P_o$ and let $T(a_{ok})= \{t_{k1}, t_{k2}, t_{k3},\dots, t_{kr}\}$ be the set of test cases which were generated to explore this assertion during the application of assertion-oriented testing [2] on the original program $P_o$. In order to ensure that faults are not introduced during the production of the modified version $P_m$, regression testing has to be performed on $P_m$ which has a set of assertions $A_m$. Given $a_{ok} \in A_o$, $T(a_{ok})= \{t_{k1}, t_{k2}, t_{k3},\dots, t_{kr}\}$, and $a_{mi} \in V$, it has been shown in [7] that the old test suit, $T(a_{ok})$, may be used to revalidate assertion $a_{mi}$ during regression testing of the modified version $P_m$. Furthermore, it has been shown that using previous test suits to revalidate assertions may uncover faults in the modified version if these revalidated assertions were violated. Especially, faults for which assertions were originally designed to guard against in the original version of the program had these faults re-introduced in the modified version $P_m$ [7].

Although the regression testing method for programs with assertions [7] has succeeded in saving the time to develop new test cases through the utilization of previous test suites that was used during the initial testing of the program, this method still consider using *all* test cases found in the previous test suit. Therefore, this method may not perform well in the present of a large previous test suit with thousands of test cases. In this paper we propose a test case prioritizing method which uses fuzzy logic concepts to select only a subset of the previous test cases. The proposed method is described in Sec. III.

### B. Test Case Prioritization

The main goal of the prioritization techniques is to increase the probability of detecting faults at an earlier stage of testing [21-24]. Additionally, test case prioritization techniques objective is the utilization of previous test cases for the purpose of future testing. As stated in [21], there may exists several goals of test cases prioritization such as: (1) to increase test suites fault detection rate; (2) to minimize the time required to satisfy a testing coverage criterion; (3) to enhance tester's confidence in the reliability of the software in a shorter time period; (4) to be able to detect risky faults as early as possible; (5) to increase the chances of detecting faults related to software modification during regression testing.

 ISBN: 978-1-61208-237-0

In [21], an extensive study of nine different test case prioritization techniques was presented and compared according to their ability in fault detecting during regression testing. During that study a detection rate function is used to reorder test cases according to their ability to reveal program faults during regression testing. In [24], Extended Finite State Machine (EFSM) system model is proposed to be used instead of real programs to apply the same technique presented by [21] in order to reduce the cost of running test cases in real programs. Bryce et al. [22] presented a test prioritization model for Event-Driven software. This model concentrates on testing those parts related to the interface in GUI applications.

### C. Assertions Revalidation

To deal with assertions in modified programs during regression testing, an assertions revalidation model was proposed in [6]. This approach is based on data dependency analysis and program slicing. In that research an algorithm is presented which is based on the computation of a static slice [29-30], for each assertion found in both the original and the modified program. These program slices are then compared to decide which assertions are to be revalidated. Although this method is very useful in identifying assertions that need to be revalidated, new test cases to revalidate assertions are generated from scratch for each assertion. For industrial size programs with a possibly large number of assertions, this approach may be very expensive.

### D. Fuzzy Logic Background

In our daily life we use words and terms which are vague or fuzzy such as:

"The server is *slow*" or

"The weather is *hot*" or

"John is *tall*."

Fuzzy Logic concepts, e.g., [25-27], give us the ability to quantify and reason with words which have ambiguous meanings such the words (*slow*, *hot*, *tall*) mentioned above. In fuzzy sets [25], an object may belong partially to a set as opposed to classical or "crisp" sets in which an object may belong to a set or not. For example, in a universe of heights (in feet) for adult people defined as μ= {5, 5.5, 6, 6.5, 7, 7.5, 8}, a fuzzy subset TALL can be defined as follows:

TALL = [0/5, .125/5.5, .5/6, .875/6.5, 1/7, 1/7.5, 1/8].

In this example, the degree of membership for the members of the universe, μ, with respect to the set TALL may be interpreted as that the value "6" belongs to the set TALL 60% percent of the time while the value 8 belongs to the set TALL all the time.

## III. A Fuzzy Test Cases Prioritization Technique

In this paper, our objective is to prioritize test cases according to their relative rate to violate a given program assertion. Note that it has been shown in [2] that violating an assertion implies revealing a programming fault. Our proposed fuzzy logic model for prioritization test cases during regression testing of programs with assertions is described as follows. Given an original program $P_o$ and a modified version of this program $P_m$, let $A_o$= {$a_{o1}$, $a_{o2}$, $a_{o3}$, … $a_{on}$} be a set of assertions found in $P_o$ and $A_m$= {$a_{m1}$, $a_{m2}$, $a_{m3}$, … $a_{mz}$} be a set of assertions found in $P_m$. Assume that we are performing regression testing for the modified version $P_m$ using the regression testing method for programs with assertions as reported in [7]. Let $T_o$={$t_1$, $t_2$, $t_3$,…, $t_q$} be a previous test suite that was used during the process of assertion-oriented test data generation [2] of the original version $P_o$. For commercial software, testers usually deal with a very large number of test cases which make running all of them impractical. Therefore, given a set of test cases, our objective is to only reorder them according to some criterion that may convince us that some test cases may have better chances in violating a given assertion than the others. In this research, our criterion is the history of the test case during the process of testing the original program $P_o$.

Our problem is stated as follows. Given an assertion $a_{ok} \in A_o$ and $T(a_{ok})$= {$t_{k1}$, $t_{k2}$, $t_{k3}$,…, $t_{kr}$} as the test suite which were generated to explore assertion, $a_{ok,}$ during the application of assertion-oriented testing [2] on the original program $P_o$. Our goal is to measure the effectiveness of a given test case, $t_{kj} \in T(a_{ok})$, in violating a given program assertion $a_{mr} \in A_m$, during the regression testing process of the modified version, $P_m$. To solve this problem we propose a fuzzy logic test cases prioritization technique shown in Fig. 1. The following paragraph describes how the proposed approach works.

Let $t_{kj} \in T(a_{ok})$, be a test case which was used to explore assertion $a_{ok} \in A_o$ during the initial testing of a program $P_o$. To measure the effectiveness of $t_{kj}$ in violating the corresponding assertion $a_{mr} \in A_m$ in the modified version, $P_m$, during the process of regression testing the program $P_m$, we create a fuzzy set [25] called Effectiveness as follow. *Effectiveness* = {low, moderate, high}. Test cases related to any assertion $a_{ok} \in A_o$ where $a_{ok}$ belongs to the set "Affected" will have *low* effectiveness in exploring the corresponding assertion in the modified version of the program. Similarly, test cases related to any assertion $a_{ok} \in A_o$ where $a_{ok}$ belongs to the set "Partially Affected" will have *moderate* effectiveness in exploring the corresponding assertion in the modified version of the program. By the same token, test cases related to any assertion $a_{ok} \in A_o$ where $a_{ok}$ belongs to the set "Not Affected" will have *high* effectiveness in exploring the corresponding assertion in the modified version of the program.

$$S(x;\ \alpha, \beta, \gamma) = \begin{cases} 0 & for\ x \leq\ \alpha \\ 2\ \left(\frac{x-\alpha}{\gamma\ -\ \alpha}\right)^2 & for\ \alpha\ \leq x \leq \beta \\ 1-2\ \left(\frac{x-\gamma}{\gamma\ -\ \alpha}\right)^2 & for\ \beta \leq x \leq \gamma \\ 1 & for\ x \geq\ \gamma \end{cases}$$

Figure. 2. The S-function

 ISBN: 978-1-61208-237-0

In order to define the membership or grade values for each test case in the fuzzy set *Effectiveness*, we apply fuzzy logic techniques as follows. Each test case is assigned a membership depending on its "effectiveness," i.e., low, moderate or high. The membership value is in the interval [0,1] and reflects the compatibility of each specific test case to the fuzzy set *Effectiveness*. The assignment of membership values (grades) is based on the S-function [27] which is shown in Fig. 2. Note that other fuzzy clustering techniques other than the S-function may be used for the purpose of building up fuzzy sets and the assignment of membership functions. S-functions may be described as follows [27].

- A mathematical function that is used in fuzzy sets as a membership function.
- A simple but valuable tool in defining fuzzy functions such as the word "*tall*".
- The objects × are elements of some universe X. In this research, × represents the set of test cases we are dealing with during our prioritization mechanism, where these test cases are elements of the universe of the program possible input data.
- $\alpha$, $\beta$, and $\gamma$ are parameters which may be adjusted to fit the desired membership data. The parameter $\alpha$ represents the minimum boundary and $\gamma$ represents the maximum boundary. The parameter $\beta$ is the middle point between $\alpha$ and $\gamma$ and is computed as $(\alpha + \gamma) / 2$.
- Depending on the application, a membership function may be controlled from different sources [27]. For example, in an expert system, the membership function will be constructed based on the experts' opinion modeled by the system.
- In this research, values of the parameters $\alpha$ and $\gamma$ are determined after extermination with the proposed approach. As described previously, the history of each test case will be monitored during this experiment with regard to the ability of this specific test case in violating a given assertion in the program under test.

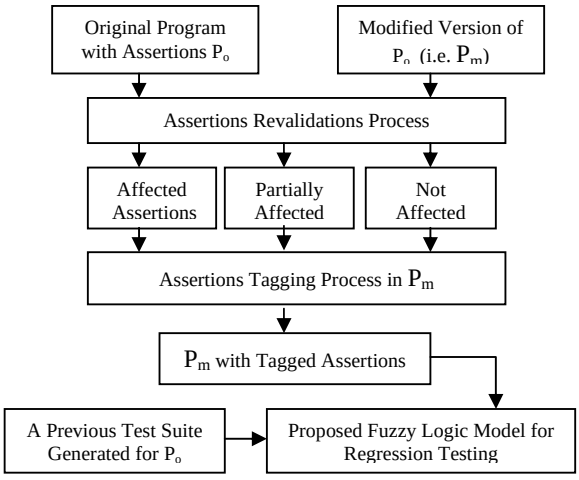



Figure 1. Fuzzy Regression Testing Model for Programs with Assertions

The model shown in Fig. 1 may be described as follows. First, we analyze both $P_o$ and $P_m$ in order to classify assertions, $A_m$, found in $P_m$ with respect to how much the modifications inflected on $P_m$ had affected those assertions. To perform this analysis, we use assertions revalidations model [6] to classify the set of assertions, $A_m$, found in $P_m$ into three different sets: "Affected," "Partially Affected" and "Not Affected." Based on the categorization of assertions in the analysis's step, the next step is to categorize test cases according to their expected effectiveness during regression testing of the modified, version of the program, i.e., $P_m$. Because the "effectiveness" of a test case is a "fuzzy" term which is very hard to measure in crisp value, we propose using fuzzy logic techniques to deal with measuring the effectiveness of a given test case as described previously.

## IV. CONCLUSION and FUTURE WORK

In this paper, we presented a new technique for test cases prioritization to be used during regression testing of programs with assertions. The proposed model employs fuzzy logic concepts to measure the *effectiveness* of a given test case in violating programs assertions during the regression testing of a modified program. Our proposed method builds upon the concepts of previous research in the fields of assertions-based software testing and assertions revalidation. Furthermore, the proposed method is intended to be used in conjunction with traditional black-box and white-box software testing methods. In order to evaluate the proposed method, we intend to perform and extensive experimental study using a variety of programs with assertions. The results of this experiment will then be compared with existing test case prioritizations techniques reported in the literature.

## REFERENCES

[1] Rosenblum, D., "Toward A Method of Programming WithAssertions," Proceedings of the International Conference on Software Engineering, pp. 92-104, 1992.

[2] Korel B. and Al-Yami A., "Assertion-Oriented Automated Test Data Generation," Proc. 18th Intern. Conference on Software Eng., Berlin, Germany, pp. 71-80, 1996.

[3] Alakeel A., "An Algorithm for Efficient Assertions-Based test Data Generation," Journal of Software, vol. 5, no. 6, pp. 644-653, 2010.

[4] Alakeel A. and Mahashi M., "Using Assertion-Based Testing in String Search Algorithms," Proceedings of The Third Int. Conf. on Advances in System Testing and Validation Lifecycle, Barcelona, Spain, pp. 1-5, 2011.

[5] Alakeel A., "A Framework for Concurrent Assertion-Based Automated Test Data Generation," European Journal of Scientific Research, vol. 46, no. 3, pp. 352-362, 2010.

[6] Korel B. , Zhang Q., and Tao L., "Assertion-Based Validation of Modified Programs," Proc. 2009 2nd Int'l Conference on Software Testing, Verification and Validation, Denver, USA, pp. 426-435, 2009.

[7] Alakeel A., "Regression Testing Method for Programs with Assertions," American Journal of Scientific Research, no. 11, pp. 111-122, 2010.

 ISBN: 978-1-61208-237-0 

[8] Hetzel W. and Hetzel B., "The Complete Guide to Software Testing," John Wiley & Sons, Inc., New York, NY, 1991.

[9] Beydeda S. and Gruhn V.,"An Integrated Testing Technique for Component-Based Software," Proc. ACS/IEEE Int'l Conference on Computer Systems and Applications, pp. 328-334, 2001.

[10] Tsai W., Bai X., Paul R., and Yu L., "Scenario-Based Functional Regression Testing," Proc. IEEE Int'l Conference on Sofware and Applications, pp. 496-501, 2001.

[11] Korel B., Tahat L., and Vaysburg B., "Model Based Regression Test Reduction Using Dependence Analysis," Proc. IEEE Int'l Conferance on Software Maintenance, pp. 214-233, 2002.

[12] Chen Y., Rosenblum D., and Vo K., "Testtube: System for Selective Regression Testing," Proc. IEEE. Int'l Conference on Software Engineering, pp. 211-220, 1994.

[13] Gupta R., Harrold M., and Soffa M., "An Approach to Regression Testing Using Slices," Proc. IEEE Int'l Conference on Software Maintenance, pp. 299-308, 1992.

[14] Korel B. and Al-Yami A., "Automated Regression Test Generation," Proc. ACM Int'l Symposium on Software Testing and Analysis, pp. 143-152, 1998.

[15] Rothermel G. and Harrold M., "A Safe, Efficient Regression Test Selection Technique," ACM Tran. on Software Eng. and Methodology, vol. 6, no. 2, pp.63-68, 1996.

[16] Beizer B., "Software System Testing and Quality Assurance," Thomson Computer Press, 1996.

[17] Rothermel G. and Harrold M., "Selecting Tests and Identifying Test Coverage Requirements for Modified Software," Proc. IEEE Int'l Conference on Software Maintenance, pp. 358-367, 1994.

[18] Masri W., Podgurski A., and Leon D., "An Empirical Study of Test Case Filtering Techniques Based on Exercising Information Flows," IEEE Trans. Software Eng., vol. 33, no. 7, pp. 454-477, 2007.

[19] Logyall J., Mathisen S. , Hurley P., and Williamson J., "Automated Maintenance of Avionics Software," Proc. IEEE Aerospace and Electronics Conference, pp. 508-514,1993.

[20] Tsai W., Bai X., Paul R., and Yu L., "Scenario-Based Functional Regression Testing," Proc. IEEE Int'l Conference on Sofware and Applications, pp. 496-501, 2001.

[21] Rothermel G., Untch R., Chu C., and Harrold M., "Prioritizing Test Cases for Regression Testing," IEEE Trans. Software Eng., vol. 27, no. 10, pp. 929-948, 2001.

[22] Bryce C, Sampath S., and Memon A., "Develping a Single Model and Test Prioritization Strategies for Event-Driven Software," IEEE Trans. Software Eng., vol. 37, no. 1, pp. 48-64, 2010.

[23] Korel B., Koutsogiannakis G., and Tahat L., "Application of System Models in Regression Test Suite Prioritization," Proc. IEEE Int'l Conference on Software Maintenance, pp. 247-256, 2008.

[24] Korel, B., Tahat L., and Harman M., "Test Prioritization Using System Models," Proc. IEEE Int'l Conference on Software Maintenance, pp. 559-568, 2005.

[25] Zadeh L., "Fuzzy Sets," Information and Control, no. 8, pp. 338-353, 1965.

[26] Kosko B., "Neural Networks and Fuzzy Systems: A Dynamical Systems Approach to Machine Intelligence," Prentice-Hall, Englewood Cliffs, NJ, 1992.

[27] Giarratano J., "*Expert Systems: Principles and Programming*," PWS-KENT Publishing Company, Boston, 1989.

[28] Xu Z., Gao K., and Khoshgoftaar T., "Application of Fuzzy Expert System in Test Case Selection for System Regression Test," IEEE International Conference on Information Reuse and Integrationon , pp. 120-125, 2005.

[29] Horowitz S., Reps. T., and Binkley D., "Interprocedural Slicing using Dependence Graphs," ACM Trasn. Programming Languages and Systems, vol. 12, no. 1, pp. 26-60, 1990.

[30] Weiser M., 1984, "Program Slicing," IEEE Trans. Software Engineering, vol. 10, no. 4, pp. 352-357, 1984.

 ISBN: 978-1-61208-237-0 

# Computer-aided Investigation of Mechanical Properties for Integrated Casting and Rolling Processes Using Hybrid Numerical-analytical model of Mushy Steel deformation

Miroslaw Glowacki
Dept. of Applied Computer Science and Modelling
AGH-University of Science and Technology
A. Mickiewicza Av. 30, 30-059 Kraków, Poland
e-mail: glowacki@metal.agh.edu.pl

Marcin Hojny
Dept. of Applied Computer Science and Modelling
AGH-University of Science and Technology
A. Mickiewicza Av. 30, 30-059 Kraków, Poland
e-mail: mhojny@metal.agh.edu.pl

***Abstract*—The main subject of the current paper is investigation of yield stress for C45 grade steel as well as development of a new methodology of such investigation. The method requires high accuracy model of semi-solid steel deformation. Hence, it requires a dedicated hybrid analytical-numerical model of deformation of steel with variable density. The newly developed methodology allows to compute curves depending on both temperature and strain rate. The experimental work has been done in Institute for Ferrous Metallurgy in Gliwice Poland using Gleeble thermo-mechanical simulator with high temperature testing equipment. A number of compression and tension tests have been done in order to verify the predictive ability of both the developed model and new, modified testing methodology. The comparison between numerical and experimental results is the supplementary subject of the presented paper, as well. The developed methodology allows reliable numerical simulation of deformation of semi-solid steel samples and calculation of realistic flow curve parameters.**

***Keywords-yield stress; semi-solid steel testing; extra-high deformation temperature; numerical analysis; inverse method***

## I. INTRODUCTION

Due to the global energy crisis in recent years, more and more new production technologies require energy preservation and environmental protection. The integrated casting and rolling technologies are newest efficient and very profitable ways of hot strip production. Only few companies all over the world are able to manage such processes. The technical staff of a plant located in Cremona, Italy is working on new methods of flat steel manufacturing for several years now. The ISP (Inline Strip Production) and AST (Arvedi Steel Technologies) technologies which are developed in Cremona are distinguished by very high rolling temperature. The main benefits of both the methods are related to very low rolling forces and favourable field of temperature. However, certain problems particular to such metal treatment arise. The central parts of slabs are mushy and the solidification is not yet finished while the deformation is in progress. This results in changes in material density and occurrence of characteristic temperatures having great influence on the plastic behaviour of the material [1, 2]. The nil strength temperature (NST), strength recovery temperature (SRT), nil ductility temperature (NDT) and ductility recovery temperature (DRT) have effect on steel plastic behaviour and limit plastic deformation. The nil strength temperature (NST) is the temperature level at which material strength drops to zero while the steel is being heated above the solidus temperature. Another temperature associated with NST is the strength recovery temperature (SRT). At this temperature the cooled material regains strength greater than 0.5 $N/mm^2$. Nil ductility temperature (NDT) represents the temperature at which the heated steel loses its ductility. The ductility recovery temperature (DRT) is the temperature at which the ductility of the material (characterised by reduction of area) reaches 5% while it is being cooled. Over this temperature the plastic deformation is not allowed at any stress tensor configuration.

The most important steel property having crucial influence on metal flow paths is the yield stress. In the literature of the past years, one can find papers regarding experimental results [3] and modelling [4] of non-ferrous metals. Both the mentioned contributions focus mainly on tixotrophy. The first results regarding steel deformation at extra high temperature were only presented in the past few years [5]. This stems from the fact that level of liquidus and solidus temperatures of steel is very high in comparison with non-ferrous metals. It causes serious experimental problems contrary to deformation tests for non-ferrous metals which are much easier. Rising abilities of thermo-mechanical simulators such as the Gleeble and development of new methods of identification of mechanical properties allow investigations leading to strain-stress relationships for semi-solid steels, as well. This problem became a subject of research done by authors of the presented paper for several years now. As a result a computer system supporting the investigation of mushy steel has been developed [6]. The current paper presents the modified methodology, which allows the calculation of the real stress-strain relationships for a wide range of temperatures and strain rate.

 ISBN: 978-1-61208-237-0 

In the second section of the paper the modified methodology of stress-strain curves is described, the second one presents a mathematical model, in the last section an example results are presented. Finally, the main conclusions and future work are shortly described.

## II. THE METHODOLOGY

Due to several serious experimental problems a special technique of testing was developed for temperatures higher than NDT. The deformation process has been divided into two main stages. The first one – a very small preliminary compression and the second one – the ultimate compression. The preliminary deformation is meant to eliminate clearances in the testing equipment. Well known Voce formula [6] has been applied to describe the yield stress functions. It takes account of a number of coefficients which are calculated using inverse analysis. This is the only acceptable method of interpretation of compression of semi-solid steel testing results. Due to strong temperature and strain inhomogeneity observed during compression of semi-solid samples the deformation causes significant barrelling of their central parts. Improvement of experimental methodology can help only to a small extent. Contrary to the old version of the method published in [6] the newly developed one allows the computation of curves depending on both temperature and strain rate. The tension tests has been replaced by compression ones and the Voce formula was replaced by more adequate equation. Fig. 1 schematically presents the modified methodology. More details concerning the experimental work were published in [7].

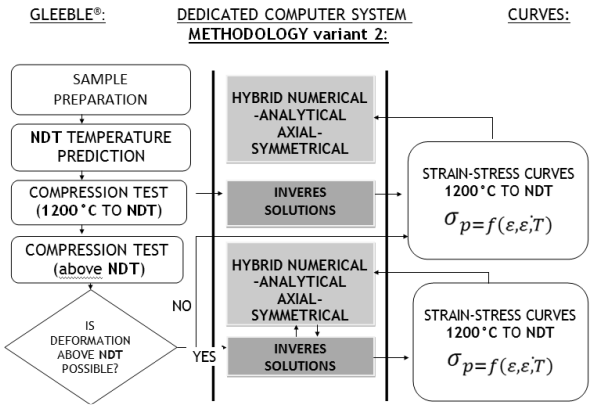



Figure 1. Flowchart of the integrated testing methodology of flow stress investigation of semi-solid steel.

The presented approach allows to compute realistic yield stress curves depending on strain, temperature and strain rate in temperature range from 1200°C to NDT and above. The objective function of the inverse analysis was defined as a square root error of discrepancies between calculated ($F_c$) and measured ($F_m$) loads at several subsequent steps of the compression process. The experimental values of the deformation forces were collected by the Gleeble equipment while the theoretical ones ($F_c$) were calculated with the help of a sophisticated solver facilitating accurate computation of strain, stress and temperature fields for materials with variable density. This solver is the least visible but the most powerful part of the computer aided testing system developed by the authors called Def_Semi_Solid (Fig. 2). The heart of the solver is based on a hybrid analytical-numerical mushy steel deformation model described in the next section.

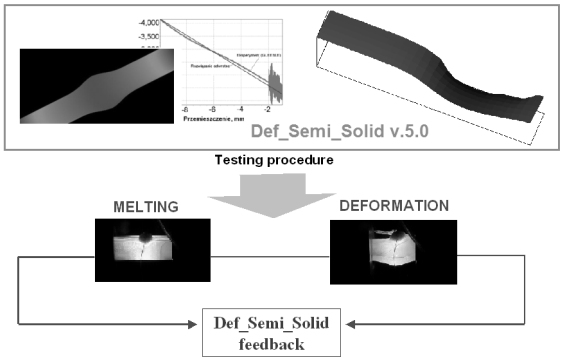



Figure 2. The Def_Semi_Solid system as a feedback unit with Gleeble simulator.

## III. MATHEMATICAL MODEL

A mathematical model of the compression process has been developed using the theory of plastic flow. The principle of the upper assessment, calculus of variations, approximation theory, optimization and numerical methods for solving partial differential equations were used [8]. The following assumptions were established:

- Deformation and stress state are axial-symmetrical;
- Deformed material is isotropic but inhomogeneous;
- The material behaviour is rigid-plastic – the relationship between the stress tensor and strain rate tensor is calculated according to the Levy-Mises flow law [8], which is given as:

$$\sigma_{ij} - \frac{1}{3}\sigma_{kk}\delta_{ij} = \frac{2}{3}\frac{\sigma_p}{\dot{\varepsilon}_i}\dot{\varepsilon}_{ij} \quad (1)$$

Rigid-plastic model was selected due to its very good accuracy at the strain field during the hot deformation and sufficient correctness of calculated deviatoric part of the stress field. Moreover, the elastic part of each stress tensor component is very low at temperatures close to solidus line and can in practice be neglected in calculations of strain distribution. The limits for plastic metal behaviour are defined according to Huber-Mises-Hencky yield criterion:

$$\sigma_{ij}\sigma_{ij} = 2\left(\frac{\sigma_p}{\sqrt{3}}\right)^2 \quad (2)$$

In (1) and (2), $\sigma_{ij}$ denotes the stress tensor components, $\sigma_{kk}$ represents the mean stress, $\delta_{ij}$ is the Kronecker delta [8],

 ISBN: 978-1-61208-237-0 

$\sigma_p$ indicates the yield stress, $\dot{\varepsilon}_i$ is the effective strain rate, and $\dot{\varepsilon}_{ij}$ denotes strain rate tensor components. The components are given by an equation:

$$\dot{\varepsilon}_{ij} = \frac{1}{2}\left(\nabla_i v_j + \nabla_j v_i\right) \tag{3}$$

In cylindrical coordinate system $Or\theta z$ the solution is a vector velocity field defined by the distribution of three coordinates $\boldsymbol{v} = (v_r, v_\theta, v_z)$ . The field is a result of optimization of a power functional, which can be written in general form as the sum of power necessary to run the main physical phenomena related to plastic deformation. Due to the axial-symmetry of the sample the circumferential component of the velocity field can be neglected and the functional is usually formulated as:

$$J[\boldsymbol{v}] = \dot{W} = \dot{W}_\sigma + \dot{W}_\lambda + \dot{W}_f \tag{4}$$

Component $\dot{W}_\sigma$ occurring in (4) represents the plastic deformation power, $\dot{W}_\lambda$ is the power which is a penalty for the departure from mass conservation condition, $\dot{W}_f$ denotes the friction power and $\boldsymbol{v} = (v_r, v_z)$ describes the reduced velocity field distribution.

Rigid-plastic formulation of metal deformation problem requires the condition of mass conservation in the deformation zone. In case of solids and liquids with a constant density, this condition can be simplified to the incompressibility condition. Such a condition is generally satisfied with sufficient accuracy during the optimization of functional (4). In most solutions a slight, but noticeable loss of volume is observed. The loss occurs because the incompressibility condition imposed on the solution is not completely satisfied in numerical form. It is negligible in case of traditional computer simulation of deformation processes although in some embodiments more accurate methods are used to restore the volume of metal subjected to the deformation. In contrast, in the presented case the density of semi-solid materials varies during the deformation process and these changes result in a physically significant change in the volume of a body having constant mass. The size of the volume loss due to numerical errors is comparable with changes caused by fluctuation in the density of the material.

A further problem specific to the variable density continuum is power $\dot{W}_\lambda$, which occurs in functional (4). It is used in most solutions and has a significant share of total power. Even when the iterative process approaches the end, this power component is still significant, especially if the convergence of the optimization procedures is insufficient. In case of discretization of the deformation area (e.g. using the finite element method) if one focuses solely on the $\dot{W}_\lambda$ a number of possible locally optimal solutions appear. They are related to a number of possible directions of movement of discretization nodes providing the volume preservation of the deformation zone. Each of these solutions creates a local optimum for $\dot{W}_\lambda$ power and thus for the entire functional (4). This makes it difficult to optimize because of lack of uniform direction of fall of total power which leads to global optimum. The material density fluctuation causes further optimization difficulties, resulting from additional replacement of incompressibility condition with a full condition of mass conservation.

The proposed solution requires high accuracy in ensuring the incompressibility condition for the solid material or mass conservation condition for the semi-solid areas. This stems from the fact that the errors resulting from the breach of these conditions can be treated as a volume change caused by the steel density variation in the semi-solid zone. High accuracy solution is required also due to large differences in yield stress for the individual subareas of the deformation zone. In the discussed temperature range they appear due to even slight fluctuations in temperature. In presented solution the second component of functional (4) is left out and mass conservation condition is given in analytical form constraining the radial $(v_r)$ and longitudinal $(v_z)$ velocity field components. The functional takes the following shape:

$$J[\boldsymbol{v}] = \dot{W}_\sigma + \dot{W}_t \tag{5}$$

In case of functional (5), the numerical optimisation procedure converges faster than the one for functional (4) due to the reduced number of velocity field parameters (only radial components are optimisation parameters) and the lack of numerical form of mass conservation condition. The accuracy of the proposed hybrid solution is higher also due to negligible volume loss caused by numerical errors, which is very important for materials with variable density.

As mentioned before, the solution of the problem is a velocity field in cylindrical coordinate system in axial-symmetrical state of deformation. Optimization of metal flow velocity field in the deformation zone of semi-variational problem requires the formulation according to equation (5). The radial velocity distribution $v_r(r, \theta, z)$ and the longitudinal one $v_z(r, \theta, z)$ are so complex that such wording in the global coordinate system poses considerable difficulties. These difficulties are the result of the mutual dependence of these velocities. Therefore the basic formulation will be written for the local cylindrical coordinate system $Or\theta z$ with a view to the future discretization of deformation area using one of the dedicated methods. In addition one will find that the deformation of cylindrical samples is characterized by axial symmetry. As demonstrated by experimental studies conducted using semi-solid samples the symmetry may be disturbed only as a result of unexpected leakage of liquid phase.

Such experiments, however, are regarded as unsuccessful and not subject to numerical analysis. Establishment of the axial symmetry, which except in cases of physical instability can be considered valid also for the process of compression or tensile test of semi-solid samples, allows one to simplify the model because of the identical strain distribution at any axial sample cross-section. Considerations will therefore be carried out in $Orz$ coordinates for the sample cross-section using one of the planes containing the sample axis. Components of power functional given by (5) have been formulated in accordance with the general theory of

 ISBN: 978-1-61208-237-0 

plasticity by relevant equations. The plastic power for the deformation zone having volume of $V$ is given by the subsequent relation:

$$\dot{W}_\sigma = \int_V \sigma_i \dot{\varepsilon}_i \, dV \tag{6}$$

where $\sigma_i$ is the effective stress and $\dot{\varepsilon}_i$ denotes the effective strain. The plastic deformation starts when the rising effective stress reaches yield stress limit $\sigma_p$ ($\sigma_i = \sigma_p$) according to yield criterion given by equation (2). Effective strain occurring in (6) is calculated on the basis of the strain tensor components $\dot{\varepsilon}_{ij}$ according to following relationship:

$$\dot{\varepsilon}_i = \sqrt{\frac{2}{3}\dot{\varepsilon}_{ij}\dot{\varepsilon}_{ij}} \tag{7}$$

The components are given by (3). The second component of functional (5) is responding for friction. To compute friction power on the boundary $S$ of area $V$ a model given by the subsequent equation was used:

$$\dot{W}_t = \int_S m \frac{\sigma_p}{\sqrt{3}} \|\overline{\boldsymbol{v}}\| dS \tag{8}$$

In (8), $m$ is the so called friction factor, which is usually experimentally selected and $\overline{\boldsymbol{v}}$ is a relative velocity vector of metal and tool $\overline{\boldsymbol{v}} = \boldsymbol{v} - \boldsymbol{v_t}$. In case of tensile test the samples are permanently fixed in jaws of a physical simulator and friction must not be taken into account. However, compression test requires sharing the friction power which is significant.

