

Zdzisław S. Hippe
Juliusz L. Kulikowski
Teresa Mroczek (Eds.)

Human – Computer Systems Interaction: Backgrounds and Applications 2

Advances in Intelligent and Soft Computing

98

Editor-in-Chief: J. Kacprzyk

Advances in Intelligent and Soft Computing

Editor-in-Chief

Prof. Janusz Kacprzyk
Systems Research Institute
Polish Academy of Sciences
ul. Newelska 6
01-447 Warsaw
Poland
E-mail: kacprzyk@ibspan.waw.pl

Further volumes of this series can be found on our homepage: springer.com

Vol. 88. Y. Demazeau, M. Pěchouček,
J.M. Corchado, and J.B. Pérez (Eds.)
*Advances on Practical Applications of Agents
and Multiagent Systems, 2011*
ISBN 978-3-642-19874-8

Vol. 89. J.B. Pérez, J.M. Corchado,
M.N. Moreno, V. Julián, P. Mathieu,
J. Canada-Bago, A. Ortega, and
A.F. Caballero (Eds.)
*Highlights in Practical Applications of Agents
and Multiagent Systems, 2011*
ISBN 978-3-642-19916-5

Vol. 90. J.M. Corchado, J.B. Pérez,
K. Hallenborg, P. Golinska, and
R. Corchuelo (Eds.)
*Trends in Practical Applications of Agents
and Multiagent Systems, 2011*
ISBN 978-3-642-19930-1

Vol. 91. A. Abraham, J.M. Corchado,
S.R. González, J.F. de Paz Santana (Eds.)
*International Symposium on Distributed
Computing and Artificial Intelligence, 2011*
ISBN 978-3-642-19933-2

Vol. 92. P. Novais, D. Preuveneers, and
J.M. Corchado (Eds.)
*Ambient Intelligence - Software and
Applications, 2011*
ISBN 978-3-642-19936-3

Vol. 93. M.P. Rocha, J.M. Corchado,
F. Fernández-Riverola, and A. Valencia (Eds.)
*5th International Conference on Practical
Applications of Computational Biology &
Bioinformatics 6-8th, 2011*
ISBN 978-3-642-19913-4

Vol. 94. J.M. Molina, J.R. Casar Corredera,
M.F. Cátedra Pérez, J. Ortega-García, and
A.M. Bernardos Barbolla (Eds.)
*User-Centric Technologies and
Applications, 2011*
ISBN 978-3-642-19907-3

Vol. 95. Robert Burduk, Marek Kurzyński,
Michał Woźniak, and Andrzej Żołnierek (Eds.)
Computer Recognition Systems 4, 2011
ISBN 978-3-642-20319-0

Vol. 96. A. Gaspar-Cunha, R. Takahashi,
G. Schaefer, and L. Costa (Eds.)
Soft Computing in Industrial Applications, 2011
ISBN 978-3-642-20504-0

Vol. 97. W. Zamojski, J. Kacprzyk,
J. Mazurkiewicz, J. Sugier,
and T. Walkowiak (Eds.)
Dependable Computer Systems, 2011
ISBN 978-3-642-21392-2

Vol. 98. Z.S. Hippe, J.L. Kulikowski,
and T. Mroczek (Eds.)
*Human – Computer Systems Interaction:
Backgrounds and Applications 2, 2012*
ISBN 978-3-642-23186-5

Zdzisław S. Hippe, Juliusz L. Kulikowski,
and Teresa Mroczek (Eds.)

Human – Computer Systems Interaction: Backgrounds and Applications 2

Part 1

Editors

Dr. Zdzisław S. Hippe
Department of Expert Systems and
Artificial Intelligence,
University of Information Technology
and Management,
35-225 Rzeszów,
Poland
E-mail: zhippe@wsiz.rzeszow.pl

Dr. Teresa Mroczek
Department of Expert Systems and
Artificial Intelligence,
University of Information Technology
and Management,
35-225 Rzeszów,
Poland
E-mail: tmroczek@wsiz.rzeszow.pl

Dr. Juliusz L. Kulikowski
Polish Academy of Sciences,
M. Nalecz Institute of Biocybernetics and
Biomedical Engineering,
4 Ks. Trojdena Str.,
02-109 Warsaw,
Poland
E-mail: juliusz.kulikowski@ibib.waw.pl

ISBN 978-3-642-23186-5

e-ISBN 978-3-642-23187-2

DOI 10.1007/978-3-642-23187-2

Advances in Intelligent and Soft Computing

ISSN 1867-5662

Library of Congress Control Number: 2011936642

© 2012 Springer-Verlag Berlin Heidelberg

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilm or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable to prosecution under the German Copyright Law.

The use of general descriptive names, registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

Typeset & Cover Design: Scientific Publishing Services Pvt. Ltd., Chennai, India

Printed on acid-free paper

5 4 3 2 1 0

springer.com

From the Editors

The history of human-system interactions is as long as this of human civilization. Human beings by natural evolution have been adapted to live in groups and to commonly fight for food and shelter against other groups or against the natural forces. The effects of this fight was depended on two basic factors: on ability to communicate among collaborating groups or persons and on capability to understand and to preview the principles and behavior of the opponent groups or forces. This, in fact, is also the main contemporary human-system interaction (**H-SI**) problem. A *system* is in this case – in a narrow sense – considered as a system created on the basis of electronic, optoelectronic and/or computer technology, in order to aid humans in reaching some of their vital goals. So-defined system is not only a passive tool in human hands; it is rather an active partner equipped with a sort of artificial intelligence, having access to large information resources, being able to adapt its behavior to the human requirements and to collaborate with the human users in order to reach their goals. The area of such systems' applications practically covers most of human activity domains and is still expanding. Respectively, the scientific and practical **H-SI** problems need a large variety of sophisticated solution methods. This is why the **H-SI** problems in the last decades became an important and extensively growing area of investigations.

In this book some examples of the **H-SI** problems and solution methods are presented. They can be roughly divided into the following groups: **a)** Human decisions supporting systems, **b)** Distributed knowledge bases and WEB systems, **c)** Disabled persons aiding systems, **d)** Environment monitoring and robotic systems, **e)** Diagnostic systems, **f)** Educational systems, and **g)** General **H-SI** problems. As usually, some papers to more than one class can be assigned and that is why the classification suits only to a rough characterization of the book contents.

The human decisions supporting systems are presented by papers concerning various application areas, like e.g.: enterprises management (A. Burda and Z.S. Hippe; T. Żabiński and T. Mączka; S. Cavalieri), healthcare (E. Zaitseva), agricultural products storage (W. Sieklicki, M. Kościuk and S. Sieklicki), visual design (E.J. Grabska), sport trainings planning (J. Vales-Alonso, P. López-Matencio, J.J. Alcaraz, et al.). The papers by I. Rejer; J.L. Kulikowski; K. Hareźlak and A. Werner; E. Nawarecki, S. Kluska-Nawarecka and K. Regulski; A. Grzech, A. Prusiewicz and M. Zięba; A. Andrushevich, M. Fercu, J. Hopf, E. Portmann and A. Klapproth to various problems of data and knowledge bases exploration in computer decision-aiding systems are devoted.

The WEB-based, including distributed knowledgebases based systems, are presented in the papers by N. Pham, B.M. Wilamowski and A. Malinowski; M. Hajder and T. Bartczak. K. Skabek, R. Winiarczyk and A. Sochan present a concept of a distributed virtual museum. An interesting concept of managing the process of intellectual capital creation is presented by A. Lewicki and R. Tadeusiewicz. A document-centric instead of data-centric distributed information processing paradigm in a paper by B. Wiszniewski is presented. New computer networks technologies by K. Krzemiński and I. Józwiak and by P. Rożycki, J. Korniak and J. Kolbusz are described. The last two Authors also present a model of malicious network traffic. Selected problems of distributed network resources organization and tagging are presented by A. Dattolo, F. Ferrara and C. Tasso as well as by A. Chandramouli, S. Gauch and J. Eno.

Various problems of disabled persons aiding by their communication with external systems improvement in a next group of papers are presented. The papers by M. Porta and A. Ravarelli and by D. Chugo, H. Ozaki, S. Yokota and K. Takase to physically disabled persons aiding systems are devoted. The spatial orientation and navigation aiding problems by P. Strumillo; A. Śluzek and M. Paradowski and by M. Popa are described. A proposal of an ubiquitous health supervising system by P. Augustyniak is presented. The problems of hand posture or motions recognition for disabled persons aiding by R.S. Choraś and by T. Luhandjula, K. Djouani, Y. Hamam, B.J. van Wyk and Q. Williams have been described while similar problems for a therapy of children supporting by J. Marnik, S. Samolej, T. Kapuściński, M. Oszust and M. Wysocki are presented.

A paper by Mertens, C. Wacharamanatham, J. Hurtmanns, M. Kronenburger, P.H. Kraus, A. Hoffmann, C. Schlick and J. Borchers to a problem of communication through a touch screen improvement is devoted. Some other problems of tactile communication by L. M. Muñoz, P. Ponsa and A. Casals are considered. J. Ruminski, M. Bajorek, J. Ruminska, J. Wtorek, and A. Bujnowski present a method of computer-based dichromats aiding in correct color vision.

In the papers by A. Roman-Gonzalez and by J.P. Rodrigues and A. Rosa some concepts of direct EEG signals using to persons with lost motor abilities aiding are presented. Similarly, some basic problems and experimental results of a direct brain-computer interaction by M. Byczuk, P. Poryzała and A. Materka are also described.

A group of papers presented by Y. Ota; P. Nauth; M. Kitani, T. Hara, H. Hanada and H. Sawada; D. Erol Barkana, and by T. Sato, S. Sakaino and T. Yakoh contains description of several new robotic systems' constructions.

The group concerning diagnostic systems consists of papers mainly to medical applications devoted (K. Przystalcki, L. Nowak, M. Ogorzałek and G. Surówka; P. Cudek, J.W. Grzymała-Busse and Z.S. Hippe; A. Świtoński, R. Bieda and K. Wojciechowski; T. Mroczek, J.W. Grzymała-Busse, Z.S. Hippe and P. Jurczak; R. Pazzaglia, A. Ravarelli, A. Balestra, S. Orio and M.A. Zanetti; M. Jaszuk, G. Szostek and A. Walczak; Gomuła, W. Paja, K. Pancierz and J. Szkoła). Besides, in a paper by R.E. Precup, S.V. Spătaru, M.B. Rădac, E.M. Petriu, S. Preitl, C.A. Dragoş and R.C. David an industrial diagnostic system is presented. K. Adamczyk and A. Walczak present an algorithm of edges detection in images which in various applications can be used.

In the papers by L. Pyzik; C.A. Dragoş, S. Preitl, R.E. Precup and E.M. Petriu and by E. Noyes and L. Deligiannidis examples of computer-aided educational systems are presented. K. Kaszuba and B. Kostek describe a neurophysiological approach to learning processes aiding.

The group concerning general **H-SI** problems consists of the papers presented by T.T. Xie, H. Yu and B.M. Wilamowski; H. Yu and B.M. Wilamowski; and G. Draľus. General problems of rules formulation for automatic reasoning are described by A.P. Rotshtein and H.B. Rakytyanska as well as by M. Paľasiński, B. Fryc and Z. Machnicka. Close to the former ones, S. Chojnacki and M.A. Kľopotek consider a problem of Boolean recommenders evaluation in decision systems. Various aspects of computer-aided decision making methods are presented in the papers by M.P. Dwulit and Z. Szymański, L. Bobrowski and by A. Puľka and A. Miľlik. A problem of ontology creation by A. Di Iorio, A. Musetti, S. Peroni and F. Vitali is described. At last, A. Maľysiak-Mrozek, S. Kozielski and D. Mrozek present a concept of proteins structural similarity describing language.

This panorama of works conducted by a large number of scientists in numerous countries shows that **H-SI** is a wide and progressive area of investigations aimed at human life conditions improvement. It also shows that between different scientific disciplines new and interesting problems arise and stimulate development on both sides of the borders.

Editors

Zdzisław S. Hippe
Juliusz L. Kulikowski
Teresa Mroczek

Contents

Part I: Decision Supporting Systems

From Research on Modeling of Uncertain Data: The Case of Small and Medium Enterprises	3
<i>A. Burda, Z.S. Hippe</i>	
Implementation of Human-System Interface for Manufacturing Organizations	13
<i>T. Zabiński, T. Mączka</i>	
Precision of an Expert Fuzzy Model	33
<i>I. Rejer</i>	
Database Access and Management with the Use of the MOODLE Platform	49
<i>K. Haręźlak, A. Werner</i>	
Human Activity Supporting by Deontological Knowledgebases	67
<i>J.L. Kulikowski</i>	
Multi-aspect Character of the Man-Computer Relationship in a Diagnostic-Advisory System	85
<i>E. Nawarecki, S. Kluska-Nawarecka, K. Regulski</i>	
Services Merging, Splitting and Execution in Systems Based on Service Oriented Architecture Paradigm	103
<i>A. Grzech, A. Prusiewicz, M. Zięba</i>	
Importance Measures in Reliability Analysis of Healthcare System	119
<i>E. Zaitseva</i>	

Towards a Formal Model of Visual Design Aided by Computer	135
<i>E.J. Grabska</i>	
Agricultural Products' Storage Control System	149
<i>W. Sieklicki, M. Kościuk, S. Sieklicki</i>	
A Dynamic Programming Approach for Ambient Intelligence Platforms in Running Sports Based on Markov Decision Processes	165
<i>J. Vales-Alonso, P. López-Matencio, J.J. Alcaraz, J.L. Sieiro-Lomba, E. Costa-Montenegro, F.J. González-Castaño</i>	
Prometheus Framework for Fuzzy Information Retrieval in Semantic Spaces	183
<i>A. Andrushevich, M. Fercu, J. Hopf, E. Portmann, A. Klapproth</i>	
Evaluating Overheads Introduced by OPC UA Specifications	201
<i>S. Cavalieri</i>	
<hr/>	
Part II: Distributed Knowledgebase's and Web Systems	
<hr/>	
A Document-Centric Processing Paradigm for Collaborative Computing	225
<i>B. Wiszniewski</i>	
Computing Utilization via Computer Networks	239
<i>N. Pham, B.M. Wilamowski, A. Malinowski</i>	
The Method of Communication Quality Improvement in Distributed Systems with Real-Time Services	253
<i>M. Hajder, T. Bartczak</i>	
An Autocatalytic Emergence Swarm Algorithm in the Decision-Making Task of Managing the Process of Creation of Intellectual Capital	271
<i>A. Lewicki, R. Tadeusiewicz</i>	
Implementation of the Progressive Meshes for Distributed Virtual Museum	287
<i>K. Skabek, R. Winiarczyk, A. Sochan</i>	
The Impact of the Control Plane Architecture on the QoS in the GMPLS Network	299
<i>P. Rozycki, J. Korniak, J. Kolbusz</i>	

On Social Semantic Relations for Recommending Tags and Resources Using Folksonomies	311
<i>A. Dattolo, F. Ferrara, C. Tasso</i>	
The Simulation of Malicious Traffic Using Self-similar Traffic Model	327
<i>J. Kolbusz, P. Rozycki, J. Korniak</i>	
A Cooperative Approach to Web Crawler URL Ordering	343
<i>A. Chandramouli, S. Gauch, J. Eno</i>	
Synergic Intranet: An Example of Synergic IT as the Goal of E-Engineering	359
<i>K. Krzemiński, I. Józwiak</i>	
<hr/>	
Part III: Impaired Persons Aiding Systems	
<hr/>	
Electronic Systems Aiding Spatial Orientation and Mobility of the Visually Impaired	373
<i>P. Strumillo</i>	
Towards Vision-Based Understanding of Unknown Environments	387
<i>A. Śluzek, M. Paradowski</i>	
Recognition of Hand Posture for HCI Systems	403
<i>R.S. Choraś</i>	
Some Eye Tracking Solutions for Severe Motor Disabilities	417
<i>M. Porta, A. Ravarelli</i>	
A Visual Hand Motion Detection Algorithm for Wheelchair Motion	433
<i>T. Luhandjula, K. Djouani, Y. Hamam, B.J. van Wyk, Q. Williams</i>	
Computerized Color Processing for Dichromats	453
<i>J. Ruminski, M. Bajorek, J. Ruminska, J. Wtorek, A. Bujnowski</i>	
Sitting Motion Assistance for a Rehabilitation Robotic Walker	471
<i>D. Chugo, H. Ozaki, S. Yokota, K. Takase</i>	
Pedestrian Navigation System for Indoor and Outdoor Environments	487
<i>M. Popa</i>	

Model Based Processing of Swabbing Movements on Touch Screens to Improve Accuracy and Efficacy for Information Input of Individuals Suffering from Kinetic Tremor	503
<i>A. Mertens, C. Wacharamanotham, J. Hurtmanns, M. Kronenbuerger, P.H. Kraus, A. Hoffmann, C. Schlick, J. Borchers</i>	
Compound Personal and Residential Infrastructure for Ubiquitous Health Supervision	523
<i>P. Augustyniak</i>	
Using Computer Graphics, Vision and Gesture Recognition Tools for Building Interactive Systems Supporting Therapy of Children	539
<i>J. Marnik, S. Samolej, T. Kapuściński, M. Oszust, M. Wysocki</i>	
EEG Biofeedback: Viability and Future Directions	555
<i>J.P. Rodrigues, A. Rosa</i>	
EEG Signal Processing for BCI Applications	571
<i>A. Roman-Gonzalez</i>	
Author Index	593
Subject Index	595

Part I

Decision Supporting Systems

From Research on Modeling of Uncertain Data: The Case of Small and Medium Enterprises

A. Burda¹ and Z.S. Hippe²

¹ University of Management and Administration in Zamość, Poland

aburda@wsz.zia.edu.pl

² Institute of Biomedical Informatics,

University of Information Technology and Management, Rzeszów, Poland

zhippe@wsiz.rzeszow.pl

Abstract. A new procedure for combined validation of learning models – built for specifically uncertain data – is briefly described. The procedure, called the *queue validation*, relies on a combination of *resubstitution* with the modified learn-and-test paradigm. In the initial experiment [Burda and Hippe 2010] the developed procedure was checked on doubtful (presumably distorted by creative accounting) data, related to small and medium enterprises (further called **SME**), displaying two concepts: *bankrupt* or *non-bankrupt*. In the current research a new set of learning models was generated for the same data using various types of optimized artificial neural networks. All learning models were evaluated using the *queue validation* methodology. It was found that error rates for *bankrupt* concept are much larger than error rates for the concept *non-bankrupt*. It is assumed that this difference in error rates discovered by the *queue validation* procedure can be probably used as a hint pointing frauds in the investigated **SME** data.

1 Introduction

Nowadays, the economic power of well-developed countries is not dependent on huge companies, but rather on small and medium enterprises which hire up to 250 people; frequently **SME** are run by very restricted number of people, say even by a family. Recently **SME** constitute 99.8% of companies functioning on the whole European Union area and they are the place of work for more than 67% people employed in the private sector [Schmiemann 2008]. The bankruptcy of **SME** in every case generates considerable threats of unemployment on their area, especially in the regions weakly urbanized. For this reason, the search for reliable and effective methods of assessment of their status bears significant importance not only for **SME** themselves, but it also plays an important social role.

2 Objectives and Scope of the Research

Data concerning the **SME** properties are usually uncertain owing to many specific reasons; the most important of them can be caused by deliberate action in the process of so called *creative accounting* [Nowak 1998]. This notion is referred to accounting practices that may follow the letter of the rules of standard accounting practices, but certainly deviate from the spirit of those rules. In other words, this term generally refers to systematic misrepresentation of the true income and assets of enterprises, corporations or other organizations. Therefore, the creative accounting can be one of the most important reasons of the unsatisfactory results of the assessment of **SME**'s status given in the available literature [Haider and Bukhari 2007; Pongsatat et al. 2004; Kim and Sohn 2010]. These findings were also confirmed in our previous statistical experiments [Burda 2009]. It should be emphasized, that in many countries (also in Poland) the annual balance sheet of a given **SME** reported to revenue departments, contains the *qualitative* declaration of his/her owner, whether the enterprise is in the state of survival or in bankruptcy. The very strict evaluation of the enterprise is made only every four years; therefore in between there is plenty of room for creative accounting. Therefore in this research – using supervised machine learning methods – the preliminary appraisal to classification of **SME** data was made, to develop learning models suitable for satisfactory recognition of the possible status of a given **SME**: *bankrupt* or *non-bankrupt*. The validation methodology discussed in details in [Burda and Hippe 2010] (also briefly outlined here) was checked on data described in section 2.2. In the extension of our previous experiments a new stream of the investigation, based on the application of various optimized artificial neural networks, was recently performed. Correctness of the new set of learning models was estimated on the basis of the error rate of classification and the quality index of prediction, Q .

2.1 The Basics of Performance Criterion and Validation Procedures

The most important performance criterion of rule induction methods is an error rate. If the number of cases is less than 100, the *leaving-one-out* method is used to estimate the error rate of the rule set. In leaving-one-out, the number of learn-and-test experiments is equal to the number of cases in the data set. During the i -th experiment, the i -th case is removed from the data set, a rule set is induced by the rule induction system from the remaining cases, and the classification of the omitted case by rules produced is recorded. The error rate is computed as the ratio of the total number of misclassifications to the number of cases.

On the other hand, if the number of cases in the data set is greater than or equal to 100, the *ten-fold cross-validation* should be used. This technique is similar to leaving-one-out in that it follows the learn-and-test paradigm. In this case, however, all cases are randomly re-ordered, and then a set of all cases is divided into

ten mutually disjoint subsets of approximately equal size. For each subset, all remaining cases are used for training, i.e., for rule induction, while the subset is used for testing. This method is used primarily to save time at the negligible expense of accuracy. Ten-fold cross validation is commonly accepted as a standard way of validating rule sets.

For large data sets (at least 1000 cases) a single application of the *train-and-test paradigm* may be used. This technique is also known as *holdout*. Two thirds of cases should be used for training, one third for testing. In yet another way of validation, *resubstitution*, it is assumed that the training data set is identical with the testing data set. In general, an estimate for the error rate is here too optimistic. However, this technique is used in many applications.

The main objective of our paper was to develop and test a new procedure for validation of uncertain data (thus, for validation of uncertain learning models), which is a combination of *resubstitution* (applied to check the i -th model, using data of the considered year (i -th year), and validation based on *modified learn-and-test paradigm*, in which i -th year model is consecutively applied for evaluation of learning models developed for the year i -th \pm 1, i -th \pm 2, i -th \pm 3, etc. Modification of the *learn-and-test paradigm* relied in our case on validation (classification) of all data contained in the year i , $i\pm 1$, $i\pm 2$, $i\pm 3$, etc. This evaluation procedure, is tentatively called by us a *queue validation*.

2.2 Investigated Datasets

Objects (individual enterprises) collected in seven datasets describe SME's from the Podkarpackie-province in Poland, in years 2000-2006. Databases consisted of an equal number of two categories of cases: *bankrupts* and *non-bankrupts*. Each object was described by means of 7 descriptive attributes, namely: (1) *Share of inventories in the overall assets*, (2) *Share of working capital in the overall financial assets*, (3) *Shortage of net working capital*, (4) *Asset productivity*, (5) *Gross financial result*, (6) *Rate of sale changes*, and (7) *Rate of employment changes*. Table 1 gathers information about the numerical taxonomy of the investigated data in subsequent years.

2.3 Research Details

Learning models described in our introductory paper [Burda and Hippe 2010] were generated using two different methods of machine learning (ML), namely, the in-house improved ID3/C4.5 algorithm for generating quasi-optimal decision trees [Hippe and Knap 2003] and the NGTS algorithm, initially described in [Hippe 1999]. The new set of models was created using optimized artificial neural networks: the Radial Basis Function Network (RBF) [Powell 2001], the Multi-layer Perceptrons MLP3 – (3 layer network) and MLP4 – (4 layer network)

[Rumelhart and McClelland 1986] and the **Linear Network (LIN)**. The methodology of selecting the optimal number of neurons in the hidden layer of the artificial neural network is described in section 3.

Table 1 Numerical taxonomy of **SME** data in the investigated span 2000-2006

File name	Year	Number of cases		
		total	<i>non-bankrupt</i>	<i>bankrupt</i>
SME_2000	2000	132	66	66
SME_2001	2001	150	75	75
SME_2002	2002	144	72	72
SME_2003	2003	130	65	65
SME_2004	2004	128	64	64
SME_2005	2005	132	66	66
SME_2006	2006	132	66	66
Together	⇒	948	474	474

3 Artificial Neural Networks – Selection of a Quasi-optimal Architecture

This entirely new part of the research was devoted to finding the most promising learning algorithm (and associated with it network architecture) for a generalized analysis of real data on small and medium enterprises.

In the case of Radial Basis Function (**RBF**) networks, assigning centers of basis functions has been carried out using the k-means method; after determining centers and deviations for the hidden layer, the output layer has been optimized using the standard linear optimization technique – pseudoinverse algorithm (singular value decomposition), whereas for the **MLP3** and **MLP4** networks, in the first step the backpropagation method has been used for 100 epochs, and next the conjugate gradient method has been used for 500 epochs.

For each topology, 50 independent learning processes on the examined **SME_2000** set, described in Table 1, has been carried out. All the tests were initialized by selecting random weights from the interval (0,1). A number of input neurons was equal to the number of descriptive attributes, i.e., 7. A number of neurons in the hidden layer of the **MLP3** and **RBF** models was selected from the range of 1 - 20. In the case of the **MLP4** model, a number of neurons in the first hidden layer was changed in the range of 1–10. A number of output neurons was equal to the number of decision attributes (1). The best model out of 50 models, selected using the *redistribution* method, has been tested on data sets concerning years 2000 – 2006.

The quality of models has been evaluated individually for each category of *bankrupt/non-bankrupt* concept by the quality index of prediction, Q , defined below:

$$Q = \frac{\sum_{i=1}^n P_i + (1 - N_i)}{2n} \quad (1)$$

where:

n – is a number of independent testing sets,

P_i – is a ratio of a number of cases correctly classified to a number of all cases of the i -th subset,

N_i – is a ratio of a number of cases incorrectly classified to a number of all cases of the i -th subset.

In the ideal cases (all cases of testing sets are correctly classified), the value of Q equals to 1. A model classifying incorrectly all cases has Q equal to 0, whereas if a generated model does not classify correctly any case or only a half of them is classified correctly, then Q accounts to 0.5.

4 Results of Experiments

In this section, for the sake of clarity of the general discussion, basic results of our previous experiments are re-cited (Table 2 and Table 3). On the other side, results of the new experiments are presented in Tables 4–7. As usual, rows are labeled with self-explanatory names (e.g., **ID3_2000** - a model created by the **ID3/C4.5** algorithm on the basis of data collected in year 2000). Columns contain classification errors of models tested on data collected in years 2000–2006.

5 Discussion

The research devoted to application of the developed *queue validation* method supplied very interesting observation. It was found, namely, that all generated learning models related to the investigated concepts are characterized by very different, typical size of error rates. The vast majority of the validated models predict with high accuracy the *non-bankrupt* enterprise concept. The quality index of prediction, Q , of the best models fluctuates between 0.95 and 0.97. It means that such models classify, in a stable and credible way, enterprises, whose states do not threaten to lose the continuity of their functioning. Such good classification results have been achieved for all neural network models, independently on assumed topologies and learning methods as well as for the **ID3/C4.5** algorithm. The worse results have been obtained for models built using the **NGTS** algorithm only.

Table 2 Validation of the uncertain SME data (ID3/C4.5 models)

Model	Concept	Error rate [%] in Year							Index
		2000	2001	2002	2003	2004	2005	2006	Q
ID3_2000	<i>bankrupt</i>	0.0	52.0	55.6	53.8	25.0	33.3	27.3	0.59
	<i>non-bankrupt</i>	1.5	10.7	19.4	12.3	10.9	9.1	9.1	0.88
ID3_2001	<i>bankrupt</i>	54.5	4.0	22.2	46.2	25.0	33.3	27.3	0.65
	<i>non-bankrupt</i>	3.0	1.3	15.3	7.7	10.9	12.1	6.1	0.91
ID3_2002	<i>bankrupt</i>	36.4	44.0	0.0	38.5	0.0	16.7	27.3	0.73
	<i>non-bankrupt</i>	4.5	8.0	6.9	3.1	4.7	3.0	6.1	0.95
ID3_2003	<i>bankrupt</i>	45.5	36.0	44.4	0.0	25.0	33.3	27.3	0.65
	<i>non-bankrupt</i>	6.1	14.7	19.4	3.1	14.1	15.2	15.2	0.86
ID3_2004	<i>bankrupt</i>	36.4	40.0	11.1	30.8	0.0	16.7	27.3	0.73
	<i>non-bankrupt</i>	4.5	6.7	12.5	3.1	4.7	3.0	6.1	0.94
ID3_2005	<i>bankrupt</i>	45.5	56.0	44.4	46.2	18.8	0.0	27.3	0.60
	<i>non-bankrupt</i>	4.5	2.7	13.9	4.6	3.1	3.0	6.1	0.94
ID3_2006	<i>bankrupt</i>	45.5	36.0	44.4	38.8	18.8	16.7	9.1	0.67
	<i>non-bankrupt</i>	7.6	10.7	20.8	10.8	9.4	10.6	9.1	0.88

Table 3 Validation of the uncertain SME data (NGTS models)

Model	Concept	Error rate [%] in Year							Index
		2000	2001	2002	2003	2004	2005	2006	Q
NGTS_2000	<i>bankrupt</i>	0.0	28.0	22.2	30.8	12.5	50.0	36.4	0.41
	<i>non-bankrupt</i>	0.0	12.0	5.6	6.2	3.1	9.1	4.5	0.78
NGTS_2001	<i>bankrupt</i>	36.4	0.0	44.4	30.8	25.0	16.7	27.3	0.47
	<i>non-bankrupt</i>	10.6	0.0	11.1	9.2	10.9	9.1	9.1	0.72
NGTS_2002	<i>bankrupt</i>	54.5	36.0	0.0	46.2	18.8	33.3	45.5	0.34
	<i>non-bankrupt</i>	6.1	10.7	0.0	9.2	7.8	10.6	13.6	0.70
NGTS_2003	<i>bankrupt</i>	36.4	16.0	22.2	0.0	18.8	50.0	45.5	0.46
	<i>non-bankrupt</i>	10.6	4.0	13.9	0.0	12.5	12.1	10.6	0.69
NGTS_2004	<i>bankrupt</i>	18.2	20.0	11.1	15.4	0.0	33.3	27.3	0.53
	<i>non-bankrupt</i>	7.6	20.0	6.9	7.7	0.0	10.6	9.1	0.70
NGTS_2005	<i>bankrupt</i>	54.5	32.0	55.6	30.8	25.0	0.0	9.1	0.37
	<i>non-bankrupt</i>	10.6	8.0	9.7	21.5	6.3	0.0	13.6	0.74
NGTS_2006	<i>bankrupt</i>	9.1	20.0	44.4	30.8	25.0	50.0	0.0	0.50
	<i>non-bankrupt</i>	7.6	10.7	16.7	10.8	9.4	7.6	0.0	0.72

Table 4 Validation of the uncertain SME data (RBF models)

Model	Concept	Error rate [%] in Year							Index
		2000	2001	2002	2003	2004	2005	2006	Q
RBF_2000	<i>bankrupt</i>	18.2	44.0	11.1	53.8	25.0	33.3	45.5	0.65
	<i>non-bankrupt</i>	15.2	14.7	12.5	13.8	9.4	10.6	16.7	0.87
RBF_2001	<i>bankrupt</i>	36.4	36.0	0.0	46.2	6.3	33.3	54.5	0.71
	<i>non-bankrupt</i>	10.6	16.0	11.1	6.2	7.8	6.1	10.6	0.91
RBF_2002	<i>bankrupt</i>	45.5	36.0	0.0	38.5	25.0	50.0	72.7	0.55
	<i>non-bankrupt</i>	1.5	5.3	8.3	3.1	4.7	3.0	6.1	0.96
RBF_2003	<i>bankrupt</i>	36.4	28.0	22.2	23.1	0.0	50.0	45.5	0.70
	<i>non-bankrupt</i>	16.7	16.0	12.5	4.6	9.4	9.1	12.1	0.87
RBF_2004	<i>bankrupt</i>	45.5	44.0	33.3	38.5	6.3	50.0	54.5	0.56
	<i>non-bankrupt</i>	13.6	17.3	9.7	6.2	7.8	10.6	10.6	0.89
RBF_2005	<i>bankrupt</i>	63.6	36.0	22.2	30.8	18.8	16.7	36.4	0.65
	<i>non-bankrupt</i>	12.1	17.3	19.4	12.3	14.1	15.2	10.6	0.86
RBF_2006	<i>bankrupt</i>	45.5	36.0	33.3	30.8	12.5	33.3	18.2	0.68
	<i>non-bankrupt</i>	19.7	26.7	20.8	10.8	12.5	16.7	19.7	0.82