For the solid zones the incompressibility condition can be described by universal operator equation independently of the mechanical state of the deformation process:

$$\nabla \mathbf{v} = 0 \tag{9}$$

Because the semi-solid zone is characterized by density change due to still ongoing progress of steel solidification, the condition of incompressibility is inadequate to reflect changes and was replaced with the mass conservation condition, which describes the following modified operational equation:

$$\nabla \mathbf{v} - \frac{1}{\rho}\frac{\partial \rho}{\partial t} = 0 \tag{10}$$

The basis for the optimization of functional (5) is the velocity field determined by appropriate system of velocity functions in the concerned area. These functions are then the source of deformation field and other physical quantities affecting the power functional formulation. Obtaining an accurate real velocity field requires the use of velocity functions depending on a number of variational parameters. The functions should be flexible enough to map the field throughout the whole volume of the deformation zone. Analytical description of each component of the velocity field with a single function in the whole area of deformation is not preferred. This approach creates difficulties especially in areas not subjected to the deformation where the velocity function should remain constant. Therefore, the solution to the problem of semi-solid metal flow was based on a specific method.

In the case of deformation of axial-symmetrical bodies, the incompressibility condition is given by following differential equation:

$$\frac{\partial v_r}{\partial r} + \frac{v_r}{r} + \frac{\partial v_z}{\partial z} = 0 \tag{11}$$

For the semi-solid area, (11) is replaced by the mass conservation condition due to existing density changes. The longitudinal velocity has been calculated as an analytical function of radial velocity using this condition. In cylindrical coordinate system the condition has been described with an equation:

$$\frac{\partial v_r}{\partial r} + \frac{v_r}{r} + \frac{\partial v_z}{\partial z} - \frac{1}{\rho}\frac{\partial \rho}{\partial \tau} = 0 \tag{12}$$

Equation (11) is a special case of (12) and therefore the proposed solution will consider the dependence (12) as more general. In (12) $\rho$ is the temporary material density and $\tau$ is the time variable. The proposed variational formulation makes the longitudinal velocity dependent on the radial one. Condition (12) allows for the calculation of $\partial v_z/\partial z$ derivative as a function of $\partial v_r/\partial r$ after analytical differentiation of radial velocity distribution function $v_r(r,z)$. Hence, the longitudinal velocity is calculated as a result of analytical integration according to following equation:

$$v_z = -\int \left(\frac{\partial v_r}{\partial r} + \frac{v_r}{r} - \frac{1}{\rho}\frac{\partial \rho}{\partial \tau}\right) dz \tag{13}$$

In this case, the velocity field depends only on one function – the radial velocity distribution.

Heat exchange between solid metal and environment, and its flow inside the metal is controlled by a number of factors. During phase transformation two additional phenomena have to be taken into account. Note that in the process of deformation of steel at temperature of liquid to solid phase transformation there are two sources of heat changes. On the one hand heat is generated due to the state transformation. On the other hand it is secreted as a result of plastic deformation. In addition, steel density variations also cause changes of body temperature.

Thermal solution has a major impact on simulation results, since the temperature has strong effect on remaining variables. This is especially evident if the specimen temperature is close to solidus line when the body consist of both solid and semi-solid regions. In such case the affected phenomena are: plastic flow of solid and mushy materials, stress evolution and density changes. The theoretical temperature field is a solution of Fourier-Kirchhoff equation with appropriate boundary conditions.

 ISBN: 978-1-61208-237-0 

The most general form of the Fourier-Kirchhoff equation in any coordinate system can be written in operator form as follows:

$$\nabla^T(\mathbf{\Lambda}\,\nabla T) + Q = c_p\rho\left(\boldsymbol{v}^{\boldsymbol{T}}\nabla T + \frac{\partial T}{\partial \tau}\right) \qquad (14)$$

where $T$ is the temperature distribution in the controlled volume and $\mathbf{\Lambda}$ denotes the symmetrical second order tensor called heat transformation tensor. In case of thermal inhomogeneity the whole tensor has to be considered. $Q$ represents the rate of heat generation (or consumption) due to the phase transformation, due to plastic work done and due to electric current flow (resistance heating of the sample is usually applied). Finally $c_p$ describes the specific heat, $\rho$ the steel density, $\boldsymbol{v}$ the velocity vector of specimen particles and $\tau$ the elapsed time.

For axial-symmetrical case, (14) can be simplified. The following form of Fourier-Kirchhoff equations for isotropic, axially-symmetric heat flow was applied in the presented solution:

$$\lambda\left(\frac{\partial^2 T}{\partial r^2} + \frac{1}{r}\frac{\partial T}{\partial r} + \frac{\partial^2 T}{\partial z^2}\right) + Q = \rho c_p \frac{\partial T}{\partial \tau} \qquad (15)$$

Equation (15) needs to be solved with appropriate initial and boundary conditions. Combined Hankel's boundary conditions have been adopted for the presented model.

## IV. EXAMPLE RESULTS

The experimental work was done in Institute for Ferrous Metallurgy in Gliwice, Poland using Gleeble thermo-mechanical simulator. The steel used for the experiments was the C45 grade steel having 0.45% of carbon content. In all cases, experiments were performed according to the following schedule:

- initial stage: sample preparation divided into several sub stages (e.g., thermocouple assembly); die selection, etc,
- stage 2: melting procedure,
- stage 3: deformation process.

It is good practice to test materials in isothermal conditions [9]. Unfortunately, this is not possible for semi-solid steel. Nevertheless, the condition should be as close to isothermal as possible due to the very high sensitivity of material rheology to even small variations of temperature. The basic reason for uneven temperature distribution inside the sample body on the Gleeble simulator is the contact with hot copper handles presented in Fig. 3.

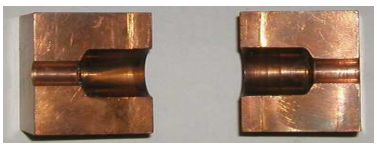


Figure 3. The short contact zone handles used in experiments (hot handle).

The estimated liquidus and solidus temperature levels of the investigated steel are: 1495°C and 1410°C, respectively. Thermal solution of the theoretical model has crucial influence on simulation results, since the temperature has strong effect on remaining parameters. The resistance sample heating and contact of the sample with cold cooper handles cause non-uniform distribution of temperature inside heated material, especially along the sample. The semi-solid conditions in central parts of the sample cause even greater temperature gradient due to latent heat of transformation. Such non-uniform temperature distribution is the source of significant differences in the microstructure and hence in material rheological properties.

During the experiments samples were heated to 1430°C and after maintaining at constant temperature were cooled down to the required deformation temperature. In case of heating the heat generated is usually not known because the Gleeble equipment uses an adaptive procedure for resistive heating controlled by temperature instead of current flow. Hence, the actual heat generated by current flow (in fact the rate of heat generation $Q$) has to be calculated using inverse procedure. In this case the objective function ($F$) was defined as a norm of discrepancies between calculated ($T_c$) and measured ($T_m$) temperatures at a checkpoint (steering thermocouple position: TC4 in Fig. 4) according to the following equation:

$$F(\mathrm{Q}) = \int_{\tau_0}^{\tau_i}[T_c(Q,r,z,T) - T_m(r,z,T)]d\tau \qquad (16)$$

where: where $\tau$ is the time variable, $Q$ is the rate of heat generation.

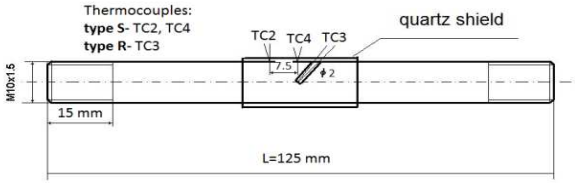



Figure 4. Samples used for the experiments. TC2, TC3 and TC4 thermocouples.

In the final stage of physical test, the temperature difference between core of the sample (TC3 thermocouple position) and its surface (TC4 thermocouple position) can be significant. In all cases the core temperature was higher than surface temperature. Differences between these two reach around 30°C for cold handle (handle with long contact zone) and about 40°C for hot handle. The results of numerical simulation are in agreement with experiments. Fig. 5 presents the temperature distributions in the cross section of sample tested at 1380°C right before deformation (variant with hot handle). One can observe major temperature gradient between die-sample contact surface. However, difference between experimental and theoretical core temperatures for hot handles was only 3°C (calculated core

temperature was equal to 1417°C and measured one was equal to 1420°C).

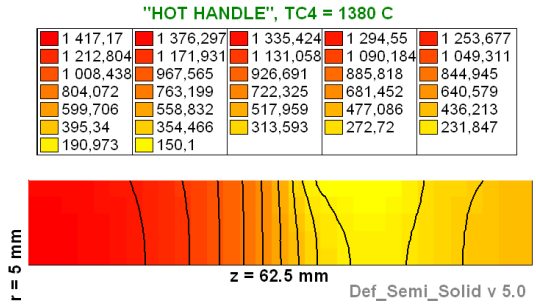



Figure 5. Distribution of temperature in the cross section of sample tested at temperature 1380°C right before deformation (variant with hot handle).

The micro and macrostructure of the tested samples were investigated as well. Fig. 6 shows microstructure right before deformation for both central and boundary regions of the heating zone. Microstructure of the cooled samples consists of pearlite (the darkest phase), bainite (grey phase mainly near the borders of grains) and the bright ferrite. This is a result of phase composition, wide melting zone and almost two times lower rate of cooling of central parts of the sample (in the case of hot handles).

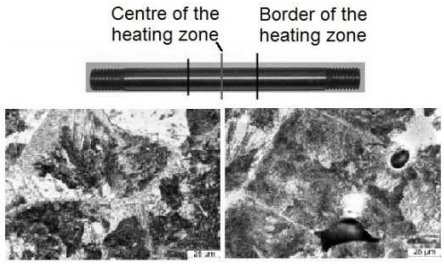



Figure 6. Microstructure of the central and boundary regions of sample right before deformation. Variant with cold handle. Magnification: 400x.

Fig. 7 shows macrostructure of the central part of cross-sections of samples right before deformation. Liquid phase particles were observed. Experimental and numerical results can be compared taking into consideration the temperature gradient within the sample. This shows that the mathematical model of resistance heating is consistent with the experimental data.

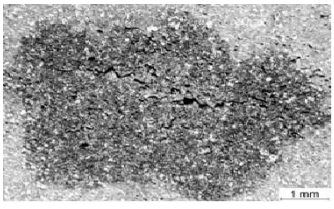


Figure 7. Macrostructure of the sample central part right before deformation. Variant with cold handle. Magnification: 10x.

Compression and tension tests were performed, according to the given methodology. During experiments die displacement, force and temperature changes in the deformation zone were recorded. The computer simulations were performed as well. All series of tests and computer simulations were done using long contact zone between samples and simulator jaws (cold handle). The deformation zone had the initial height of 62.5 mm. The sample diameter was 10 mm. The samples were melted at 1430°C, and then cooled to deformation temperature. During the tests each sample was subjected to 10 mm reduction of height. Results of each test were used for inverse analysis to compute yield stress curve parameters. Fig. 8 shows strain-stress curves at several strain rate levels for temperature 1300°C. The relationships were calculated using presented experimental methodology.

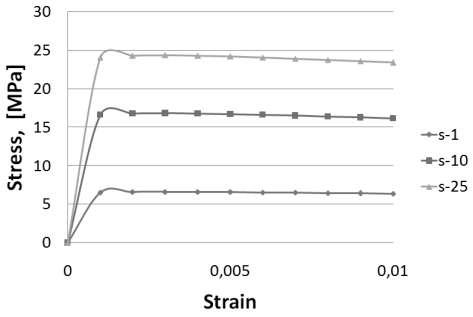



Figure 8. Stress-strain curves at several strain rate levels for temperature 1300°C.

Comparison between the calculated and measured loads are presented in Fig. 9, showing quite good agreement.

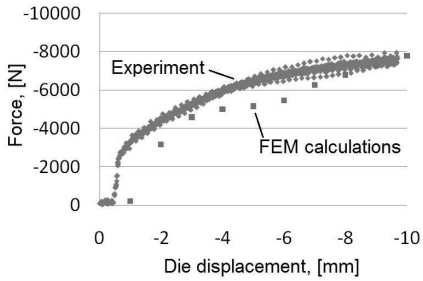



Figure 9. Comparison between measured and predicted loads at temperature 1380°C (new methodology).

Shape of a sample after experiment at 1300°C is presented in Fig. 10.

Figure 10. Example final shape of the sample after deformation at 1300oC.

Comparison between the measured and calculated maximal diameters of samples allow rough verification of

the developed computer aided experimental methodology. Results of such comparison are presented in Table 1. The table shows results for samples which have been subjected to deformation at several levels of temperature, i.e., 1300°C, 1350°C and 1380°C. Good agreement between the real diameter and its calculated value is observed. The relative mean square error between both the values is equal to 2.76%.

TABLE I. COMPARISION OF THE MEASURED AND CALCULATED MAXIMAL DIAMETERS OF SAMPLES DEFORMED AT DIFFERENT TEMPERATURE

| Test | Experiment | Simulations |
|---|---|---|
| 1300°C | 15.3 mm | 15.6 mm |
| 1350°C | 15.3 mm | 15 mm |
| 1380°C | 17.8 mm | 18.2 mm |
| **Relative mean square error: 2.76%** | | |

Fig. 11 compares measured loads with those computed using the methodology previously used by the authors. One can see that the mean square error in this case is significantly greater than its equivalent for the newly developed method.

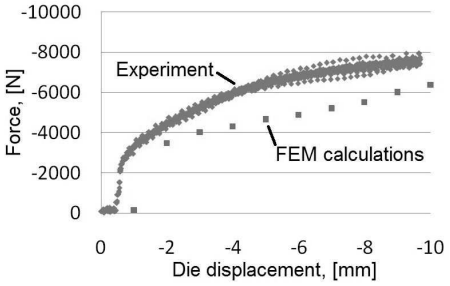



Figure 11. Comparison between measured and predicted loads at temperature 1380°C (old methodology).

The main reason for that is the lack of strain rate dependency of yield stress in the old model. The results obtained taking into account the strain rate as a parameter of the flow curve are more accurate for temperatures exceeding the NDT level.

## V. CONCLUSION AND FUTURE WORK

The investigation reported in the current paper has shown, that temperature distribution inside the controlled semi-solid volume is strongly heterogeneous and non-uniform. Axial-symmetrical model does not take into account all the physical phenomena accompanying the deformation. Finally, the error of the predicted strain-stress curves can still be improved. The proposed solution of the presented problem is application of both fully three-dimensional solution and more adequate solidification model taking into consideration evolution of forming steel microstructure. Therefore, the study of multiscale modelling of mechanical properties is the main target of the future work. Contrary to the current model the new approach should allow to better capture the physical principles of semi-solid steel deformation in micro-scale. Additionally, the methodology should allow to transfer the characteristics of the material behaviour between the micro- and macro-scale. As a consequence the final results should be more precise and accurate. Modelling of deformation of steel samples at extra-high temperatures involves a number of issues. One of them is the difficulty of calculating of thermal and mechanical material properties. Another most important problem is the right interpretation of the results of compression tests that provide data for flow stress calculation. The presented testing methodology allows reliable numerical simulation of deformation of semi-solid steel samples and calculation of realistic flow curve parameters. The presented research was focused on mechanical properties of investigated semi-solid steel. Compression tests carried out for semi-solid materials could only be interpreted using inverse analysis. Temperature strain and strain rate as a parameters of the flow curve provide accurate results of computer simulation of semi-solid steel behaviour.

## ACKNOWLEDGMENT

The work has been supported by the Polish Ministry of Science and Higher Education Grant N N508 585539

## REFERENCES

[1] D. Senk, F. Hagemann, B. Hammer, and R. Kopp, “Umformen und Kühlen von direkt gegossenem,” Stahlband, Stahl und Eisen, vol. 120, 2000, pp. 65-69.

[2] H.G. Suzuki and S. Nishimura, “Physical simulation of the continuous casting of steels,” Proceedings of Physical Simulation of Welding, Hot Forming and Continuous Casting, Canmet Canada, May 2-4, 1988, pp. 166-191.

[3] R. Kopp, J. Choi, and D. Neudenberger, "Simple compression test and simulation of an Sn–15% Pb alloy in the semi-solid state," J. Mater. Proc. Technol.,vol. 135, 2003, pp. 317-323, doi: 10.1016/S0924-0136(02)00863-4.

[4] M. Modigell, L. Pape, and M. Hufschmidt,"The Rheological Behaviour of Metallic Suspensions," Steel Research Int., vol. 75, 2004, pp. 506–512.

[5] Y.L. Jing, S. Sumio, and Y. Jun, "Microstructural evolution and flow stress of semi-solid type 304 stainless steel," J. Mater. Proc. Technol., vol. 161, 2005, pp. 396–406, doi: 10.1016/j.jmatprotec.2004.07.063.

[6] M. Hojny and M. Glowacki," Computer modelling of deformation of steel samples with mushy zone," Steel Research Int., vol. 79, 2008, pp. 868-874. doi: 10.1007/1-4020-5370-3_537

[7] M. Hojny and M. Glowacki, "Modeling of strain-stress relationship for carbon steel deformed at temperature exceeding hot rolling range," Journal of Engineering Materials and Technology, vol. 133, 2011, pp. 021008.1-021008.7, doi: 10.1115/1.4003106.

[8] Z. Malinowski and R.Szyndler, “Axial transformation method in the analysis of the axisymmetrical plastic-flow,” Steel Research Int., vol. 58, 1987, pp. 503–507.

[9] M. Hojny and M. Glowacki, "The methodology of strain – stress curves determination for steel in semi-solid state," Archives of Metallurgy and Materials, vol. 54, 2009, pp. 475–483.

[10] M. Hojny and M. Glowacki, "The physical and computer modeling of plastic deformation of low carbon steel in semi-solid state," Journal of Engineering Materials and Technology, vol. 131, 2009, pp. 041003.1-041003.7, doi: 10.1115/1.3184034.

 ISBN: 978-1-61208-237-0 

# Parallelization on Heterogeneous Multicore and Multi-GPU Systems of the Fast Multipole Method for the Helmholtz Equation using a Runtime System

Cyril Bordage
*CEA/CESTA*
*Le Barp, France*
*INRIA*
*Bordeaux, France*
*cyril.bordage@inria.fr*

***Abstract*—The Fast Multipole Method (FMM) is considered as one of the top ten algorithms of the 20th century. The FMM can speed up solving of electromagnetic scattering problems. With N being the number of unknowns, the complexity usually $O(N^2)$ becomes $O(N \log N)$ allowing a problem with hundreds of millions of complex unknowns to be solved. The FMM applied in our context has a serious drawback: the parallel version is not very scalable. In this paper, we present a new approach in order to overcome this limit. We use StarPU, a runtime system for heterogeneous multicore architectures. Thus, our aim is to have good efficiency on a cluster with hundreds of CPUs, and GPUs. Much work have been done on parallelization with advanced distribution techniques but never with such a runtime system. StarPU is very useful, especially for the multi-level algorithm on a hybrid machine. At present, we have developed a multi-core and a GPU version. The techniques for distributing and grouping the data are detailed in this paper. The first results of the strategy used are promising.**

***Keywords-Fast multipole method (FMM); Helmholtz equation; heterogeneous architecture; parallel algorithm.***

## I. INTRODUCTION

The main aim is the simulation of the electromagnetic behavior of 3D complex objects in the frequency domain. For that, we use standard numerical methods such as Boundary Integral Equations [1], [2] based on a classical Finite Element approximation of surface Integral Equations such as EFIE and CFIE formulations [3].

These formulations lead to a linear system with a full matrix, which is complex non Hermitian but symmetric. It is solved by an iterative method, which has a complexity of $O(N^2)$, with $N$ being the number of unknowns The complexity comes from the matrix-vector products computed at each iteration. The Fast Multipole Method (FMM) [4] is able to reduce the complexity of these matrix-vector products, and so of the global problem, to $O(N \log(N))$ [5]. In this paper, we will study the FMM only in the context of electromagnetic scattering problems, with the kernel from the Helmholtz equation.

With modern parallel architectures, the parallelization of the FMM is essential, if we want to solve very large problems, which need a lot of memory. The different parallelizations are not efficient on distributed memory architectures and unfortunately, architectures nowadays are becoming more complex, by integrating accelerators like GPUs. With these new architectures, load balancing is more complicated and calculations have to be fitted.

This paper is organized as follows. In Section II, we outline the FMM. In Section III, we briefly describe the different strategies in the parallelization for distributing the computations. Our approach and its justification are explained in Section IV. Finally, in Section V, we present some results.

## II. THE FMM

The FMM was introduced by Greengard and Rokhlin in 1987 [4]. In the 90s, the method was applied to electromagnetism by Rokhlin [6] and Chew [5] in its diagonal version. We will present briefly the FMM algebraically from [7],[8]. A good analytic presentation can be found in [9] or [10].

### A. Principle

The FMM computes the matrix-vector product:

$$\vec{v} = \mathbb{G}.\vec{u} \tag{1}$$

with: $\mathbb{G}_{i,j} = G(|x_i - x_j|)$.

In our context, the Helmholtz equation, the Green function $G$, is defined by:

$$G(|x_i - x_j|) = \frac{\mathrm{e}^{ik|x_i - x_j|}}{4\pi|x_i - x_j|}$$

The FMM is based on a space partitioning. First, a partitioning $\mathcal{P}$ of the points is created, based on a geometrical criterion. The partitioning is made up of boxes. If $\mathcal{B}$ is a box of the partition, we define:

$$\begin{cases} \vec{u}^{\mathcal{B}} = (u_i)_{x_i \in \mathcal{B}} \\ \mathbb{G}^{\mathcal{B}_t, \mathcal{B}_s} = (G_{x_i,x_j})_{(x_i,x_j) \in (\mathcal{B}_t \times \mathcal{B}_s)} \end{cases} \tag{2}$$

With the Gegenbauer theorem [11], we have an approximate factorization of $\mathbb{G}^{\mathcal{B}_t, \mathcal{B}_s}$ if $\mathcal{B}_t$ and $\mathcal{B}_s$ are not neighbours, which means that they do not share any vertex.

 ISBN: 978-1-61208-237-0

$$\mathbb{G}^{\mathcal{B}_t,\mathcal{B}_s} \simeq \left(\mathbb{A}^{\mathcal{B}_t}\right)^\star \ \mathbb{T}^{\mathcal{B}_t,\mathcal{B}_s} \ \mathbb{A}^{\mathcal{B}_s}, \qquad \text{with } \mathcal{B}_s \notin \mathcal{V}(\mathcal{B}_t) \quad (3)$$

where:

- $\mathcal{V}(\mathcal{B}_t)$, the set of the neighbours of $\mathcal{B}_t$.
- $\mathbb{A}^{\mathcal{B}}$ is a $P \times N^{\mathcal{B}}$ matrix, called the aggregation matrix.
- $\mathbb{T}^{\mathcal{B}_t,\mathcal{B}_s}$ is a diagonal $P \times P$ matrix, called the translation matrix.
- $\mathbb{A}^\star$ is the conjugate transposition of $\mathbb{A}$.
- $N^{\mathcal{B}}$ is the number of elements in $\mathcal{B}$
- $P$ is inversely proportional to the squared number of boxes, called the number of directions.

The factorization is valid only if $\mathcal{B}_t$ and $\mathcal{B}_s$ are not neighbours. So, we have to split the computation of $\vec{v}^{\mathcal{B}_t}$ in two parts:

$$\vec{v}^{\mathcal{B}_t} = \vec{v}^{\mathcal{B}_t}_{\text{far}} + \vec{v}^{\mathcal{B}_t}_{\text{near}} \quad (4)$$

$\vec{v}^{\mathcal{B}_t}_{\text{near}}$ is computed directly, but for $v^{\mathcal{B}_t}_{\text{far}}$, the factorization yields:

$$\vec{v}^{\mathcal{B}_t}_{\text{far}} \simeq \left(\mathbb{A}^{\mathcal{B}_t}\right)^\star \sum_{\mathcal{B}_s \notin \mathcal{V}(\mathcal{B}_t)} \mathbb{T}^{\mathcal{B}_t,\mathcal{B}_s} \ \mathbb{A}^{\mathcal{B}_s} \vec{u}^{\mathcal{B}_s} \quad (5)$$

$\mathbb{A}^{\mathcal{B}_s} \vec{u}^{\mathcal{B}_s}$ is computed once by $\mathcal{B}_s$, for every $\mathcal{B}_t$. This is the key point of the FMM.

With all these elements, we have a method to quickly compute the product.

### B. Single-Level Multipole Method

First, we have to split our object in boxes. Then, the algorithm requires three steps:

1) Aggregation:

$$\vec{F}^{\mathcal{B}_s} = \mathbb{A}^{\mathcal{B}_s} \vec{u}^{\mathcal{B}_s}, \quad \forall \mathcal{B}_s \in \mathcal{P} \quad (6)$$

$\vec{F}^{\mathcal{B}_s}$ is a $P$ vector, called the vector associated with box $\mathcal{B}_s$.

2) Translation:

$$\vec{N}^{\mathcal{B}_t} = \sum_{\mathcal{B}_s \notin \mathcal{V}(\mathcal{B}_t)} \mathbb{T}^{\mathcal{B}_s,\mathcal{B}_t} \ \vec{F}^{\mathcal{B}_s} \quad (7)$$

$\vec{N}^{\mathcal{B}_t}$ is a $P$ vector, it represents the contribution on $\mathcal{B}_t$ from its non neighbours.

3) Disaggregation:

$$\vec{v}^{\mathcal{B}_t}_{\text{far}} = \left(\mathbb{A}^{\mathcal{B}_t}\right)^\star \ \vec{N}^{\mathcal{B}_t} \quad (8)$$

Finally, from (2) and (4, we obtain $v$:

$$\vec{v} = \sum_{\mathcal{B}_s \in \mathcal{P}} \vec{v}^{\mathcal{B}_s} \quad (9)$$

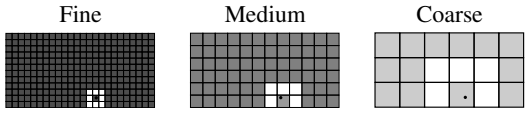



Figure 1. Translations to the gray boxes, of the box with the black point, depending on three partitionings. White boxes are direct calculations.

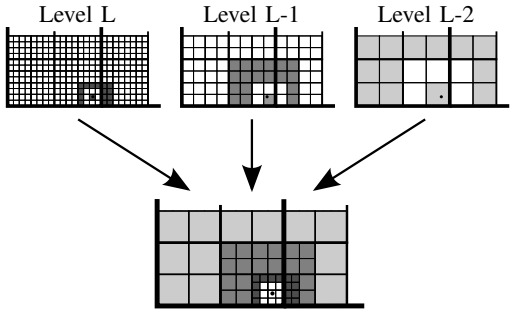



Figure 2. The translations in the ML-FMM.

### C. Multi-Level Multipole Method

The most time consuming step in the FMM is the translation step. Indeed, there are many translations, their total amount being equal to the squared number of boxes. By increasing the size of the boxes, the amount of translations is decreased, but at the same time, the amount of direct computations is increased, Figure 1.

In order to reduce the total amount of translations, the Multi-Level Multipole Method (ML-FMM) uses several levels of interleaved partitions. The first level, called highest level is a partition of just one box. The next level is built by splitting the partition of the upper level. The last level is called the lowest one. We call the parent of a box, the box in the previous level, including this box. We define the children of a box reciprocally.

In the ML-FMM, the translation between two points is done in the lowest level, in which the boxes containing the points are not neighbours. Thus, the translations are carried out where they involve the smallest number of boxes. In Figure 2, the translations of the black point are done in three levels depending on the targets. The boxes are translated only to their non neighbour boxes that is to say the children of the neighbours of their parent. The explanation is simple: the non neighbour boxes of the father, and so their children too, are processed by the parent. We call these boxes far neighbours. Finally, the point is translated to each box besides the neighbours of the box, as in the SL-FMM.

The multilevel method involves knowing the vectors in the boxes at each level. A box can be computed by aggregating its child. That corresponds to the sum of all the child vectors multiplied with a shifting matrix $\mathbb{E}$. We remind that $P$, the size of the aggregated vector $\vec{F}$, depends on the number of boxes and so, on the level. Therefore, the size $\mathcal{P}^l$ of a child vector is lower than $P^{l-1}$, the size of its parent vector.

 ISBN: 978-1-61208-237-0 

The father vector has to be interpolated, it is done by the multiplication with the interpolation matrix $\mathbb{I}$. These two operations are merged in the downward pass.

We also have to fetch the values from the higher levels for the disaggregation step. This operation is the upward pass.

In conclusion, the algorithm differs from the SL-FMM in the translation step, which is replaced by a loop on the levels of 3 steps: the downward pass, the translation and the upward pass. We define $\mathcal{B}^l$, a box on the level $l$, and $\mathcal{P}^l$, the partition of the level $l$. The highest level is level 1 and the lowest, level $L$. The algorithm has five types of operations:

1) Aggregation:

$$\vec{F}^{\mathcal{B}_s^L} = \mathbb{A}^{\mathcal{B}_s^L} \vec{u}^{\mathcal{B}_s^L}, \quad \forall \mathcal{B}_s^L \in \mathcal{P}^L \tag{10}$$

2) Upward pass: $\forall \mathcal{B}_s^l \in \mathcal{P}^l,\ L-1 \leq l \leq 3$,

$$\vec{F}^{\mathcal{B}_s^l} = \mathbb{I}^{l,l+1} \sum_{\mathcal{B}_s^{l+1} \subset \mathcal{B}_s^l} \mathbb{E}^{\mathcal{B}_s^l, \mathcal{B}_s^{l+1}} \vec{F}^{\mathcal{B}_s^{l+1}} \tag{11}$$

where:

- $\mathbb{E}^{\mathcal{B}^l, \mathcal{B}^{l+1}}$ is a diagonal $P^{l+1} \times P^{l+1}$ matrix.
- $\mathbb{I}^{l,l+1}$ is a $P^l \times P^{l+1}$ matrix.

3) Translation: $\forall \mathcal{B}_s^l \in \mathcal{P}^l,\ L \leq l \leq 3$,

$$\vec{N}_T^{\mathcal{B}_t^l} = \sum_{\mathcal{B}_s^l \in \mathcal{V}_{\text{far}}(\mathcal{B}_t^l)} \mathbb{T}^{\mathcal{B}_t^l, \mathcal{B}_s^l}\ \vec{F}^{\mathcal{B}_s^l} \tag{12}$$

4) Downward pass: $L-1 \leq l \leq 3$,

$$\vec{N}^{\mathcal{B}_t^{l+1}} = \begin{cases} \vec{N}_T^{\mathcal{B}_t^{l+1}} & \text{if } l = 3 \\ \vec{N}_T^{\mathcal{B}_t^{l+1}} + \left(\mathbb{E}^{\mathcal{B}_t^l, \mathcal{B}_t^{l+1}}\right)^\star \left(\mathbb{I}^{l,l+1}\right)^\star \vec{N}^{\mathcal{B}_t^l} & \\ & \text{if } l \in [\![4, L]\!] \end{cases} \tag{13}$$

5) Disaggregation: $\forall \mathcal{B}^L \in \mathcal{P}^L$,

$$\vec{v}_{\text{far}}^{\mathcal{B}_t^L} = \left(\mathbb{A}^{\mathcal{B}_t^L}\right)^\star \mathbb{N}^{\mathcal{B}_t^L} \tag{14}$$

## III. Parallelization of the FMM

The parallelization of the FMM depends on the context. For example, for Laplace or Stokes kernels, good performances have been achieved with hundreds of thousands cores [12]. Unfortunately, in our context, the parallelization of the FMM is not so efficient. The reason is the size of vectors which increases when we go up in the tree. Thus, the amount of computations is nearly the same at each level unlike for Laplace.