Table 5 Validation of the uncertain SME data (MLP3 models)

Model	Concept	Error rate [%] in Year							Index
		2000	2001	2002	2003	2004	2005	2006	Q
MLP3_2000	<i>bankrupt</i>	27.3	36.0	33.3	46.2	12.5	33.3	54.5	0.64
	<i>non-bankrupt</i>	12.1	16.0	13.9	9.2	14.1	10.6	10.6	0.88
MLP3_2001	<i>bankrupt</i>	36.4	20.0	22.2	46.2	18.8	33.3	45.5	0.66
	<i>non-bankrupt</i>	6.1	10.7	23.6	9.2	9.4	9.1	6.1	0.89
MLP3_2002	<i>bankrupt</i>	54.5	56.0	0.0	46.2	31.3	66.7	72.7	0.45
	<i>non-bankrupt</i>	4.5	9.3	5.6	3.1	1.6	3.0	4.5	0.96
MLP3_2003	<i>bankrupt</i>	63.6	44.0	22.2	15.4	18.8	50.0	54.5	0.58
	<i>non-bankrupt</i>	4.5	12.0	11.1	6.2	7.8	6.1	7.6	0.92
MLP3_2004	<i>bankrupt</i>	54.5	44.0	44.4	46.2	0.0	66.7	63.6	0.47
	<i>non-bankrupt</i>	1.5	5.3	4.2	3.1	3.1	0.0	4.5	0.97
MLP3_2005	<i>bankrupt</i>	63.6	52.0	55.6	46.2	18.8	0.0	27.3	0.56
	<i>non-bankrupt</i>	13.6	8.0	15.3	6.2	3.1	6.1	7.6	0.91
MLP3_2006	<i>bankrupt</i>	45.5	48.0	44.4	38.5	6.3	16.7	18.2	0.67
	<i>non-bankrupt</i>	15.2	17.3	25.0	15.4	15.6	15.2	13.6	0.83

Table 6 Validation of the uncertain SME data (MLP4 models)

Model	Concept	Error rate [%] in Year							Index
		2000	2001	2002	2003	2004	2005	2006	\bar{Q}
MLP4_2000	<i>bankrupt</i>	27.3	52.0	44.4	53.8	12.5	33.3	63.6	0.57
	<i>non-bankrupt</i>	1.5	10.7	6.9	3.1	3.1	3.0	3.0	0.95
MLP4_2001	<i>bankrupt</i>	36.4	32.0	22.2	46.2	31.3	33.3	63.6	0.61
	<i>non-bankrupt</i>	1.5	8.0	16.7	7.7	7.8	7.6	6.1	0.92
MLP4_2002	<i>bankrupt</i>	54.5	56.0	0.0	46.2	37.5	66.7	72.7	0.44
	<i>non-bankrupt</i>	3.0	9.3	4.2	3.1	1.6	4.5	4.5	0.96
MLP4_2003	<i>bankrupt</i>	63.6	44.0	22.2	15.4	25.0	50.0	63.6	0.55
	<i>non-bankrupt</i>	3.0	12.0	6.9	3.1	6.3	4.5	6.1	0.94
MLP4_2004	<i>bankrupt</i>	36.4	44.0	11.1	46.2	6.3	50.0	54.5	0.60
	<i>non-bankrupt</i>	1.5	5.3	5.6	3.1	3.1	1.5	4.5	0.96
MLP4_2005	<i>bankrupt</i>	45.5	52.0	66.7	53.8	18.8	0.0	9.1	0.59
	<i>non-bankrupt</i>	16.7	13.3	13.9	7.7	15.6	7.6	16.7	0.86
MLP4_2006	<i>bankrupt</i>	54.5	48.0	55.6	53.8	31.3	33.3	18.2	0.54
	<i>non-bankrupt</i>	9.1	5.3	8.3	6.2	6.3	3.0	3.0	0.94

Table 7 Validation of the uncertain SME data (LIN models)

Model	Concept	Error rate [%] in Year							Index
		2000	2001	2002	2003	2004	2005	2006	\bar{Q}
LIN_2000	<i>bankrupt</i>	27.3	24.0	33.3	38.5	12.5	33.3	45.5	0.69
	<i>non-bankrupt</i>	21.2	37.3	26.4	26.2	23.4	28.8	37.9	0.70
LIN_2001	<i>bankrupt</i>	36.4	32.0	44.4	46.2	12.5	33.3	45.5	0.64
	<i>non-bankrupt</i>	3.0	6.7	19.4	7.7	9.4	7.6	6.1	0.91
LIN_2002	<i>bankrupt</i>	54.5	48.0	0.0	53.8	37.5	50.0	72.7	0.47
	<i>non-bankrupt</i>	4.5	9.3	9.7	3.1	1.6	4.5	6.1	0.95
LIN_2003	<i>bankrupt</i>	45.5	40.0	0.0	15.4	6.3	50.0	45.5	0.69
	<i>non-bankrupt</i>	4.5	9.3	13.9	3.1	10.9	4.5	6.1	0.92
LIN_2004	<i>bankrupt</i>	45.5	48.0	44.4	38.5	12.5	33.3	36.4	0.59
	<i>non-bankrupt</i>	3.0	8.0	12.5	4.6	7.8	3.0	6.1	0.94
LIN_2005	<i>bankrupt</i>	63.6	40.0	44.4	23.1	18.8	16.7	36.4	0.62
	<i>non-bankrupt</i>	10.6	25.3	23.6	20.0	20.3	16.7	10.6	0.82
LIN_2006	<i>bankrupt</i>	45.5	56.0	55.6	61.5	31.3	33.3	27.3	0.53
	<i>non-bankrupt</i>	9.1	4.0	11.1	4.6	3.1	1.5	3.0	0.94

The classification results of **SME** in the *bankrupt* category obtained using the same machine learning models are not satisfying and they diverge from results obtained for the opposite category. A ratio of the quality index of prediction for the *non-bankrupt* category to the quality index of prediction for the *bankrupt* category for all the models is between 1.21 (for the **RBF_2006** model) to 2.15 (for the **MLP4_2002** model). Additionally, obtained learning models, tested using a specific technique called the queue validation, revealed a lack of stability of results in consecutive years spanned by a procedure applied for this category of cases.

It is worth noting that the models, generated using the **ID3/C4.5** algorithm and the **Radial Basis Function** network architecture, classify the analyzed **SME** data better than the other ones, especially in the *bankrupt* category.

7 Conclusions

Research carried out using selected machine learning methods fully validated hypotheses included in [Burda and Hippe 2010] that classifying enterprises into the *non-bankrupt* category can be recognized as credible. Simultaneously, results of the undertaken experiment indicate that it is not possible to build a learning model with comparable classification quality, using the same machine learning methods, for companies belonging to the *bankrupt* category. This observation may have a crucial meaning in interpretation of results of classification of **SME**, at least for companies functioning in south-east Poland.

It can be only assumed, on the basis of our experiences, that intentional lies in data presented in financial reports by majority of companies after declaring one-self bankrupt can be possible reasons of an observed occurrence. Therefore, further research will be devoted to developing methods for generation of learning models adjusted better to classifying uncertain data.

Acknowledgment

The authors made polite thanks to Dr. Marek Cierpiął-Wolan, director of the Statistical Office in Rzeszów, for providing the necessary data to conduct our research.

References

- [Burda 2009] Burda, A.: Multicategory evaluation of prediction models for small and medium enterprises. *Barometr Regionalny* 15, 77–84 (2009)
- [Burda and Hippe 2010] Burda, A., Hippe, Z.S.: Uncertain data modeling: The case of small and medium enterprises. In: Pardela, T., Wilamowski, B. (eds.) 3rd International Conference on Human System Interaction, pp. 76–80. e-Book, Rzeszów (2010)
- [Haider and Bukhari 2007] Haider, S.A., Bukhari, A.S.: Evaluating Financial sector firm's creditworthiness for south-asian countries. *Asian Journal of Information Technology* 6, 329–341 (2007)

- [Hippe 1999] Hippe, Z.S.: Data mining and knowledge discovery in business: past, present, and future. In: Abramowicz, W., Orłowska, M. (eds.) *Business Information Systems 1999*, pp. 158–169. Springer, Heidelberg (1999)
- [Hippe and Knap 2003] Hippe, Z.S., Knap, M.: Research on development of certain and possible decision trees. In: Krawczyk, H., Kubale, M. (eds.) *Informational Technologies*, pp. 189–194. Gdańsk Univ. Edit. Office, Gdańsk (2003)
- [Kim and Sohn 2010] Kim, H.S., Sohn, S.Y.: Support vector machines for default prediction of SMEs based on technology credit. *European Journal of Operational Research* 201, 838–846 (2010)
- [Nowak 1998] Nowak, M.: Practical evaluation of the enterprise financial condition, p. 89. *Foundation of Accountancy Development in Poland*, Warsaw (1998)
- [Pongsat et al. 2004] Pongsat, S., Ramage, J., Lawrence, H.: Bankruptcy prediction for large and small firms in asia: A comparison of ohlson and altman. *Journal of Accounting and Corporate Governance* 2, 1–13 (2004)
- [Powell 2001] Powell, M.J.D.: Radial basis function methods for interpolation to functions in many variables. Report DAMPT 2001/NA11, Department of Applied Mathematics and Theoretical Physics, University of Cambridge (2001)
- [Rumelhart and McClelland 1986] Rumelhart, D.E., McClelland, J. (eds.): *Parallel distributed processing*, 1st edn. MIT Press, Cambridge (1986)
- [Schmiemann 2008] Schmiemann, M.: Enterprises by size class-overview of SMEs in the EU (2008), http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-SF-08-031/EN/KS-SF-08-031-EN.PDF (accessed January 20, 2011)

Implementation of Human-System Interface for Manufacturing Organizations

T. Żabiński and T. Mączka

Department of Computer and Control Engineering,
Rzeszów University of Technology, Poland
{tomz, tmaczka}@prz-rzeszow.pl

Abstract. In the paper current results of a project devoted to the development of a software and hardware platform for a holonic based Intelligent Manufacturing System dedicated to small and medium metal component production companies are presented. Recent trends in factory automation which create a promising perspective for industrial implementations of Intelligent Manufacturing Systems including extended Human-System Interfaces, are briefly discussed. It was shown that modern factory automation controllers facilitate implementations of extended Human-System Interface as well as Human-Machine Interface and enable the integration of human operators and manufacturing resources into an Intelligent Manufacturing System. In the paper two industrial testbeds used in daily production processes in two different companies are described and future directions of the system development are discussed.

1 Introduction

Nowadays and in the last few decades global competition and constantly growing market demands have imposed a significant pressure on the manufacturing sector. It is foreseen [Institute for Prospective Technological Studies 2003; National Research Council 1998] that this phenomenon is to be even more intensive in the future. Manufacturing industry needs to face a significant change in order to increase competitiveness and meet the demands of society and sustainability. Manufacturing companies must undergo many changes in order to cope with intensive competition, unpredictable markets, constantly growing demands for products quality, customization and variety as well as the necessity in the decrease of time-to-market, life-cycles, batch sizes, delivery times and prices of products.

Most production systems currently used in the manufacturing industry, especially in small and medium companies, are characterized by centralized solutions and lack of direct communication with operators and machines operated on the shop-floors of factories. These solutions are no longer appropriate, as they were

designed to perform high volume, low variety and low flexibility production processes. In traditional centralized manufacturing systems the fulfillment of the current market demands would create an unacceptable decrease in efficiency due to, e.g. high replacements costs. For this reasons, the current challenge is to implement a new generation of innovative manufacturing control and monitoring concepts that exhibit intelligence, robustness and adaptation to environment changes and disturbances [Institute for Prospective Technological Studies 2003; National Research Council 1998]. The new generation of manufacturing systems is referred to as Intelligent Manufacturing System (IMS). The IMS concept requires an intensive use of Information and Communication Technologies (ICT) to support reliable management and control of production processes.

IMS should utilize Artificial Intelligence (AI) techniques [Oztemel 2010; Żabiński 2010] to:

- minimize human involvement in manufacturing activities;
- automatically arrange material, tools and production compositions;
- monitor, control and diagnose machines and production processes;
- recommend and perform actions to prevent faulty production, performance reduction and machines breakdowns;
- automatically discover and provide knowledge about manufacturing process, equipment efficiency and condition;
- provide knowledge and tools for reliable management decisions;
- support techniques for production process optimization,
- etc..

IMS implementation requires computer and factory automation systems characterized by the distributed structure, direct communication with manufacturing resources and the application of sophisticated networked embedded devices on the shop-floor [Żabiński 2010]. It must also be emphasized that IMS is devoted to support human operators, not to replace them [Oztemel 2010]. As human operators play an even more important role in manufacturing systems nowadays than they did in the past, the research in the field of Human-System Interface (HSI) design and implementation, for this systems, is still an active area [Gong 2009; Cummings et al. 2010; Żabiński and Mączka 2010]. HSI is responsible for the efficient cooperation between operators and computer systems and can significantly improve overall production effectiveness. It seems to be clear that convenient and reliable human system interaction, especially at shop-floor level, is an important factor for a successful and evolutionary industrial IMS implementation.

In the paper, the project devoted to the development of a software and hardware platform for holonic based IMS dedicated to small and medium metal component production companies is presented. The project goal determines three main assumptions for the selection of hardware and software elements. The first assumption is a possibility to include modern machines with advanced control equipment as well as older ones not equipped with controllers in the system. The second one is a reasonable cost of the system and a possibility for a gradual system

implementation in a real manufacturing process. The third one is high flexibility and high communication capability of the system components which will enable holonic manufacturing concept implementation, i.e. software agents running directly on the devices on the shop-floor. Due to the following reasons, most of the software system elements have been created using general programming languages and open-source software tools instead of applying commercially available dedicated solutions like, e.g. Supervisory Control And Data Acquisition (SCADA) or Manufacturing Execution Systems (MES). At the shop-floor system level modern Programmable Automation Controllers (PAC) [Żabiński 2010] with extended functionality were chosen.

In the paper, previously published [Żabiński et al. 2009; Żabiński 2010; Żabiński and Mączka 2010] project results have been reviewed and new achievements have been presented, i.e. tools management module and the platform integration with production scheduling system.

The project has been made by Department of Computer and Control Engineering in cooperation with student scientific circle ROBO, Green Forge Innovation Cluster and WSK "PZL-Rzeszów" and Bernacki Industrial Services companies.

2 Programmable Automation Controllers as a Platform for IMS and HSI Implementation

An interesting trend which can be observed currently in the field of factory automation is the direct integration of automation and computer tools, methods and devices [Żabiński 2010]. An important tendency is to substitute classical Programmable Logic Controllers (PLC) for Programmable Automation Controllers (PAC), by some vendors called the embedded PC. The most interesting for IMS and HSI purposes are those PACs which are equipped with operating systems like, e.g. Windows CE, Windows XP Embedded or Linux. PACs meet the complex demands of modern manufacturing control systems as they combine features of traditional PLCs and personal computers. The main feature of PACs is the ability to use the same device for various tasks simultaneously:

- classical and intelligent real-time process or motion control;
- data collection, processing and temporary storage;
- communication with databases or other system components using standard computer protocols and technologies (Ethernet, TCP/IP, web services) as well as typical fieldbus networks;
- running Graphical User Interface (GUI);
- etc.

Applications for PACs can be developed using various programming languages, i.e. typical PLC like ST, FBD, IL, SFC, LD (IEC-61131-3) as well as general programming languages like C, C++, C#, Visual Basic, Java, Delphi, etc. Suppliers of

software tools for PACs currently tend to provide one Integrated Development Environment (IDE) for various tasks:

- PLC and motion controllers configuration and programming;
- data processing and visualization;
- GUI development;
- database and fieldbus communication;
- integration with Rapid Control Prototyping environments like Matlab;
- condition monitoring;
- advanced measurements;
- robotics and vision.

It is also important that the communication between different software modules which run on PACs is relatively easy to perform. In Beckhoff systems communication between PLC and C# applications can be done, for example, via direct reading and writing variables from PLC programs, using their names or notification mechanism with callback functions [WWW-1]. Additional important features of PACs are multitasking and flexible and modular structure. They simplify implementation of complex software solutions and enables the system scalability in accordance with the needs, e.g. the amount and type of inputs and outputs or fieldbus communication devices can be successively modified.

The PACs features enumerated above enable the use of PACs not only as devices to machines or processes control, but also as platforms for the development of HSIs and sophisticated software systems like, e.g. multi-agent based ones, which can be used to create holonic based IMS.