That is why much work has been carried out on its parallelization since the 90s. Research works has been focused on the ML-FMM, but recent years, the SL-FMM has drawn attention with a better scalability. In addition to parallelization on many CPUs, there is a trend towards using GPUs to carry out calculations.

### A. The ML-FMM

The first parallelization is based on the distribution of the boxes among the processors at each level .

Unfortunately, as we can see in [13], this parallelization on 16 processors gives poor results in terms of scalability. The reason is that at a given level, there are not enough boxes to share between processors so a good load balancing is impossible. It seems not to be important because it concerns only the boxes on the levels, which have less computations. But the problem is that, in the FMM applied to electromagnetism, each level needs the same amount of calculations, since $P_l$, the size of the vectors in (11) (12) (13), is in inverse proportion to the squared number of boxes. As a result, the bad parallelization of these levels has a great negative impact on the global scalability.

Velamparambil and Chew discovered [13] another strategy to overcome that. The previous distribution is kept for the fine levels, called the distributed layer. For the coarse levels, the vector attached to a box is split into blocks and distributed among the processors. The computations are carried out by block on each processor. In the levels with the new distribution, called the shared layer, data have to be replicated.

ErgÃ$\frac{1}{4}$l and GÃ$\frac{1}{4}$rel [14], presented another distribution called the hierarchical distribution. It consists of distributing the fields not only for the coarse levels but at each level . With this distribution 60% of efficiency can be achieved with 128 processors, whereas we have only 40% with the hybrid distribution and 20% with the simple distribution.

### B. The SL-FMM

The ML-FMM has a better complexity but its parallelization is not efficient enough and all the possible strategies seem to have been considered. That is why Waltz et al. [15] take an interest in the SL-FMM in the context of the parallelization. Indeed, in the single-level method, the block of the vector linked to the boxes, can be computed independently. We just have to gather them at the end of the algorithm. So, the parallelization over the samples of field will be very efficient. It is not the same for the ML-FMM where the blocks have to be gathered for downward pass.

Moreover, in [16], Wagner et al. proposed the FMM-FFT, which reduces the complexity of the translation stage. The complexity becomes $\mathrm{O}(N^{4/3} \log^{2/3} N)$, which represents an important improvement compared to the $\mathrm{O}(N^{3/2})$ [15]. In [17], we have a proof of the efficiency of the FMM-FFT, with a perfect scalability up to 512 processors and a very good one for 1024.

Last year, Taboada et al. [18] combined the FMM-FFT and the ML-FMM. When the number of processors is bigger than the number of nodes, they use the FMM-FFT. For the next levels down, they use independent ML-FMM on each node. With this new method, they have solved a problem with 620 million unknowns.

 ISBN: 978-1-61208-237-0

### C. With GPUs

Work have been done in the FMM, but only for Laplace, or other non oscillatory kernels. Good efficiencies have been achieved on GPUs thanks to the BLAS [19].

The important points for using a GPU in scientific applications are the consistency of the computations and enough computations compared to the data transfers. The data have to be well-sorted to benefit from the coalescing accesses.

## IV. Our strategy

Many efforts have been done on the distribution of the computations in the last fifteen years. There has also been research on computation scheduling with tasks queues [20]. We have chosen the hierarchical distribution, and for the top level, a simple SLFFM, which can be upgraded in a FMM-FFT. We have decided to focus on computation scheduling because a good scheduling is the key in modern machines, with heterogeneous processing units. A good scheduling depends on the machine: speed of the processing units, bus speed, network speed, etc. The schedule has to fit the algorithm but also the machine, as the computations have to be adapted to the processing units. Another important point for a good scalability is to hide the communications by computations.

### A. The dynamic scheduling

Our aim is to compute the FMM on a supercomputer with shared and distributed memory, thousands of CPUs and GPUs. For that we use the same distribution between the nodes as in the combination of the FMM-FFT and the ML-FMM. In a node, we use the dynamic scheduler StarPU [21]. It can handle the scheduling on CPUs and GPUs with different strategies: greedy, work stealing, minimal termination time, priority, etc. It automates transfers throughout heterogeneous machines and favours data locality. Tasks and data dependencies must be declared and StarPU does the rest.

The strategy, which has been considered, is the minimal termination time. It takes a task execution model and data transfer model into account to know where a task will end the soonest. The models can be provided for StarPU or it can build them.

### B. Efficient operations

To be efficient, the tasks handled by the task scheduler should imply enough computations to hide the costs of the scheduling and of the data transfers. This is especially the case for the GPUs where the data transfers are more costly and because they can do simultaneous computations.

*1) The data:* Many computations have to be grouped in a same task. For that, the same strategy as in parallelization III-B is used: groups are made with many directions for many boxes.

To avoid calculation starvation and deadlocks, the granularity must be low. But high enough to permit the GPUs to be efficient. The number and the size of the blocks should be tuned depending on the machine and the input data. For that, we can set the number of tasks by level, depending on the numbers of computing workers.

*2) Dependencies:* All the operations except the translation and the upward pass can be done just by using the data contained in one direction block. Thus, there is no communication.

For the upward pass, the computation of a parent box needs all the directions of all its children. Transposed to our blocks, that becomes: a direction block needs all its child box blocks with all their direction blocks. This can be done by using one task for each direction block in the child level but by reducing the out data. StarPU can deal with reduce operations itself.

The translations to a box use all its far neighbours. Consequently, for translating all the boxes in a block, the far neighbours of all the boxes are needed. Although most of the far neighbours are in the same block, some are external to the block. Therefore, the translations in a block need data from other blocks. To limit the memory accesses between blocks, the data of the external far neighbours are copied to the block; see Figure 3. Thus, all the translations will be internal to the block. That adds synchronization because a translation can occur in a block only if all its neighbours have been computed. In fact, this is not a problem because the translations are useful only for the downward pass.

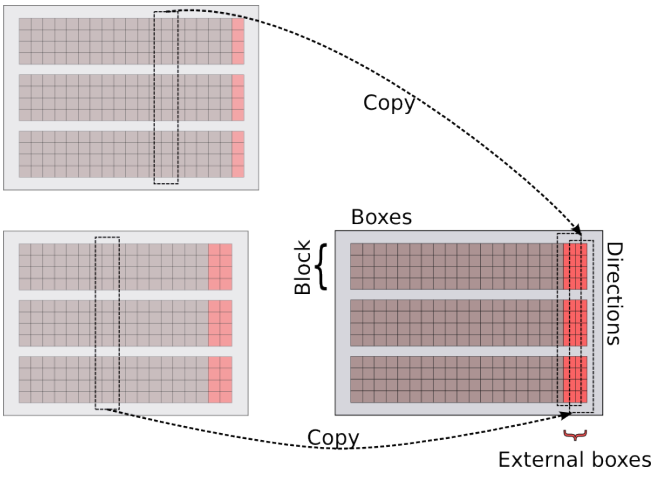



Figure 3. The distribution of the data

*3) With GPUs:* All the operations in the FMM are simple to implement on GPUs. There are mainly matrix vector multiplications. This is relatively efficient on the GPUs when the sizes of the matrix are not to small. We just have make use of coalescent accesses and avoid bank conflicts.

 ISBN: 978-1-61208-237-0 

## V. RESULTS

For the time being, the tests have only been done for the shared memory and for the GPUs but not for the distributed memory yet.

### A. On shared memory

The test was done on a 2 Hexa-core Westmere Intel Xeon X5650 2.67 GHz (10.664 GFlops by core) with a sphere of 2 million points at 500 MHz. The results are presented in the Table I. The scalability is strong with 12 processors but

| # cores | # blocks by level | Time (s) | Efficiency |
|---|---|---|---|
| 1 | 1 | 80.1 | 100% |
| 2 | 8 | 41.1 | 97% |
| 4 | 8 | 21.4 | 94% |
| 6 | 10 | 14.5 | 92% |
| 8 | 10 | 10.9 | 92% |
| 10 | 10 | 8.7 | 92% |
| 12 | 10 | 7.3 | 91% |

Table I
ON SHARED MEMORY WITH A 2 MILLION SPHERE

we have to do some tests on a machine with more CPUs. When we look at the scheduling, Figure 4, we find that the copies of the far neighbours (called block sharing) do not represent much time. Therefore, the cost of the parallelism is insignificant. The other important point is the waste time of the processors (called blocked): instead of executing a task, a processor is waiting for a new task.

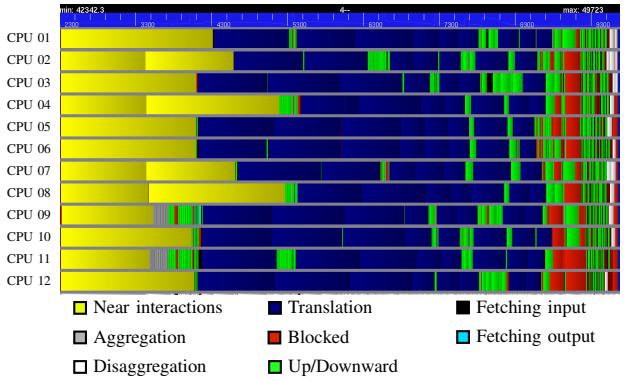



Figure 4. Gantt diagram for the execution on 12 cores

This is the result of an insufficient number of tasks, but here, if we increase the number of tasks, the cost of the parallelism becomes significant and the global time increases. Nevertheless, to avoid this kind of situation, we can favour the tasks that create other tasks and so parallelism. These tasks are the aggregations and the upward passes. In our scheduler, favouring a task can be done easily by adding a priority to this task.

### B. With GPUs

The aim of our approach is to use GPUs. We have only done preliminary tests on the same machine with 3 NVIDIA Tesla M2070 (1 TFlops). 3 processors are dedicated to the handling of the 3 GPUs by StarPU. The test case is a 10 meter sphere with 2 million points at 1 GHz. At present we have only implemented the aggregation, the translations and the near interactions.

The aggregation is computed at the same speed on the GPUs, Figure 5. The near interactions are 7 times faster. The translation is 50 times faster. This last result is very good if we look at the flops of the processing units.

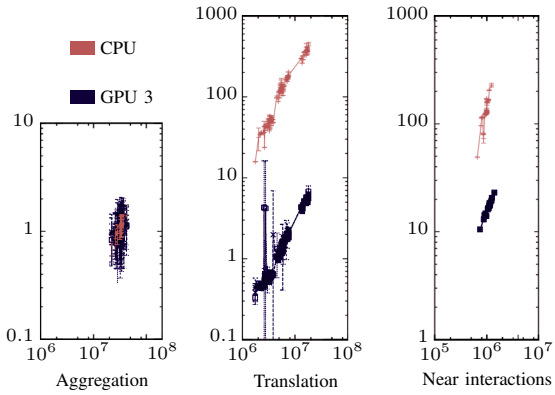



Figure 5. Execution time ($\mu s$) depending on the input size (B)

For having this efficiency, we must have enough directions (at least 100). But we have planned to develop kernels for small numbers of directions. StarPU will choose the better kernel depending on the size of the inputs.

The scheduling, Figure 6, is good on the GPUs but poor on CPUs. This is due to the fact that the blocks are too big for the CPUs. But if we decrease the size, we will loose our efficiency on the GPUs. That is why we want to create tasks with different sizes.

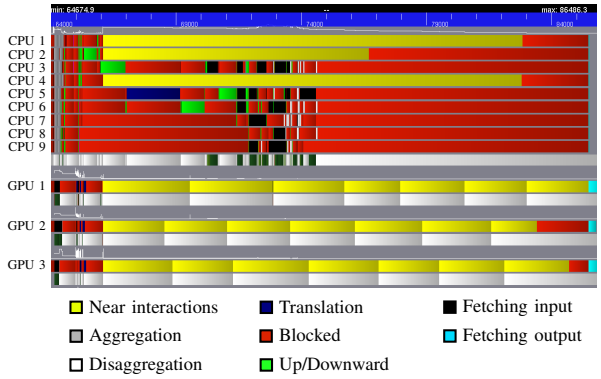



Figure 6. Gantt diagram for the execution on 9 cpu cores and 3 gpus

Without the GPUs, the global time of the computations is 4 times greater. It is not much as compared to the acceleration of the tasks separately. We hope for better results with

 ISBN: 978-1-61208-237-0 

a better block creation and with the implementation of all the tasks on GPUs.

## VI. Conclusion and future work

In this paper, we have studied the parallelization of the FMM, applied to scattering problems, with a dynamic scheduler. We have also seen how to arrange the computations in order to permit an efficient scheduling. Thus, on shared memory, the strong scalability is good. On GPUs,our first tests encourage us to continue. Our work is promising; but, to make conclusions we still have a lot of work to do: upward pass on GPUs, improvement of the kernels, strategies for the tasks, adaptive method, and MPI.

## References


[1] D. Jones, "Acoustic and electromagnetic waves," *Oxford/New York, Clarendon Press/Oxford University Press, 1986, 764 p.*, vol. 1, 1986.

[2] J. Stratton, *Electromagnetic theory*. Wiley-IEEE Press, 2007, vol. 33.

[3] D. L. Colton and R. Kress, *Integral equation methods in scattering theory*, ser. Pure and Applied Mathematics (New York). New York: John Wiley & Sons Inc., 1983, a Wiley-Interscience Publication.

[4] L. Greengard and V. Rokhlin, "A fast algorithm for particle simulations* 1," *Journal of Computational Physics*, vol. 73, no. 2, pp. 325–348, 1987.

[5] J. Song and W. Chew, "Multilevel fast-multipole algorithm for solving combined field integral equations of electromagnetic scattering," *Microwave and Optical Technology Letters*, vol. 10, no. 1, pp. 14–19, 1995.

[6] R. Coifman, V. Rokhlin, and S. Wandzura, "The fast multipole method for the wave equation: A pedestrian prescription," *Antennas and Propagation Magazine, IEEE*, vol. 35, no. 3, pp. 7–12, 2002.

[7] E. Darve, "The fast multipole method. I. Error analysis and asymptotic complexity," *SIAM J. Numer. Anal.*, vol. 38, no. 1, pp. 98–128 (electronic), 2000.

[8] X. Sun and N. Pitsianis, "A matrix version of the fast multipole method," *Siam Review*, vol. 43, no. 2, pp. 289–300, 2001.

[9] G. Sylvand, "La méthode multipôle rapide en électromagnétisme. Performances, parallélisation, applications," 2002.

[10] W. Chew, E. Michielssen, J. Song, and J. Jin, *Fast and efficient algorithms in computational electromagnetics*. Artech House, Inc. Norwood, MA, USA, 2001.

[11] M. Abramowitz and I. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. National Bureau of Standards Applied Mathematics Series 55. Tenth Printing*, 1972.

[12] A. Rahimian, I. Lashuk, S. Veerapaneni, A. Chandramowlishwaran, D. Malhotra, L. Moon, R. Sampath, A. Shringarpure, J. Vetter, R. Vuduc, D. Zorin, and G. Biros, "Petascale direct numerical simulation of blood flow on 200k cores and heterogeneous architectures," *SC Conference*, pp. 1–11, 2010.

[13] S. Velamparambil and W. Chew, "Analysis and performance of a distributed memory multilevel fast multipole algorithm," *Antennas and Propagation, IEEE Transactions on*, vol. 53, no. 8, pp. 2719–2727, 2005.

[14] Ö. Ergül and L. Gürel, "Hierarchical parallelization strategy for multilevel fast multipole algorithm in computational electromagnetics," *Electron. Lett.*, vol. 44, pp. 3–5, Jan. 2008.

[15] C. Waltz, K. Sertel, M. Carr, B. Usner, and J. Volakis, "Massively parallel fast multipole method solutions of large electromagnetic scattering problems," *Antennas and Propagation, IEEE Transactions on*, vol. 55, no. 6, pp. 1810–1816, 2007.

[16] R. Wagner, J. Song, and W. Chew, "Monte Carlo simulation of electromagnetic scattering from two-dimensional random rough surfaces," *Antennas and Propagation, IEEE Transactions on*, vol. 45, no. 2, pp. 235–245, 2002.

[17] J. Mouriño, A. Gómez, J. Taboada, L. Landesa, J. Bértolo, F. Obelleiro, and J. Rodríguez, "High scalability multipole method. Solving half billion of unknowns," *Computer Science-Research and Development*, vol. 23, no. 3, pp. 169–175, 2009.

[18] J. Taboada, M. Araujo, J. Bertolo, L. Landesa, F. Obelleiro, and J. Rodriguez, "Mlfma-fft parallel algorithm for the solution of large-scale problems in electromagnetics," *Progress In Electromagnetics Research*, vol. 105, pp. 15–30, 2010.

[19] N. Gumerov and R. Duraiswami, "Fast multipole methods on graphics processors," *Journal of Computational Physics*, vol. 227, no. 18, pp. 8290–8313, 2008.

[20] G. Sylvand, "Performance of a parallel implementation of the FMM for electromagnetics applications," *International Journal for Numerical Methods in Fluids*, vol. 43, no. 8, pp. 865–879, 2003.

[21] C. Augonnet, "Scheduling Tasks over Multicore machines enhanced with Accelerators: a Runtime System's Perspective," Ph.D. dissertation, Université Bordeaux 1, 351 cours de la Libération — 33405 TALENCE cedex, Dec. 2011.


 ISBN: 978-1-61208-237-0 

# Profile-based Recruiting of New Students

Ray R. Hashemi[1], Louis A. Le Blanc[2], Azita Bahrami[3], Kevin Willett[1], and Xaunna J. Krehn[1]

[1]Department of Computer Science
Armstrong Atlantic University
Savannah, GA 31419, USA
Ray.Hashemi@armstrong.edu

[2]Campbell School of Business
Berry College
Mount Berry, GA 30149-5024, USA
lleblanc@berry.edu

[3]IT Consultation Company
Savannah, GA, USA
Azita.G.Bahrami@gmail.com

*Abstract* – **A profile-based recruiting of new students for an institution of higher education is more efficient, financially sound, and more successful. In this paper, two different methodologies, Aprioi Algorithm and Modified Rough Sets, are used to create profiles from historical data collected by an admissions office of a college in the southeast United States. The first approach delivered two and the second approach delivered five profiling rules. The profiling rules were evaluated against a test set. The success rate of the Apriori Algorithm and Modified Rough Sets were 87% and 75%, respectively. The first approach had the false positive of 6% a false negative of 4%. The second approach had a false positive of 10% and false negative of 2%.**

***Keywords–Profiling; Profiling Rules; Recruiting; Apriori Algorithm; Modified Rough Sets.***

## I. INTRODUCTION

An institution of higher education receives many applications from prospective students. After a lengthy process of screening applications, a select number of new students are offered seats at the institution. Often students who have been offered these seats withdraw their applications and they do not show up for classes. These students have strong GPAs, good standardized tests scores, and therefore they receive multiple offers and then choose their favorite. As a result, the respective admissions offices accept more than their capacities and develop "alternate" or "wait" lists of applicants.

The money spent on recruiting, advertising, and long hours for screening of the applications consumes a sizable chunk of the admissions budget [1, 2]. These costs can be reduced substantially if a profile of potential or likely new students was known. That is, advertising will be tailored toward this targeted group of students and the advertisements will appear only in locations that reach potential students. The number of students who do not matriculate should drop, suggesting a more successful recruiting process.

The goal of this research is to generate profiles of those students who might attend the institution once accepted. Upon establishing such a profile, all the recruiting activities are channeled to those students who meet the profile.

The organization of the remainder of this paper is: relevant background in Section 2, the methodology in Section 3, empirical results in Section 4, and conclusion in Section 5.

## II. RELEVANT BACKGROUND

Both approaches, the Apriori Algorithm and the Rough Sets, are introduced in the next two subsections, respectively. (The concept of the modified Rough Sets is discussed in sub-section 3.3.)

### *A. Apriori Algorithm*

A record with n attributes consists of n predicates of $A_i(x_{ij})$ (for i = 1to n and j = 1 to m), where $A_i$ is the i-th attribute with m possible values and $x_{ij}$ is the j-th possible value for $A_i$. In reference to the goal of the study, a predicate's value represents one piece of information about a student who applies for a seat in a university. In addition, one predicate is designated as the decision predicate. Its value represents the admissions office's action on the student. The predicates other than the decision predicate are referred to as condition predicates. We use the Apriori Algorithm [3] to establish the association(s) between the decision predicate and the condition predicates. The algorithm identifies the predicate sets that appear together most frequently. If k predicates frequently

 ISBN: 978-1-61208-237-0 

appear together in a dataset, they make a k-dimensional predicate set (k-D predicate set).

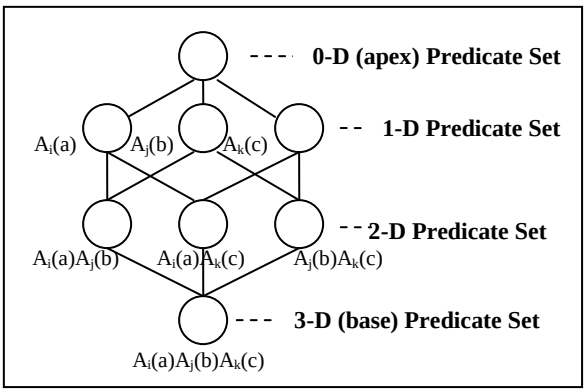



Figure 1. A predicates' lattice for a dataset with three most frequent predicates of $A_i(a)$, $A_j(b)$ and $A_k(c)$

The predicates of a predicate set reflect a stronger bond among participant predicates. The predicate set which includes a decision predicate is of interest because it shows a strong bond between condition predicates and decision predicate. These predicate sets are used to build association rules.

The algorithm starts by checking the frequency of appearance of each individual predicate in all records of a dataset to identify those predicates whose frequency is greater than a threshold (*support count*) t. The outcome makes 1-D predicate sets. As the second step, the most frequent 2-D predicate sets are identified. To do so, all 2 by 2 possible combinations of the 1-D predicate sets are built; and, those with support count less than threshold t are filtered. The process continues until the most frequent k-D predicate set is identified. The value for k is decided when (k+1)-D predicate set is empty because all of them are filtered out. All the frequent predicate sets for a dataset may be shown in form of a predicates' lattice, Fig. 1. The number of possible predicates for n attributes is :

$$\sum_{j=1}^{n} \frac{n!}{j!(n-j)!} \qquad (1)$$

For a relatively large n, the number of predicates is too many. The use of a support count threshold weeds out a large number of the predicates in each predicate set.

*B. Rough Sets*

The Rough Sets approach was introduced by Pawlack in 1984 [4]. The details can be found in [5, 6, 7]. First, the Rough Set approach is defined, and then it is put in perspective to the problem at hand.

*Definition 1*: An approximate space P is an ordered pair P(U, R), where U is the universe of objects and R is a binary equivalence relation over U.

*Definition 2:* Let R* be a family of subsets of R and let $A \subseteq U$. If for some $Y \subseteq R^*$, A is equal to the union of all sets in Y, then A is definable in P; otherwise, A is non-definable or A is *a rough set*.

*Definition 3*: Any Rough Set A has a lower approximation space Low(A), an upper approximation space Up(A), and a boundary B(A). And they are defined as follows:

$$\text{Low}(A) = \{a \subseteq U \mid [a]_R \subseteq A], \qquad (2)$$

$$\text{Up}(A) = \{a \subseteq U \mid [a]_R \cap A \neq \varnothing], \qquad (3)$$

$$B(A) = \text{Up}(A) - \text{Low}(A). \qquad (4)$$

Let the relatively large rectangle, Fig. 2, represent a dataset (universe), and each small rectangle represent a student record. Small rectangles of the same shade are records with the same condition predicates. There are four such sets in Fig. 2, namely S1 (all black rectangles), S2 (all gray rectangles), S3 (all rectangles with pattern), and S4 (all the white rectangles). Let the possible values for a decision predict be m and one of these values is $x_a$. The records that have the same decision predicate of decision($x_a$) are shown bordered by the broken line and make a Rough Set, because it cannot be created by any possible union of the four sets. The lower approximation of the Rough Set includes those four sets that are totally inside the rough set (i.e., S1). The upper approximation of the Rough Set includes those four sets that are totally or partially inside the Rough Set (i.e., S1, S2, and S3). The boundary of the Rough Set includes S2 and S3. The condition predicates belonging to the lower approximation of a decision predicate have a stronger bond with the decision value than those in the upper approximation space.

Since there are more than one value for the decision predicate, there are more than one Rough Set for the dataset. Thus, the methodology is named Rough Sets (plural).

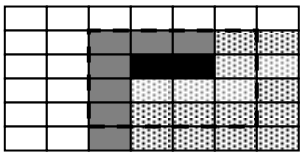


Figure 2. Rough sets visualization

 ISBN: 978-1-61208-237-0

## III. METHODOLOGY

To create profiles of new students who are highly likely to attend the university after they are accepted, we use the Apriori Algorithm and the Modified Rough Sets approach separately for building the profiles. The profiling process is completed using the following steps:

1. Cleaning historical data.
2. Discovery of association rules from the cleaned historical data.
3. Discovery of approximation rules from the cleaned historical data using Modified Rough Sets approach.
4. Building the profiles.

The resulting profiles are evaluated on a sample set obtained from the original dataset to compare the performance of the two approaches.

### A. Cleaning Historical Data

Historical data was collected by an admissions office of a college in the southeast United States. The historical data is cleaned vertically and horizontally. The vertical reduction is done by removing (a) the duplicate records (objects) from the historical dataset and (b) records with missing data.

The horizontal reduction is done by removing the redundant attributes from the vertically reduced dataset. The entropy approach [3] is used to identify the redundant attributes. To explain further: Let one or a set of attributes of the dataset be the *decision attributes* and the rest of them be *condition attributes*. In addition, let a decision attribute have m distinct values (classes). (In the case that the decision attribute is made up of more than one attribute, i.e. complex attribute, the classes for the complex decision attribute are all the possible combinations of the classes of the constituents.)

To determine the redundant condition attributes:

1. The entropy of the set of condition attributes, C, is calculated as follows:

$$E(C) = -\sum_{i=1}^{m} p_i * log_2 * p_i \quad (5)$$

Where, $p_i$ is frequency of class i in the dataset.

2. For each condition attribute q in C which has $v_1$, $v_2$, . . ., $v_n$ possible values, the *information gain* is calculated using formulas 6 and 7.

$$B(q) = \sum_{i=1}^{n} N_{v_i} * E(v_i) \quad (6)$$

Where, $N_{v_i}$ is the number of records in the dataset with $q = v_i$ and $E(v_i)$ is the entropy of these records.

$$Gain(q) = E(c) - B(q) \quad (7)$$

3. If Gain(q) is less than a chosen threshold value, the attribute q is considered redundant and it is removed from the dataset.

### B. Discovery of Strong Association Rules

Consider a record of a given dataset. This record is composed of a set of condition predicates and decision predicates. A predicate is composed of an attribute and its value. The condition predicates are considered to be the conditions under which a process takes place. The decision predicates are considered to be the outcome of the process.

For example, the record in Fig. 3 represents an application of a student who has applied for admission along with the respective decisions made by the institution and applicant. The first five attributes make the condition predicates and attributes, Accepted and Matriculated make the decision predicates. To explain further, Accepted(1) and Accepted(0) mean the student was or was not admitted; and Matriculated(1) and Matriculated(0) mean the student enrolled or did not enroll at the university.

We apply the Apriori Algorithm to the dataset and obtain all the frequent k-predicate sets. A k-predicate set is composed of k predicates. Let $k_1$-predicate sets and $k_2$-predicate sets be two subsets of a k-predicate set such that $k_1 \cup k_2 = k$ (condition 1) and $k_1 \cap k_2 = \varnothing$ (condition 2). The association rules of $k_1 \Rightarrow k_2$ and $k_2 \Rightarrow k_1$ are generated out of k-predicates.

There are several subsets of k that satisfy the conditions (1) and (2), therefore, several association rules are generated from one frequent k-predicate set.

To each associate rule two measurements of *support* and *confidence* are assigned using the following formulas:

$$\text{Support } (k1 \Rightarrow k2) = P(k1 \cup k2)/M, \quad (8)$$

$$\text{Conf } (k1 \Rightarrow k2) = P(k1 \cup k2)/P(k1). \quad (9)$$

For the above formulas, $P(k_1 \cup k_2)$ and $P(k_1)$ are the number of records with $(k_1 \cup k_2)$ and $(k_1)$ in the dataset, respectively. M = |dataset|.

| ID | Zip Code | GPA | SAT | Income | Accepted | Matriculated |
|---|---|---|---|---|---|---|

Figure 3. A record layout

Filtering of the association rules in form of $k_i \Rightarrow k_j$ are done based on the following set of principles:

 ISBN: 978-1-61208-237-0 

*Principle 1*: Rules with confidence less than a selected threshold are pruned.

*Principle 2*: If $k_i \cap$ (Decision attributes) $\neq \varnothing$, then rule is pruned.

*Principle 3*: If $k_j \cap$ (Condition attributes) $\neq \varnothing$, then rule is pruned.

Principle 1 ensures that the higher the confidence, the stronger the rule. Principles 2 and 3 deliver the *inter-dimension* association rules for the decision attributes.

### C. Modified Rough Sets

In Rough Set nomenclature, an information system, S, is a quadruple (U, Q, V, d )
Where,

U is a non-empty finite set of objects, u.

Q is a finite set of attributes, q.

V $= \cup_{q \in Q} Vq$, and Vq is the domain of attribute q.

d is a mapping function such that d (a,q) $\in$ Vq for every q $\in$ Q and a $\in$ U.

All the objects who have the same values for their condition attributes constitute one *class*. And all the objects who have the same values for their decision attributes constitute one *partition*. The number of partitions is equal to the number of decision values. Consider partition $\lambda_i$, and classes $c_1, \ldots, c_n$. The objects of those classes that are totally contained within $p_1$ make the *lower approximation space* of $\lambda_i$, Low($\lambda_i$). The objects of all the classes that are either "totally" or "partially" contained in $\lambda_i$ make the *upper approximation space* of $\lambda_i$, Up($\lambda_i$ ). The objects of all the classes that are "partially" contained in $p_1$ make the *boundary* of $\lambda_i$, B($\lambda_i$ ). The objects in boundary of $\lambda_i$ have the same set of condition attributes but different decisions.

In any statistical model, these objects are removed because they are conflicting. However, we use the Modified Rough Sets approach to salvage the conflicting objects. Conflicting objects are part of life. For example, two patients (objects) with the same symptoms may be diagnosed differently (conflicting objects). Or, two prospective students with the same set of conditions, one decides to attend the college and the other one does not.

One of the decision values in B($\lambda_i$) is designated as the *dominant decision* using Bayes' Theorem [5]. The decision values for all the subjects in B($\lambda_i$) are then changed to the dominant decision. Such modification makes B($\lambda_i$) = $\varnothing$. And, therefore the Rough Sets are changed into Modified Rough Sets. The rules that are generated from the objects of a Modified Rough Set are called *approximate rules* [5]. Each approximate rule has a certainty factor that is the same as the probability assigned to its decision value by Bayes' theorem, Formula 10.

$$P[d_i|Class_j)\frac{P[Class_j|d_{ij}]}{\sum_{i=1}^{m} P[Class_j|d_{ij}]*Q_j]} \qquad (10)$$

Where, $d_i$ is the i-th decision value and $Class_j$ is all the records with the same set of condition values.

In Modified Rough Sets, those classes of data in which all subjects have the same decision values, the dominant decision is the common decision value and the probability assigned to such a dominant decision is 1.

### D. Building the Profiles

Profiles are meta-rules that are built from a set of rules. This is completed through a collapsing process that integrates and generalizes rules. The following three guidelines govern the collapsing process.

Guideline1:
The following rules of r1 and r2 are given:

r1: $ATT_1$=a1 $\wedge$ $ATT_2$ = b3 $\wedge$ $ATT_3$ = c4 $\wedge$ $ATT_4$ = d2 $\rightarrow$ $ATT_d$ = f2

r2: $ATT_1$=a1 $\wedge$ $ATT_2$ = b3 $\wedge$ $ATT_3$ = c4 $\wedge$ $ATT_4$ = d2 $\wedge$ $ATT_5$ = e1 $\rightarrow$ $ATT_d$ = f2

The rule r1 reads "If Attribute#1 is equal to a1 and Attribute#2 is equal to b3 and Attribute #3 is equal to c4 and Attribute#4 is equal to d2, then Attribute decision = f2".

All the objects that fire rule r2 are a subset of objects firing rule r1. Therefore, r1 and r2 are collapsed into a new rule that is the same as the rule r1.

Guideline 2:
The following rules of r1 and r2 are given:

r1: $ATT_1$=a1 $\wedge$ $ATT_2$ = b3 $\wedge$ $ATT_3$ = c4 $\wedge$ $ATT_4$ = d2 $\rightarrow$ $ATT_d$ = f2

r2: $ATT_1$=a2 $\wedge$ $ATT_2$ = b3 $\wedge$ $ATT_3$ = c4 $\wedge$ $ATT_4$ = d2 $\rightarrow$ $ATT_d$ = f2

The rules r1 and r3 may collapse into a new rule

r': $ATT_1$ = (a1$\vee$ a2) $\wedge$ $ATT_2$ = b3 $\wedge$ $ATT_3$ = c4 $\wedge$ $ATT_4$=d2 $\rightarrow$ $ATT_d$= f2

If a1, and a2 are the only possible values for attribute $ATT_1$, then r' changes into

r": $ATT_2$ = b3 $\wedge$ $ATT_3$ = c4 $\wedge$ $ATT_4$=d2 $\rightarrow$ $ATT_d$ = f2

Guideline 3 (heuristic rule):
The following rules of r1 and r2 are given:

r1: $ATT_1$=a1 $\wedge$ $ATT_2$ = b3 $\wedge$ $ATT_3$ = c4 $\wedge$ $ATT_4$ = d2 $\rightarrow$ $ATT_d$ = f2

r2: $ATT_1$= k3 $\wedge$ $ATT_2$ = b3 $\wedge$ $ATT_3$ = c4 $\wedge$ $ATT_4$ = d2 $\rightarrow$ $ATT_d$ = f2

a = |r1.conditions| and
b = |r2.conditions|

 ISBN: 978-1-61208-237-0

If $|r1 \cap r2| > Thcommon \wedge$
$|r1 - r2| < Thdifference \wedge$
$|r2 - r1| < Thdifference$
Then r1 and r2 can be collapsed into a new rule that is the same as r1 (if $a \leq b$) or the same as r5 (if $a > b$).
The Thcommon and Thdifference are two thresholds decided by the analyst.

## IV. EMPIRICAL RESULTS

Three different JAVA programs were developed to implement data cleaning, Apriori Algorithm and Modified Rough Sets. All three programs were executed on an Hewlet Packard laptop.

The historical dataset had over 30,000 records and each record had 20 attributes. Only 1,577 records survived the vertical cleaning and ten attributes survived the horizontal cleaning. The ten attributes along with their values and meanings are described in Table 1.

TABLE I. NON-REDUNDANT ATTRIBUTES AND THEIR VALUES

| Attribute | Values |
|---|---|
| *Parent's Zip* | 1 (All zip codes belong to Atlanta),<br>2 (All GA zip codes of Atlanta Suburbs), and<br>3 (All other zip codes) |
| *Gender* | 1(Female) and 2( Male) |
| *Ethnicity* | 1 (Caucasian), 2 (Others) |
| *State* | 1(Georgia), 2(Other States) |
| *SAT Score* | 1 ($\leq$ 1000),<br>2 (> 1000 and $\leq$ 1200), and<br>3 (> 1200) |
| *GPA:* | 1 (<3.00),<br>2 ( $\geq$3.00 and <3.5),<br>3 ($\geq$3.50 and <4.00), and<br>4 ($\geq$4.00) |
| *Family Contribution* | 1(= 0),<br>2 (>0 and <=5000),<br>3(>5000 and $\leq$ 15000),<br>4(>15000 and $\leq$ 25000), and<br>5 (>25000) |
| *Accepted* | 0 (No), 1 (Yes) |
| *Matriculated* | 0 (No), 1 (Yes) |
| *Cancelled* | 0 (No), 1 (Yes) |

The numbers of association rules obtained by applying Apriori Algorithm, their minimum and maximum confidence levels along with the number of profiling rules are shown in Table 2. The numbers of approximate rules generated by the Modified Rough Sets approach along with the profiling rules are displayed in Table 3.

To check the validity of the profiles, we have applied the set of profiles on the test set. The test set has 159 records (roughly 10% of the total records). The test results are shown in Table 4.

The profiling results, Table 4, using profile rules of the Apriori Algorithm has 87% correct profiling with false positive of 6% and false negative of 4%. Using the profile rules of the Modified Rough Sets approach has 75% correct profiling with false positive of 10% and false negative of 2%.

TABLE II. ASSOCIATION RULE STATISTICS

| Decision Attribute | Association Rules | | | No. Profiling Rules |
|---|---|---|---|---|
| | No. | Min Conf | Max Conf | |
| Matriculated | 43 | 75% | 79% | 1 |
| Not Matriculated | 8 | 68% | 71% | 1 |

Profile 1:
If Parent Zip = 3 $\wedge$ SAT =3, Then Matriculation = 0 (Conf = 68%)
Profile 2:
If Parent Zip = 2 $\wedge$ Gender = 1 $\wedge$ SAT $\geq$ 2 $\wedge$ Family contribution = 5, Then Matriculation = 1 (Conf= 75%)

TABLE III. APPROXIMATE RULE STATISTICS

| Decision Attribute | Approximate Rules | | | No. Profiling Rules |
|---|---|---|---|---|
| | No. | Min Conf | Max Conf | |
| Matriculated | 107 | 60% | 100% | 4 |
| Not Matriculated | 159 | 60% | 100% | 2 |

Profile 1:
If Parent Zip = 1 $\wedge$ Family Contribution = 0 $\wedge$ GPA>=4.00 $\wedge$ SAT >1200
Then Matriculation =1 (100%)
Profile 2
If Parent Zip = 1 $\wedge$ Family Contribution = 4 $\wedge$ 3.00 <=GPA <=3.5
Then Matriculation =1 (60%)
Else Matriculation =0 (40%)
Profile 3:
If Parent Zip = 1 $\wedge$ Family Contribution = 5 and GPA >=3.00
Then Matriculation =1 (100%)
Profile 4:
If Parent Zip = 3 $\wedge$ Family Contribution = 5 and GPA <3.00,
Then Matriculation =1 (66%)
Profile 5:
If Parent Zip = 3 $\wedge$Family Contribution = 0,
Then Matriculations = 0;

 ISBN: 978-1-61208-237-0 

TABLE IV. RESULTS

| | No. Record Match the Profiles | No. Record with Correct Profiling | No. Record with Incorrect Profiling | No. Record Match No Profiles |
|---|---|---|---|---|
| Apriori Algorithm | 152 | 138 | 14 | 7 |
| Modified Rough Set | 137 | 119 | 18 | 22 |

## V. CONCLUSION

The results in Table 4 reveal that 81% and 75% of the records have been correctly profiled for the predicates of Decision(Matriculated) and Decision(Not Matriculated) by the Apriori Algorithm. For the Modified Rough Sets approach the success rate is 50% for Decision(Matriculated) and 100% for Decision (Not Matriculated). The results seem rather diverse between the two algorithms. The reason may stem from the fact that the Apriori Algorithm acts at the condition predicates level, whereas the Modified Rough Sets approach acts at the record level. In other words, the first approach builds a set of the most frequent predicates regardless of the concern about record boundaries, but the Modified Rough Sets build the approximate rules with rigid concern about the record boundaries.

For the current dataset hand, results show profile-based recruiting of new students may save time and money. The saving is accomplished by concentrating the recruiting efforts on specific geographical areas with potential students who will matriculate.

## REFERENCES

[1] Sun N. and Z. Yang, ”Equilibria and indivisibilities: gross substitutes and complements”, Econometrica, 2004; 74-5: 1385-1402.

[2] Abizada A., “Pairwise stability and strategy-proofness for college admissions with budget constraints”, Social Science Electronic Publishing, Inc. 2012.

[3] Han J. and Kamber M, *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers, 2001.

[4] Pawlak Z. “Rough Classification”, *Journal of Man-Machine Studies* 1984; 20: 469-83.

[5] Hashemi R., Pearce B., Arani R., Hinson W., Paule M. “A Fusion of Rough Sets, Modified Rough Sets, & Genetic Algorithms for Hybrid Diagnostic Systems”, In: Lin TY, Cercone N Editors. *Rough Sets & Data Mining: Analysis of Imprecise Data*. Kluwer Academic Publishers, 1997. pp. 149-76.

[6] Hashemi R., Tyler A., Bahrami A., "Use of Rough Sets as a Data Mining Tool for Experimental Bio-Data”, In *Computational Intelligence in Biomedicine and Bioinformatics: Current Trends and Applications*, Tomasz G. Smolinski, Mariofanna G. Milanova, and Aboul Ella Hassanien, Editors, Springer-Verlag Publisher, June 2008, pp. 69-91.

[7] Hashemi R., Choobineh F., Slikker W., and Paule M., "A Rough-Fuzzy Classifier for Database Mining", The *International Journal of Smart Engineering System Design*, No. 4, 2002, pp.107-114.

 ISBN: 978-1-61208-237-0 

# Engineering Adaptation: A Component-based Model

**Nikola Šerbedžija**
Fraunhofer FIRST
Berlin, Germany
nikola.serbedzija@first.fraunhofer.de

***Abstract*—The novel concept of user-centric pervasive adaptive systems has been designed to deliver services adapted to our needs and wishes according to the context of use. Engineering adaptation is a cross disciplinary endeavour requiring synergy of computer and human sciences as well as the practice. This work describes a novel *reflective approach* for development and deployment of pervasive adaptive systems. Special focus is on *reflective architecture* which uses component and service based programming model for developing the reflective framework as a generic support for the pervasive adaptive systems. A strong pragmatic orientation of the component-based approach is illustrated by a prototype named affective music player.**

***Keywords–Adaptive Systems; Autonomous Behaviour; Component-based Systems.***

## I. INTRODUCTION

Seamless and implicit human-computer interaction is an important characteristic of smart technology [1]. The 'smart' attribute is achieved by intuitive system control, based on the context assessment. This allows the system to function autonomously, without the requirement for explicit user intervention. In other words, the user becomes a part of the control loop, making system reaction adaptive and appropriate to users' behaviour.

Most of the present systems that deal with personal human experience are poorly engineered [1]. As a consequence, maintenance and modification of smart applications are very difficult. Furthermore, re-usability as capability of re-deploying the same software structures in different application domains is not possible. The presented reflective approach adds complexity to the current pervasive adaptive[2] systems by introducing seamless and implicit man-machine interaction based on emotional, cognitive and physical experience. At the same time it strives at generic, flexible and re-usable solutions that should overcome the poor engineering problems.

In effort to mimic the adaptation process, as it appears in the nature, and to apply it within man-machine interaction, reflective approach deploys the biocybernetic loop to make users' psychophysiological data a part of computer control logic [2][3]. The function of the loop is to monitor changes in user's state in order to initiate an appropriate computer response. This approach also takes results of affective/physiological computing [3] and combines it with high level understanding of social and goal–oriented situations. Biocybernetic loop [4] is implemented with the help of sense-analyze-react control troika. Firstly, reflective ontology [5] classifies numerous factors that determine user's states, social situation and application goals, defining elements for decision making. The ontology is then expressed in a number of XML-based taxonomies that allow for a uniform deployment in data acquisition, user's state diagnoses and activation of corrective actions. Finally the component based framework is developed with a goal to support adaptation process and deployment of adaptive applications [6].

The rest of the paper focuses on software engineering strategies of reflective approach. Firstly, the adaptation concept is presented, followed by technical blueprints of the component based architecture for adaptive man-machine interaction. Finally, the application of this approach is illustrated by the prototype music player, a system that controls the player according to the listener emotional state. The conclusion summarizes the work described and indicates further challenges and research topics in the domain of adaptive systems.

## II. ADAPTATION STRATEGY

The overall goal of reflective systems [7] is to create a software framework that controls and adapts the environment (a home, an office or an automotive environment) according to the users' situation. To be able to perform this task, a system must be able to perceive its environment through sensors and influence it through actuators. Therefore, a reflective application always consists of hardware (sensors and actuators) and software (reading sensor values, controlling the application and operating actuators) that together with users build the application context.

Controlling the environment through software can be done in two different ways: as feed-forward system (Figure 1a, also called open control loop), or as feedback system (Figure 1b, also called closed control loop).

A feed-forward system does not take into account the reactions of the system under control, but only the environment under which it operates: in the reflective domain, this would amount to the control of the environment without observing the reactions of the user.

 ISBN: 978-1-61208-237-0 

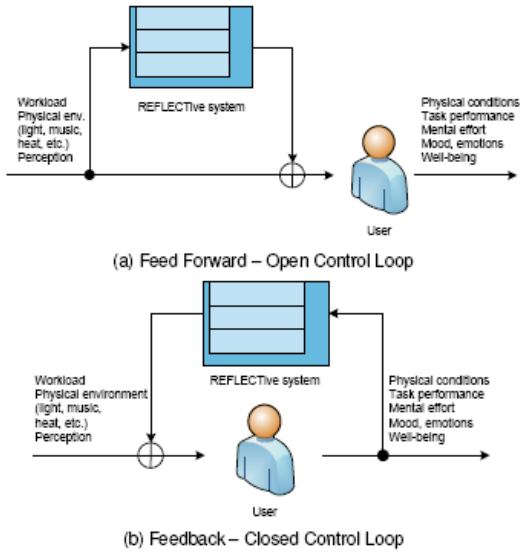



Figure 1. User centric system

Feed-forward systems offer quick response and high performance with few needed sensors – if they can be implemented. The main difficulty in creating feed-forward systems is that the impact and effects of the environment and the controlled actuators must be precisely known, as they are not measured. E,g. the effects of changing the lighting in a room must be completely known under all conditions – obviously, a task too challenging and far-fetched for now. Instead, this approach focuses on creating feedback control systems, measuring and reasoning on its effects and the current behaviour of the user. By doing this, a reflective system may also become able to counter the effects of unknown factors influencing the users' well-being.

Feed-forward systems are sufficient for a user-friendly behaviour in situation where satisfying an average user goals are needed (e.g., word processing systems, etc.). However, as reflective systems should be sensitive to a personal user state (emotional, cognitive and physical) policy of "satisfying an average user" never works, as each user is different and the behaviour of the single user often differs in different circumstances. If a system is to meet more personal user needs, a self-tuning and self-correcting strategy is a *conditio sine qua non.*

Even if building of closed control loops instead of open control loops seems to be a more promising approach, it brings along additional challenges. First of all, in the setting that is investigated, it is never guaranteed that an action performed on the environment will show the desired effects or any effects at all: preferences vary greatly over persons and over time. The reactions to hard rock and rap music differ between individuals, but also the reaction to a certain lighting condition may differ depending on the user's context. Reflective systems will therefore have to self-assess their effects on their environment, and hold several alternative strategies ready to achieve their goal. Furthermore, extracting the user's conditions from low-level psycho-physiological measures or computing it through image processing is a challenging task with varying success depending on the user's context; detecting the facial expression under bad lighting conditions for example is very difficult, if not impossible.

These challenges call for a well-structured software solution that allows for flexible response to the user's environment, and self-assessment of its performance. In any case (both feed-forward and feed-back control), a troika "sense-analyse-react" needs a special consideration. The closed loop control for implicit human-machine interaction, based on user psycho-physiological state is often denoted as a biocybernetic loop [2][4]. The function of the loop is to monitor changes in user state in order to initiate an appropriate adaptive response. The biocybernetic loop is designed according to a specific rationale, which serves a number of specific meta-goals. For instance, the biocybernetic loop may be designed to: (1) promote and sustain a state of positive engagement with the software/task and (2) minimise health or safety risks inherent within the human-computer interaction.

Both biocybernetic loop deployment and reflective ontology that supports high level reasoning have been described elsewhere [4][5], as well as some of the reflective applications which deploy this technology [6][7]. This paper focuses on reflect component model and its organization.

## III. Reflective Component Architecture

The programming paradigm for building reflective applications is component-based [8]. Software components are units of software that make their communication capabilities and requirements explicit by means of ports: provided ports describe what communication the component can accept and process, while required ports describe the communication the component requires to perform its work. Ports are given types that describe the set of messages that can be received or sent. A component based system is comprised of a set of components and an assembly that describes the way they are connected; required ports can be connected to the provided ports with the same type by a connector.

Having components that make all communication requirements and capabilities explicit helps in several ways: this provides a simple framework for reusability of components. At the same time, components can be written without considering one specific application. A component is written with just its own communication in mind, and, at a later point and possibly by a different person, it is made part of an application. Hence, components are usually more generic than special-purpose algorithms and can be employed in different applications. Furthermore, a component-based system can be reconfigured at run-time [9].

 ISBN: 978-1-61208-237-0 

This reconfiguration can be parametric, i.e., the reconfiguration is performed by changing parameters of components that affect their behaviour. Reconfiguration can be also structural, i.e. the change of behaviour is achieved by removing components, adding components, or changing the interconnections between components.

Both kinds of reconfiguration allow to realize short-, mid-, and long-term adaptation of the system – as needed for different biocybernetic loops (as described elsewhere) [4][5]. Some components do not require communication partners, but merely provide communication. This especially holds true for wrappers of hardware devices like sensors or actuators. In accordance with common terminology, such components are called services.

Since services do not require communication partners, suitable services can be chosen by considering only the ports they provide. Especially for sensors and actuators, this enables automated discovery as well as dynamic response to the availability of new services on behalf of the architecture.

### A. Reflective Framework

Reflective systems are structured as feedback loops that sense the environment, analyse the gathered data, and react according to the analysed results. Consequently, applications are structured in three layers:

- The tangible layer,
- The reflective layer, and
- The application layer.

The detailed description of the reflective layered architecture can be found in [7][10]. The focus here is on reflective layer and its component-based organization. It forms a pool of reusable software components that can be used to analyse the set of features exposed by sensors, and components to coordinate actuators well as a generic reasoner and learning algorithm components. On the one hand, reasoners are used to provide a mapping from the current user and environmental state to plans. Learners on the other hand are used to adjust component parameters to the given user, and by this personalize the closed loop.

Considering the abstraction level hierarchy, the reflective layer consists of analysis and coordination components providing abstractions from sensor data (i.e., features), and coordinating the operation of multiple actuators. Analysis components create abstract representation of the user's state and his context from available data: among others the user's current mood, cognitive workload, and physical conditions. In the opposite direction, coordination components map high-level plans to operations of – possibly several – actuators. In the abstraction level hierarchy, the application layer consists of components reasoning only on the abstract representation of the user and his context created by the reflective layer components, and sends abstract plans to coordinators in the reflective layer.

In a perfect reflective application, the two hierarchies (reuse and abstraction) will coincide: the analysing and coordinating components will be reusable among several applications, while application-specific components will solely define high-level goals using high-level representations of state. For some reflective applications, however, there is still a gap between both hierarchies. This gap is due to the fact that it is yet not always possible to create a common high-level abstraction of the user state and environment on which application components can reason, while still yielding a satisfactory behaviour [10]. Therefore, the reuse hierarchy allows for separating reflective layer and application layer where both hierarchies diverge. Nevertheless, the ultimate goal to merge reuse and abstraction hierarchy remains.

### B. Reflective Component Model

Components are the major building elements of the reflective layered middleware. They are used to implement reflective system functionality at any level of abstraction and complexity (e.g., sensor input services, higher level diagnoses, and goal-based reasoning components are all represented by the same structural elements). In order to make a uniform and versatile component-based support, a comprehensive meta model has been designed, programmed in Java language and deployed in practice.

The structure of the reflective component is depicted in Figure 2. The class is declared to be a component by extending the "AComponent" class, and declaring a dependency through required ports. In order to satisfy those dependencies components have to be instantiated and connected with each other. In the configure method (in BundleContainer class), which is called on system start up, the developer can specify creation and connection rules. Connection rules may be either very precise or loose, as required by the system design. The functionality of a reflect entity is encapsulated within the component and function groups are placed within component containers. In order to communicate components can offer or require functionality, depending on whether a component sends or receives data (e.g., sensor services are modeled with components that only provide functionality, i.e., sensors' measurements). A same component may provide multiple functionality, in which case multiple ports are to be used (either different in case of different functionalities or instantiated, in the case of the same functionality). Required functionalities are obtained by connecting to the ports of other components that provide them. Connectors are responsible for tracking and binding the components. Component and bus manager represent core functions of the model, providing the access to other components, connectors and containers. With such a model, encapsulation, clear inter-component communication and a sound software composition are ensured. All these features contribute to faster development, testing and deployment of reflective software. Reflective framework offers all the needed functionality for this component model and serves as its run-time environment.

 ISBN: 978-1-61208-237-0 

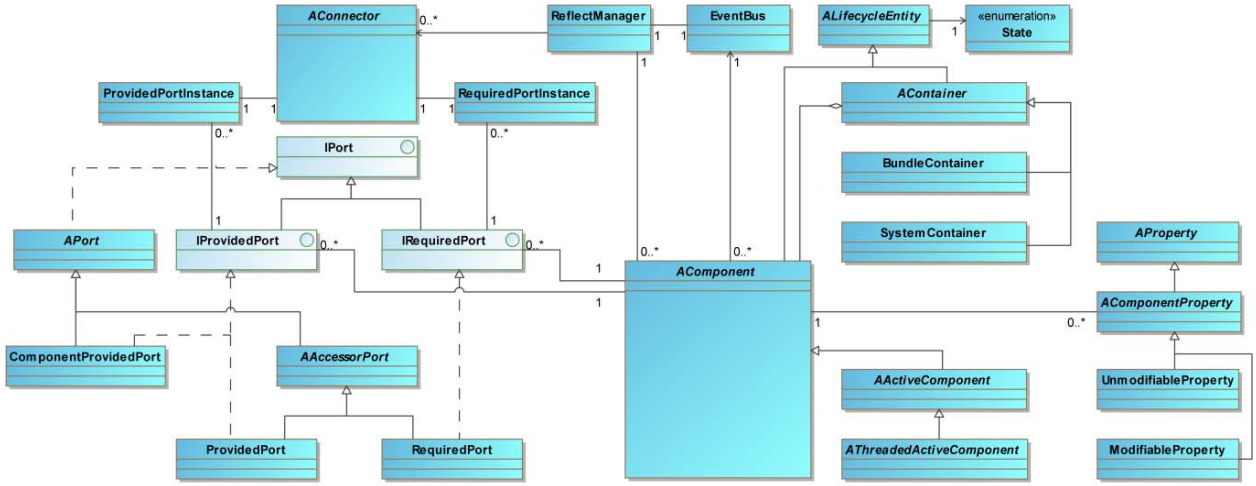



Figure 2. Reflective component model

### C. Implementation Details

The reflect component model is a Java-based component framework running on top of OSGi [11]. Using OSGi allows to dynamically load and unload Java classes at run-time, adding to the flexibility of the reflective approach. One of the main features of OSGi is the registry mechanism allowing to register objects as services based on interfaces and properties, and performing query services in the same way. In the reflect framework, these facilities are used to add capabilities to the reflect core component framework: user-level OSGi bundles can add component and connector factories, thereby enabling the core framework to create instances of these components and connectors.

Connectors are not predefined by the component framework, but are "fat" user-level entities [12][13]. In this way, a software developer has the flexibility to define the communication patterns between components using the user-defined connector, and the way the communication is affected by reconfiguration (e.g., whether the state of communication must be preserved under reconfiguration, or whether a communication protocol must be ended before reconfiguration may take place).

The reflective component framework defines three types of components:

**Passive components** offer functionality to other components by declaring one or more provided ports. They might call other components via their required ports, but only do so within the processing of communication received from their own provided ports.

**Active components** may also provide ports towards other components, but, unlike passive components, they also implement autonomous behaviour. This behaviour is realized by a thread running concurrently to the threads processing the received communication, as well as the threads of other components or the framework. An active component may issue calls to required ports both during processing of calls issued to its own provided ports and in the course of the autonomous loop.

**Composite components** are containers for other components. A composite component owns an assembly of components and connectors and encloses them with a facade that makes the composite component act like a normal active component. Provided and required ports of the composite component are delegated to compatible ports of enclosed components.

In the reflect component framework, components may additionally declare properties that can be used to query and manipulate its configuration or its state. These properties can be associated with constraints that are automatically enforced by the framework when the property is manipulated. Though it would also be possible to make the component state accessible via provided ports, properties offer a generic and very comfortable way for developing components that offer means for monitoring and adaptation.

Furthermore, the framework offers a message bus facility. Components can register to topics on the bus by method annotations declaring to topic listened to. This message bus facility can be used by components to listen to system events or as a way to easily realize broadcast-style communication between components.

The reflect architecture is realized with the help of OSGi and the reflect component framework defined on top of it. The reflect architecture therefore comes in a set of OSGi bundles.

- A reflect core bundle defining the core concepts.
- A reflect generic bundle defining the generic extensions of the core.
- A reflect ontology bundle containing standardized interfaces and data types from the ontology.
- A set of reflective layer bundles containing reusable components for analysis, coordination, and other purposes.

 ISBN: 978-1-61208-237-0 

By registering the reusable components on bundle start-up, they can be created by the reflect component framework. Then, application-specific assemblies can instruct the reflect component framework to create reusable components and connect them with sensors, actuators, and application-specific components.

## IV. Application Example

The reflect system has been tested on several prototype demonstrations, home ambient [14], vehicular support [6] and mood player [15].


Figure 3. PC as a mood player

The mood player, implemented on Samsung ultra PC as shown on Figure 3, deploys so called “music directs your mood” concept. The psychological background and the control strategy of the music player are described elsewhere [15]. A music player functions as a closed loop repeatedly measuring the current mood state of a user and selecting music from the user’s own music database depending on the current mood and a predetermined target mood state. Since a positive mood enhances several cognitive processes, the ability to improve mood is particularly interesting in driving or working situation [15], but also at home in a more relaxed setting.

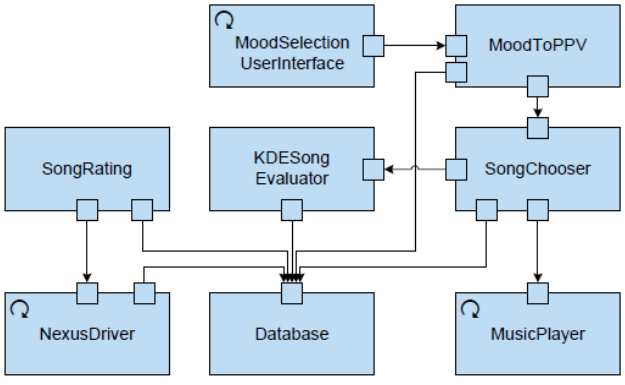



Figure 4. Component model of the mood player

Figure 4 illustrates the component model of the mood player. It embraces eight major components that are connected to each other. The horizontal layout indicates major architectural abstractions: (1) tangible layer (components controlling Nexus driver i.e., sensor devices, database and the actuator i.e., music player); (2) reflective layer (components performing song evaluation, rating and selection) and application layer (components performing user interface and the mood selection).

Table 1 illustrates some of the performance figures and implementation details of the mood player prototype. It can be seen that even on a low-performance computer, the system runs effectively and is flexible for further extensions. To our knowledge, there are no similar systems that could be used for the performance comparison.

Table 1: Mood player system figures

| ***Mood Player*** | | ***Comment*** |
|---|---|---|
| Size (source Java code) | 8,6K lines | Extra requirements:<br><br>1. Reflect Framework: 28,4K lines and<br>2. Monitoring tools: 19,4K lines |
| Run time memory use | 30 MB | Most of memory is used by OSGI and external libraries |
| External devices | Nexus 10 | Used for physiological measurements |
| PU load at an ultra PC | 60% | Min. Requirements:<br><br>Windows XP, 128 MB RAM, 60MB hard disk, TCP/IP |
| Development Environment | Java 1.6, Eclipse RCP Tool, Equinox OSGi, MySQL, Reflect custom developers tool (for interfacing external devices and rule engine) | |

## V. Conclusion

Developing pervasive adaptive computing applications is still a domain not well understood by classical software engineers. This often leads to poorly engineered and not well defined system structure. The inevitable result is difficult maintenance and poor extensibility of present systems.

To overcome these problems, reflective approach investigates software infrastructures and patterns for pervasive adaptive systems, characterised by seamless integration in the everyday environments. Reflective systems make use of available physical devices to sense and derive the state of their environment and their user, infer the user’s current context and conditions, and finally, try to improve the overall users’ conditions accordingly.

The paper discussed the requirements and challenges that software engineers have to deal with when implementing the adaptation phenomenon. A generic component framework has been introduced, geared towards ease of use and flexibility. It consists of three layered architecture. The goal of the three layered architecture is to

 ISBN: 978-1-61208-237-0 

provide the software engineer with prefabricated, generic components for analysis, coordination and dynamic re-configuration. It also encourages early prototyping as many parts of the systems can be simulated [16] by dummy services/components (having the same interface to the rest of the system) and later on substituted by the real parts. The reflect component framework offers an easy-to-use development environment that can be picked up easily by software engineers familiar with Java and OSGi. Reflective approach has been successfully tested in home ambient [7][14], public advertising [7] and vehicular domains [6,7].

The further work is oriented towards improving the communication modules of the reflective framework that should allow for the exchange of the information among different reflective applications in a pervasive manner. Other research topics cross the disciplines, as the techniques for diagnosing different human psychophysiological states need to be further improved, enlarging the application spectrum.

## ACKNOWLEDGEMENT

The author expresses his thanks to Dr. Andreas Schroeder from LMU Munich, who read the previous versions of this paper and plaid significant role in design and development of the reflective component model. Most of the work presented here has been done under the REFLECT project [7] (project number FP7-215893) and ASCENS project [17] (project number FP7- 257414), both funded by the European Commission within the 7th Framework Programme.

## *REFERENCES*

[1] D.A. Norman. *The Design of Future Things*. 2007, New York: Basic Books.

[2] A.T. Pope, E.H. Bogart and D.S. Bartolome. Biocybernetic system evaluates indices of operator engagement in automated task. *Biological Psychology*, 40, 1995, 187-195.

[3] S.H. Faircloug. Fundamentals of physiological computing. *Interacting with Computers*, 21, 2009, pp. 133-145.