PACs are becoming more and more popular in the industry due to the fact that their prices are currently comparable to traditional PLCs but their functionality is significantly richer. It can be stated that PACs has overcome one important barrier for the industrial implementation of IMS, i.e. the absence of industrial controllers with capabilities to run software agents directly on the controller in parallel with the PLC or motion control programs [Żabiński 2010]. It all together constitutes a promising perspective for IMS and advanced HSI industrial implementation even in small and medium manufacturing companies.

3 Industrial Testbeds Structure

Up to now, two real testbeds have been constructed, the first one was installed in a screws production company which is a member of the Green Forge Innovation Cluster. The second one was installed at the WSK “PZL-Rzeszów” in the department which produces major rotating parts for the aviation industry. It should be emphasized that the department size is similar to medium sized Polish companies but on the other hand WSK “PZL-Rzeszów” is one of the biggest companies in Poland which produces parts for the aviation industry. The first testbed consists of two machines sections formed by pushing machines for cold forging. The first

section includes machines without PLC controllers but the second one consists of six modern machines equipped with PLCs and advanced cold forge process monitoring devices. The WSK testbed includes one production line with four CNC vertical turning lathes and two CNC machining centers.

The machines in the testbeds have been operated by experienced operators who interact with the system using RFID cards, barcode readers, electronic calipers and industrial touch panels [Zabiński and Mączka 2010]. Due to a reasonable cost and broad range of hardware platforms with different computational power capabilities PACs from Beckhoff [WWW-1], i.e. embedded PCs controllers were chosen. There are two kinds of Windows systems available for the controllers, i.e. Windows CE and Windows XP Embedded. Windows CE is equipped with .Net Compact Framework, Windows XP Embedded is equipped with .Net Framework. There are benefits of using the XP Embedded platform, e.g. homogeneity of the software platform for controllers and PC stations as well as availability of network and virus protection. Due to the financial reasons Ethernet network for communication and controllers with Windows CE were chosen for the two testbeds.

Current system structure implemented in the testbeds is shown in Fig. 1. There is one controller with an industrial 15" touch panel in each machine section or production line installed. The controller is connected with six machines via distributed EtherCAT communication devices equipped with digital or analog inputs and outputs.

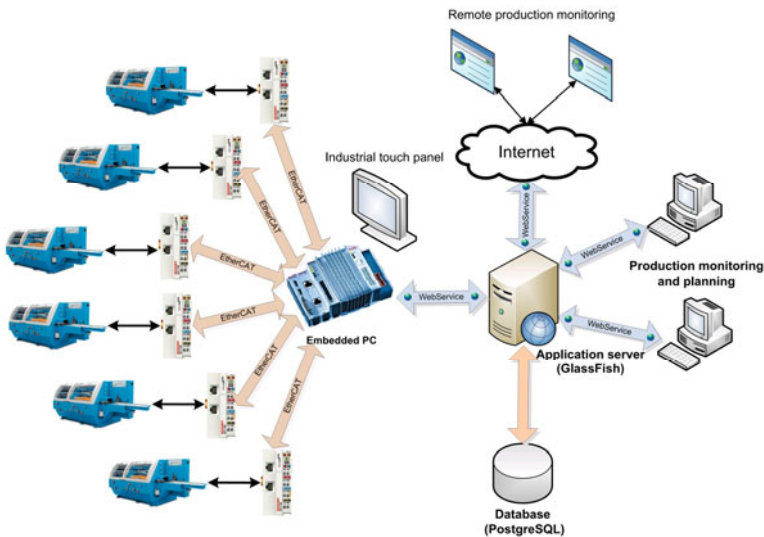


Fig. 1 Testbeds system structure

Embedded PCs are equipped with Windows CE 6.0 operating system, real-time PLC subsystem TwinCAT [WWW-1], UPS, Ethernet as well as RS-232/485 interfaces for communication and DVI/USB interfaces for touchable monitors connection.

The software part of the system consists of three layers: a shop-floor level software for embedded PC, a data and application server and www client stations. Detailed communication software structure for the shop-floor level is shown in Fig. 2.

In the software for an embedded PC four layers can be distinguished:

- the PLC program written in the ST language;
- the middleware module for communication between the PLC program and other system parts, written in C# with utilization of the Automation Device Specification (ADS) protocol [WWW-1];
- the HSI module - operator's GUI written in C#;
- the communication module written in C# for communication with database using web services technology.

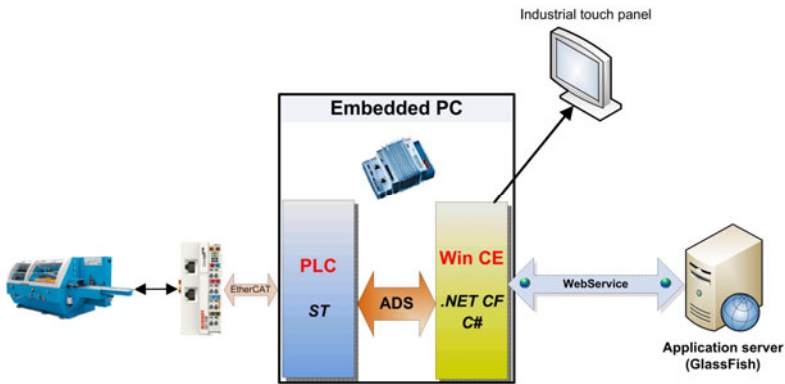


Fig. 2 Software structure for the shop-floor system level

PLC control programs run in the PLC layer on the same device simultaneously with GUI, data processing and database communication modules which run in Windows CE (Fig. 2). The ADS protocol enables C# programs to read and write data directly from and to PLC programs via names of PLC variables. It significantly simplifies the communication between PLC and C# applications. In the data and application server layer the PostgreSQL database has been used together with the GlashFish application server and web services written in Java. In the www client stations layer websites written in JSF, JSP, Ajax and JavaScript are used. Communication between the controllers and the database and between the presentation layer and the database is performed using web services or EJB technology. The presented approach simplifies communication inside the system in a heterogeneous software environment.

4 Human-System Interface

The main elements of HSI developed for the project were presented in the paper [Żabiński and Mączka 2010]. In this section the elements are reviewed as well as the newest prototype modules which have been created recently in cooperation with Bernacki Industrial Services company, i.e. GUI desktop application for closed-loop production scheduling and tools-management module are presented.

In general, HSI consists of two main layers, i.e. www and shop-floor. The www layer is a web page accessed through a web browser from the factory intranet or the Internet. The shop-floor layer is a GUI application which runs on embedded PCs installed on the factory shop-floor. In this layer the communication between an operator and the system is done via a 15" touchable monitor, RFID and barcode readers and electronic calipers. The Polish language is used in GUI as the system was installed in two Polish factories. Due to this reason GUI language presented in the figures below is Polish.

The www layer for each system field includes two main sections, i.e. an on-line view and statistics. Different structures of the web page have been tested and at the moment the most promising structure consists of four panels (see Fig. 3): a title panel, a toolbar, a left panel with a tree control for choosing elements inside the main sections and a right panel for data and graphs presentation. The structure was chosen on the basis of commercial MES systems and web sites examination as well as in cooperation with the testbeds users.

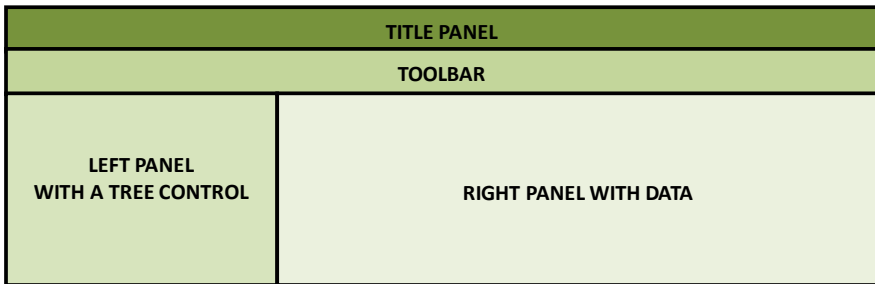


Fig. 3 Web page main structure

The on-line view enables real-time parameters monitoring of the chosen system field, e.g. in the production monitoring field, machines operation modes like: production, stoppage, lack of operator and also other information like: operator ID, order ID, shift production quantity, daily machine operation structure or detailed history of events are presented. The on-line view of the screws production company shop-floor in the production monitoring system field is presented in Fig. 4. Each square symbol represents one machine, the square color reflects the machine current state, e.g. green – production, yellow – stoppage. The tree control in the left panel reflects the factory hierarchical structure with shop-floors and machine sections. Information presented in the right panel depends on the chosen element in the left panel tree control. For example, if a particular machine section is chosen

in the tree control, then in the right panel a detailed view with graphical machines symbols is presented (see Fig. 5). Mouse cursor move towards the rectangle associated with a particular machine shows a tooltip with detailed information (Fig. 4). A double mouse click on the rectangle shows a separate web page with detailed information of the machine, e.g. daily operation structure or detailed history of events. The statistics section has a structure similar to the on-line view. In the left panel there is a tree control which shows hierarchical structure of the available statistics.

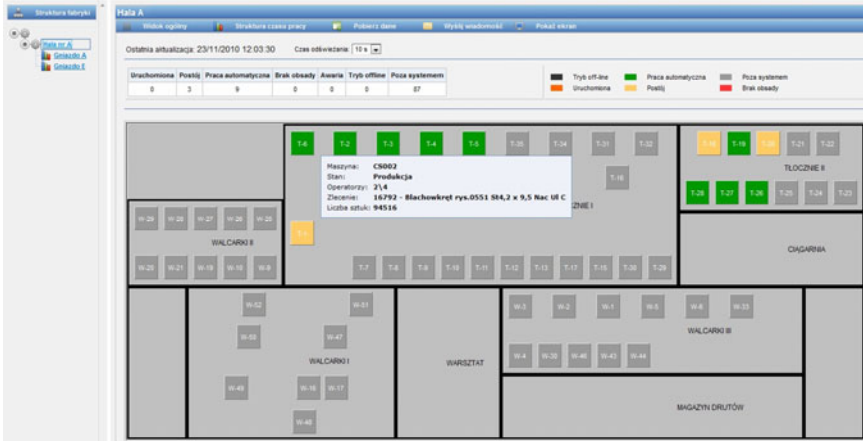


Fig. 4 On-line view of the factory shop-floor in the production monitoring system field

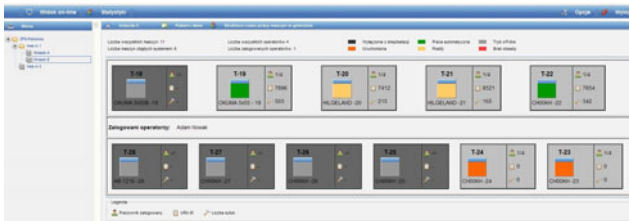


Fig. 5 On-line view of a machine section in the production monitoring system field

HSI for a shop-floor layer is an application written in C# for .Net CF. It has two main operation modes, i.e. locked and unlocked. The locked mode is a read-only mode with limited access to information. In the unlocked mode an operator can interact with the system. An operator can change HSI mode using his RFID card. Thanks to the RFID operator's badges a security policy was implemented at the shop-floor level of the system. According to the policy, suitable rights have been granted to the factory employees in order to establish the access levels in interactions with the system. Currently, four access levels were implemented in the shop-floor system level, i.e. operator, supervisor, quality inspector and maintenance department staff. All important actions and data inputting performed by users must be confirmed by their RFID badges. Thanks to the use of contactless cards, the

confirmation actions are convenient and not time consuming. In the locked mode visual information of machines operation modes, production plan, plan realization and the necessity of an operator interaction with the system is presented. The GUI was designed using the principle “*see rather than calculate*”. In the locked and unlocked mode the necessity of an operator interaction with the system is indicated by blinking of a signal tower and blinking of a GUI panel associated with the machine which needs intervention. The schematic structure of information for one machine in the locked mode is shown in Fig. 6.

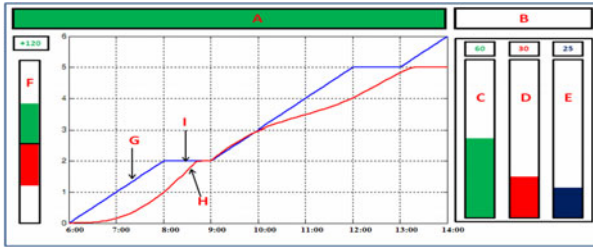


Fig. 6 Shop-floor HSI locked mode - section for one machine

The main part of the screen is a chart which shows planned and current production quantity (in items) as a function of shift time. The letters in Fig. 6 denotes: A – the main machine panel with a machine name, the panel blinks when an intervention is needed, B – a current operation name, e.g. setup, C – planned time for the current operation, D – the current operation time for the present operator, E – the current operation time for the leader, F – global plan realization progress for the machine obtained from a scheduling system, G – a production plan for the shift, H – real realization of the production plan for the shift, I – planned setup or adjustment. The locked mode screen for a machine section with six machines is presented in Fig. 7.

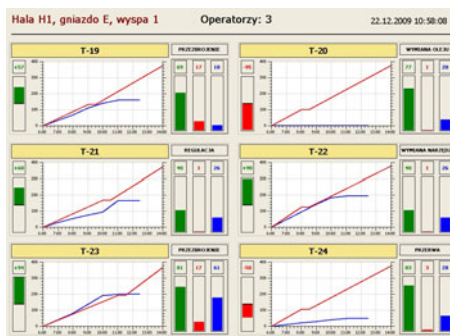


Fig. 7 Shop-floor HSI locked mode screen for six machines

In the unlocked HSI mode an operator can perform various tasks connected with the system, e.g. login, logout, taking up shift, order selection or confirmation,

stop reason inputting, quality control data inputting, tools management, etc. The functionality of the HSI which has been implemented so far is schematically presented in Fig. 8.

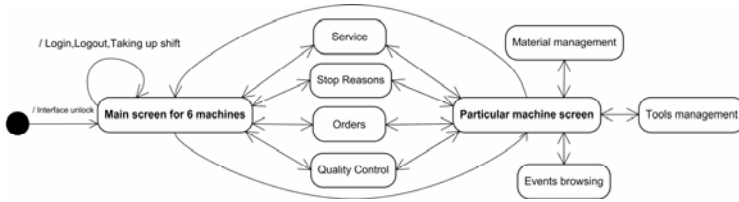


Fig. 8 Shop-floor HSI functionality in the unlocked mode

The main unlocked mode screen for a machine section with six machines is presented in Fig. 9.

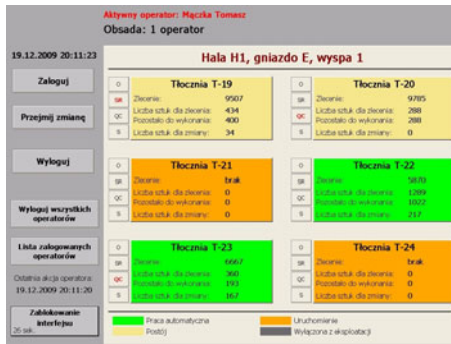


Fig. 9 Shop-floor HSI unlocked mode screen for six machines

The unlocked mode screen for a particular machine with chosen stop reason data inputting interface consisted of a set of hierarchical grid controls is presented in Fig. 10.



Fig. 10 Shop-floor HSI unlocked mode screen for stop reason data inputting

The main shop-floor GUI consists of two main sections, i.e. the system section (dark gray space – it allows login, logout, etc.) and the machines section (light gray space – it allows data inputting for particular machines) (Fig. 9). Small rectangles with letters O, SR, QC, S, associated with each machine panel (a large rectangle with machine name, e.g. Tłocznia T-19), indicate the action which should be performed for the particular machine, i.e. O – order selection or confirmation, SR – stop reason inputting, QC – quality control data inputting and S – service. When the machine panel is blinking, an operator can quickly determine the operation which should be performed, the color of the letters O, QC, SR or S becomes red for the active action.

The first attempt to the integration of a scheduling software used in the screws production factory and the monitoring system has been recently done. The goal of the integration is to create a closed-loop production scheduling system with a real-time data exchange between monitoring and scheduling subsystems. In the prototype software version, production orders for particular machines are automatically sent to the controllers by the scheduling software due to the production plan. Basic orders details are presented to operators using simple GUI in section orders (Fig. 8). Information about particular orders processing states and progress levels are updated in the database in real-time and presented to a production planner using prototype GUI desktop application (Fig. 11).