[4] N.S. Serbedzija and S. Fairclough. Reflective Pervasive Systems. ACM Transactions on Autonomous and Adaptive Systems (TAAS), Vol. 7 (1), April 2012.

[5] G. Kock, M. Ribaric and N. Serbedzija. Modelling User-Centric Pervasive Adaptive Systems - the REFLECT Ontology. In: *Intelligent Systems for Knowledge Management*, Vol. 252. Nguyen, Ngoc Thanh; and Szczerbicki, Edward Eds. Series: "Studies in Computational Intelligence", Springer 2009, ISBN: 978-3-642-04169-3.

[6] N. Serbedzija, A. Calvosa and A. Ragnoni. Vehicle as a Co-Driver, *Proc. 1st Annual International Symposium on Vehicular Computing Systems - ISVCS 2008*, Dublin, Ireland.

[7] REFLECT. REFLECT project - Responsive Flexible Collaborating Ambient, available at, http://reflect.first.fraunhofer.de (2012).

[8] C. Szyperski. *Component Software: Beyond Object-Oriented Programming*. ACM Press and Addison-Wesley, 1998.

[9] J.B. Bradbury. Organizing definitions and formalisms of dynamic software architectures. *Technical Report 2004-477*, Queen's University, 2004.

[10] A. Schroeder, M. Zwaag and M.A. Hammer. Middleware Architecture for Human-Centred Pervasive Adaptive Applications, *Proc.1st PerAda Workshop at SASO 2008*, Venice, Italy, Oct. 21th 2008.

[11] OSGi Alliance. OSGi service platform release 4. http://www.osgi.org/, 2005.

[12] K.K. Lau, P.V. Elizondo and Z. Wang. Exogenous connectors for software components. *8th International SIGSOFT Symposium on Component-based Software Engineering, volume 3489 of Lecture Notes in Computer Science*, Springer, 2005, pp. 90–106.

[13] T. Bures, P. Hnetynka and F. Plasil. Runtime concepts of hierarchical software components. *International Journal of Computer & Information Science*, 2007, 8:454–463.

[14] N. Serbedzija, Reflective Assistance for Eldercare Environments, *SEHC'10, Proceedings of Second Workshop on Software Engineering in Health Care*, Cape Town, May 2010.

[15] J.H. Janssen, E.L. Broek, and J. Westerink. Personalized affective music player. *Proceedings of the 2009 International IEEE Conference on Affective Computing and Intelligent Interaction (ACII)*, September 10-12, 2009, Amsterdam, pp. 472-477.

[16] N.S. Serbedzija. MACS – Modular Affective Computing Simulator, *Proc. of Applied Simulation and Modelling*, 2008, June 23 - 25, 2008, Corfu, Greece.

[17] ASCENS. ASCENS project – Autonomic Service Component Ensembles, available at, http://www.ascens-ist.eu/, 2011.

 ISBN: 978-1-61208-237-0 

# Minmax Regret 1-Center on a Path/Cycle/Tree

Binay Bhattacharya, Tsunehiko Kameda, and Zhao Song
School of Computing Science, Simon Fraser University
University Drive, Burnaby, Canada V5A 1S6
Email: {binay, tiko, zhaos}@sfu.ca

***Abstract*—In a facility location problem in computational geometry, if the vertex weights are uncertain one may look for a "robust" solution that minimizes "regret." The best previously known algorithm for finding the minmax regret 1-center in a tree with positive vertex weights is due to Yu *et al.*, and runs in sub-quadratic asymptotic time in the number of vertices. Assuming that the minimum weight of at least one vertex is non-negative, we present a new, conceptually simpler algorithm for a tree with the same time complexity, as well as an algorithm that runs in linear (respectively sub-quadratic) time for a path (respectively cycle).**

***Index Terms*—facility location; 1-center; minmax regret optimization**

## I. Introduction

Deciding where to locate facilities to minimize the communication or transportation costs is known as the *facility location problem*. For a recent review of this subject, the reader is referred to [1]. The cost of a vertex is formulated as the distance from the nearest facility weighted by the vertex weight. In the *minmax regret* version of this problem, there is uncertainty in the weights of the vertices and/or edge lengths, and only their ranges are known. Chen and Lin (Theorem 1 in [2]) proved that in solving this problem, the edge lengths can be set to their maximum values, when the vertex weights are non-negative. Our model assumes that the edge lengths are fixed and uncertainty is only in the weights of the vertices.

A particular *realization* (assignment of a weight to each vertex) is called a *scenario*. Intuitively, the minmax regret 1-center problem can be understood as a 2-person game as follows. The first player picks a location $x$ to place a facility. The opponent's move is to pick a scenario $s$. The payoff to the second player is the cost of $x$ minus the cost of the 1-center, both under $s$, and he wants to pick the scenario $s$ that maximizes his payoff. Our objective (as the first player) is to select $x$ that minimizes this payoff in the worst case (i.e., over all scenarios).

The rest of the paper is organized as follows. Section II reviews work relevant to our problem. In Section III, we define the terms that are used throughout the paper, and cite or prove some center-related properties. Section IV first discusses how to compute the centers and their costs under a certain set of scenarios. We then present a linear time algorithm for computing the minmax regret 1-center of a path. In Section V, we present an algorithm for a cycle. Finally, Section VI presents a simplified algorithm for a tree, and Section VII concludes the paper.

## II. Related Work and Our Objectives

The problem of finding the minmax regret 1-center on a network, and a tree in particular, has been attracting great research interest in recent years. Several researchers have worked on this problem. The classical $p$-center problem is discussed by Kariv and Hakimi in [3]. Megiddo [4] computes the classical 1-center of a tree with non-negative vertex weights in $O(n)$ time, where $n$ is the number of vertices. Averbakh and Berman [5] proved some basic results in the minmax regret 1-center problem. More recently, Yu et al. [6] solved the problem in $O(mn \log n)$ (resp. $O(n \log^2 n)$) time for a general network (resp. tree network) with positive vertex weights, where $m$ is the number of edges.

Since the known algorithm for a general network is relatively inefficient, we look for more efficient algorithms for special families of networks. Yu et al. [6] proposed an $O(n \log^2 n)$ time algorithm for a tree with positive vertex weights. Their algorithm is rather involved, so we want to come up with a conceptually simpler algorithm, under the more relaxed assumption that at least one vertex has the maximum weight that is positive. A cycle is the simplest building block of any network that is more general than a tree, but to the best of our knowledge, nothing is known about the complexity of finding the minmax regret 1-center on a cycle. We want to design an efficient algorithm that is specifically tailored to a cycle network.

## III. Preliminaries

As stated above, we assume that there is at least one vertex whose minimum weight is non-negative. Then it is clear that all vertices whose weights are negative can be ignored, because they cannot influence the 1-center. Therefore, we assume without loss of generality that the minimum weights of all vertices are non-negative.

Let $G = (V, E)$ be a path, cycle or tree network with $n$ vertices. We also use $G$ to denote the set of all points (vertices and points on edges) on $G$. Each vertex $v \in V$ is associated with an interval of integer weights $W(v) = [\underline{w}_v, \overline{w}_v]$, where $0 \leq \underline{w}_v \leq \overline{w}_v$, and each edge $e \in E$ is associated with a positive length (or distance). We assume that the distances between a point on an edge and its end vertices are prorated fractions of the edge length. For any pair of points $p, q \in G$, the shortest distance between them is denoted by $d(p, q)$. Let $\mathcal{S}$ be the Cartesian product of all $W(v),\ v \in V$:

$$\mathcal{S} \triangleq \prod_{v \in V} [\underline{w}_v, \overline{w}_v].$$

 ISBN: 978-1-61208-237-0 

The *cost* of a point $x \in G$ *with respect to* $v \in V$ under scenario $s$ is $d(v,x)w_v^s$, where $w_v^s$ denotes the weight of $v$ under $s$, and the *cost* of $x$ under $s$ is defined by

$$F^s(x) \triangleq \max_{v \in V} d(v,x) w_v^s. \quad (1)$$

The point $x$ that minimizes $F^s(x)$ is called a *(classical) 1-center* under $s$, and is denoted by $c(s)$. Throughout this paper the term *center* refers to this weighted 1-center. The difference

$$R^s(x) \triangleq F^s(x) - F^s(c(s)) \quad (2)$$

is called the *regret* of $x$ under $s$. We finally define the *maximum regret* of $x$ as

$$R^*(x) \triangleq \max_{s \in \mathcal{S}} R^s(x). \quad (3)$$

The scenario that maximizes $R^s(x)$ for a given $x \in G$ is called the *worst case scenario* for $x$. Note that $R^*(x)$ is the maximum payoff with respect to $x$ that we mentioned in the Introduction. We seek location $x^* \in G$, called the *minimum regret 1-center*, that minimizes $R^*(x)$.

Let $V = \{v_1, v_2, \ldots, v_n\}$. For simplicity, we use $w_i = w_{v_i}$ in what follows. The *base scenario*, $s_0$, is defined by $w_i^{s_0} = \underline{w}_i$ for all $i$. A vertex $v$ that maximizes (1) is said to be a *critical* vertex for $x$ [5]. For $i = 1, 2, \ldots, n$, let us define the *single-max scenario* $s_i$ by $w_i = \overline{w}_i$ and $w_j = \underline{w}_j$ for all $j \neq i$, and let $\mathcal{S}^*$ denote the set of all single-max scenarios. Averbakh and Berman proved the following fundamental theorem for the trees, but the same proof is valid for the general network.

*Theorem 1:* [5] For any point $x$ in a network, there is a worst case scenario $s_i \in \mathcal{S}^*$. Vertex $v_i$ is a critical vertex for $x$. ■

Theorem 1 implies

$$R^*(x) = \max_{s \in \mathcal{S}^*} R^s(x). \quad (4)$$

Therefore, we need to consider only the single-max scenarios in looking for the minmax regret location $x^*$.

The cost of $x \in G$ with respect to $v_i$ is given by

$$f_i(x) \triangleq d(x, v_i) w_i. \quad (5)$$

It will turn out that we need to consider either

$$\underline{f}_i(x) = d(x, v_i)\underline{w}_i \text{ or } \overline{f}_i(x) = d(x, v_i)\overline{w}_i, \quad (6)$$

which is called the *min-weight cost* and *max-weight cost* with respect to $v_i$, respectively. To evaluate (2) for $s = s_j$, we need to find $c(s_j)$ first. Most of our effort will be on how to compute $c(s_j)$ efficiently.

## IV. Path Network

We start with a path, which is the simplest network.

### A. Computing the Upper Envelope of a Set of Line Segments

Assume that the vertices of a path $\mathcal{P} = (V, E)$ are laid out horizontally, in the order $v_1, v_2, \ldots, v_n$ from left to right. Let $L_i(x)$ represent an increasing function, such as $\underline{f}_i(x)$, defined for $x$ to the right of $v_i$. Its value gets larger as $x$ moves right from $v_i$. To find $c(s_j)$ efficiently, we construct the upper envelope of $\{L_i(x) \mid i = 1, 2, \ldots, n\}$, assuming $w_1 \geq 0$. If $w_i \geq w_j$ for $i < j$, then we have $L_i(x) > L_j(x)$ for all $x$ (to the right of $v_j$, where $L_j(x)$ is defined), and $L_j(x)$ won't be a part of the upper envelope. In this case, we say that $L_i(x)$ *dominates* $L_j(x)$, and can ignore $L_j(x)$. Therefore, the following algorithm assumes that $w_i < w_j$ holds for any $i$ and $j$ such that $i < j$.

**Algorithm** `Find-Envelope/Path`

1) Push $[1 : (v_1, 0)]$ in stack $\Sigma$. (Meaning: $L_1(x)$, starting at $(v_1, 0)$, is the initial part of the upper envelope.)
2) For $k = 2, 3, \ldots, n$, carry out steps 3 to 5.
3) Let $[i : (x_i, y_i)]$ be the item at the top of $\Sigma$. If line $L_k(x)$ passes above $(x_i, y_i)$ pop up the top item from $\Sigma$ and repeat this step.
4) Compute the intersection $(x_k, y_k)$ of $L_k(x)$ and $L_i(x)$. If $x_k$ is to the left of $v_n$, then push $[k : (x_k, y_k)]$ into $\Sigma$.

■

*Lemma 1:* When Algorithm `Find-Envelope/Path` completes, an entry $[i : (x_i, y_i)]$ in stack $\Sigma$ means $L_i(x)$ forms a segment of the upper envelope of $\{L_j(x) \mid j = 1, 2, \ldots, n\}$, starting at point $(x_i, y_i)$. `Find-Envelope/Path` runs in $O(n)$ time.

*Proof:* The first part is obviously true of the entry $[1 : (v_1, 0)]$, which is the bottom entry of $\Sigma$ that is never removed. Suppose $[j : (x_j, y_j)]$ (resp. $[i : (x_i, y_i)]$) was the entry at the top (resp. 2nd from the top) of $\Sigma$, when line $L_k(x)$ is processed. See Figure 1. If $L_k(x)$ is like the dotted line, then step 3

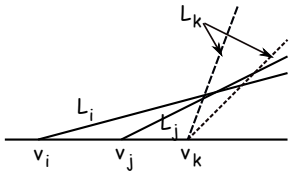



Fig. 1. Computing the upper envelope of lines $L_i(x)$, $L_j(x)$, and $L_k(x)$.

computes the intersection of $L_k(x)$ and $L_j(x)$ and pushes $[k : (x_k, y_k)]$ into $\Sigma$, indicating that $L_k(x)$ is the next segment starting at $(x_k, y_k)$. If $L_k(x)$ is like the dashed line, then step 4 computes the intersection $L_k(x)$ and an earlier line $L_i(x)$, and pushes $[k : (x_k, y_k)]$ into $\Sigma$, while step 3 discards $L_j(x)$. In this case $[i : (x_i, y_i)]$ and $[k : (x_k, y_k)]$ become adjacent entries in $\Sigma$, indicating that the line segment from $(x_i, y_i)$ to $(x_k, y_k)$ is a part of the upper envelope. It is easy to show that the algorithm runs in $O(n)$ time. ■

For the base scenario, $s_0$ (defined in Section III), let $F_r^{s_0}(x)$ denote the upper envelope of the set of functions, $\{\underline{f}_i^r(x) \mid 1 \leq i \leq n\}$, where $\underline{f}_i^r(x) = \underline{f}_i(x)$ for $x$ to the right of $v_i$, and 0 otherwise. A typical function $F_r^{s_0}(x)$ is plotted in

Figure 2. Symmetrically, let $F_l^{s_0}(x)$ denote the upper envelope

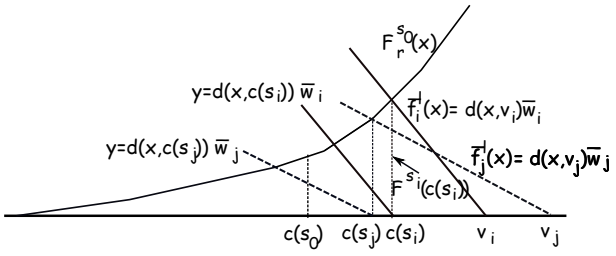



Fig. 2. Intersection of $\overline{f}_i^l(x)$ and $F^{s_0}(x)$ gives $c(s_i)$.

of $\{\underline{f}_i^l(x) \mid 1 \le i \le n\}$, where $\underline{f}_i^l(x) = \underline{f}_i(x)$ for $x$ to the left of $v_i$, and 0 otherwise. The value of $F_l^{s_0}(x)$ gets smaller as $x$ approaches $v_i$ from left. We have

$$F^{s_0}(x) = \max\{F_l^{s_0}(x), F_r^{s_0}(x)\},$$

and the 1-center under $s_0$, $c(s_0)$, is clearly at the lowest point of $F^{s_0}(x)$, namely at the intersection of $F_l^{s_0}(x)$ and $F_r^{s_0}(x)$. From Lemma 1 it follows that

*Lemma 2:* [4] The 1-center under $s_0$, $c(s_0)$, and $F_l^{s_0}(c(s_0)) = F_r^{s_0}(c(s_0))$ can be computed in $O(n)$ time. ■

### B. Regret Function

Our next objective is to find $R^*(x)$. But to do so, we need to know *some* of the centers in $\{c(s_i) \mid i = 1, 2, \ldots, n\}$. Center $c(s_i)$ can be found by computing the intersection of $\overline{f}_i(x) = d(x, v_i)\overline{w}_i$ and $F^{s_0}(x)$. Recall that $F^{s_0}(x) = F_l^{s_0}(x)$ (resp. $= F_r^{s_0}(x)$) for $x$ to the left (resp. right) of $c(s_0)$. Let $c(s_0)$ be on edge $(v_m, v_{m+1})$. Let $\overline{f}_i^l(x) = \overline{f}_i(x) = d(v_i, x)\overline{w}_i$ for $x$ to the left of $v_i$, and 0 otherwise. Figure 2 shows the intersection of $\overline{f}_i^l(x)$ and $F_r^{s_0}(x)$, where $m+1 \le i \le n$. If $\overline{f}_i^l(x)$ passes below the lowest point of $F^{s_0}(x)$ at $x = c(s_0)$, then we have $c(s_i) = c(s_0)$. For $1 \le i \le m$, we can similarly compute $c(s_i)$, as the intersection of $\overline{f}_i^r(x)$ and $F_l^{s_0}(x)$. Again, if $\overline{f}_i^r(x)$ passes below the lowest point of $F^{s_0}(x)$ at $x = c(s_0)$, then we have $c(s_i) = c(s_0)$.

Function $R^*(x)$ will be convex, and the optimal location $x^*$ will correspond to its lowest point. By definition, $s_i$ is obtained from $s_0$ by changing $w_i$ from $\underline{w}_i$ to $\overline{w}_i$. From (1), we have

$$\begin{aligned} F^{s_i}(x) &= \max_{v \in V} d(v, x) w_v^{s_i} \\ &= \max\{\overline{f}_i(x), F^{s_0}(x)\}. \end{aligned} \tag{7}$$

Cf. Lemma 3.5 in [6]. From (2), (4) and (7), we get

$$\begin{aligned} R^*(x) &= \max_{s_i \in \mathcal{S}^*} \{\max\{\overline{f}_i(x), F^{s_0}(x)\} - F^{s_i}(c(s_i))\} \\ &= \max_{s_i \in \mathcal{S}^*} \{\max\{d(c(s_i), x)\overline{w}_i, F^{s_0}(x) - F^{s_i}(c(s_i))\}\} \\ &= \max\{\max_{s_i \in \mathcal{S}^*} \{d(c(s_i), x)\overline{w}_i\}, F^{s_0}(x) - F^{s_k}(c(s_k))\}, \end{aligned} \tag{8}$$

where $F^{s_k}(c(s_k)) = \min_{s_i \in \mathcal{S}^*} F^{s_i}(c(s_i))$. Therefore, $R^*(x)$ can be computed as the upper envelope of $\{d(c(s_i), x)\overline{w}_i \mid s_i \in \mathcal{S}^*\}$ and $F^{s_0}(x) - F^{s_k}(c(s_k))$.

Figure 2 shows $F_r^{s_0}(x)$ and two cost functions $\overline{f}_i^l(x)$ and $\overline{f}_j^l(x)$, where $i < j$ and $\overline{w}_i > \overline{w}_j$. If $c(s_i)$ lies to the right of $c(s_j)$, then $\overline{f}_i^l(x) - F^{s_i}(c(s_i)) = d(c(s_i), x)\overline{w}_i$ dominates $\overline{f}_j^l(x) - F^{s_j}(c(s_j)) = d(c(s_i), x)\overline{w}_j$, and thus $s_j$ can be discarded. In this case, $c(s_j)$ need not even be computed, saving time.

**Algorithm** `Find-Nondominated`
Push point $p = (c(s_0), F_r^{s_0}(c(s_0)))$ into stack $\mathcal{ND}$.
For $i = m+1, \ldots, n$, do the following:

1) Find if $\overline{f}_i^l(x)$ passes above point $p = (x_p, y_p)$ at the top of $\mathcal{ND}$.
2) If it does, compute the intersection of $\overline{f}_i^l(x)$ and $F_r^{s_0}(x)$, and push it into $\mathcal{ND}$. Otherwise, $c(s_i)$ is to the left of $x_p$, and $d(c(s_i), x)\overline{w}_i$ could not be a part of the upper envelope. In Figure 2, $s_j$ satisfies this condition. ■

*Lemma 3:* Algorithm `Find-Nondominated` runs in $O(n)$ time.

*Proof:* By Lemma 2, we can compute $(c(s_0), F_r^{s_0}(c(s_0)))$ in $O(n)$ time. Step 1) of the above algorithm entails evaluating $\overline{f}_i^l(x_p)$, and takes constant time. In Step 2), we check successive segments of $F_r^{s_0}(x)$ upwards, starting at $p = (x_p, y_p)$. We never backtrack along $F_r^{s_0}(x)$, so that the total time required is $O(n)$. ■

Let us rewrite $\{d(c(s_i), x)\overline{w}_i \mid s_i \in \mathcal{S}^*\}$ in (8) as follows:

$$\begin{aligned} \{d(c(s_i), x)\overline{w}_i \mid s_i \in \mathcal{S}^*\} &= \{d(c(s_i), x)\overline{w}_i \mid 1 \le i \le m\} \\ &\quad \cup \{d(c(s_i), x)\overline{w}_i \mid m+1 \le i \le n\} \end{aligned} \tag{9}$$

In order to find the upper envelope of the functions in $\{d(c(s_i), x)\overline{w}_i \mid m+1 \le i \le n\}$ and identify its linear sections, we use Algorithm `Find-Envelope/Path`, with left and right reversed, starting at the right end of the path. This envelope is a continuous, piecewise linear, decreasing function of $x$, if we move $x$ from $v_1$ towards $v_n$. We combine it with $F^{s_0}(x) - F^{s_k}(c(s_k))$, according to (8), to find the left half of $R^*(x)$. Analogously, we can find the right half of $R^*(x)$, based on $\{d(c(s_i), x)\overline{w}_i \mid 1 \le i \le m\}$, which is a continuous, piecewise linear, increasing function of $x$. The lowest point of $R^*(x)$ can be found in $O(n)$ time.

*Theorem 2:* The minmax regret 1-center of a path network can be computed in $O(n)$ time. ■

## V. Cycle Network

Let $\mathcal{C} = (V, E)$ be a cycle with length $L_{\mathcal{C}}$. For $p, q \in \mathcal{C}$, the part of $\mathcal{C}$ clockwise, from $p$, to $q$ is denoted by $\mathcal{C}(p, q)$, and its length is denoted by $d(p, q)$. The *cost* of a point $x \in \mathcal{C}$ under $s$ is given by

$$F^s(x) = \sum_{v \in V} \min\{d(v, x), d(x, v)\} w_v^s. \tag{10}$$

Suppose that the vertices of $\mathcal{C}$ are numbered cw as $v_1, \ldots, v_n$ from the *origin* $o$ that lies in the interior of edge $(v_n, v_1)$. If $d(p, q) = d(q, p)$, we say that points $p$ and $q$ are *antipodal* to each other, and $p$ (resp. $q$) is the *antipode* [7] of $q$ (resp. $p$), denoted by $p = \alpha(q)$ (resp. $q = \alpha(p)$).

Recall the base scenario $s_0$ and the set $\mathcal{S}^*$ of single-max scenarios from Section III. To evaluate (2) for different scenarios and points, our first task is to find the centers $\{c(s_i) \,|\, s_i \in \mathcal{S}^*\}$.

### A. Finding Upper Envelope

The cost line of vertex $v_i$ under base scenario $s_0$, is $\underline{f}_i(x) = d(x, v_i)\underline{w}_i$. Figure 3 shows $\underline{f}_2(x)$, $\underline{f}_3(x)$, and $\underline{f}_4(x)$. They are plotted over two periods of a cycle, so that for any point $p \in \mathcal{C}$, one of the two occurrences of $p$ in it has the property that $\mathcal{C}(\alpha(p), p)$ and $\mathcal{C}(p, \alpha(p))$ appear continuously in the diagram. This property becomes useful when we process "queries" later. Observe that $\underline{f}_i(x)$ takes the minimum (resp. maximum) value

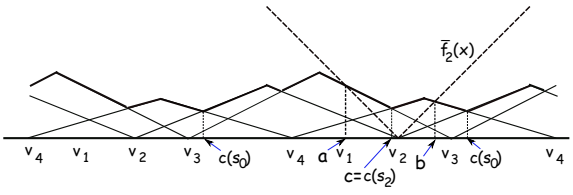



Fig. 3. The upper envelope is shown in thick line segments.

at $x = v_i$ (resp. $x = \alpha(v_i)$). The center $c(s_0)$ is at the lowest point of the upper envelope of $\{\underline{f}_i(x) \,|\, i = 1, 2, \ldots, n\}$.

The following algorithm constructs the *clockwise (cw) upper envelope*, $\mathcal{E}_{cw}$, of $\{L_i(x) \mid i = 1, 2, \ldots, n\}$, where $L_i(x) = \underline{f}_i(x)$ for $x \in \mathcal{C}(v_i, \alpha(v_i))$ and $L_i(x) = 0$ for $x \in \mathcal{C}(\alpha(v_i), v_i)$. It is very similar to `Find-Envelope/Path`.

**Algorithm** `Find-Envelope/Cycle`

1) Put $[1 : (v_1, 0)]$ in stack $\Sigma$. (Meaning: $L_1(x)$, starting at $(v_1, 0)$ is the initial part of the upper envelope.)
2) For $k = 2, 3, \ldots, n$, carry out steps 3 and 4.
3) Let $[i : (x_i, y_i)]$ be the item at the top of $\Sigma$. If line $L_k(x)$ passes above $(x_i, y_i)$ pop up the top item from $\Sigma$ and repeat this step.
4) Compute the intersection $(x_k, y_k)$ of $L_k(x)$ and $L_i(x)$. If the intersection lies within $L_{\mathcal{C}}/2$ clockwise from $v_i$, then push $[k : (x_k, y_k)]$ into $\Sigma$. ■

Clearly, there are $O(n)$ linear segments in $\mathcal{E}_{cw}$. The following lemma can be proved similarly to Lemma 1.

*Lemma 4:* When `Find-Envelope/Cycle` applied to a cycle completes, an entry $[i : (x_i, y_i)]$ in stack $\Sigma$ means $L_i(x)$ forms a segment of the upper envelope of $\{L_j(x) \mid j = 1, 2, \ldots, n\}$, starting at point $(x_i, y_i)$. `Find-Envelope/Cycle` runs in $O(n)$ time. ■

Similarly, we can compute the *counterclockwise (ccw) upper envelope*, $\mathcal{E}_{ccw}$, in $O(n)$ time.

*Lemma 5:* The 1-center, $c(s_0)$, of the base scenario on a cycle network can be found in $O(n)$ time.

*Proof:* After computing $\mathcal{E}_{cw}$ and $\mathcal{E}_{ccw}$ in $O(n)$ time, we compute their upper envelope, $\mathcal{E}$, in linear time. The lowest point of the merged envelope corresponds to $c(s_0)$. ■

For each single-max scenario $s_j \in \mathcal{S}^*$, we consider $\overline{f}_j(x)$ as its *query function*. In the example shown in Figure 3, the query function $\overline{f}_2(x)$ intersects upper envelope $\mathcal{E}$ at two points, $a$ and $b$. Point $c$ is the lowest point in $\mathcal{C}(a, b)$, so $c(s_2)$ is given by point $c$. In general, let $a$ (resp. $b$) be the closest point to $v_j$, if any, where query line $\overline{f}_j(x)$ intersects $\mathcal{E}$ in the one-period interval $\mathcal{C}(\alpha(v_j), \alpha(v_j))$. If there is no such intersection point, set $a = \alpha(v_j)$ that lies on the ccw side of $v_j$ in the 2-period diagram of $\mathcal{E}$, and $b = \alpha(v_j)$ that lies on the cw side of $v_j$ in the 2-period diagram of $\mathcal{E}$. Then $c(s_j)$ is given by the lowest point on $\mathcal{E}$, clockwise from $a$ to $b$.

*Lemma 6:* The centers $\{c(s_i) \,|\, s_i \in \mathcal{S}^*\}$ and the costs of the centers $\{F^{s_i}(c(s_i)) \,|\, s_i \in \mathcal{S}^*\}$ can be computed in $O(n \log n)$ time.

*Proof:* The intersection points $a$ and $b$ mentioned above, if any, can be determined in $O(\log n)$ time, using the query algorithm in [8] twice. It yields both $c(s_i)$ and $F^{s_i}(c(s_i))$. We repeat this for every scenario $s_i \in \mathcal{S}^*$. ■

### B. Finding the Optimal Location

Let $G$ be a network, which is a cycle in our current situation. Averbakh and Berman [9] convert $G$ to its *auxiliary network* $G'$ in such a way that the minmax regret 1-center problem on $G$ becomes the classical 1-center problem on $G'$. Assume that $F^s(c(s))$ for all $s \in \mathcal{S}^*$ are available, and let $M$ be a positive integer that is sufficiently large [9]. They append an edge $(v_i, v_i')$ of length $\{M - F^{s_i}(c(s_i))\}/\overline{w}_i$ to every vertex $v_i \in V$, where $v_i'$ is a new vertex of degree 1, called the *dummy vertex* corresponding to $v_i$. The vertices of $G'$ that are present in $G$ have weights equal to zero. For any vertex $v_i' \in G'$ its weight is set to $\overline{w}_i$. Clearly, $G'$ is a weighted cactus graph. Averbakh and Berman prove

*Theorem 3:* [9] The minmax regret 1-center problem on network $G$ can be solved by computing the classical weighted 1-center problem on $G'$. ■

Now that we have computed $\{F^{s_i}(c(s_i)) \mid s_i \in \mathcal{S}^*\}$ (Lemma 6), we can invoke Theorem 3. It is known that the classical weighted 1-center problem on a cactus network can be computed in $O(n \log n)$ time [10]. We thus have

*Theorem 4:* The minmax regret 1-center in a cycle network can be found in $O(n \log n)$ time. ■

## VI. Tree Network

Since our algorithm for a tree network needs a balanced tree, we first review *spine decomposition* [11], [12] that can convert any tree into a structure that has properties of a balanced tree.

### A. Review of Spine Decomposition

Many efficient tree algorithms preprocess a given tree, by aggregating information on its subtrees, and storing the condensed information at their root vertices. In *spine decomposition*, the structure that stores the aggregate information is separated from the given tree, and the height of this structure is bounded by $O(\log n)$, so that it can provide benefits of a balanced tree.

Given tree $T$ with $n$ vertices, we first find its 1-center under the base scenario $r_T = c(s_0)$, and assume that $T$ is a binary tree rooted at $r_T = c(s_0)$. If $T$ is not binary, it can be transformed into a binary tree by adding no more than

 ISBN: 978-1-61208-237-0 

$n$ zero weight vertices and zero length edges [13]. In spine decomposition, we select a path from $r_T$ to a leaf in $T$ such that the next vertex on this path always follows the child vertex that leads to the largest number of leaves from it. More formally, let $N_l(v)$ denote the number of leaves that have $v$ as an ancestor. If $v_0(=r_T), v_1, \ldots, v_k$ are the vertices on the path, then $N_l(v_{i+1}) \geq N_l(v'_i)$ always holds, where $v'_i$ is the other child vertex of $v_i$, if any. We call path $\langle v_0, v_1, \ldots, v_k \rangle$ the *top spine*, if $k \geq 1$. Now, for each $i = 1, \ldots, k$, we consider $v_i$ as the root of the subtree containing $v'_i$, and find other *spines* recursively. Note that a vertex belongs to at most two spines. Between these two spines, the one that is closer to the root of $T$ is considered the *parent spine* of the other. For the tree $T$ in Figure 4(a), its spine decomposition is shown in Figure 4(b), where the vertices of each spine are aligned either horizontally or vertically.

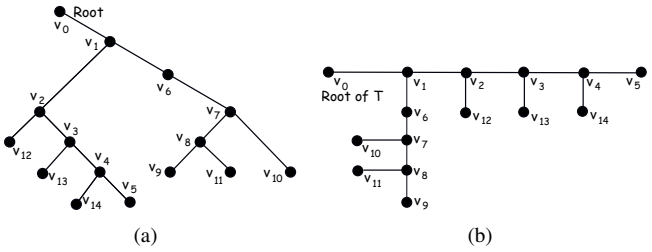



Fig. 4. (a) Tree $T$; (b) Spine decomposition of $T$.

A spine could be of length $O(n)$. To gather information efficiently from the vertices of spine $\sigma$, we build an auxiliary binary tree, called the *spine tree* and denoted by $\tau(\sigma)$, on top of it as follows. For spine $\sigma = \langle v_0, v_1, \ldots, v_k \rangle$ with root $r(\sigma) = v_0$, let $N(v_i)$ denote the number of descendant vertices of $v_i$ in $T$, excluding the vertices on $\sigma$, but including $v_i$ itself. Therefore, we have $N(v_0) = 1$. Under the root of $\tau(\sigma)$, we create two child nodes. (We use the term "nodes" to distinguish them from the vertices of the original tree $T$.) We partition $\{v_0, \ldots, v_k\}$ into two subsets $L = \{v_1, \ldots, v_h\}$ and $R = \{v_{h+1}, \ldots, v_k\}$ in such a way that $\sum_{i=0}^{h} N(v_i)$ and $\sum_{i=h+1}^{k} N(v_i)$ are as close as possible, and associate $L$ (resp. $R$) with the left (resp. right) child node. We keep partitioning until each node is associated with just one vertex, which becomes a leaf node of $\tau(\sigma)$. We then identify the root of $\tau(\sigma)$ with $r(\sigma)$ in the parent spine. In Figure 5(a), for example, the nodes at the two ends of each arrow are really just one node. The resulting structure, together with the spine tree for each spine, is denoted by $SD(T)$.

*Lemma 7:* [12] $SD(T)$ can be constructed in $O(n \log n)$ time, where $n$ is the number of vertices in the given tree $T$. ■

### B. Envelope Tree

We remove the edges belonging to the original tree $T$ from $SD(T)$ as in Figure 5(b), and call the resulting tree the *envelope tree*, denoted by $\mathcal{E}(T)$. An argument in [12] (Subsection 2.2) directly implies that

*Lemma 8:* [12] The height of $\mathcal{E}(T)$ is $O(\log n)$, where $n$ is the number of vertices in $T$. ■

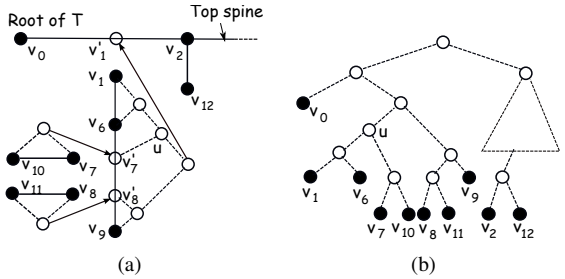



Fig. 5. (a) Part of $SD(T)$; (b) $\mathcal{E}(T)$.

Note that Lemma 8 does not imply that the sum of the heights of all spine trees is $O(\log n)$. For our purpose, we slightly modify spine decomposition. *As the top spine, we choose a maximal path from the root that contains the center* $c(s_0)$. Lemma 8 is still valid with this modification.

### C. Algorithm

For a path and a cycle, instead of computing the intersection of $\overline{f}_j(x)$ with $\underline{f}_i(x)$ individually for all $i$, which is too time-consuming, we constructed the upper envelope for $\{\underline{f}_i(x) \mid v_i \in V\}$, and computed the intersection of $\overline{f}_j(x)$ with it. However, it is difficult to make this approach work for a tree.

For a tree, we first compute the upper envelopes for various subsets of $\{\underline{f}_i(x) \mid v_i \in V\}$ instead. Namely, in Phase 1, we compute an upper envelope at each node (called an *envelope node*) $u$ in $\mathcal{E}(T)$. Such an upper envelope is for the cost functions of all vertices of $T$ that have $u$ as an ancestor in $\mathcal{E}(T)$, based on the vertex weights under $s_0$. For example, at $u$ in Figure 5(b), we compute and store the upper envelope for its descendant vertices i.e., $v_1, v_6, v_7$ and $v_{10}$. In Phase 2, for each vertex $v \in V$, we extract a set of $O(\log n)$ upper envelopes from those computed in Phase 1, compute the intersection of $\overline{f}_j(x)$ with each of them, and determine the most costly one among them, which gives $c(s_j)$. Here are the details.

*1) Phase 1:* We assume that envelope tree $\mathcal{E}(T)$ has already been constructed for a given $T$ (Lemma 7). At a leaf node of $\mathcal{E}(T)$, which is a vertex ($v_i$) of $T$, the upper envelope is $\underline{f}_i(x) = d(x, v_i)\underline{w}_i$, which is composed of either one line (if $v_i$ is a leaf vertex of $T$) or two lines (if $v_i$ is a non-leaf vertex of $T$) forming the shape of letter V. At the 2nd level from the bottom of $\mathcal{E}(T)$, each envelope node has two end vertices of an edge of $T$ as its descendants. Eventually, we reach an envelope node that is the root of a spine tree. Figure 6 illustrates how to propagate the cost contributions from the vertices of a lowest spine $\sigma_1$ to its parent spine $\sigma_2$. The horizontal line in Figure 6(a) represents $\sigma_1$. It is connected to $\sigma_2$ at its root vertex $v = r(\sigma_1)$. The information concerning the upper envelope from $\sigma_1$ that we need to propagate to $\tau(\sigma_2)$ is shown by the thick line segments, labeled "Contribution to parent spine" in Figure 6(a). It, together with its mirror image, is stored as envelope $E_v(x)$ at $v$. This is because the cost contribution from a vertex on $\sigma_1$ is the same at any two points on $\sigma_2$ that are at equal distance from $v$, one closer to $r(\sigma_2)$ and the other away from it. See Figure 6(b).

 ISBN: 978-1-61208-237-0 

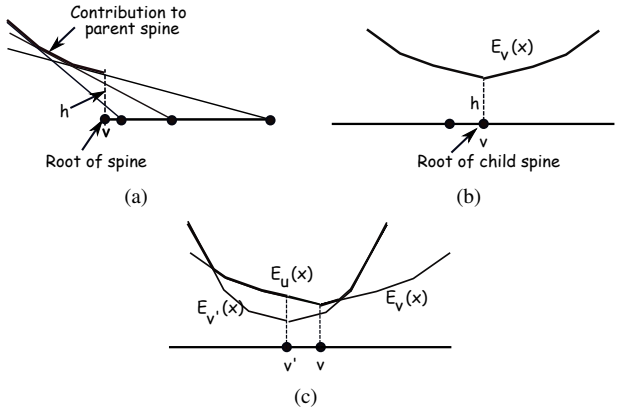



Fig. 6. (a) Spine $\sigma_1$ with $v = r(\sigma_1)$; (b) $E_v(x)$ stored at $v$ on spine $\sigma_2$; (c) $E_u(x)$ stored at $u$ (parent of $v$ and $v'$).

Suppose that two vertices on $\sigma_2$, $v$ and $v'$, have an envelope node $u$ as their parent node. We now construct $E_u(x)$ from $E_v(x)$ and $E_{v'}(x)$ by computing their upper envelope. In Figure 6(c), $E_u(x)$ is shown by thick line segments. We can similarly construct the upper envelopes at higher envelope nodes in $\mathcal{E}(T)$. Since $\mathcal{E}(T)$ is of height $O(\log n)$ (Lemma 8), the total time required for computing the envelopes at all the nodes of $\mathcal{E}(T)$ is $O(n \log n)$.

*2) Phase 2:* Let $\pi(v_j)$ denote the path from the leaf node $v_j$ (a vertex of $T$) to the root of $\mathcal{E}(T)$. We compute $c(s_j)$, tracing $\pi(v_j)$ upward from $v_j$ as follows:

**Algorithm** `Find-Center/Tree`

1) At each node $u$ visited, do the following:
   (a) If the left (resp. right) child node of $u$ is on $\pi(v_j)$, let $u'$ denote $u$'s right (resp. left) child node.
   (b) Find the lowest intersection point, if any, of $\overline{f}_j(x)$ with $E_{u'}(x)$, where they have opposite slopes.
2) From among the intersection points computed in step 1(b), find the highest point. If this point has a higher cost than $c(s_0)$, then the corresponding location gives $c(s_j)$. Otherwise $c(s_j) = c(s_0)$. ■

Since $c(s_0)$ is on the top spine, if $v_j$ is on spine $\sigma$, $c(s_j)$ can lie either on $\sigma$ or an anscester spine. Thus, the spine trees for the other spines can be ignored. Suppose, for example, that we want to compute $c(s_{10})$ in Figure 5. We first visit the parent node of $v_{10}$, and check its other child node, where the upper envelope (i.e., $\underline{f}_7(x)$) for $v_7$ is stored as $E_{v_7}(x)$. We now carry out step 1(b), i.e., compute the intersection of $\overline{f}_{10}(x)$ and $\underline{f}_7(x)$ , to find the first candidate for $c(s_{10})$. We then trace $\mathcal{E}(T)$ upwards to $u$, and repeat step 1(a)(b). In this case, it entails computing the intersection of $\overline{f}_{10}(x)$ with $E_{u'}(x)$ stored at the left child $u'$ of $u$, where $E_{u'}(x)$ is the upper envelope of $\underline{f}_1(x)$ and $\underline{f}_6(x)$.

*Theorem 5:* The minmax regret 1-center of a tree network can be found in $O(n \log^2 n)$ time.

*Proof:* In executing `Find-Center/Tree`, the number of envelope nodes encountered on $\pi(v_j)$ is $O(\log n)$ by Lemma 8. Step 1(b) can be carried out, using binary search, since $E_{u'}(x)$ is convex with a number of corner points. Thus the processing time per vertex ($v_j$) is $O(\log^2 n)$, and the total time for all 1-centers $\{c(s_j) \mid s_i \in \mathcal{S}^*\}$ is $O(n \log^2 n)$. Once we have computed $\{c(s_i) \mid s_i \in \mathcal{S}^*\}$, we can evaluate $\{F^{s_i}(c(s_i)) \mid s_i \in \mathcal{S}^*\}$ easily, and invoke Theorem 3. ■

## VII. Conclusion and Future Work

We have presented an algorithm for finding the minmax regret 1-center of a path (resp. cycle) network that runs in $O(n)$ (resp. $O(n \log n)$) time. No algorithms specifically tailored to these families of networks were previously known. We also presented a new algorithm for a tree that runs in $O(n \log^2 n)$ time. For a tree, an algorithm with the same complexity was already known [6], but its description takes up several pages.

Lemma 5 implies that the classical weighted 1-center in a cycle network can be computed in $O(n)$ time, which is a new result and is optimal. It is valid for any cycle network that contains at least one vertex whose weight is non-negative. As future work, we want to see if the $O(n \log^2 n)$ time complexity for a tree can be improved, and also work on a cactus network. A more challenging problem is finding the minmax regret $p$-center for $p \geq 2$.

## Acknowledgment

This work was partially supported by Discovery Grants from the Natural Science and Engineering Research Council of Canada.

## References

[1] T. S. Hale and C. R. Moberg, "Location science research: A review," *Annals of Operations Research*, vol. 123, pp. 21–35, 2003.
[2] B. Chen and C.-S. Lin, "Minmax-regret robust 1-median location on a tree," *Networks*, vol. 31, pp. 93–103, 1998.
[3] O. Kariv and S. Hakimi, "An algorithmic approach to network location problems, part 1: The p-centers," *SIAM J. Appl. Math.*, vol. 37, pp. 513–538, 1979.
[4] N. Megiddo, "Linear-time algorithms for linear-programming in $R^3$ and related problems," *SIAM J. Computing*, vol. 12, pp. 759–776, 1983.
[5] I. Averbakh and O. Berman, "Algorithms for the robust 1-center problem on a tree," *European Journal of Operational Research*, vol. 123, no. 2, pp. 292–302, 2000.
[6] H.-I. Yu, T.-C. Lin, and B.-F. Wang, "Improved algorithms for the minmax-regret 1-center and 1-median problem," *ACM Transactions on Algorithms*, vol. 4, no. 3, pp. 1–1, June 2008.
[7] A. Goldman, "Optimal center location in simple networks," *Transportation Science*, vol. 5, pp. 212–221, 1971.
[8] B. Chazelle and L. J. Guibas, "Fractional cascading: II. Applications," *Algorithmica*, vol. 1, pp. 163–191, 1986.
[9] I. Averbakh and O. Berman, "Minimax regret p-center location on a network with demand uncertainty," *Location Science*, vol. 5, pp. 247–254, 1997.
[10] B. Ben-Moshe, B. Bhattacharya, Q. Shi, and A. Tamir, "Efficient algorithms for center problems in cactus networks," *Theoretical Compter Science*, vol. 378, no. 3, 2007.
[11] R. Benkoczi, B. Bhattacharya, M. Chrobak, L. Larmore, and W. Rytter, "Faster algorithms for $k$-median problems in trees," *Mathematical Foundations of Computer Science, Springer Verlag*, vol. LNCS 2747, pp. 218–227, 2003.
[12] R. Benkoczi, "Cardinality constrained facility location problems in trees," Ph.D. dissertation, School of Computing Science, Simon Fraser University, Canada, 2004.
[13] A. Tamir, "An $O(pn^2)$ algorithm for the $p$-median and the related problems in tree graphs," *Operations Research Letters*, vol. 19, pp. 59–64, 1996.

 ISBN: 978-1-61208-237-0 

# The use of Bioinformatics Techniques for Time-Series Motif-Matching: A Case Study

Mark Transell and Carl Sandrock
Department of Chemical Engineering
University of Pretoria
Pretoria
marktransell@gmail.com

***Abstract*—Process engineers have more access to historical plant data than ever before. Finding recurring patterns in process data, also referred to as motif-matching, may reveal diagnostic information to engineers and operators. Dynamic Time Warping (DTW) is one of the most widely used techniques for performing these motif matches. Sequence matching is also an important part of bioinformatics; a field which has received a marked increase in research funding and attention in recent times. Therefore, the techniques developed in bioinformatics may be beneficial to the field of time-series motif matching. In this study, a combination of the Symbolic Aggregate Approximation (SAX) algorithm and the PSI-BLAST bioinformatics algorithm is compared to DTW as a potential method to perform time-series matches. Preliminary results suggest that this combination may be faster than global DTW techniques for large datasets. Details of the implementation are given, along with preliminary results confirming that this method is feasible. Due to implementation difficulties, accuracy and robustness remain uninvestigated. More research is recommended into the potential for this technique as an alternative to Dynamic Time Warping techniques.**

***Index Terms*—Dynamic Time Warping; BLAST; motif-matching; time series; PAA***

## I. Introduction

It is often necessary for chemical and control engineers to diagnose a recurring plant behaviour, and for operators to receive alerts when undesirable plant behaviour is occuring. Dynamic Time Warping (DTW) was first applied to speech-recognition [1], but has since been applied to many fields [2] [3].

Sequence matching is also an important part of bioinformatics; a field which has received a marked increase in research funding and attention in recent times, helped in part by the publicity received by the Human Genome Project. This has led to the development of highly efficient algorithms and automated software implementations.

These freely avalable algorithms promise easier implementation, benefits to computational load and improved matching accuracy than DTW and similar methods. Particular attention is paid to the applicability of the PSI-BLAST algorithm combined with Symbolic Aggregate Approximation (SAX) to match process data. Time series need to be converted into character strings before Bioinformatics techniques can be applied to them. In this work, Symbolic Aggregate Approximation is used to convert the time series into strings, and PSI-BLAST is used to match these sequences.

## II. Theory

### A. Piecewise Aggregate Approximation (PAA)

In order to reduce the search space, it is often required to resample an original time series of length $n$ to a reduced length $w$. A simple approach, known as Piecewise Aggregate Approximation (PAA), is to divide the entire time range into blocks and average over the blocks, as in Equation 1 [4].

$$\bar{c_i} = \frac{w}{n} \sum_{j=\frac{n}{w}(i-1)+1}^{\frac{n}{w}i} c_j \tag{1}$$

### B. Symbolic Aggregate Approximation (SAX)

These continuous values need be quantized into character strings. The Symbolic Aggregate Approximation (SAX) method [4] uses breakpoints which will result in an equal probability of letters if the data were normally distributed. This is a highly desirable characteristic for sequence-matching algorithms as it makes the scoring matrices easy to calculate. Although the full procedure is outlined in [4] and [5], Figure 1 illustrates the basic concept of SAX: using the average value for a window of time series data to allocate a specific character.

PAA may acheive similar results to the more complex and computationally demanding methods [5], so it and SAX are used as the baseline technique for converting time-series data into character strings.

### C. Wavelets

It has been shown that Haar Wavelet Transforms (HWT) can outperform discrete fourier transforms (DFTs) when reducing dimensionality in time series [6]. Although wavelets have the helpful multiresolution property, they are only defined for time series which are an integer power of two in length [4].

One important feature of wavelets and DFTs is that they are real valued. This limits the algorithms, data structures and definitions available for them. In anomaly detection, we cannot meaningfully define the probability of observing any particular set of wavelet coefficients, since the probability of observing any real number is zero [4].

### D. Triangular Episodic Representation

Complete, correct, robust and compact models can be constructed using a small base of primitive representations for

 ISBN: 978-1-61208-237-0 

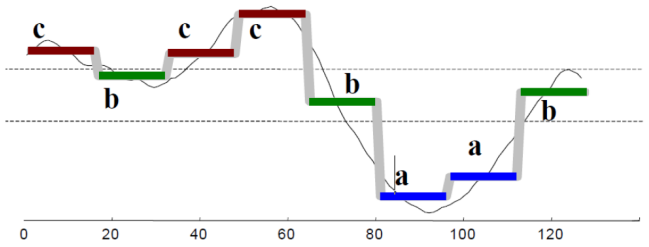



Fig. 1. An illustration of the SAX technique. The different colours are added for clarity, and characters are assigned to each window in the time series. Datapoints are numbered on the x-axis [5].

process behaviours [7]. The qualitative representations can be defined by any time over which the qualitative state of the process variable $x$ is constant. This qualitative state is defined as in Equations 2 to 5.

$$\mathrm{QS}(x,t) = \begin{cases} \text{undefined if } x \text{ is discontinuous at } t \\ < [x(t)], [\partial x], [\partial^2 x] > \text{elsewhere} \end{cases} \quad (2)$$

Where:

$$[x(t)] = \begin{cases} + \text{ if } x > 0 \\ - \text{ if } x < 0 \\ 0 \text{ if } x = 0 \end{cases} \quad (3)$$

$$[\partial x(t)] = \begin{cases} + \text{ if } \partial x > 0 \\ - \text{ if } \partial x < 0 \\ 0 \text{ if } \partial x = 0 \end{cases} \quad (4)$$

$$[\partial^2 x(t)] = \begin{cases} + \text{ if } \partial^2 x > 0 \\ - \text{ if } \partial^2 x < 0 \\ 0 \text{ if } \partial^2 x = 0 \end{cases} \quad (5)$$

This triangular episodic representation gives seven basic types of episodes that can describe an interval of process behaviour [8].

This method is also currently being investigated, but preliminary results show no improvement over SAX techniques.

### E. Shapelets

Time series shapelets [9] are used in image recognition, and rely on libraries of stored prominent motifs found within time-series data. Matches are classified using decision trees which compare subsequences that are maximally representative of a class.

Shapelets are selected based either on learning sets, or by using knowledge of the process in question to manually build decision trees. Process data may not necessarily be classified by this method appropriately; online chemical processes do not always have large enough learning sets to ensure accurate shapelet identification. Due to the unpredictable nature of process disturbances, operator knowledge of which motif to select as a relevant shapelet may be incomplete.

### F. Dynamic Time Warping

Dynamic Time Warping [1] is often used as a faster and more robust method than Euclidian Distance to quantify the similarity of two time series [10]. It can also be used as a method to match subsequences [11], as implemented in the Machine Learning Python library (mlpy) [12].

It was selected as the baseline for time-series motif-matching, since there is a large amount of literature which employs this technique and it is popular in several applications, including speech and image recognition [1].

This algorithm employs a lower bounding technique based on a warping window or envelope. The most commonly used warping constraints are the Sakoe-Chiba band [1] and the Itakura parallelogram [13] [14].

### G. BLAST (PSI-BLAST)

The BLAST algorithm [15] was developed to match strings representing nucleotide or protein fragments to large databases more efficiently than the FASTP [16] algorithm. The most recent iteration of the BLAST library includes the PSI-BLAST program [17], which matches queries in a gap-tolerant fashion. This makes the matching process more robust, and therefore better suited to finding *similar* strings as opposed to exact matches.

The PSI-BLAST algorithm uses scoring matrices to determine the similarity of two strings, the most popular being the BLOSUM62 substitution matrix [18].

### H. Scoring matrices

The BLOSUM62 matrix [18] is based on amino-acid substitution; the probability that one amino-acid would replace another in a particular protein string. This scoring matrix is therefore not suitable for normally distributed time-series data. However, a scoring matrix can be calculated directly during the SAX procedure, based on the assumption that all characters are equiprobable [5].

### I. Performance Metrics

The metrics used to quantify the performance of the matching algorithms compared are:

- Computing time taken to complete a search, and the ability of the algorithm to scale with database size.
- The accuracy of the matches found, with respect to Euclidean distance and DTW cost for the matches [10], [11], [19].
- The ease by which each algorithm can be implemented. (Does an algorithm require system file access? How many hardware and software dependencies are there?)
- Tolerance to datasets with noise, time-axis shifting, vertical shifting or time warping.

## III. Implementation

Figure 2 shows how a library of time-series and query data is processed by each algorithm to produce a match with quantifiable accuracy.

The development framework necessitated the use of the NCBI BLAST+ executable library, available on their website [17]. The language selected for the coding framework was Python 2.7, due to the availability of the Biopython library

 ISBN: 978-1-61208-237-0

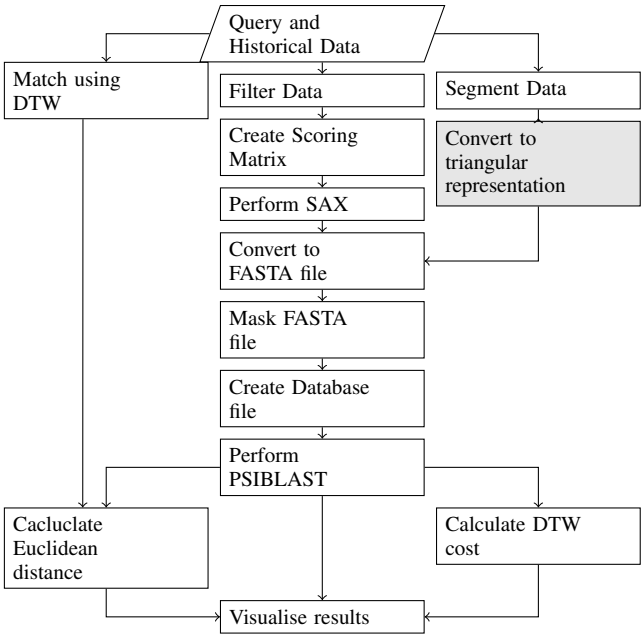



Fig. 2. A simplified overview of the data flow for testing

for use with bioinformatics programs such as BLAST [20]. Additional python libraries were used:

- numpy [21] for data handling
- matplotlib [22] for plotting purposes
- mlpy [12] for the DTW algorithm
- scipy [23] for data filtering.

The breakpoints for SAX conversion are calculated based on the cumulative distribution of the database data, thus ensuring that every character in the database string is equiprobable. [5] The implementation of BLAST matching requires the operating system used to be a Linux-based system, as it is necessary to overwrite the BLOSUM scoring matrix directly with a custom scoring matrix generated during the SAX process. [4]

The SAX-PSIBLAST method has four paramaters, namely the size of time window and number of breakpoints used for SAX, and the expected error value and search word size given as inputs to PSI-BLAST. These parameters enable the user to adjust the degree to which fuzzy matches can be found, increase or decrease noise tolerance in the dataset, or to adjust the number of potential matches listed for any given query.

The SAX-PSIBLAST procedure does not make use of training sets to perform matches. This allows non-mutated test data to be used for speed tests. In order to test whether a matching algorithm is performing correctly, two base cases are examined. These cases are:

1) A query time-series is matched to itself.
2) A query time-series is matched to a time-series which includes, but is not restricted to the query itself.

## IV. Results and Discussion

The PSI-BLAST technique is faster than the global DTW algorithm on the self-matching task for large datasets. Figure 3

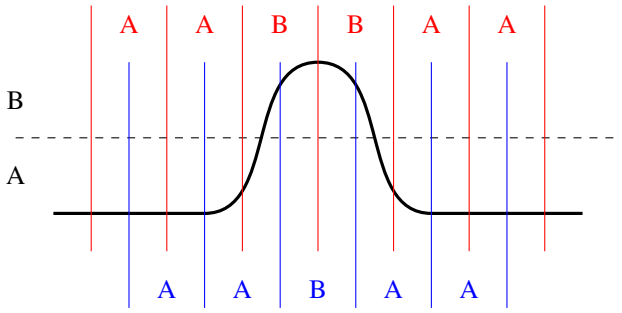



Fig. 4. An illustration of how shifting a PAA time window can alter the string sequence produced by SAX. The upper sequence output is AABBAA, but the lower sequence is AABAA, even though they represent the same time series. This can cause the PSI-BLAST algorithm not to match these series correctly.

shows the processing times for global DTW and the SAX-PSIBLAST method for different PAA average-value windows. The mlpy DTW algorithm fails on the two largest test data sets.

However, the PSI-BLAST algorithm does not reliably find a full match for the second base case after dimensionality reduction. In some cases a partial match is returned, or the match is erroneously extended to include adjacent data. Possible reasons for this failure include:

- The internal workings of the PSI-BLAST executable do not use the modified scoring matrix correctly.
- The modified scoring matrix is in a form which, for statistical reasons, does not fit the specifications for PSI-BLAST scoring matrices. (This could possibly be remedied by intelligent selection of the expected error value and word-size)
- The queried data is shifted during the PAA step for any time window, which may cause the string representation of the query and dataset to be slightly different.

Figure 4 illustrates how the string representation may be altered by shifts in the time-windows for PAA.

## V. Conclusions and Recommendations

Preliminary results suggest that a motif-matching algorithm which employs a combination of Symbolic Aggregate Approximation (SAX) and the use of the PSI-BLAST algorithm is faster than conventional Dynamic Time Warping (DTW) techniques. Additional metrics remain to be compared.

It is currently not known whether the partial failure of the SAX-PSIBLAST method to find embedded sequences when large time windows are used is due to implementation error, or to inherent limitations in the PSI-BLAST algorithm when processing dimensionally reduced sequences.

## References

[1] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 26, pp. 43–49, 1978.

[2] T. M. Rath and R. Manmatha, "Word image matching using dynamic time warping," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '03)*, vol. 2, p. 521.

 ISBN: 978-1-61208-237-0 

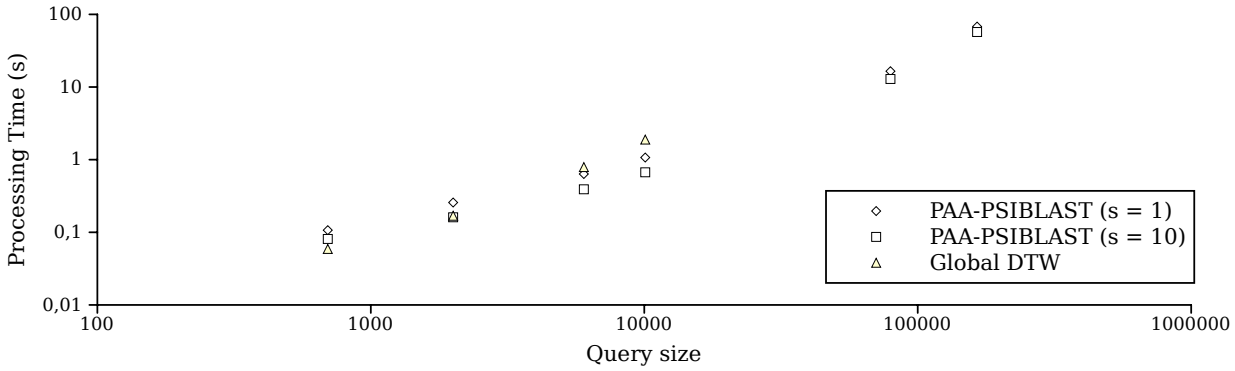



Fig. 3. Base case processing time comparison on Intel Core i5 processor

[3] K. Santosh, "Use of dynamic time warping for object shape classification through signature," *Kathmandu University Journal of Science, Engineering and Technology*, vol. 6, pp. 33–49.

[4] J. Lin, E. Keogh, S. Lonardi, and B. Chiu, "A symbolic representation of time series, with implications for streaming algorithms," in *Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, ser. DMKD '03. New York, NY, USA: ACM, 2003, pp. 2–11. [Online]. Available: http://doi.acm.org/10.1145/882082.882086

[5] B. Chiu, E. Keogh, and S. Lonardi, "Probabilistic discovery of time series motifs," in *The 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, vol. 1, 2003, pp. 493–498.

[6] K. Chan and A. Fu, "Efficient time series matching by wavelets," in *Proceedings of the 15th IEEE International Conference on Data Engineering*, Sydney, Australia, March 1999, pp. 126–133.

[7] J.-Y. Cheung and G. Stephanopoulos, "Representation of process trends part i. a formal representation framework," *Computers and Chemical Engineering*, vol. 14 (4/5), pp. 495–510, 1990.

[8] S. Kivikunnas, "Overview of process trend analysis methods and applications," in *Proceedings of the ERUDIT Workshop on Applications in Pulp and Paper Industry*, University of Oulu, Finland, 1998.

[9] L. Ye and E. Keogh, "Time series shapelets: A new primitive for data mining," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, ser. DMKD '03. New York, NY, USA: ACM, 2009.

[10] E. J. Keogh and M. J. Pazzani, "Scaling up dynamic time warping for datamining," University of California, Tech. Rep., 2000.

[11] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series," Stern School Of Business, New York University, Tech. Rep., 1994.

[12] D. Albanese, S. Merler, G. Jurman, R. Visintainer, and C. Furlanello, "Mlpy machine learning py," 2012, http://mloss.org/software/view/66/.

[13] F. Itakura, "Minimum prediction residual principle applied to speech recognition," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. ASSP-23, pp. 52–72, 1975.

[14] C. A. Ratanamahatana and E. Keogh, "Everything you know about dynamic time warping is wrong," in *3rd Workshop on Mining Temporal and Sequential Data, in conjunction with 10th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, 2004.

[15] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of Molecular Biology*, vol. 215, pp. 403–410, 1990.

[16] D. Lipman and W. Pearson, "Rapid and sensitive protein similarity searches," *Science, New Series*, vol. 227(4693), pp. 1435–1441, 1985.