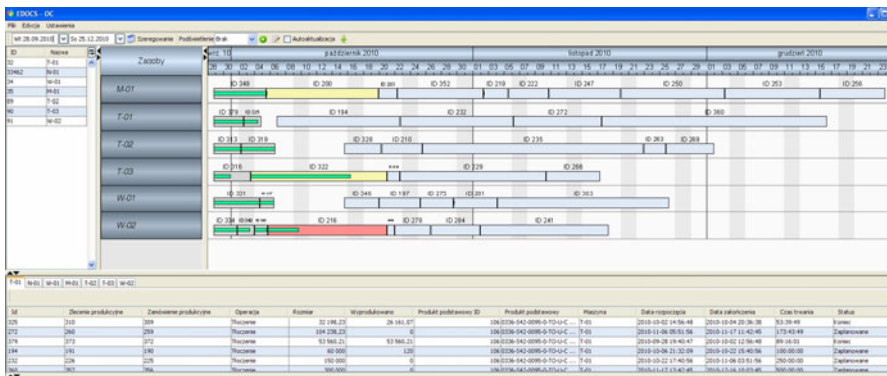


Fig. 11 Scheduling software HSI

A production plan is presented in the form of a Gantt chart with production orders represented by rectangles. Grey color rectangles represent completed production orders, light blue ones represent planned orders, yellow ones represent orders currently processed and light red ones represent orders for which there is a pause in production. A rectangle length represents a nominal processing time for a particular order. A thin green rectangle placed inside an order rectangle represents the actual progress in the order completion, which is calculated on the basis of the amount of produced pieces. After an order completion, planned start and completion dates of following orders are updated on the basis of real processing times of the order.

The tools-management module enables precise tools management including information about real usage of particular tools in production processes, e.g. number of pieces produced with the use of the particular tool. The module simplifies communication between machines operators and tools supply department workers and prevents mistakes in using inappropriate tools in the production process. The on-line view for the tools-management module is shown in Fig. 12.

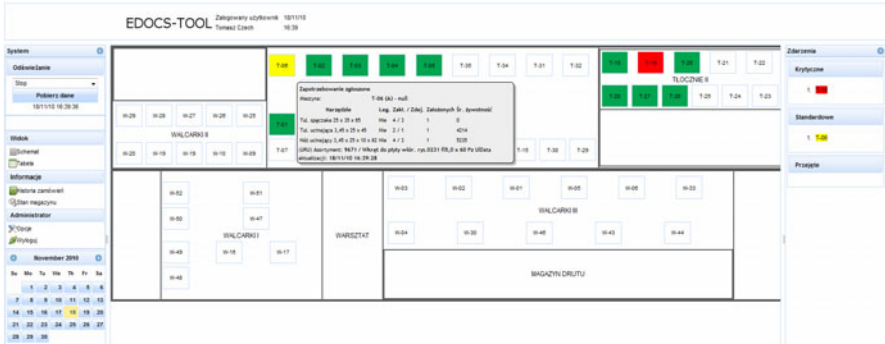


Fig. 12 On-line view of the factory shop-floor in the tools-management system field

The on-line view has a similar structure to the on-line view for production monitoring (Fig. 4) but shows information concerning tools management, i.e. elements of the current tools set used at the particular machine. The squares represent machines, squares colors reflect current states of the tools sets, e.g. necessity for making up the tools set. At the shop-floor level operators can interact with the module using GUI. The main screen for the tools-management module is shown in Fig. 13.



Fig. 13 Shop-floor main HSI screen for the tools-management module

The HSI presented in this section has been intensively tested in a real production processes and is continuously improved in accordance with users suggestions.

5 Current Results of the System Operation

The testbed installed in the screws production factory has been included in a daily production process since May 16th, 2009. The testbed installed at the WSK company has been a part of the real shop-floor since September 21th, 2010.

At the moment, the system is mainly used for data collection concerning the production process, machine operation and operators' work. The PLC layer is responsible for detecting and registering events which occurred in the machines as well as gathering chosen machines work parameters, e.g. the oil pump and the main motor starts and stops, failure and emergency signals, the machine operational mode (manual, automatic or for CNC machines MDA, cycle), signals from diagnostics modules (process monitoring devices), spindle current value, etc. The PLC program in the screws factory testbed also registers the amount of the produced pieces. Information about events, including timestamps, machine and operator identifiers and other additional parameters, is stored in the database. Two mechanisms are used to store data in the database, i.e. an asynchronous event driven method and a synchronous one with 10 sec. period time for diagnostics purposes. The system also detects and stores information of breakdowns, setup and adjustments, small stops, production speed in particular machines, spindle overloads, tools usage, material consumption, etc. Every production stoppage must be assigned to an appropriate reason, some like tool failure are automatically detected, while others must be manually chosen by operators via HSI.

On the server side there are software modules used for calculations of different KPIs (Key Performance Indicator), e.g. production efficiency, equipment and operators efficiency, etc. The real screws production quantity report (stated in items) calculated for the time interval from December 1st, 2009 to December 17th, 2009 for one machine included in the screws factory testbed, is shown in Fig. 14. During the analyzed period, the planned production time for the machine was 16 hours per day (2 shifts), as it is shown in Fig. 15, there were large fluctuations of production quantity.

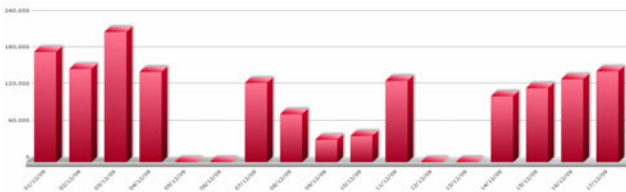


Fig. 14 Production quantity report – number of produced items as a function of days

A machine operation time structure can be analyzed and can be shown as a horizontal graph (Fig. 15). At the moment, it is possible to analyze machine operation time data from three points of view, i.e. a general view, a view with stop reasons and a detailed view. The general view divides machine operational time into three categories, i.e. the operator's absence, the automatic production and the stoppage. In the stoppage view, each stoppage time period is associated with the

appropriate stop reason. In the detailed view, periods of the manual machine operation are distinguished in each stoppage time. Different colors are designated to appropriate time intervals (Fig. 11), e.g. the operators absence – dark gray, the stoppage – light brown, the automatic production – light green, the manual operation – dark green, the start-up time – gray, the electrical breakdown – red, etc.

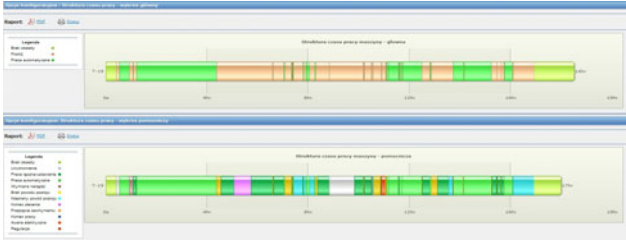


Fig. 15 Machine operation structure – general view and detailed view

Within 544 days of the system operation in the screws production factory, 508846 events were registered in the machine which was the first one included in the system. The whole number of events registered in the testbed equals 2226553. It should be emphasized that the testbed was gradually constructed. At the first stage of the project only one machine was included in the system [Żabiński et al. 2009], at the second stage of the project six machines constituted the testbed [Żabiński and Mączka 2010] and since May 26th, 2010 twelve machines have been connected to the system [Żabiński 2010]. Currently the system covers the following fields: production monitoring, quality control, material and tools management and fundamental support for maintenance department .

Within 59 days of the system operation at the WSK company, 57230 events were registered in the testbed. Up to know, the limited functionality system version has been running in the testbed. At present, the system provides fundamental tools for production monitoring and maintenance department. The testbed is currently under the development and the system is adapted to the company requirements and characteristics.

Due to a huge number of data gathered in the system there is a need to employ an artificial intelligence and data mining technology to supply the factories management with reliable knowledge of the production processes. During the long term test, when the system was included in the regular daily production, it was experimentally proven that the selected hardware and software platform is suitable for industrial implementation of the IMS [Żabiński et al. 2009]. The software modules (HSI, communication, web services, data acquisition) for Windows CE have been successfully running on the embedded PC controllers in parallel to PLC programs.

6 Future Work

There are six main fields in which the future work is planned to be done in parallel:

- more flexible system structure development by applying more intensively the service-oriented architecture (SOA) paradigm;
- industrial testbeds development;
- intelligent condition monitoring subsystem development;
- data mining and artificial intelligence production management support;
- Petri Nets based manufacturing system modeling and intelligent production process scheduling;
- multi-agent software structure development.

The first and the second field deals with the modification of the current system structure combined with the industrial testbeds development. The system structure must be modified in order to attain more flexible architecture which should satisfy different demands of diverse metal component production companies. The new structure should simplify the system implementation in companies with different levels of ICT employment and various production and data resources. The new software structure should also enable its easy adaptation to the needs of other production sectors. It is planned to include additional 58 machines in the screws production factory testbed with a minimized number of embedded PCs in order to reduce the system costs. During the development of the testbed, additional diagnostics and process monitoring equipment will be included in the system, e.g. quality measurement devices, current and force sensors, etc. In accordance with the WSK testbed, additional 64 machines are planned to be included in the system. Due to the customer demands one embedded PC and one PC computer for every machine is to be installed. The system is to be integrated with SAP business software and will be used for delivering electronic versions of technical and quality control documents directly to operators' workstations on the shop-floor. The system should support production management as well as support maintenance department by the usage of advanced and intelligent software tools devoted to machines condition monitoring. The new testbed structure is shown in Fig. 16.

In the third field, automated and intelligent condition monitoring subsystem is to be developed specially for the WSK testbed. Vibration and current sensors and AI methods (support vector machines, neural networks) for alarms detection and machines breakdowns predictions are planned to be employed.

The fourth field concerns the development of data mining and artificial intelligence support tools for the management and control of the manufacturing system. Continuously discovered knowledge will support everyday production process management and control, providing the answers to numerous questions as follow:

- what are the bottlenecks in the production system;
- what factors influence the production process and in what manner;
- what problems occur and under what circumstances;

- what should be done to increase manufacturing process productivity;
- what should be done to increase product quality;
- etc.

Moreover, it is planned that the system will automatically:

- identify connections and relations in the production system;
- discover possibilities for more effective usage of production resources;
- aid work planning and scheduling process using data mining and artificial intelligence support, including discovered knowledge of workers skills, machines conditions and performance, etc.;
- detect the possibility of problem occurrence and suggest the best solution, e.g. task allocation to reduce the probability of the stoppages;
- initiate corrective and preventive actions and inform appropriate persons who manage processes, production or company to prevent or limit occurrences of the problems in the future;
- use Statistical Process Control (SPC) with artificial intelligence support for early detection of possible problems in production systems;
- continuously calculate Key Performance Indicators (KPI), including Overall Equipment Effectiveness (OEE);
- etc.

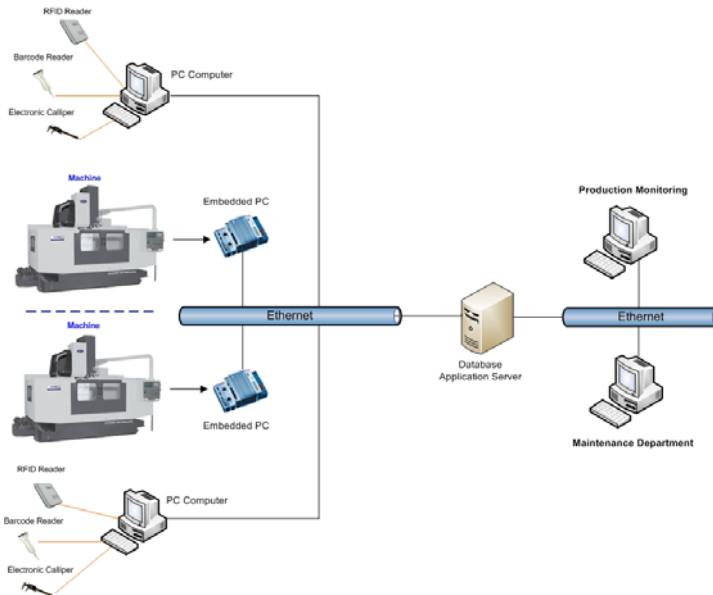


Fig. 16 New WSK testbed structure

Data mining techniques like classification and association rules, decision trees, etc. are going to be employed. So far, the research has been devoted to automatic patterns (rules) generation to discover knowledge and connections between the collected data [Żabiński et al. 2009]. For model reduction the automated query generation method is to be considered. Waikato Environment for Knowledge Analysis (WEKA) is planned to be employed to develop data preprocessing and to build mining models. WEKA is a popular suite written in Java and available under the GNU General Public License.

The fifth field concerns the development of Hierarchical Timed Coloured Petri Net based manufacturing system modeling and production process scheduling software tools. A scheduling method employs the production process model (Flexible Job Shop) and takes into account [Božek and Żabiński 2010]:

- human, tools, material and transportation resources availability;
- setup times;
- transportation times;
- release time restrictions;
- batch processing.

The fully automated software module devoted to define and simulate the production process in the screws production factory is under development. The scheduling module will acquire data from the monitoring system in real time and will generate production plans by the model simulation with data mining and artificial intelligence support. So far, the CPN software has been used to develop and test the Petri Nets production system model. The destination software module which will be integrated with the currently developed system will be based on Java technology and will use web services for data exchange.

The sixth field concerns a multi-agent software structure development for a holic-based IMS. The analysis of the requirements and design of an appropriate system topology will be performed. It is considered to combine two approaches for an agent encapsulation, i.e. the functional and the physical decomposition. Due to the functional decomposition, software agents for system support, process modeling and task scheduling (contact agent, order agent, supply agent) will be defined. Due to the physical decomposition, agents related to factory floor equipment, e.g. machine agent, machine section agent, etc. are to be used. Currently available multi-agents frameworks (e.g. FIPA-OS, April Agent Platform, JADE, Comtec Agent Platform) will be examined in order to specify a possibility for using an existing framework for the system development.

New system versions, especially multi-agent ones, are going to be developed and tested with the usage of the laboratory Flexible Manufacturing System (FMS) testbed (Fig. 17) consisting of an integrated CNC milling machine, robot and vision system.

For every field in which the future work is going to be performed an appropriate HSI will be designed and implemented. So far, the developed HSI is to be constantly improved and developed due to the real production testbeds operation results and users feedback.

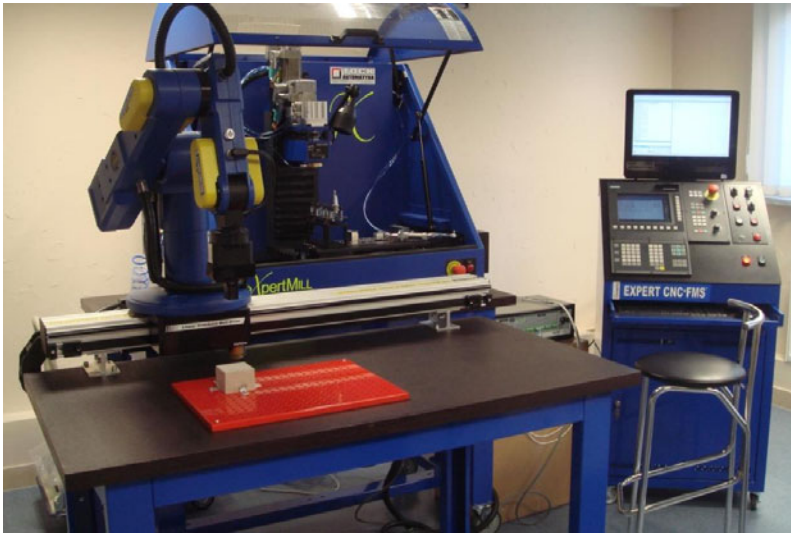


Fig. 17 Laboratory FMS testbed

7 Conclusions

Most of the currently operating manufacturing systems suffer from the lack of flexibility, robustness, re-configurability, machine knowledge discovery and direct communication with operators and machines operated on the shop-floors of factories. Holonic organizational concept combined with multi-agent system architecture and AI techniques has been widely recognized as an appropriate platform for the implementation of next-generation innovative manufacturing systems referred to as Totally Intelligent Manufacturing Systems [Oztemel 2010].

Up to now, in the project devoted to industrial implementation of holonic based Intelligent Manufacturing System the basic prototype hardware and software platform has been developed. The platform has been tested in two different metal component production factories. The extended HSI for users on a factory-floor level as well as for users from management has been designed, implemented and tested.

Current project results constitute promising perspective for IMS and advanced HSI industrial implementation even in small and medium manufacturing companies.

Acknowledgment

FMS testbed was bought as a part of the project No POPW.01.03.00-18-012/09 from the Structural Funds, The Development of Eastern Poland Operational Programme co-financed by the European Union, the European Regional Development Fund.

References

- [Bożek and Żabiński 2010] Bożek, A., Żabiński, T.: Colored timed Petri Nets as tool of off-line simulating for intelligent manufacturing systems. *Electrical Review*, Association of Polish Electrical Engineers SEP 86(9), 101–105 (2010) (in Polish)
- [Cummings et al. 2010] Cummings, M.L., Sasangohar, F., Thornburg, K.M.: Human-system interface complexity and opacity part i: literature review. MIT Humans and Automation Laboratory, Cambridge (2010), <http://web.mit.edu/aeroastro/labs/halab/index.shtml> (accessed November 26, 2010)
- [Gong 2009] Gong, C.: Human-machine interface: Design principles of visual information in human-machine interface design. In: *Proc. IEEE Conference on Intelligent Human-Machine Systems and Cybernetics*, San Antonio Texas, USA, pp. 262–265 (2009)
- [Institute for Prospective Technological Studies 2003] Institute for Prospective Technological Studies, Technical report: The future of manufacturing in Europe 2015-2020 - The challenge for sus-tainability, European Commission's Joint Research Centre (2003)
- [National Research Council 1998] National Research Council.: Visionary manufacturing challenges for 2020. Committee on visionary manufacturing challenges, board on manufacturing and engineering design, commission on engineering and technical systems. National Academy Press, Washington D.C (1998), <http://www.nap.edu>
- [Oztemel 2010] Oztemel, E.: Intelligent manufacturing systems. In: Benyoucef, L., Grabot, B. (eds.) *Artificial Intelligence Techniques for Networked Manufacturing Enterprises Management*, pp. 1–41. Springer, London (2010)
- [WWW-1] Beckhoff Information System, <http://infosys.beckhoff.com/> (accessed November 26, 2010)
- [Żabiński 2010] Żabiński, T.: Implementation of programmable automation controllers - promising perspective for intelligent manufacturing systems. *Management and Production Engineering Review*, Polish Academy of Sciences 1(2), 56–63 (2010)
- [Żabiński and Mączka 2010] Żabiński, T., Mączka, T.: Human system interface for manufacturing control - industrial implementation. In: *3rd Int. Conf. on Human System Interaction*, Rzeszow, pp. 350–355 (2010)
- [Żabiński et al. 2009] Żabiński, T., Mączka, T., Jędrzejec, B.: Control and monitoring system for intelligent manufacturing – hardware and communication software structure. In: *Computer Methods and Systems*, Cracow, pp. 135–140 (2009)

Precision of an Expert Fuzzy Model

I. Rejer

Department of Information Technology, West Pomeranian University of Technology
in Szczecin, Szczecin, Poland
irejer@wi.zut.edu.pl

Abstract. The aim of this paper is to present some difficulties which can be met when a fuzzy model is built with a domain expert help. The paper discusses methods used at succeeding steps of expert modeling process and also presents some of them on a real example. The main concern of the paper is a high quality of a fuzzy expert model which can be easily missed when not all aspects of expert modeling are carefully tackled. Practical experiments presented in the last part of the paper shown that most popular methods used in the process of interviewing an expert are not always appropriate and can be a reason of creating a completely useless expert model.

1 Introduction

There are two general approaches to fuzzy modeling - to create model automatically on the basis of numeric data or to build model manually with an assistance of a domain expert. Both approaches are of the same importance for practical applications but in scientific literature mostly only first of them is considered. Expert fuzzy modeling is regarded by most researchers as much easier to apply and thereby – less interesting. In reality, however, the situation is quite opposite – there are so many tools for automatic fuzzy modeling that when a data set is prepared correctly, it is enough to use one of them to obtain a fuzzy model of a high quality. Unfortunately, such automatic tools do not exist in case of expert modeling. Of course there are computer programs which assist an expert in the process of a fuzzy model creation but their only task is to combine knowledge provided by the expert to the form of a fuzzy model – they do not automatically derive knowledge from expert heads.

Meanwhile, the problem of extracting expert knowledge (regardless of manual or automatic) is just considered as the biggest problem of a fuzzy expert modeling. It is a common situation that, in case of a very complicated system, expert cannot give an unambiguous answer for the question why he took a specific decision or advised a specific course of action. This is due to the fact that an expert solves

problems not only by applying some general rules but also by using his internal intuition which cannot be easily represented in a form of logic rules.

In order to create a fuzzy expert model, an expert has to provide information about membership functions and rules which should be applied in this model. The process of defining membership functions is commonly regarded as a difficult one because the knowledge of functions parameters is hidden deep insight experts heads [Piegat 1999]. On the contrary, the process of defining rules is regarded as quite effortless and therefore it is often carried out without enough care. In practice, however, the quality of the fuzzy expert model depends straightly on the rule extraction process – to be more specific, it depends on the order of rules presented to an expert during an interview with a model creator.

Why the rule order is so important? In separate situations a human expert takes decisions which are not connected to each other. So when these separate decisions are formed to the shape of rules and then joined together, they mostly constitute a non-linear model. However, when the expert is presented with all possible situations, given in the form of fuzzy rules, at once, he is very prone to assign conclusions to each rule so that a linear model was created. In this way a non-linear model used by an expert in his actions changes into a linear one. This of course brings a rapid drop in a model quality.

The aim of the paper is to present the process of creating a fuzzy expert model and to show on a real example that a quality of this type of fuzzy model depends directly on both - the method used for determining membership function parameters and the order of rules presented to the expert during an interview. The content of the paper is as follows. Section 2 shortly characterizes some main aspects of a fuzzy model; Section 3 describes main steps of fuzzy expert modeling process and Section 4 presents a set of fuzzy expert models built for the same relation existing between *car price*, *engine size* and *highway mpg*.

2 Fuzzy Model

A fuzzy model is a model which is defined in terms of membership functions and rules. The mathematic form of a fuzzy model depends on applied mathematics engine. For example, assuming:

- Larsen model,
- grid partitioning of an input space,
- singleton membership functions of output variable,
- PROD-MAX inference mechanism,
- weighted average sum defuzzification method,

the equation of a fuzzy model takes the following form [Rutkowska et al. 1999]:

$$y_0 = \frac{\sum_{i=1}^m y_i \left(\mu_{B_i}(y_i) \prod_{j=1}^s \mu_{A_{ij}}(x_j) \right)}{\sum_{i=1}^m \left(\mu_{B_i}(y_i) \prod_{j=1}^s \mu_{A_{ij}}(x_j) \right)}, \quad (1)$$

where: y_0 - output variable, x_j - input variable j ($j=1, \dots, s$), y_i - conclusion of i rule ($i=1, \dots, m$), $\mu_{A_{ij}}(x_j)$ - degree of activation of j premise of i rule, $\mu_{B_i}(y_i)$ - degree of activation of i rule conclusion.

In order to create a fuzzy model given by (1), the following scheme should be performed:

- define one set of membership functions per each model variable (input and output),
- create rule net of the fuzzy model by joining cores of fuzzy sets of succeeding input variables,
- define a conclusion per each physically sound node of the rule net,
- apply a chosen mathematic engine.

There are two general approaches which can be used to perform steps of the above algorithm – to create model automatically on the basis of numeric data or to build model manually with an assistance of a domain expert. The first approach is based on the knowledge derived from a data set describing the analyzed relation. This knowledge is acquired during an optimization process which formal definition is as follows [Chen and Linkens 2004]: given the n input-output pairs $P(x, y)$ and the specified model error $\epsilon > 0$ obtain the minimal number of rules and optimal parameters of membership functions of the fuzzy model such that the error function $E = \|y - F\|$ satisfies the inequality $E(\theta, w) < \epsilon$. The process of optimizing fuzzy model parameters can be carried out with different methods [Yager and Dimitar 1994] e.g.: gradient algorithms, genetic algorithms, fuzzy relations, methods based on decision trees etc.

A fuzzy model built automatically is in most cases much more precise than an expert fuzzy model and its creation is not as time consuming as interviewing a domain expert. However, the overall quality of this type of fuzzy models depends on the quality of a data set which in reality is often insufficient to build a model producing credible results in the whole physically sound input domain of the analyzed relation. As a result, models built for most practical applications give precise results but only in a very small part of the whole input domain – this part which is properly covered by data points.

The second approach to fuzzy model creation – it is by interviewing a domain expert – omits this drawback. Fuzzy models created with an assistance of a

domain expert are in most cases credible in the whole physically sound problem domain. Of course their creation needs more time but the range of applicability of the final model is incomparably greater than in case of models generated automatically on the basis of a data set.

3 Expert Knowledge Extraction

In order to build a fuzzy model for a given relation with a domain expert help, a model creator has to obtain from an expert information about membership functions and rules essential for the model. To deal with this task he has to carry out a series of interviews during which an expert is asked [Baetge and Heitmann 2000]:

- Which factors should be used to evaluate the output variable of the relation?
- How should the potential values of each factor be interpreted?
- How to aggregate the factors to form an overall evaluation?

The first question is a question about input variables of the fuzzy model, it is about factors which influence the output variable. The second question is a question about membership functions of each input variable. Answering this question, an expert has to provide information about:

- numeric domain of each variable,
- number of linguistic terms describing each variable,
- names of these terms,
- numeric intervals assigned to each linguistic term (or intervals/points characteristic for each term),

And the last question is a question about fuzzy model rules. Addressing this question, an expert has to define:

- which linguistic terms of succeeding input variables of the analyzed relation should be joined together to form a rule premise and
- which linguistic term of output variable (rule conclusion) should be assigned to each premise to form the whole rule.

Sometimes in order to simplify the process of deriving rules from an expert, rules premises are created in an automatic way, by building all possible combinations of linguistic terms of all input variables. When such approach is applied, one additional condition has to be fulfilled in order to keep the high model quality – an expert should determine whether all combinations of linguistic terms of input variables are physically sound. If some of them are not, they should be removed from the model rule base.

3.1 Extraction of Membership Functions Parameters

Four main methods for creating membership functions with an expert assistance can be pointed out [Rejer 2006]:

- method of discrete points,
- method of modal values,
- method of α_{50} -cuts,
- method of 100% membership.

To define a membership function with the method of discrete points, an expert has to assign membership degrees to a given set of points. Expert answers constitute a discrete membership function; if a continuous function is needed, succeeding membership degrees are joined together, mostly with straight lines. The method of discrete points is the most laborious one but under the condition that an expert is capable of giving more or less precise membership degrees, it provides the most precise mapping of expert knowledge. It is due to the fact that an expert provides not only one or two characteristic points of the membership function (like in other methods) but indirectly decides also on the function shape. Fig. 1a presents this method in practice – black points correspond to membership degrees of succeeding values of variable x_1 presented to an expert.

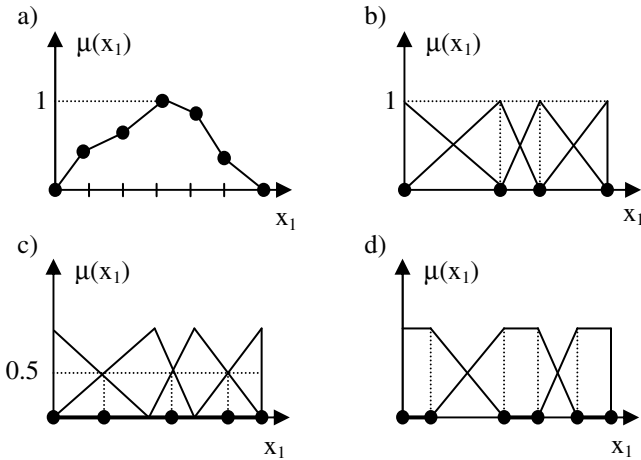


Fig. 1 Membership functions defined by an expert with: method of discrete points (a), method of modal values (b), method of α_{50} -intersections (c), method of 100% membership (d)

In the second method an expert task is to give a set of one-element cores of all fuzzy sets defined over the whole variable domain. That means that this time an expert is not asked about membership degrees of given values of the analyzed variable, but he is asked about the typical values of each fuzzy set. As a result of the

interview, a set of modal values of membership functions is obtained. The given modal values are then used by a model creator to create a set of membership functions (mostly triangular or Gaussian). Fig. 1b presents a set of four asymmetrical triangular membership functions defined with this method over variable x_I (black points correspond to four modal values provided by a domain expert).

The method of α_{50} -cuts is similar to the previous one in respect to its output which is also a set of one-element cores of all fuzzy sets existing in an analyzed variable domain. This time, however, an expert task is not to define modal values of each fuzzy set but to divide the domain of the analyzed variable into a set of separable intervals which, in his opinion, contain values belonging to succeeding fuzzy sets. Since these intervals are treated as α_{50} -cuts of fuzzy sets, their borders define crossing points of neighboring membership functions. Like in the previous method, mostly triangular or Gaussian functions are defined over the given set of α_{50} -cuts. Fig. 1c presents a set of four asymmetrical triangular membership functions defined with this method over variable x_I (bold lines on the x_I axis correspond to four intervals provided by a domain expert).

The last method is the only one in which membership functions of multi-element cores are created. In most cases this method utilizes trapezoid functions but it is also possible to use Gaussian functions with flat top. The method is similar to the previous one in respect to the expert task which is to give a set of intervals defined over the analyzed variable domain. This time, however, the interpretation of intervals is different – an interval is not interpreted as an α_{50} -cut of a fuzzy set but as a core of this set. So, in this method an expert is asked for defining intervals of typical values of the analyzed variable, it is intervals of values which membership degrees to corresponding fuzzy sets are equal to 100%. Fig. 1d presents a set of three asymmetrical trapezoid membership functions defined with this method over the variable x_I (bold lines on the x_I axis correspond to three intervals provided by a domain expert).

The choice of the method for a practical application is always an individual matter and depends only on an expert. The model quality (defined in terms of model precision) will be the highest not when the most precise method (it is a method of discrete points) is applied but when the method corresponding to expert mental abilities and preferences is chosen. Therefore, if the model quality is a crucial factor of the fuzzy modeling, an expert should be first tested (on the testing problem) in respect to all four methods and the method best suited to him should be chosen. This method should be then applied in the process of determining the parameters of membership functions for a given problem.

3.2 Extraction of Rules

During this step of the fuzzy expert modeling, an expert task is to assign suitable conclusions to all physically possible combinations of fuzzy sets defined over domains of succeeding input variables. In other words, an expert is presented with rule premises and his task is to provide rule conclusions. At the end of this step a set of rules of a form (2) is obtained:

if x_1 is A and x_2 is B then y is C, (2)

where: x_1, x_2 – input variables; y – output variable; A, B, C – fuzzy sets defined over domains of corresponding variables.

In most practical applications, all rule premises are presented to the expert simultaneously (Table 1). Due to this an expert gains a global view on the whole rule base. Theoretically, such approach simplifies the process of defining rules because an expert can make corrections to all rule conclusions simultaneously but in practice it has a big negative influence on the model quality. Why is it so?

Table 1 Rule base of a fuzzy model

		X1		
		A1	A2	A3
X2	B1			
	B2			
	B3			

First of all, the classic approach forces the global process of reasoning – an expert does not think how the model should look like in separate subregions of the input domain but thinks about the shape of the whole model at once. This causes that first he creates in his mind the general view of the model and then tries to fit rule conclusions so that they map this model.

And what is the most realistic shape of a human expert model? Of course a linear one. A human expert behaves in a non-linear way but mostly only on a level of his subconsciousness. When comes to express his opinion or when comes to define rules of his behavior, his way of thinking changes to a linear one. This is nothing strange or new because a lot of researches on human behavior show that people (consciously or subconsciously) prefer symmetry and order [Jaśkowski 2009]. So even if they think in a non-linear way, when they are forced to vocalize their believes, they occur to be linear.

All these mean that when an expert is presented with all possible rules at once, he is very prone to create a rule base of a linear characteristic instead of providing rule conclusions mapping his real rules of behavior. Therefore, when a quality of a fuzzy expert model is important, it is much better to choose the second possible option for presenting rule premises to the expert – it is presented them one by one. Of course also this time it is important not to present rules in a linear order (e.g. by presenting neighboring rules from Table 1) but to disturb linearity by picking out rules randomly from different positions of the table.

Defining conclusions for randomly chosen rules is a much more difficult and mind-demanding task for the expert. However, this effort is profitable because provided that the expert is a real one, the quality of the final model built with this approach can be significantly higher than the quality of the model built with a classic approach. Some would say that in order to make the task simpler for an expert,

rule premises could be presented to him not in a total random way but could be ordered in a way shown in Table 2 (beginning with rules coding the clearest situations and ending with rules coding the most vague situations). According to the author of the paper this simplification, however, introduce too much linearity to rule premises presentation and hence can be a reason of an unnecessary drop in a model quality.