[17] S. Altschul, T. Madden, A. Schffer, J. Zhang, Z. Zhang, W. Miller, and D. Lipman, "Gapped blast and psi-blast, a new generation of protein database search programs," *Nucleic Acids Research*, vol. 25(17), pp. 3389–3402, 1997.

[18] J. Setubal and R. Braeuning, *Similarity Search*, A. Gruber, A. Durham, and C. H. et al., Eds. National Center for Biotechnology Information (US), 2006.

[19] H. Ding, G. Trajcevski, P. Scheuermann, X. Wang, and E. Keogh, "Querying and mining of time series data: Experimental comparison of representations and distance measures," *VLDB*, 2008.

[20] P. J. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski1, and M. J. L. de Hoon, "Biopython: freely available python tools for computational molecular biology and bioinformatics," *Bioinformatics*, vol. 25, pp. 1422–1423, 2009.

[21] P. F. Dubois, K. Hinsen, and J. Hugunin, "Numerical python," *Computers in Physics*, vol. 10, no. 3, May/June 1996.

[22] J. D. Hunter, "Matplotlib: A 2d graphics environment," *Computing In Science & Engineering*, vol. 9, no. 3, pp. 90–95, May-Jun 2007.

[23] E. Jones, T. Oliphant, P. Peterson *et al.*, "SciPy: Open source scientific tools for Python," 2001–. [Online]. Available: http://www.scipy.org/

 ISBN: 978-1-61208-237-0 

# Bringing Viability to Service-Oriented Enterprises in Cloud Ecosystems

Nizami Jafarov, Edward Lewis and Gary Millar
School of Engineering and Information Technology, University of New South Wales at Australian Defence Force Academy
nizami.jafarov@student.adfa.edu.au, e.lewis@adfa.edu.au, g.millar@adfa.edu.au

***Abstract* - Cloud Computing is the next evolution of computing systems, which shifts the whole application infrastructure away from old monolithic architectural models to more open, interoperable, modular, reusable and agile building blocks. Many Enterprises realise the benefits of Cloud Computing and position it as the Information Technology (IT) outsourcing solution, which is cheaper to operate and maintain. However, Cloud Computing is not just a technology behind an infrastructure, but also a new model, which needs to be properly merged with legacy paradigms of an Enterprise. In our work, we are focused on the challenges Cloud Computing is bringing to an Enterprise along with the promising benefits. The central issue in the transformation of IT infrastructure of an Enterprise into the Cloud is the uncertainty it is able to bring to a well established Business Architecture (BA) and Business Models (BM) of a system. This disruptive uncertainty, which can be driven by an introduction of new technologies or modification of already integrated ones, needs to be addressed properly and governed accurately to guarantee a viability of an Enterprise. This work proposes solutions to the issues an Enterprise might face in its Cloud transformation initiative.**

*Keywords–cloud computing; enterprise architecture; technological innovations; uncertainty management.*

## I. INTRODUCTION

Cloud and Cloud Computing are no more buzzwords in the world of Enterprises. Corporations and organisations slowly realise what Cloud-based solutions can offer and what benefits they can bring. Researchers worldwide, in academia and industry, are involved into analysis of impacts of Cloud Computing on IT infrastructures, organisational changes, business operations, system architectures to mention a few. Research challenges in Cloud Computing are systematically reviewed and addressed in academia [1, 2, 3, 4, 5, 6]. Many white papers are also trying to give a clear definition of the Cloud and its capabilities in system architecture [7, 8, 9]. Market analysts suggest that the Cloud Computing capitalisation is already, almost three times more in 2012 than that of in 2008 [10]. Though, Cloud has slowly became a trend in the world of IT, active research and wide range of definitions clearly indicate that Cloud Computing is still immature as a paradigm. Thus, its integration with complex systems might lead to unpredictable consequences in systems architectures which position Cloud as a source of uncertainty which needs a clear realisation and governance.

In our work, we are looking at the challenges in Cloud ecosystems from the Enterprise Architecture (EA) perspective.

Amongst the typical domains of EA such as Data Architecture, Application Architecture and Technology Architecture we identify the Business Architecture (BA) as the domain of interest, which is purely addressed in academia, however vital for an Enterprise to remain viable and competitive on the market in times of insertions of innovative technologies. In our research, we position Cloud Computing as the *disruptive technology which is capable of bringing uncertainty to a BA of an Enterprise in times of technological insertions*, in one or more of the following scenarios:

1. *Modifications in already integrated technologies*

In times, when an Enterprise faces the changes in the scope of operations of already integrated technologies consumed over the Cloud (e.g., Public Cloud).

2. *Introduction of new technologies*

In times, when new technologies got introduced to the public, but their benefits are not yet realised by an Enterprise.

3. *Parallel Innovations*

In times, when an introduction of new technologies and changes to already integrated ones are happening at the same time.

In our work, we aim to overcome the above mentioned challenges by bringing Cybernetic concepts to the domain of EA to model a system accompanied with the decision models, uncertainty management and risk remedy mechanisms, which would help it remain viable in times of technological innovations.

This paper is structured as follows: Section 2 gives a detailed overview on the risks and challenges Cloud Computing brings to the EA; Section 3 proposes various solutions to overcome the issues Enterprises face in the Cloud and Section 4 summarises the findings and gives a glimpse on our future work.

## II. BACKGROUND AND CHALLENGES

Nowadays, reliance of IT systems on remote computational power is accepted as a matter of fact and service providers such as Google, Citrix, Microsoft, VMWare, etc. offer a wide

 ISBN: 978-1-61208-237-0 

range of diverse solutions which can be integrated into architectures of any IT-enabled Enterprises over the Cloud. The National Institute of Standards and Technology (NIST) defines Cloud Computing, as "*a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. This cloud model promotes availability and is composed of five essential characteristics, three service models, and four deployment models.*" [11]

Amongst the deployment models of Cloud Computing such as Community Cloud, Private Cloud, Public Cloud and Hybrid Cloud, we consider the latter models and Public Cloud in particular as more challenging, since in the Public Cloud, Enterprises leverage on remote computational power, which might be out of their direct governance. Absence of direct governance of IT can be challenging to the BA of an Enterprise, since for an Enterprise whose mission is to remain competitive and viable on the market in the long run, BA is considered to be the core unit of its business operations and Enterprise identity. BA is "*a blueprint of the Enterprise that provides a common understanding of the organisation and is used to align strategic objectives and tactical demands.* " [12] In a typical Enterprise, such a blueprint might require continuous monitoring, revision of its artefacts and relevant BMs to remedy possible risks at global and local scales. However, in nowadays IT-enabled Enterprises these types of risks can possibly be part of BMs of BA, which can arise from a reliance on a remote computational power.

### *A. Risks of modifications in already integrated technologies*

Amongst the main challenges of an Enterprise is preserving its identity in a long run and therefore continuously governing and maintaining its BA, which represents corporate values. Enterprises of modern age are trying to be pro-active, rather than reactive to possible changes in their business operations and this in fact requires significant investment of time and money into the management of relevant business units of the EA. One of the widely used system design paradigms, such as Service Oriented Architecture (SOA) is able to bring agility to a BA of an Enterprise through the decomposition of complex BMs into small ones and decoupling of business logic from operational logic. However, this might not be enough when an Enterprise is running its operations in the Cloud. Impacts of SOA and effects of introduction of Cloud to an Enterprise were addressed in range of papers in academia [2, 5, 13]. However, none of the existing works comprehensively analysed the impacts of Cloud Computing on BA of an Enterprise.

The core issue with the Public Cloud deployment model integrated with the IT infrastructure of an Enterprise is that the external services can be part of internal BMs. Though, this is one of the main missions of the Cloud Computing, such integration can also become a source of unexpected risks, which might question the viability of an entire Enterprise. In a Cloud SOA-enabled ecosystem services can be provided and consumed by a range of participants (i.e., service providers and service consumers). In fact, a service consumer on one side is only aware of the end-point on the other, whereas the end-point itself can be a consumer of services external to its Enterprise scope and therefore play a role of a medium. Such role, though promotes the core benefits of service-oriented systems, can still be challenging for an Enterprise when a change in an operational scope of any service in the service consumption chain is happening independently at different levels of service consumers hierarchy (Fig. 1).

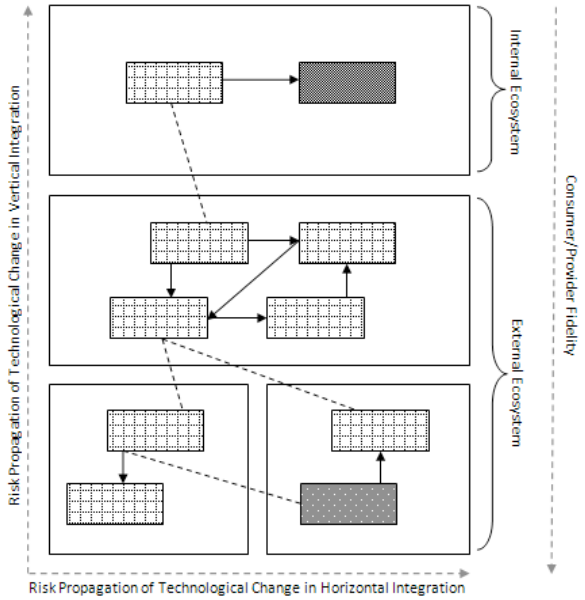



Fig. 1. Impacts of technological modifications

Though, the best practices of SOA teach us the ways to avoid possible risks through the use of dynamic and redundant service contracts and other relevant SOA patterns, they are unable to remedy risks associated with the propagation of technological changes in service consumers hierarchy by the time they got introduced to the end-point Enterprise. Which as a result can affect the BA of an Enterprise in times of *Vertical* and *Horizontal* integrations.

### *B. Risks of introductions of new technologies*

Another challenge to the BA of an Enterprise is the introduction of new technologies. For Enterprises innovations of any kind can be disruptive and shall be managed according to organisational objectives and tactical demands. In modern IT-enabled Enterprises sources of innovations, and thus disruption, can be various Cloud deployment models integrated with the IT infrastructure. Various works [2, 5] in academia position Cloud Computing as a disruptive technology, but no solutions are proposed and much work done towards analysing and overcoming possible risks Cloud ecosystem can generate.

 ISBN: 978-1-61208-237-0 

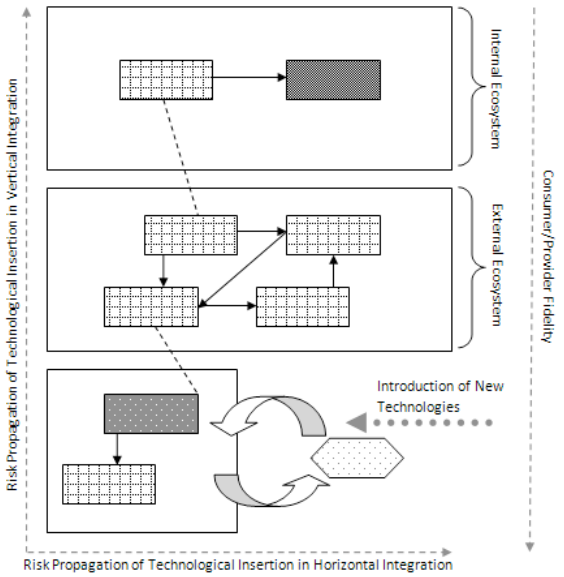



Fig. 2. Impacts of new technologies introductions

Scenarios that need deeper understanding and scientific analysis are those associated with the merge of an infrastructure of an Enterprise with the Public or Hybrid Cloud deployment models. In such scenarios an Enterprise does not have direct governance over some parts of its IT infrastructure and thus any introduction of a new technology can confuse decision makers (Fig. 2). In other words, time required for an Enterprise to realise possible opportunities or dangers of a newly introduced technology might not be enough and as a result increase or decrease its competitiveness on the market as well as question its viability.

### C. *Risks of Parallel Innovations*

Lastly, a challenge, which needs careful analysis, is the one associated with the decision making in times of Parallel Innovations. We position Parallel Innovations as *the process of simultaneous modifications in existing technologies and introductions of new ones to an Enterprise*. In the context of Parallel Innovations we divide an introduction of a new technology into two processes: *1) introduction of a technology closely related to the one already integrated into the infrastructure of an Enterprise*; *2) introduction of a technology which is not yet integrated and unrelated to any existing technology of an Enterprise*; whereas innovation in already integrated technologies is a sole *process of modifications in the scope of services functionalities in the hierarchy of service consumers*. All these processes are capable of bringing uncertainty to an Enterprise and require relevant tools to support the management in making relevant decisions based on the possible events. For example, a new technology, closely related to the already integrated one, can substitute the old technology and bring financial benefits to an Enterprise, whereas a technology unrelated to the existing can bring uncertainty or destruction to the BA and Enterprise itself. There are of course various other possible scenarios of BA evolution in such integration initiatives that require deeper research and analysis.

## III. METHODS AND SOLUTION SEEDS

### A. *Using Ontologies*

One of the solution seeds, which can help in handling possible issues in propagation of risks from IT to a BA of an Enterprise, is to decouple these domains from each other. Such an approach was addressed in a number of works in academia [14, 15] and suggests the use of ontologies as mediums between these domains. In Information Science (IS), ontology is a "*formal, explicit specification of a shared conceptualisation*" [16]. Aier and Winter [14] propose a method for the identification of alignment artefacts using the example of domain clustering which is reflected in their Meta model. Kalogeras et al. [15] suggest using Web-Services technology as universal interfaces to achieve the decoupling of IT from BMs due to their wide utilization of open standards. Both approaches claim that through ontologies any modifications to the IT infrastructure of an Enterprise will have less impact on its BMs and therefore bring more agility to the BA. However, both methods have flaws and it was identified that the decoupling through ontologies can aid the BA, but is not enough for an Enterprise to become pro-active to possible technological modifications in the chain of service providers.

The core issue with the use of the methods is associated with the lexicon terms in system ontologies, which in times of technological innovations can sometimes encompass wider meanings and system might not necessary be aware of this (e.g., when there is change in a technology, somewhere in the chain of service providers). In most cases, a careful design of ontologies may not help, since the core issue is the - *unexpected change in a service scope*. So, for instance a service, which was proving radar services, will still be providing radar services, but the radius it covers might change from 5km to 10km, which will eventually affect a BM that consumes it. Another significant issue is that, it is not that easy to change system ontologies once they are established, more than that if such a decision was made it might lead to over-cluttering of system taxonomy. Semantically-aware approach seem to be a solution, but once it got integrated into a system it acts as a black-box and interpreting inputs just from outputs might not be so effective, since a *dangerous* output can only be statistically detected after a particular amount of inputs, period of time, which means: lose of opportunities, decrease of competitiveness, viability and so on. Both works [14, 15] suggest automation, which makes sense in such complex systems, but this automation lacks a feedback mechanism. More importantly, feedback mechanism is not enough and such systems need a more comprehensive tool. We see a need for the Decision Support System (DSS), a tool that can aid overall Enterprise design and governance of its underlying systems. In our work, this tool is inspired by Cybernetics

 ISBN: 978-1-61208-237-0 

concepts and therefore called the Cybernetic Decision Model (CDM).

### B. Modeling Decision Support System

The goal of the CDM is to become a comprehensive EA tool that will bridge the gap between what is needed for planning for the viable systems in times of disruptive change and what is available. Its main application is within a DSS and it based on the GRAI Grid Decisional Model [17] developed in the University of Bordeaux. The CDM incorporates the GRAI's concepts of *horizons* and *periods* to support continuous improvement of overall EA. In CDM ontologies can play a role of seeds with terms, which can form the pieces of decisional solutions (e.g., sentences). Domains of ontologies can be modeled as decision centres and impacts of changes in IT domain can be initially analysed at these centres and then decisions can be made to either approve or decline the integration of a particular service into the BMs (in a manual and automated fashions). GRAI can also be used to decompose each process based on various possible criteria, such as: resource structure, steps of transformation, etc. CDM can help in the analysis of possible changes in system evolution through *horizons*, which are used in the context of future planning. Future planning is associated with services that are part BMs of a system, and relevant analysis can be conducted on tasks basis to emulate likelihood of possible events in times of technological innovations. As in GRAI, in CDM *horizons* are quantified and aim to represent decisions scopes in short-term, mid-term and long-term time scales. That is, if a plan was made for six months then a *horizon* is six months. *Horizons* can be key instruments of decision-making and are also useful in the assessment of likelihood of ungoverned system shifts (e.g. technological, management). *Periods* of CDM aim to re-evaluate the time allocated to a particular plan and can either shrink the time or extended it. *Periods* allow managers to take into account dynamic changes in the environment of a DSS. This change might include internal (e.g. change of system managers, change of system architecture) and external (e.g. service modification, introduction of new platforms) disruptions.

### C. Building Uncertainty Management Engine

There is also a need for the tool that would help in computation of conditional probabilities. In our system, this tool is based on the Bayesian Belief Network (BBN). The aim of this tool is to analyse posterior probability distribution to aid the governance and course of evolution of EA. One of the core advantages of the Bayesian Network (BN) over the most of predictive models, such as neural networks, is its capability to explicitly represent the interrelationships between the dataset attributes [18]. *Periods* of the CDM can be transformed into forecasting centres that can use Bayesian Belief Network for the assessing the circumstances and conditions that influence system performance. This in turn can create system buffers where possible deviations from the forecasted system evolution path are calculated.

### D. Bringing Viable System Model into the Cloud

The heart of the CDM is built upon the Cybernetic model developed by Stafford Beer in 1972 [19]. This model, called the Viable System Model (VSM), consists of 5 sub-systems that help an entire system to operate autonomously, make strategic decisions and remain viable in continuously changing environments. The VSM has a feedback mechanism that escalates alarms and rewards through different levels of recursion when there is a change in a system performance. The CDM incorporates the VSM to aid the viable design of an Enterprise and handle the impacts of technological innovations on the performance of a BA. Various sub-systems of the VSM are different roles in the decision-making. That is, System-1 and System-2 are used in the analysis of current states of a system, whereas System-3 and System-4 are used in the analysis and forecasting of future states.

### E. Risk Remedy System Design

Contemporary Enterprises deliver their goods through services, which in fact are their end products. Shift towards service delivery requires understanding of complex system processes and designs to secure investments and guarantee increasing scalability of a system in the future. SOA [20] and Service Oriented Enterprise Architecture (SOEA) [21] paradigms are the most modern approaches that can handle and meet these needs. The philosophies of SOA and SOEA are based on open system architecture and can help in building systems with time buffers so critical for the smooth integration of innovative technologies and management of uncertainty in times of integrations.

## IV. CONCLUSION, DISCUSSION AND FUTURE WORK

In this work, we highlighted various challenges Cloud Computing can bring to the world of EA in the context of BA and proposed the methods that can be used to overcome them. These challenges are associated with the modifications in technologies integrated into existing BMs of an Enterprise; introductions of new technologies; as well as challenges in decision-making when both modifications and introductions are happening at the same time.

There are various tools that can assist, the transition process of moving an IT system into the Cloud; however, no tools are provided to support Information Systems Management (ISM) once in the Cloud. Our future work aims to develop a method that can analyse relationships between resources across business processes, forecast their evolution at various time scales, and improve the overall quality of systems management. The method is the practical implementation of the CDM that is based on the GRAI Decisional Model and amalgamates the Design Structure Matrices and the Cybernetic concepts of the VSM. It is to be used to determine shifts in system states, detect deviations in the course of system evolution, and aid decision makers with clear schemata of all system entities.

 ISBN: 978-1-61208-237-0 

## REFERENCES


[1] L. Mei, W. Chan, and T. Tse. A Tale of Clouds: Paradigm comparisons and some thoughts on research issues. *In Asia-Pacific Services Computing Conference, 2008*. APSCC'08. IEEE, pages 464–469. IEEE, 2008.

[2] I. Sriram and A. Khajeh-Hosseini. Research agenda in Cloud technologies. Arxiv preprint arXiv:1001.3259, 2010.

[3] M. Vouk. Cloud computing–issues, research and implementations. *Journal of Computing and Information Technology*, 16(4):235–246, 2004.

[4] A. Fox, R. Griffith, et al. Above the Clouds: A Berkeley view of Cloud computing. *Dept. Electrical Eng. and Comput. Sciences, University of California, Berkeley*, Rep. UCB/EECS, 28, 2009.

[5] A. Khajeh-Hosseini, I. Sommerville, and I. Sriram. Research challenges for Enterprise Cloud Computing. Arxiv preprint arXiv:1001.3257, 2010.

[6] L. Youseff, M. Butrico, and D. Da Silva. Toward a unified ontology of Cloud computing. *In Grid Computing Environments Workshop, 2008. GCE'08*, pages 1–10. Ieee, 2008.

[7] J. Carolan, S. Gaede, J. Baty, G. Brunette, A. Licht, J. Remmell, L. Tucker, and J. Weise. Introduction to Cloud computing architecture. White Paper. 2009. http://www.gtsi.com/eblast/corporate/cn/09_09_2009/PDFs/Sun.pdf, [retrieved: May 2012].

[8] J. Viega. Cloud computing and the common man. *Computer*, 42(8):106–108, 2009.

[9] D. Chappell. Introducing the Azure services platform. White Paper. David Chappell & Associates. 2010. http://www.davidchappell.com/writing/white_papers/Introducing_the_Windows_Azure_Platform,_v1.4-Chappell.pdf, [retrieved: May 2012].

[10] E. Gleeson. Computing industry set for a shocking change, 2009. http://www.moneyweek.com/investment-advice/computing-industry-set-for-a-shocking-change-43226, [retrieved: May 2012].

[11] NIST. Definition of cloud computing v15. 2011. http://www.nist.gov/itl/csd/cloud-102511.cfm, [retrieved: May 2012].

[12] O. M. Group. Business Architecture Working Group definition of Business Architecture, 2010. http://bawg.omg.org/, [retrieved: May 2012].

[13] P. Patrick. Impact of SOA on Enterprise Information Architectures. *In Proceedings of the 2005 ACM SIGMOD International.*

[14] S. Aier and R. Winter. Virtual decoupling for IT/Business alignment–conceptual foundations, architecture design and implementation example. *Business & Information Systems Engineering*, 1(2):150–163, 2009.

[15] A. Kalogeras, J. Gialelis, C. Alexakos, M. Georgoudakis, and S. Koubias. Vertical integration of Enterprise industrial systems utilizing web services. *Industrial Informatics, IEEE Transactions on*, 2(2):120–128, 2006.

[16] T. Gruber. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2):199–220, 1993.

[17] G. Doumeingts, B. Vallespir, and D. Chen. GRAI Grid Decisional Modelling. *Handbook on Architectures of Information Systems*, pages 321–346, 2006.

[18] J. Cheng and R. Greiner. Learning Bayesian Belief Network classifiers: Algorithms and System. *Advances in Artificial Intelligence*, pages 141–151, 2001.

[19] S. Beer. Brain of the firm: A Development in Management Cybernetics. *Herder and Herder*, 1972.

[20] T. Erl. Service-oriented Architecture: Concepts, Technology, and Design. *Prentice Hall PTR*, 2005.

[21] R. Knippel. Service Oriented Enterprise Architecture. IT University of Copenhagen, 2005.


 ISBN: 978-1-61208-237-0 

# A Hybrid Method for Extraction of Low-Order Features for Speech Recognition Application

Washington Luis Santos Silva*
*Laboratory of Eletronics Instruments
Federal Institute of Education, Science and Technology
São Luis, Maranhão, Brazil
e-mail:washington.wlss@ifma.edu.br

Ginalber Luiz de Oliveira Serra†
†Laboratory of Computational Intelligence Applied to Technology
Federal Institute of Education, Science and Technology
São Luis, Maranhão, Brazil
e-mail:ginalber@ifma.edu.br

***Abstract*—The concept of fuzzy sets and fuzzy logic is widely used in the proposal of several methods applied to systems modeling, classification and pattern recognition problem. This paper proposes a genetic-fuzzy system for extraction of low-order features for speech recognition application. In addition to pre-processing, with mel-cepstral coefficients, the Discrete Cosine Transform (DCT) is used to generate a two-dimensional time matrix with the features of low-order for each pattern to be recognized. A genetic algorithm is used to optimize a Mamdani fuzzy inference system in order to obtain the best model for final recognition. The proposed method used in this paper was named Hibrid Method for Extraction of Low-Order Features for Speech Recognition Application (HMFE). Experimental results for speech recognition applied to Brazilian language show the efficiency of the proposed methodology compared to methodologies widely used and cited in the literature.**

***Keywords*–Discrete Cosine Transform; Speech Recognition; Fuzzy Systems; Genetic Algorithm.**

## I. INTRODUCTION

Parameterization of an analog speech signal is the first step in speech recognition process. Several popular signal analysis techniques have emerged as standards in the literature. These algorithms are intended to produce a perceptually meaningful parametric representation of the speech signal: parameters that can emulate some behavior observed in human auditory and perceptual systems. Actually, these algorithms are also designed to maximize recognition performance [1][2]. The selection of best representation for parametric speech signal is a very important task of developing any speech recognition system. The problem of pattern recognition might be formulated as follows: Let $S_k$ classes, where $k \in \{1, 2, 3...K\}$, and $S_k \subset \Re^n$. If any pattern space is taken with dimension $\Re^x$, where $x \leq n$, it should transform this space into a new pattern space with dimension $\Re^a$, where $a < x \leq n$. Then assuming a statistical measure or second order model for each $S_k$, through a covariance function represented by $\left[\Phi_x^{(k)}\right]$, the covariance matrix of the general pattern recognition problem becomes:

$$[\Phi_x] = \sum_{k=1}^{K} P(S_k) \left[\Phi_x^{(k)}\right] \quad (1)$$

where $P(S_k)$ is a distribution function of the class $S_k$, *a priori*, with $0 \leq P(S_k) \leq 1$. A linear transformation operator through the matrix $\mathbf{A}$ maps the pattern space in a transformed space where the columns are orthogonal basis vectors of this matrix $\mathbf{A}$. The patterns of the new space are linear combinations of the original axes as structure of the matrix $\mathbf{A}$. The statistics of second order in the transformed space are given by:

$$\Phi_{\mathbf{A}} = \mathbf{A}^T[\Phi_x]\mathbf{A} \quad (2)$$

where $\Phi_{\mathbf{A}}$ is the covariance matrix which corresponds to the space generated by the matrix $\mathbf{A}$ and the operator $[\cdot]^T$ corresponds to the transpose of a matrix. Thus, it can extract features that provide greater discriminatory power for classification from the dimension of the space generated [3]. One of the most widespread techniques for pattern speech recognition is the "Hidden Markov Model" (HMM)[4]. A well known deficiency of the classical HMMs is the poor modeling of the acoustic events related to each state. Since the probability of recursion to the same state is constant, the probability of the acoustic event related to the state is exponentially decreasing. A second weakness of the HMMs is that the observation vectors within each state are assumed uncorrelated, and these vectors are correlated [5]. To overcome these drawbacks, robust recognizer has been proposed, since it has been experimentally shown that spectral variations are discriminant features for similar sounds. Several errors occur because an observation sequence is decoded by a few states typically absorbing low-energy frames [6]. The other states, instead, are quickly crossed because their distribution do not adapt well to the rest of the observation. Therefore, these errors do not depend on the intrinsic confusion of the words with similar sound, but on the poor modeling of the acoustic event which produces hypothesis weakly related to the acoustics of the correct word [7]. In order to justify the dynamic structure of the observation vectors, including global and local variations, this paper proposes a speech recognition system for isolated digits that are not based directly on the modeling of the state/word, but based on the global changes in the spectral characteristics of each word and their correlation in time, two important features partially explored by classical HMM [8][9].

Recently several works on digit recognition has been presented using MFCC classifiers and Neural Networks [10][11][12], Hybrid HMM-Suport Vector Machine (HMM-SVM) [13], Sparse Systems for Speech Recognition [14],

 ISBN: 978-1-61208-237-0 

Hybrid Robust Voice Activity Detection System [15], Wolof Speech Recognition with Limited vocabulary Based HMM and Toolkit [16], Real-Time Robust Speech Recognition using Compact Support Vector Machines [17], Digit Recognition with Confidence [18], and others.

### A. Proposed Methodology

In this proposal, a speech signal is encoded and parameterized in a two-dimensional time matrix with four parameters of the speech signal. After coding, the mean and variance of each pattern are used to generate the rule base of Mamdani fuzzy inference system. The mean and variance are optimized using genetic algorithm in order to have the best performance of the recognition system. This paper consider as patterns the Brazilian locutions (digits): $'0', '1', '2', '3', '4', '5', '6', '7', '8', '9'$. The Discrete Cosine Transform (DCT) [19][20] is used to encoding the speech patterns. The use of DCT in data compression and pattern classification has been increased in recent years, mainly due to the fact its performance is much closer to the results obtained by the Karhunen-Loève transform which is considered optimal for a variety of criteria such as mean square error of truncation and entropy [21]. This paper demonstrates the potential of DCT and fuzzy inference system in speech recognition [22]. These two tools have shown good results in the temporal modeling of speech signal [23].

## II. Speech Recognition System

The proposed recognition system HMFE block diagram is depicted in Fig. 1.

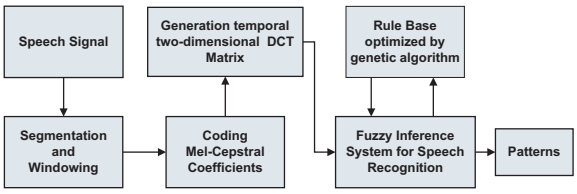



Fig. 1. Block diagram of the proposed recognition system HMFE.

### A. Pre-processing Speech Signal

Initially, the speech signal is digitizing, so it is divided in segments which are windowed and encoded in a set of parameters defined by the order of mel-cepstral coefficients (MFCC). The DCT coefficients are computed and the two-dimensinal time DCT matrix is generated, based on each speech signal to be recognized.

*1) Segmentation and windowing of the speech signal:* When a window is applied to a given signal, it selects a small portion of this signal, named frame, to be analyzed. The duration of the frame $T$ is defined as the total of time over which a set of parameters is considered valid. The duration of the frame is used to determine the total of time from successive calculations of parameters [1]. It is necessary to use a process called overlap to control how quickly the signal parameters may change from frame to frame because the windows at the ends of the analyzed signal have an excessive smoothing in their samples. In speech processing, the Hamming window is widely. This paper uses the hamming window with duration time (frames) of 10ms with 50% overlap between frames, thus, only a fraction of the signal is changed for each new frame .

*2) Mel-Cepstrais Coefficients Coding:* Experiments on human perception have shown that complex sound frequencies within a certain bandwidth of a nominal frequency should not be individually identified. When one of the components of this sound is out of bandwidth, this component can not be distinguished. Normally, it is considered a critical bandwidth for speech from 10 % to 20 % of the center frequency of the sound. One of the most popular way to map the frequency of a given sound signal for perceptual frequencies values, i.e., to be capable of exciting the human hearing range is the Mel-Scale [2].

### B. Two-Dimensional Time Matrix DCT Coding

The two-dimensional time matrix as the result of DCT in a sequence of $T$ mel-cepstral coefficients observation vectors on the time axis, is given by:

$$C_k(n,T) = \frac{1}{N}\sum_{t=1}^{T} mfcc_k(t) cos\frac{(2t-1)n\pi}{2T} \qquad (3)$$

where $mfcc$ are the mel-cepstral coefficients, and $k, 1 \leq k \leq K$, is the $k$-th (line) component of $t$-th frame of the matrix and $n, 1 \leq n \leq N$ (column) is the order of DCT. Thus, the two-dimensional time matrix [24], where the interesting low-order coefficients $k$ and $n$ that encode the long-term variations of the spectral envelope of the speech signal is obtained [7].