Table 2 Possible order of rule premises presentation

		X1				
		A1	A2	A3	A4	A5
X2	B1	1	3	2	3	1
	B2	3	4	3	4	3
	B3	2	3	2	3	2
	B4	3	4	3	4	3
	B5	1	3	2	3	1

One more aspect should be touched when the process of creating fuzzy expert rule base is described. Equation (2) and Table 1 presents rules located in a two input domain. It is not just a common simplification of a bigger multi-input case but the only possibility of creating a reasonable expert rule base. The problem is that a human expert is not able to reason correctly in a multi-dimensional case. His abilities ends with problems of three dimensions, in case of fuzzy rule bases – with two input dimensions and one output dimension. That means that an expert is capable for creating credible rules only when their premises contain two inputs. When number of inputs in rule premises increases, the quality of rules provided by an expert dramatically decreases.

Hence, in a multi-input case, the rule base has to be reconfigured before it is presented to an expert. In the reconfiguration process input variables are gradually aggregated and the whole rule base is decomposed to a set of 3D rule bases [Facchinetti and Mastroleo 2005]. Assuming five input rule base (of inputs: $x1$, $x2$, $x3$, $x4$, $x5$), the reconfiguration process could look like in Fig. 3. According to this figure at first input variables are group in pairs and two rule bases (RB1 and RB) are obtained. Premises of rules form both rule bases are presented to an expert to evaluate rules conclusions. Next, conclusions of rules from rule bases RB1 and RB1 are joined together and rules premises of a new rule base (RB3) are composed. Also this time rules are presented to an expert who defines their conclusions. And at the end rules conclusions from RB3 are joined with the last input variable $x5$ and the last rule base (RB4) is created and evaluated by an expert.

The main benefit of the rule base reconfiguration is that the rule base (after the reconfiguration – a set of rule bases) can be easily defined by a domain expert. Without such reconfiguration, an expert could not have been able to create any sensible rules. Of course, there is no free lunch – the simplification of the rule base can sometimes result in a slight drop in an expert model precision. This is a

result of a smaller number of rules existed in reconfigured rule base in comparison to the original rule base – in the example presented in Fig. 2, the original rule base was composed of 243 rules (five inputs described by three fuzzy sets) and the re-configured one of 36 rules (four rule bases of nine rules).

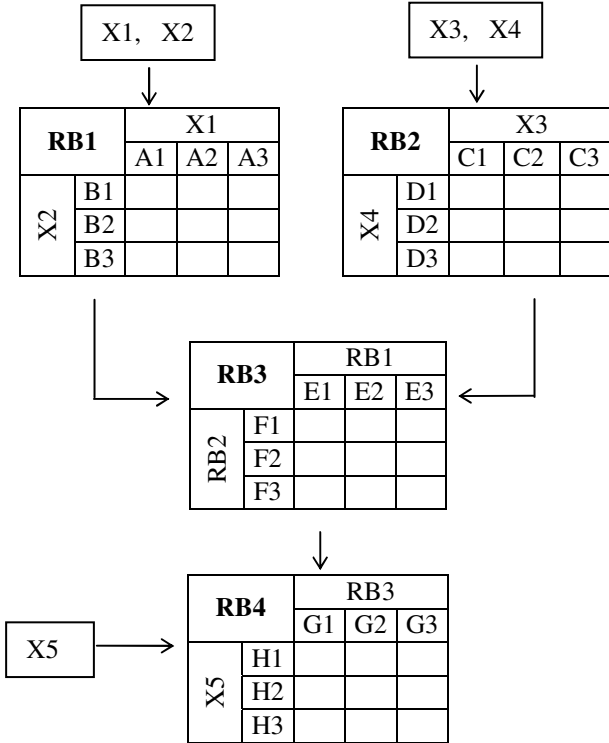


Fig. 2 The process of a rule base reconfiguration

4 Case Study: Car Price Evaluation

In order to present how the choice of method used for defining membership functions and the order of rule premises presentation can influence the model quality, two experiments were conducted. In both experiments a set of expert fuzzy models for car price evaluation were built. A car price was evaluated with regard to two attributes *engine size* and *highway mpg* (miles per gallon). The models performance was tested on a data set coming from UCI Machine Learning Repository (file: Automobile) [Asuncion and Newman 2007]. The same data set was used to establish numeric domains of the model variables:

- car price <5118, 45400>,
- engine size <61, 326>,
- highway mpg <16, 54>.

4.1 Experiment I: Defining Membership Functions

The aim of the first experiment was to examine the influence of the method used for defining fuzzy membership functions on the model quality. Because of insufficient expert availability, only three methods were compared:

- method of modal values (model A),
- method of α_{50} -cuts (model B),
- method of 100% membership (model A).

Due to the verification procedure, described later, only membership functions of input variables were under consideration. It was assumed that in sake of simplicity both input variables (*car price*, *engine size*) would be described only by three fuzzy sets. Rules conclusions for all models were established automatically on the basis of the data set described earlier in this section. The whole experiment took five days, a break between acquiring each set of membership functions was two days.

At first, the method of modal values was taken into account and the expert was asked to determine the smallest, medium and the biggest *engine size* and *highway mpg*. The expert answers are presented in Table 3. As it could be presumed the expert divided domains of all variables into more or less equal intervals.

Table 3 Modal values of fuzzy sets defined over domains of input variables

	The Smallest (TS)	Medium (M)	The Highest (TH)
Engine size	61	200	226
Highway mpg	16	35	54

Modal values from Table 3 were then used to built two sets of triangular membership functions (Fig. 3). In order to gain the homogeneous scale of both model variables, membership functions cores were normalized to the interval $\langle 0, 1 \rangle$, according to the formula:

$$x_n = \frac{x_r - x_{\min}}{x_{\max} - x_{\min}}, \quad (3)$$

where: x_n – normalized value of x variable, x_r – real value of x variable, x_{\min} – minimal value of x variable, x_{\max} – maximal value of x variable.

In order to build a complete fuzzy model, membership function from Fig. 3 were introduced to a fuzzy-neural network (of architecture matched to Larsen model given by (1)). Next the network was trained during 200 epochs according to the backpropagation algorithm with momentum rate. The training process was applied only to rules conclusions and was based on the data set from UCI Machine Learning Repository. The average model error, calculated according to (4) [Aczel 1993] was equal to 6.30%.

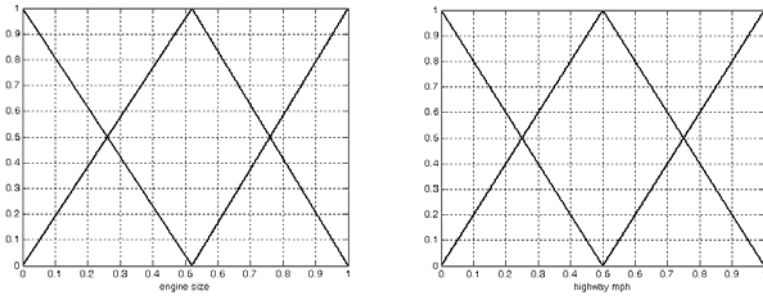


Fig. 3 Membership functions of input variables of model A after the normalization process

$$MAE = \frac{\sum_{k=1}^n |y_k^* - y_k|}{n} * 100\% , \tag{4}$$

where: y_k^* - real values, y_k – theoretical values.

The same algorithm was repeated two more times to built models B and C. In the process of creating model B method of α_{50} -cuts was used. Since this method in its classic form is very prone to create abnormal fuzzy sets, the expert was asked only to give intervals for the smallest values of both input variables, remaining intervals were calculated automatically. Expert answers together with automatically calculated intervals are given in Table 4 and membership functions created on the basis of these answers are presented in Fig. 4. MAE of the complete model B was equal to 6.08%.

Table 4 Intervals defined over domains of input variables

	The Smallest (TS)	Medium (M)	The Highest (TH)
Engine size	<61; 150)	<150; 282.5)	<282.5; 326>
Highway mpg	<16; 25)	<25; 44)	<44; 54>

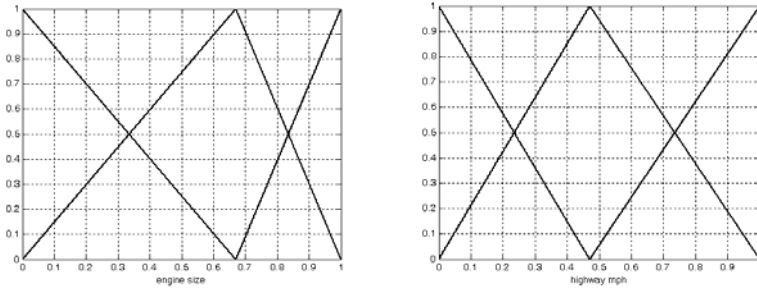


Fig. 4 Membership functions of input variables of model B after the normalization process

During the process of creating model C, the expert was ask for defining intervals of typical values of both analyzed variables. Expert answers are given in Table 5 and membership functions created on the basis of these answers are presented in Fig. 5. MAE of the complete model C was equal to 6.58%.

Table 5 Intervals of typical values defined over domains of input variables

	The Smallest (TS)	Medium (M)	The Highest (TH)
Engine size	<61; 80)	<180; 200)	<300; 326>
Highway mpg	<16; 20)	<32; 37)	<50; 54>

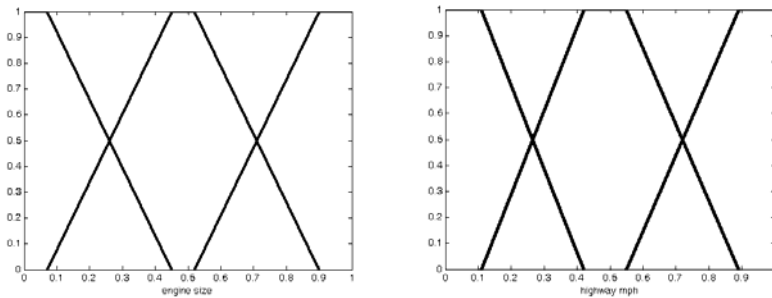


Fig. 5 Membership functions of input variables of model C after the normalization process

Summarizing the experiment described in this section it should be noticed that mean absolute error of all three models built in the experiment are on the same

level (model A - 6.30, model B - 6.08, model C - 6.58). That means that in the analyzed problem the choice of the method of creating expert membership functions had almost none influence on the model quality.

4.2 Experiment II: Order of Rule Premises Presentation

The aim of the second experiment was to examine the influence of order of rule premises presentation on the quality of a fuzzy model. These time only two fuzzy models were created:

- model D - rule premises were presented in a random way,
- model E - rule premises were presented in a tabular way.

Since the first experiment shown that all three analyzed methods of creating membership functions gave almost the same results, the simplest method (method of modal values) was used for creating membership functions for both models. It should be underlined here, that in a real application an expert preferences (about the method of membership functions creation) should be tested with a special testing problem which differs from this which is about to be solved. This was not done in the paper because the given problem is just a testing problem.

In the first part of the experiment rules premises were presented to the expert in a random way. The rule order (together with expert conclusions) is presented in Table 6. In order to built a complete fuzzy model, rules from Table 6 and membership functions from Table 3 were join together with a mathematic engine given by (1). The model performance was tested over the data set from UCI Machine Learning Repository. The average model error, calculated according to (4) was equal to 6.36%. The model surface together with data from the testing set are presented in Fig. 6.

Table 6 Expert rule base obtained after random rules premises presentation

No	Engine size	Highway mpg	Car price
1	M	TS	TS
2	M	M	S
3	TB	TB	TS
4	TS	M	M
5	M	TB	M
6	TB	TS	TS
7	TS	TS	S
8	TB	M	TS
9	TS	TB	TH

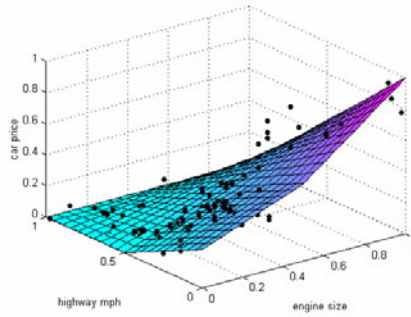


Fig. 6 Fuzzy expert model for car price evaluation obtained in the first part of the second experiment

Two weeks after the first part of the second experiment, the same expert was once again asked for defining conclusions for the same rules premises (columns 2 and 3 from Table 6). This time, however, he was presented with the whole rule base at once in a tabular form. The rule base, together with the expert answers, is presented in Table 7 and the shape of the fuzzy model built with this rule base – in Fig. 7. The model error, calculated with (4) was equal to 23.66%.

Table 7 Expert rule base obtained in the second survey

		Engine size		
		TS	M	TB
Highway mpg	TS	M	H	TH
	M	S	M	H
	TB	TS	S	M

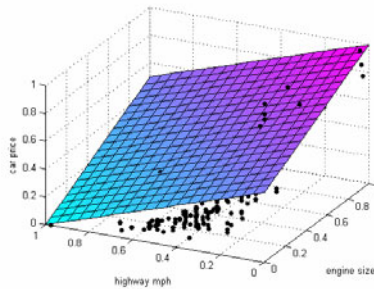


Fig. 7 Fuzzy expert model for car price evaluation obtained in the second part of the second experiment

Comparing Fig. 6 and Fig. 7 it can be noticed that model surfaces drawn in them are very similar in respect to the general shape and slopes of the surfaces towards succeeding axis. This is nothing strange because both models were prepared by the same expert which means that both of them map the same knowledge. However, the comparison of testing errors of both models reveals that their quality is completely different. While the error of the first model was equal to 6.36%, the error of the second one was almost four times greater and was equal to 23.66%. Since the only difference in the process of both model creation was the order of rule premises presentation, the errors comparison shows that the rule order presentation is really important for the model quality.

5 Conclusion

The aim of the paper was to discuss some aspects of the fuzzy expert modeling. The main issue touched in the paper was the influence of the methods used in the process of creating a fuzzy model on its final quality, measured as the model precision. Experiments carried out in the paper showed that while the method of creating expert membership functions had none influence on the model quality in the analyzed problem, the order of rules conclusion definition was a crucial factor, responsible for creating a reasonable or unreasonable model.

References

- [Aczel 1993] Aczel, A.D.: Complete business statistics. Richard D. Irwin Inc., Sydney (1993)
- [Asuncion and Newman 2007] Asuncion, A., Newman, D.J.: UCI Machine Learning Repository. University of California, School of Information and Computer Science, Irvine, CA (2007), <http://www.ics.uci.edu/~mllearn/MLRepository.html>
- [Baetge and Heitmann 2000] Baetge, J., Heitmann, C.: Creating a fuzzy rule-based indicator for the review of credit standing. *Schmalenbach Business Review* 52, 318–343 (2000)
- [Chen and Linkens 2004] Chen, M., Linkens, D.A.: Rule-base self-generation and simplification for data-driven fuzzy models. *Fuzzy Sets and Systems* 142 (2004)
- [Facchinetti and Mastroleo 2005] Facchinetti, G., Mastroleo, G.: A fuzzy way to evaluate the qualitative attributes in bank lending creditworthiness. In: Saeed, K., Pejaš, J. (eds.) *Information Processing and Security Systems*. Springer, Heidelberg (2005)
- [Jaškowski 2009] Jaškowski, P.: *Cognitive neuroscience. How the brain creates the mind*. Vizja Press&It, Poland (2009) (in Polish)
- [Piegat 1999] Piegat, A.: *Fuzzy modeling and control*. Physica-Verlag, New York (1999)
- [Rejer 2006] Rejer, I.: *Integration of knowledge sources in fuzzy models of economic dependencies*. Scientific Publishing House of Szczecin University (2006) (in Polish)
- [Rutkowska et al. 1999] Rutkowska, D., Rutkowski, P.M.: *Neural networks, genetic algorithms and fuzzy systems*. Scientific Publishing House Ltd., Warsaw (1999) (in Polish)
- [Yager and Dimitar 1994] Yager, R.R., Dimitar, P.E.: *Essentials of fuzzy modeling and control*. John Wiley & Sons, Inc, Chichester (1994)

Database Access and Management with the Use of the MOODLE Platform

K. Hareźlak and A. Werner

Silesian University of Technology, Gliwice, Poland
{katarzyna.harezlak, aleksandra.werner}@polsl.pl

Abstract. The possibility of using MOODLE e-learning platform as an environment for teaching database issues, concerning database objects and users creating, analysis and management of a transaction has been analyzed in the paper. All of the needed extensions were made using PHP scripts and were tested in the chosen database servers: MySQL, SQL Server and Oracle DBMS. Results of these tests were satisfactory, confirming preliminary assumptions.