For a given spoken word *P* (digit), ten examples of utterances of *P* are gotten. This way it has itself $P_0^0, P_1^0, ..., P_9^0, P_0^1, P_1^1 ..., P_9^1, P_0^2, P_1^2, ..., P_9^2, ..., P_m^j$, where $j \in \{0, 1, 2, ..., 9\}$ and $m \in \{0, 1, 2, ..., 9\}$. Each frame of a given example of the word ***P*** generates a total of $K$ mel-cepstral coefficients and the significant features are taken for each frame along time. The *N-th* order DCT is computed for each mel-cepstral coefficient of same order within the frames distributed along the time axis, i.e., $c_1$ of the frame $t_1$, $c_1$ of the frame $t_2, ...$, $c_1$ of the frame $t_T$, $c_2$ of the frame $t_1$, $c_2$ of the frame $t_2, ...,$, $c_2$ of the frame $t_T$, and so on, generating elements $\{c_{11}, c_{12}, c_{13}, ..., c_{1N}\}$, $\{c_{21}, c_{22}, c_{23}, ..., c_{2N}\}$, $\{c_{K1}, c_{K2}, c_{K3}, ..., c_{KN}\}$ of the matrix given in equation (3). Therefore, a two-dimensional time matrix DCT is generated for each example of the word ***P***. In this paper, the two-dimensional time matrices generated has order $(K = 2) \times (N = 2)$.

Finally, the matrices of mean $CM_{kn}^j$ (4) and variances $CV_{kn}^j$(5) are generated. The parameters of $CM_{kn}^j$ and $CV_{kn}^j$ are used to produce Gaussians matrices $C_{kn}^j$ which will be used as fundamental information for implementation of the fuzzy recognition system. The parameters of this matrix will

 ISBN: 978-1-61208-237-0 

be optimized by genetic algorithm.

$$CM_{kn}^{j} = \frac{1}{M}\sum_{m=0}^{M-1} C_{kn}^{jm} \quad (4)$$

$$CV_{kn}^{j}(var) = \frac{1}{M-1}\sum_{m=0}^{M-1}\left[C_{kn}^{jm} - \left(\frac{1}{M}\sum_{m=0}^{M-1} C_{kn}^{jm}\right)\right]^{2} \quad (5)$$

where M=10.

### C. Rule Base Used for Speech Recognition

Given the fuzzy set $A$ input, the fuzzy set $B$ output, should be obtained by the relational max-t composition [25]. This relationship is given by.

$$B = A \circ Ru \quad (6)$$

where $Ru$ is a fuzzy relational rules base.

The fuzzy rule base of practical systems usually consists of more than one rule. There are two ways to infer a set of rules: Inference based on composition and inference based on individual rules [26][27]. In this paper the compositional inference is used. Generally, a fuzzy rule base is given by:

$$Ru^{l} : IF\ \ x_1\ \ is\ \ A_1^{l}\ \ and...and\ \ x_n\ \ is\ \ A_n^{l}\ \ THEN\ \ y\ \ is\ \ B^{l} \quad (7)$$

where $A_i^l$ and $B^l$ are fuzzy set in $U_i \subset \Re$ and $V \subset \Re$, and $x \in \{x_1, x_2, ..., x_n\}^T \in U$ and $y \in V$ are input and output variables of fuzzy system, respectively. Let $M$ be the number of rules in the fuzzy rule base; that is, $l \in \{1, 2, ...M\}$.

From the coefficients of the matrices $\mathbf{C}_{kn}^{j}$ with $j \in \{0, 1, 2, ..., 9\}$, $k \in \{1, 2\}$ and $n \in \{1, 2\}$ generated during the training process, representing the mean and variance of each pattern $j$ a rule base with $M = 40$ individual rules is obtained and given by:

$$Ru^{j} : IF\ \ C_{kn}^{j}\ \ THEN\ \ y^{j} \quad (8)$$

In this paper, the training process is based on the fuzzy relation $Ru^j$ using the Mamdani implication. The rule base $Ru^j$ should be considered a relation $R(X \times Y) \rightarrow [0, 1]$, computed by:

$$\mu_{Ru}(x, y) = I(\mu_A(x), \mu_B(y)) \quad (9)$$

where the operator $I$ should be any t-norm [28][29][30]. Given the fuzzy set $A'$ input, the fuzzy set $B'$ output might be obtained by ***max-min*** composition, [26]. For a minimum t-norm and max-min composition it yields:

$$\mu_{(Ru)}(x, y) = I(\mu_A(x), \mu_B(y)) = min(\mu_A(x), \mu_B(y)) \quad (10)$$

$$\mu_{(B')} = max_x min_{x,y}(\mu_{A'}(x), \mu_{(Ru)}(x, y)) \quad (11)$$

### D. Generation of Fuzzy Patterns

The elements of the matrix $C_{kn}^{j}$ were used to generate Gaussians membership functions in the process of fuzzification. For each trained model $j$ the Gaussians memberships functions $\mu_{c_{kn}^{j}}$ are generated, corresponding to the elements $c_{kn}^{j}$ of the two-dimensional time matrix $\mathbf{C}_{kn}^{j}$ with $j \in \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$, where $j$ is the model used in training. The training system for generation of fuzzy patterns is based on the encoding of the speech signal ***s(t)***, generating the parameters of the matrix $C_{kn}^{j}$. Then, these parameters are fuzzified, and they are related to properly fuzzified output $y^j$ by the relational implications, generating a relational surface $\mu_{(Ru)}$, given by:

$$\mu_{Ru} = \mu_{c_{kn}^{j}} \circ \mu_{y^j} \quad (12)$$

This relational surface is the fuzzy system rule base for recognition optimized by genetic algorithm to maximize the speech recognition. The training system is shown in Fig. 2.

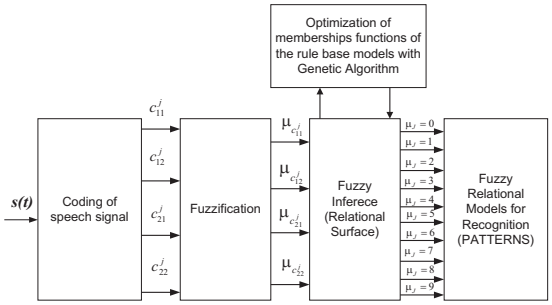



Fig. 2. Generation Systems Fuzzifieds Models.

### E. Fuzzy Inference System for Speech Recognition Decision

The decision phase is performed by a fuzzy inference system based on the set of rules obtained from the mean and variance matrices of two dimensions time of each spoken digit. In this paper, a matrix with minimum number of parameters $(2 \times 2)$ in order to allow a satisfactory performance compared to pattern recognizers available in the literature. The elements of the matrices $\mathbf{C}_{kn}^{j}$ are used by the fuzzy inference system to generate four Gaussian membership functions corresponding to each element $c_{kn}^{j}\Big|^{k\in\{1,2\};n\in\{1,2\}}$ of the matrix. The set of rules of the fuzzy relation is given by:

***Rule Bases***

$$\textbf{\textit{IF}}\ \ c_{kn}^{j}\Big|^{k\in\{1,2\};n\in\{1,2\}}\ \ \textbf{\textit{THEN}}\ \ y^{j} \quad (13)$$

***Modus Ponens***

$$\textbf{\textit{IF}}\ \ c_{kn}^{'j}\Big|^{k\in\{1,2\};n\in\{1,2\}}\ \ \textbf{\textit{THEN}}\ \ y^{'j} \quad (14)$$

From the set of rules of the fuzzy relation between antecedent and consequent, a data matrix for the given implication is obtained. After the training process, the relational surfaces is generated based on the rule base and implication method

 ISBN: 978-1-61208-237-0 

presented in Section II.D. The speech signal is encoded to be recognized and their parameters are evaluated in relation to the functions of each patterns on the surfaces and the degree of membership is obtained. The final decision for the pattern is taken according to the $max - min$ composition between the input parameters and the data contained in the relational surfaces. The process of defuzzification for the pattern recognition is based on the ***mean of maxima (mom)*** method given by:

$$\mu_{y'^j} = \mu_{c'^j_{kn}} \circ \mu_{(Ru)} \quad (15)$$

$$y^{'} = mom(\mu_{y'^j}) = mean\{y|\mu_{y'^j} = max_{y \in Y}(\mu_{y'^j})\} \quad (16)$$

### *F. Optimization of Relational Surface with Genetic Algorithm*

The continuous genetic algorithm [31][32] is configured with a population size of 100, generations of 300, with mutations probability of 15% and two chromosomes, with 40 genes each, to optimize a cost function with 80 variables, which are the mean and variances of the patterns to be recognized by the proposed fuzzy recognition system. The genetic algorithm was used to optimize the variations of mean and variances of each pattern in order to maximize the successful recognition process. For example, for the pattern of the spoken word "*zero*" is generated ten two-dimensional time matrix. For each element of the matrix $C^j_{kn}$ coefficients are determined with variations minimum and maximum , and the coefficient $c_{11} \in [c_{11(minimum)}\ c_{11(maximum)}]$, $c_{12} \in [c_{12(minimum)}\ c_{12(maximum)}]$, $c_{21} \in [c_{21(minimum)}\ c_{21(maximum)}]$, $c_{22} \in [c_{22(minimum)}\ c_{22(maximum)}]$. Thus, it has eight time varying parameters for each pattern which correspond to eighty parameters to be optimized by genetic algorithm [33].

## III. Experimental Results

### *A. System Training*

The patterns to be used in the recognition process were obtained from ten speakers who are speaking the digits 0 until 9. After pre-processing of the speech signal and fuzzification of the matrix $C^j_{kn}$, its fuzzifieds components $\mu_{c^j_{kn}}$ had been optimized by the GA that maximize the total of successful recognition. The optimization process was performed with 16 realizations of the genetic algorithm.The best result of the recognition processing by HMFE is shown in Fig. 3. The total number of hits using GA was 92 digits correctly identified in the training process. The relational surface generated for this result was used for validation process. The best individual in the first generation of the GA is shown in Fig. 4. In this case the total number of correct answers was 46 digits correctly identified. The relational surface of the best individual in the first generation of the GA is shown in Fig. 5.

In Fig. 6 are shown the features of the Gaussians membership functions of the optimum individual after the training process. This figure also shows a better distribution of the Gaussians membership functions organized by the GA during the training process. Fig. 7 presents the relational surface generated by Gaussians membership functions of the optimum individual after the training process, already organized by the GA. The better distribution of the Gaussians membership functions, made by the GA shown in Fig. 6 and Fig. 7 improved results due to reduction intrinsic confusion of the Gaussians membership functions shown in Fig. 4 and Fig. 5.

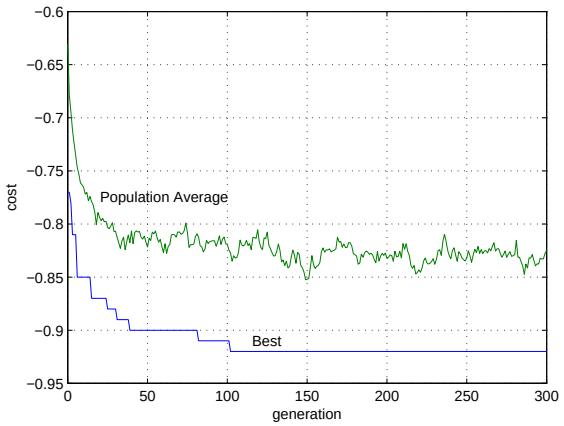



Fig. 3. Plot of the best results obtained in the training process.

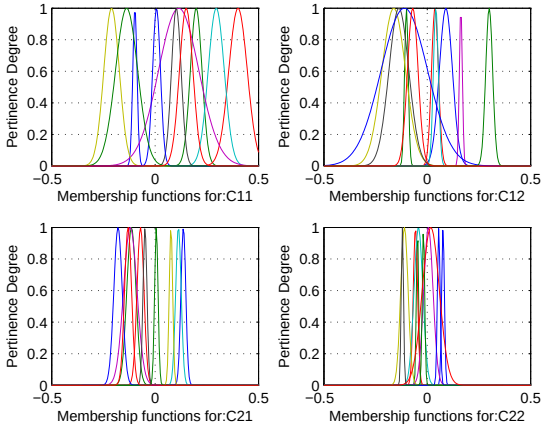



Fig. 4. Membership functions for $c^j_{kn}$ in the 1st generation.

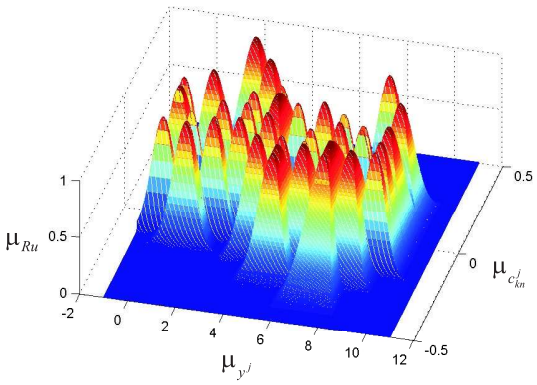



Fig. 5. Relational surface ($\mu_{Ru}$) in the 1st generation.

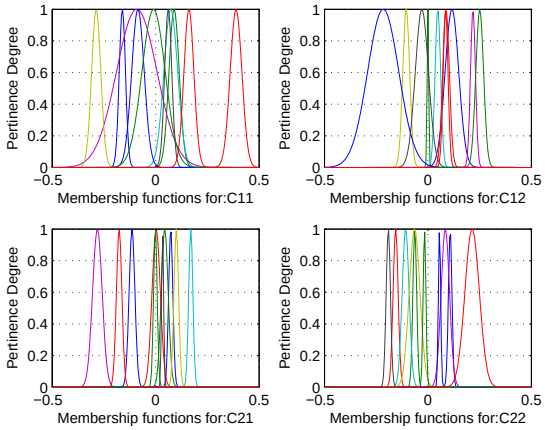



Fig. 6. Membership functions for $c_{kn}^{j}$ optmized by GA.

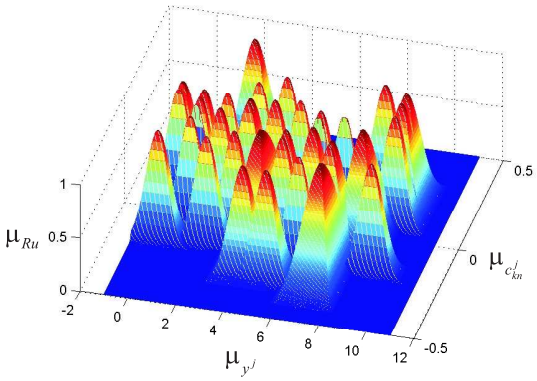



Fig. 7. Relational surface ($\mu_{Ru}$) optmized by GA.

## B. System Test - Validation

In this step, 100 locutions uttered in a room with controlled noise level and 500 locutions uttered in an environment without any kind of noise control were used. For every ten examples of each spoken digit, was generated two-dimensional time matrix cepstral coefficients $\mathbf{C}_{kn}^{j}$ and they were used in the test procedure. Six types of tests where performed:

Training: Recognition Optimized by HMFE (5 Female and 5 Male Speakers)

TEST 1: Validation - Strictly speaker dependent recognition, where the words used for training and testing were spoken by a same group of 10 speakers(5 Female and 5 Male Speakers).

TEST 2: Validation test- Recognition based on the partial dependence of the speaker with two examples for each ten examples of each digit(Female Speaker).

TEST 3: Validation test- Recognition based on the partial dependence of the speaker with two examples for each ten examples of each digit(Male Speaker).

TEST 4: Validation test- Recognition independent of the Speaker, where the speaker does not have influence in the training process(Female Speaker).

TEST 5: Validation test- Recognition independent of the Speaker, where the speaker does not have influence in the training process(Male Speaker).

Figures 8 - 13 present the comparative analysis of the HMM with two, three and four states, two, three and four Gaussians mixtures by state and order analysis, i.e., the number of mel-cepstral parameter equal 12 and HMFE with two, three and four parameteres for speech recognition.

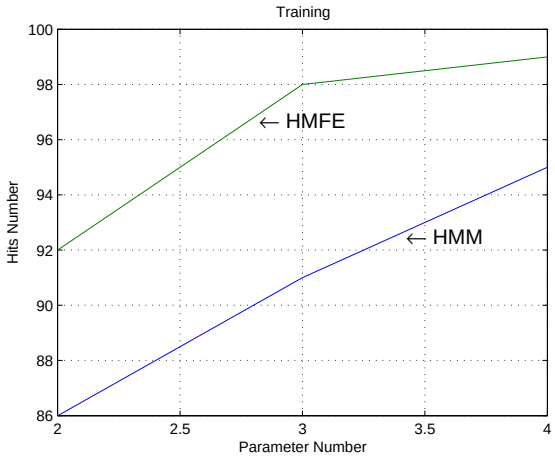



Fig. 8. Results for the digits used in the training.

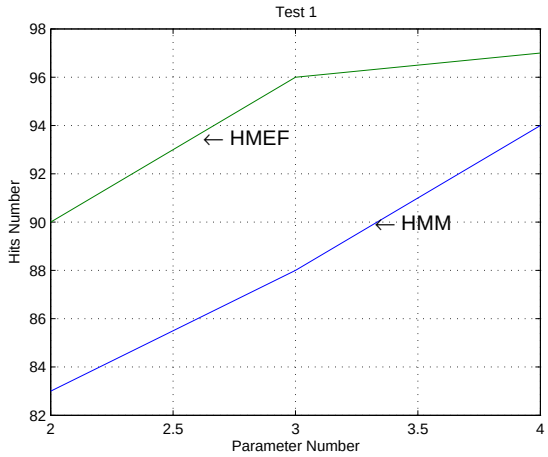



Fig. 9. Validation Test 1.

 ISBN: 978-1-61208-237-0 

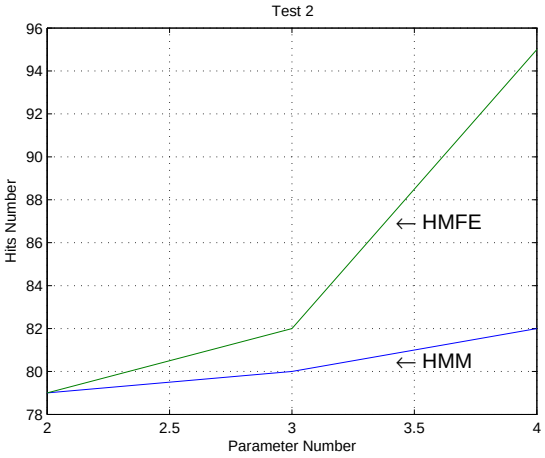



Fig. 10. Validation Test 2.

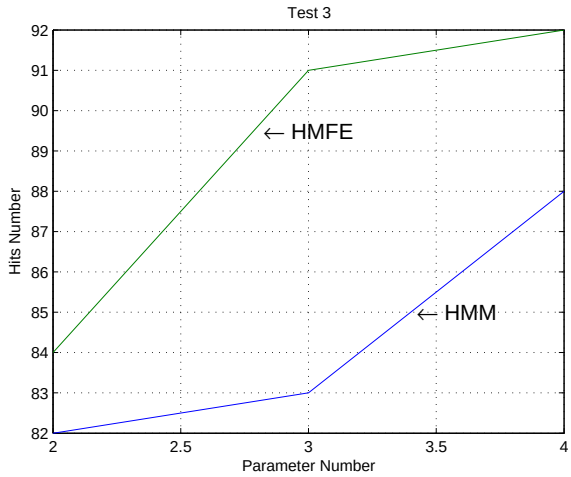



Fig. 11. Validation Test 3.

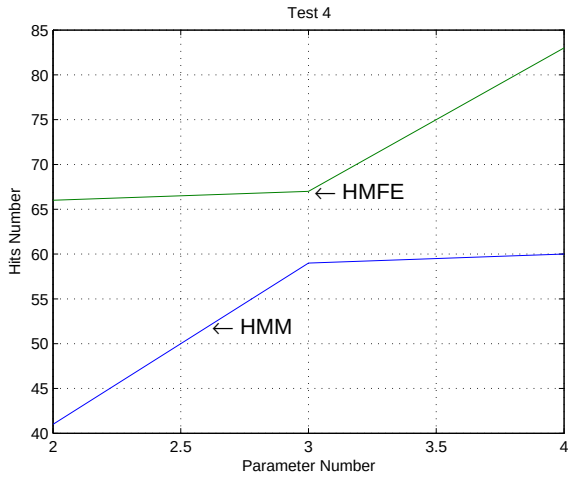



Fig. 12. Validation Test 4.

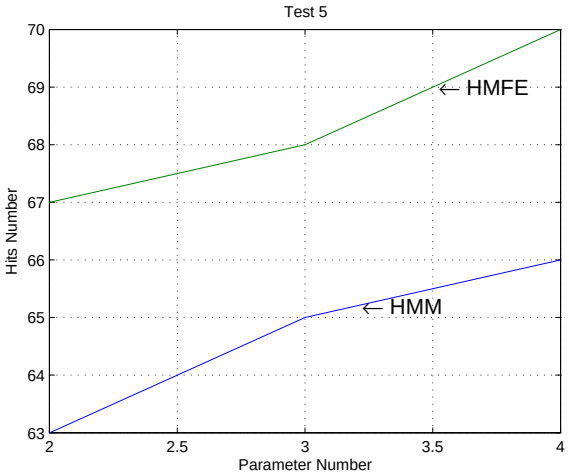



Fig. 13. Validation Test 5.

## IV. Conclusion and Future Work

Evaluating the results, it is observed that the proposed method for extraction of low-order features for speech recognition application (HMFE), even with a minimal parameters number in the generated patterns was able to extract more reliably the temporal characteristics of the speech signal and produce good recognition results compared with the traditional HMM. To obtain equivalent results with HMM is necessary to increase the state number and/or mixture number. An increase in the order of the analysis above 12 does not improve significantly the performance of HMM. Any particular technique of noise reduction, such as those commonly used in HMM-based recognizers, was not used during the development of this paper. It is believed that with proper treatment of the signal to noise ratio in the process of training and testing, the HMFE Recognizer may improve its performance:

1) Increase the speech bank with different accents;
2) Use Nonlinear Predictive Coding for feature extraction in speech recognition;
3) Use Digital Filter in the speech signal to be recognized.
4) Increase the parameters number used.

## Acknowledgment

The authors would like to thank FAPEMA for financial support, research group of computational intelligence applied to technology at the Federal Institute of Education, Science and Technology of the Maranhão by its infrastructure for this research and experimental results.

## References

[1] J.W. Picone, “Signal Modeling Techniques in Speech Recognition”, IEEE Transactions on Computer, vol. 81, 9th edition, Apr. 1993, pp. 1215-1247, doi: 10.1109/5.237532.

[2] L. Rabiner and J. Biing-Hwang, “Fundamentals of Speech Recognition”, Prentice Hall, New Jersey, 1993.

[3] H. C. Andrews, “Multidimensional Rotations in Feature Selection”, IEEE Transaction on Computers, Sep. 1971, pp. 1045-1051, doi: 10.1109/T-C.1971.223400.

 ISBN: 978-1-61208-237-0 

[4] A. A. M. Abushariah, T. S. Gunawan, O. O. Khalifa and M. A. M. Abushariah, “English Digits Speech Recognition System Based on Hidden Markov Models”, International Conference on Computer and Communication Engineer (ICCCE 2010), Kuala Lumpur, Malaysia, May 2010, pp. 1-5, doi:10.1109/ICCCE.2010.5556819.

[5] M.Wachter, M. Matton, K. Demuynck, P.K. Wambacq, R. Cools and D. Compernolle, “Template-Based Continuous Speech Recognition”, IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, no. 4, May 2007, pp. 1377-1390, doi: 10.1109/TASL.2007.894524.

[6] Y. Ariki, S. Mizuta, M. Nagata and T. Sakai, “Spoken-Word Recognition Using Dynamic Features Analysed by Two-Dimensional Cepstrum”, IEEE Proceedings, vol. 136, no. 2, Apr. 1989, pp. 133-140.

[7] P. L. L. Fissore and E. Rivera, “Using Word Temporal Structure in HMM Speech Recognition”, ICASSP 97, vol. 2, Munich, Germany, Apr. 1997, pp. 975-978, doi: 10.1109/ICASSP.1997.596101.

[8] A. Revathi and Y. Venkataramani, “Speaker Independent Continuous Speech and Isolated Digit Recognition using VQ and HMM”, International Conference on Communications and Signal Processing (ICCSP), Calicut, India, Feb. 2011, pp. 198-202, doi: 10.1109/ICCSP.2011.5739300.

[9] J. Deng, M. Bouchard and T. H. Yeap, “Feature Enhancement for Noisy Speech Recognition with a Time-Variant Linear Predictive HMM Structure”, IEEE Transactions on Audio, Speech, and Language Processing, vol. 16, no. 5, Jul. 2008, pp. 891-899, doi: 10.1109/TASL.2004.924593.

[10] D. B. Hanchate, M. Nalawade, M. Pawar, V. Pohale and P. K. Maurya, “Vocal Digit Recognition Using Artificial Neural Network”, 2nd International Conference on Coumputer Engineering and Technology, vol. 6, Chendu, China, Apr. 2010, pp. 88-91, doi: 10.1109/ICCET.2010.5486314.

[11] R. K. Aggarwal and M. Dave, “Application of Genetically Optimized Neural Networks for Hindi Speech Recognition System”, World Congress on Information and Communication Technologies (WICT), Mumbai, India, Dec. 2011, pp. 512-517, doi: 10.1109/WICT.2011.6141298.

[12] S. M. Azam, Z. A. Mansor, M. S. Mughal and S. Moshin, “Urdu Spoken Digits Recognition Using Classifield MFCC and Backpropagation Neural Network”, 4th International Conference on Computer Graphics, Imaging and Visualization (CGIV), Bangkok, Thailand, Aug. 2007, pp. 414-418, doi: 10.1109/CGIV.2007.85.

[13] S. A. Hejazi, R. Kazemi and S. Ghaemmaghami, “Isolated Persian Digit Recognition Using a Hybrid HMM-SVM”, International Symposium on Intelligent Signal Processing and Communications Systems (ISPACS), Bangkok, Thailand, Dec. 2008, pp. 1-4, doi: 10.1109/ISPACS.2009.4806757.

[14] M. Mohammed, E. Bijov, C. Xavier, A. K. Yasif and V. Supriya, “Robust Automatic Speech Recognition Systems:HMM Vesus Sparse”, Third International Conference on Intelligent Systems modelling and Simulation, Kinabalu, Malaysia, Feb. 2012, pp. 339-342, doi: 10.1109/ISMS.2012.66.

[15] C. Ganesh, H. Kumar and P. T. Vanathi, “Performance Analysis of Hybrid Robust Automatic Speech Recognition System”, IEEE International Conference on Signal Processing, Computing and Control (ISPCC), Solan, India, Mar 2012, pp. 1-4, doi: 10.1109/ISMS.2012.66.

[16] J. K. Tamgo, E. Barnard, C. Lishou and M. Richome, “Wolof Speech Recognition Model of Digits and Limited-Vocabulary Based on HMM and ToolKit”, 14th International Conference on Computer Modelling and Simulation (UKSim), Cambridge, United Kingdom, Mar. 2012, pp. 389-395, doi: 10.1109/UKSim.2012.118.

[17] R. Solera, A. Moral, C. Moreno, M. Ramon and F. Maria, “Real-Time Robust Automatic Speech Recognition Using Compact Support Vector Machines”, IEEE Transactions on Audio, Speech, and Language Processing, vol.20, no. 4, May 2012, pp. 1347-1361, doi: 10.1109/TASL.2011.2178597.

[18] G. E. Sakr and I. H. Elhajj, “Digit Recognition with Confidence”, IEEE Workshop on Signal Processing Systems (SiPS), Beirut, Lebanon, Oct. 2011, pp. 299-304, doi: 10.1109/SiPS.2011.6088993.

[19] T. N. N. Ahmed and K. Rao, “Discrete Cosine Transform”, IEEE Transaction on Computers, vol.c-24, 2th edition, Jan. 1974, pp. 90-93, doi: 10.1109/T-C.1974.223784.

[20] P. C. J. Zhou, “Generalized Discrete Cosine Transform”, Pacific-Asia Conference on Circuits, Communications and System, Chegdu, China, May 2009, pp. 449-452, doi: 10.1109/PACCS.2009.62.

[21] M. Effros, H.Feng and K. Zeger, “Suboptimality of the KarhunenLove Transform for Transform Coding”, IEEE Transactions on Information Theory, vol. 50, no. 8, Aug. 2004, pp. 293-302, doi: 10.1109/DCC.2003.1194020.

[22] J. Zeng and Z. Q. Liu, “Type-2 Fuzzy Hidden Markov Models and their Application to Speech Recognition”, IEEE Transactions on Fuzzy Systems, vol. 14, no. 3, Jun. 2006, pp. 454-467, doi: 10.1109/TFUZZ.2006.876366.

[23] W. L. S. Silva and G. L. O. Serra, “Proposta de Metodologia TCD-Fuzzy para Reconhecimentos de Voz”, X SBAI Simposio Brasileiro de Automacao Inteligente, Sao Joao del-Rei, Brasil, Sep. 2011, pp. 1054-1059.

[24] M.Y. Azar and F. Razzazi, “A DCT Based Nonlinear Predictive Coding for Feature Extraction in Speech Recognition Systems”, IEEE International Conference on Computational Intelligente for Measurement Systems and Applications, Istanbul, Turkey, Jul. 2008, pp. 19-22, doi: 10.1109/CIMSA.2008.4595825.

[25] M. Mas, M. Monserrat, J. Torrens and E. Trillas, “A Survey on Fuzzy Implication Functions”, IEEE Transactions on Fuzzy Systems, vol.15, no.6, Dec. 2007. pp. 1107-1121, doi: 10.1109/TFUZZ.2007.896304.

[26] L.-X. Wang, “A Course in Fuzzy Systems and Control”, Prentice Hall, 1994.

[27] C. Gang, “Discussion of Approximation Properties of Minimum Inference Fuzzy System”, Proceedings of the 29th Chinese Control Conference, Beijing, China, Jul. 2010, pp. 2540-2546.

[28] R. Babuska, “Fuzzy Modeling for Control”, Kluwer Academic Publishers, 1998.

[29] H. Seki, H. Ishii and M. Mizumoto, “On the Monotonicity of Fuzzy-Inference Methods Related to TS Inference Method”, IEEE Transactions on Fuzzy Systems, vol. 18, no. 3, Jun. 2010, pp. 629-634, doi: 10.1109/TFUZZ.2010.2046668.

[30] G. Gosztolya, J. Dombi and A. Kocsor, “Applying the Generalized Dombi Operator Family to the Speech Recognition Task”, Journal of Computing and Information Technology - CIT, vol. 17, no. 9, 2009, pp. 285-293, doi: :10.2498/cit.1001284.

[31] R. L. Haupt and S. E. Haupt, “Pratical Genetic Algorithms”, John Wiley & Sons, Inc, 2004.

[32] C. Tang, E. Lai and Y. C. Wang, “Distributed Fuzzy Rules for Preprocessing of Speech Segmentation with Genetic Algorithm”, Fuzzy-IEEE Conference 1997, vol. 1, Barcelona, Spain, Jul. 1997, pp. 427-431, doi: 10.1109/FUZZY.1997.616406.

[33] K. Tang, K. Man, Z. Liu and S. Kwong, “Minimal Fuzzy Memberships and Rules Using Hierarchical Genetic Algorithms”, IEEE Transactions on Industrial Eletronics, vol. 45, no. 1, Feb. 1998. pp. 427-431, doi: 10.1109/FUZZY.1997.616406.

 ISBN: 978-1-61208-237-0 