1 Introduction

The usage of e-learning platforms has become more and more popular throughout the recent years. They allow teachers to build effective online courses and to manage learning on the web. Owing to them, faster knowledge access and learning costs reduction have been achieved. There are many mechanisms of such platforms, allowing theory presentation and verification of the knowledge gained by trainees, in a simple way [Bylina et al. 2008; Gumińska and Madejski 2007]. Nevertheless, in some knowledge branches training plays much more important role than theory. Especially such database issues, as: database designing, querying and management, require intensive practice. Analyzing various database problems, SQL (Structured Query Language) query training seems to be the most important challenge in e-learning database knowledge teaching. If the necessity of practicing in real database environment – preferably environment, in which course beneficiaries will work in future – is taken into account, this challenge becomes a critical in a database e-learning.

As far as fundamentals of SQL are concerned, they can be presented on an e-learning platform in simple html webs, while quizzes, lessons or chats might be used for knowledge verification. But still there is a problem how to provide students with online connection to a database server in order to practise SQL. In many cases, the only applicable solution is requesting students to install chosen database server in their private computers, creating a database and load data,

available on html page. Actually, the guidelines for these tasks can be provided via e-learning platform, but configuration of database environment must be performed by students on their own. If we take into consideration that a group of course participants has just started studying database problems, this process can be too difficult for them.

In [Hareźlak and Werner 2010] an innovative approach to tackling this problem was proposed, where several students databases were replaced by one database managed by a teacher (Fig. 1). Besides, the new activity mechanism – SQL – of e-learning platform was introduced, which made a database interactive querying possible.

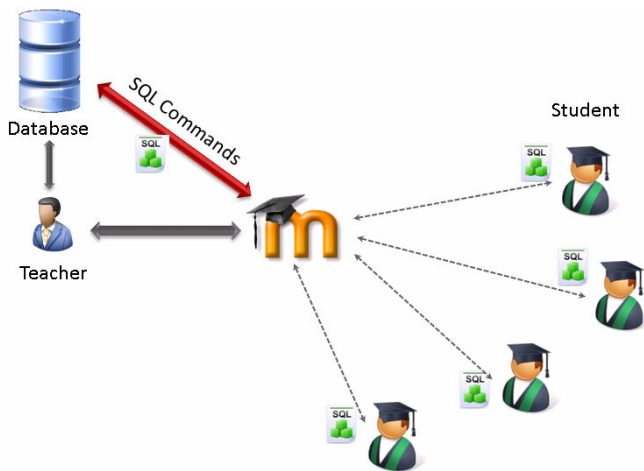


Fig. 1 The idea of the implemented architecture in an e-learning environment

The opportunity of connecting to a database server via software for collaborative learning, removes mentioned inconveniences – it means: exempt students from the obligation of database environment self-configuration. The next advantage is the centralization of the process of preparing a database for various tasks realization. If a teacher has to change database schema or structure, it can be done in one, centralized database and it will not have to be immediately propagated to students local databases.

For new database activities implementation, the MOODLE e-learning platform was chosen. Extension of its functionality was possible owing to open-source software and specific modules structure, and was resolved by creating and configuring a new component, very similar to already existing ones (in [Hareźlak and Werner 2010] the composition of each built-in MOODLE module was described in detail). Because each module (for example: quiz, glossary, forum) has its own folder containing collections of analogous files, the process of preparing a new activity in practice involves the creation of a new folder, named the same as the title of new

activity, in which all needed files are collected. The most important files, in context of the research, are: *mod_form.php* and *view.php*, which are responsible for displaying – respectively – form to teacher, where task conditions are defined and the window for students, which enables them to execute designed task.

2 SQL Modules

Taking the SQL subtypes and the database e-learning needs into account, new activities – such, as: SQL DCL, SQL DDL and SQL DML were required to be implemented. Additionally, the SQL ACID and SQL TSQL were also realized, in order to allow more advanced users to practice more complex database issues (Fig. 2).

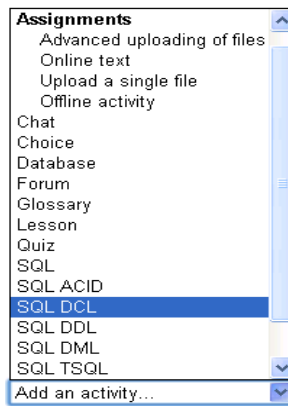


Fig. 2 List of activities connected with a given course

For this purpose, a set of new modules, destined for every task specified above, was coded. Although they differ in functionality and require separate programming, all of them define the parameters for database server communication in *mod_form.php* file.



Fig. 3 List of available database servers

Because of the fact that it was assumed that central database engine may vary depending on the teacher needs (Fig. 3), the first step of changing the MOODLE source was activating needed PHP language extensions, by clicking on WAMP server tray icon – Fig. 4.

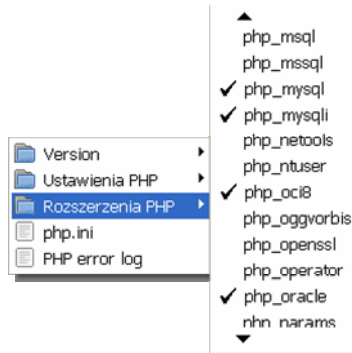


Fig. 4 List of accessible PHP extensions

The subsequent stage was connected with implementing desired functionality in PHP scripts, simultaneously with taking into account the specific characteristics of the chosen database engines (for example database backup and restore methods).

2.1 SQL

The first module, SQL, was designed to provide an environment for the realization of SELECT...FROM...WHERE... queries. As a main database server MySQL – one of components cooperating with the MOODLE platform – was chosen [Welling and Thomson 2005], but proposed solutions were also tested in other database servers like Oracle, MS SQL Server and IBM DB2. In order to enable connection to Database Management System (DBMS), a few teacher-depend parameters (database host, name, user and password) had to be set (Fig. 5). The methods of connection establishing and result set displaying were presented in details in the paper [Hareźlak and Werner 2010].

Despite the fact that simple database querying is the mostly performed action in a database, it isn't the only task realized in a database [Garcia-Molina and al. 2003]. To have a possibility of data querying, first of all, tables, which will store records of data, must be created. Subsequently, these tables must be filled with records. Besides, very often database users face the problem of records modification and databases protection against software and hardware failure, by backup execution.

Analyzing the limitations, resulting from the form of training, an underlying inconvenience was found, directly associated with the nature of the tasks performed by students. If the number of participants of the course is not settled beforehand, executing the commands in the same database can lead to some undesirable effects. Among them we can enumerate: the possibility of objects names duplication and uncontrolled objects removal, deadlock, hindering or making given tasks impossible to perform.

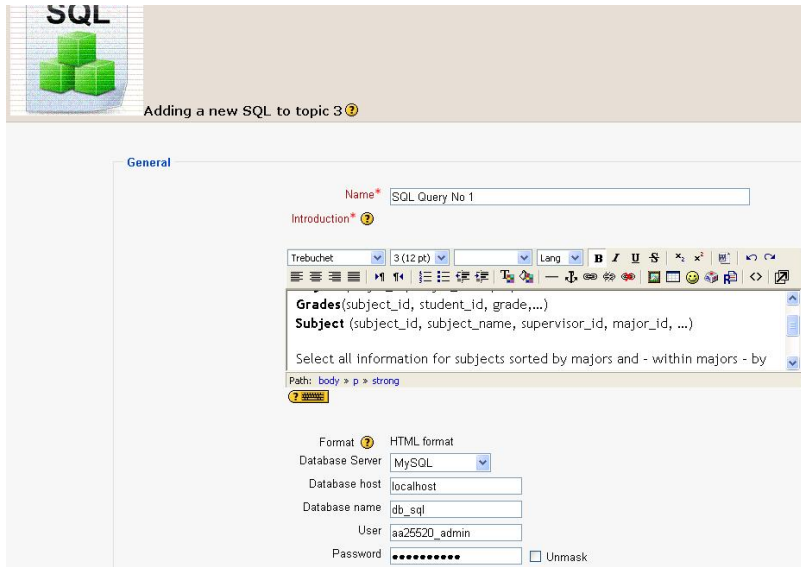


Fig. 5 View of SQL activity edition window

To avoid such problems, the mechanisms for creation a copy of a processing database (*Source Database*) for every student were implemented. The name of the database created for a particular user is constructed dynamically according to the pattern:

$$\text{SourceDatabaseName} + \text{UserId} + \text{ModuleId}.$$

Student's source database is specified in *Source Database Path* and *Database Path* parameters of teacher's window. This strategy provides a unique database naming and allows unlimited number of students to perform an SQL command. So, each course member works on his own copy of the source database.

Additionally, owing to this, the teacher obtains control over the activities of a particular student. Trainer gains a possibility for the detailed analysis of tasks performed by a student or for their evaluation.

Implementation of the source database duplication assumes, that after clicking a task of a given module, only its content and button **Create Database** are initially visible. Choosing this element results in starting of database creation process, that consists of the template database backup and retrieving the contents of its copy to the appropriate, new student's database. This is possible by an activity window configuration, where teacher fills the fields of database path, server, name and user name with enough privileges to connect to the server. These information are forwarded to the php code performing the task.

Right after that, the form for SQL query performing appears (Create Database button became inactive) – Fig. 6.

Fig. 6 Final view of the DCL task realization window

Presented mechanisms were used in every created module, described later.

The possibility of verification of the correctness of the tasks performed by student was achieved through creation of additional tables on the server. They store such information as: module identifier, sent query content, needed timestamps, whether a copy of the module has already been executed, etc. The names of these tables (Fig. 7) were constructed in the following way:

- for commands being the tasks solutions:
 - a. *DatabasePrefix_SQLSubsetName*,
 - b. *DatabasePrefix_SQLSubsetName_answare*,
- for source database copies created by students:
 - c. *DatabasePrefix_SQLSubsetName_load*,

where:

- *DatabasePrefix* means the database prefix defined in the MOODLE configuration file `config.php`,
- *SQLSubsetName* is a name of the given module.

Table Name	Collation	Engine
<input type="checkbox"/> ixcr_db_load		InnoDB
<input type="checkbox"/> ixcr_dcl		InnoDB
<input type="checkbox"/> ixcr_dcl_answare		InnoDB
<input type="checkbox"/> ixcr_dcl_load		InnoDB
<input type="checkbox"/> ixcr_ddl		InnoDB
<input type="checkbox"/> ixcr_ddl_answare		InnoDB
<input type="checkbox"/> ixcr_ddl_load		InnoDB
<input type="checkbox"/> ixcr_dmi		InnoDB
<input type="checkbox"/> ixcr_dmi_answare		InnoDB
<input checked="" type="checkbox"/> ixcr_dmi_load		InnoDB

Fig. 7 List of supplementary tables stored in the database of the MOODLE platform

For example, after pressing the Create Database button in the DML module window, the table *DatabasePrefix_DML_load* is updated by the following code in the file *view.php*:

```
switch ($_POST['bsubmit']){
case 'Create Database':
$dataobject = new Stdclass;
$dataobject->who_load_db=$USER->id;
$dataobject->is_db_load = 1;
$dataobject->module_id=$dml->id;
$dataobject->db_name=$dml->dbname;
insert_record('dml_load', $dataobject);
}
```

These settings mean that a database copy was created (value 1 in field *is_db_load*) for the module (field *module_id*) by the user, whose ID is stored in the *who_load_db* field.

In addition, the function, checking whether the user has already created his own copy of the database in the appropriate module, was written.

If a query in the general form:

```
$query = 'SELECT * FROM
DatabasePrefix_SQLSubsetName_load
WHERE who_load_db = '.$USER->id. ' and modul_id = '.$o;
```

returns a row of data, pressing the Create Database does not take any action. Such protection ensures that creating private copy of a database for each participant is performed only once. This applies for the entire set of queries using the given database source.

2.2 SQL DDL and SQL DML Modules

Each of the designed database e-course modules corresponds to specific group of database issues and requires knowledge of selected category of SQL statements. In particular, module M04 (Fig. 8) explains the basics of the Data Definition Language used to build and modify the structure of tables and other objects in the database.

M04 - Database Objects Creation

M04 is dedicated to students who want to learn relation data definition language (DDL) and ensure accuracy and consistency of data in a relational database. In this module, students gain knowledge about another important concept of a relational database – virtual tables (i.e. views) as well.

Fig. 8 Link to DDL module in designed database course

Process of the DDL module creation was influenced by a modular construction of the e-learning platform. It consisted of creating a new DDL folder, with its standard components, and placing it in the *mod* subfolder of the MOODLE project home directory. Next, the *mod_form.php* and *view.php* files were adapted to needed functionality. In the first of the mentioned files, the definitions of the fields that allow the teacher to enter the relevant parameters of the database during the activity edition were specified. The code below:


```

$mform =& $this->_form;

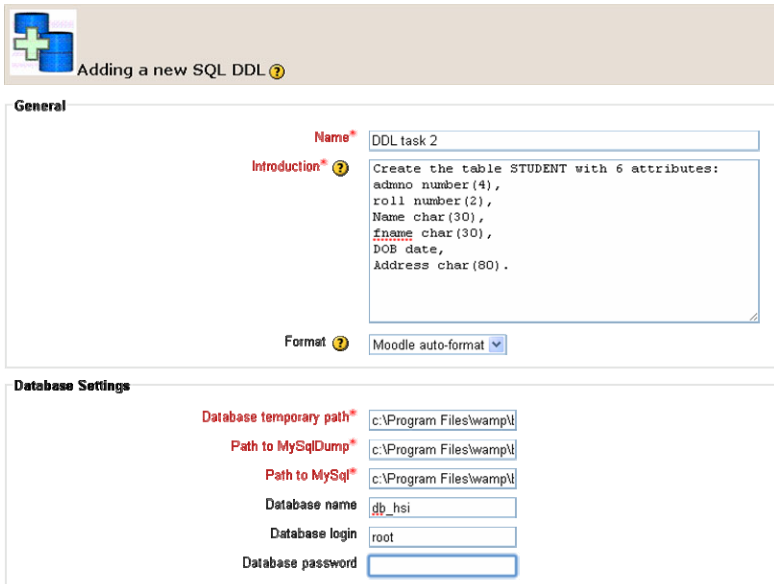
// Adding the "general" fieldset, where all the common settings are showed
$mform->addElement('header', 'general', get_string('general', 'form'));
...
// Adding the rest of ddl settings
...
$mform->addElement('text', 'mysqldump', get_string('mysqldump', 'ddl'));
$mform->setType('mysqldump', PARAM_TEXT);
$mform->addRule('mysqldump', null, 'required', null, 'client');
$mform->addRule('mysqldump', get_string('maximumchars', 255), 'maxlength', 255, 'client');

$mform->addElement('text', 'mysql', get_string('mysql', 'ddl'));
$mform->setType('mysql', PARAM_TEXT);
$mform->addRule('mysql', null, 'required', null, 'client');
$mform->addRule('mysql', get_string('maximumchars', " , 255), 'maxlength', 255, 'client');

$mform->addElement('text', 'dbname', get_string('dbname', 'ddl'));
$mform->addElement('text', 'dblogin', get_string('dblogin', 'ddl'));
$mform->addElement('text', 'dbpassword', get_string('dbpass', 'ddl'));
...

```

corresponds to the teacher's form presented on Fig. 9.



Adding a new SQL DDL

General

Name*

Introduction*

Format

Database Settings

Database temporary path*

Path to MySqDump*

Path to MySql*

Database name

Database login

Database password

Fig. 9 List of activities connected with a given course

The second file, *view.php*, contains code that:

- allows to perform DDL commands,
- manages the content of the window that student executes his tasks through.

The designed module was used – inter alia – to examine the relationships of the primary key. These include the following steps:

1. Create table:

```
CREATE TABLE Employee (emp_no int, emp_name text);
```

2. Define primary key in a table:

```
ALTER TABLE Employee ADD PRIMARY KEY (emp_no);
```

3. Insert 3 rows into a table:

```
INSERT INTO Employee VALUES (1,'Wolf'), (2,'Barry'), (3,'Smith');
```

4. Insert 3 subsequent rows into a table, but one of the key column value should be repeated (should be the same as the value previously specified):

```
INSERT INTO Employee VALUES (4,'Jackstraw'), (5,'Terry'), (3,'Kelly');
```

Message shown in the Fig. 10 allows to state that the task is not performed correctly, because it violated the integrity constraints. Besides, analysis of the results returned by the query `SELECT * FROM Employee` will show, that all changes made in point 4 have been rolled back since they were treated as a one indivisible unit – e.g. a transaction.

```
Error :
Duplicate entry '3' for key 1,
Query: :insert into Employee values (4, 'Jackstraw'), (5, 'Terry'), (3, 'Kelly')
```

Fig. 10 Error report

2.3 SQL DCL Module

Like other modules, SQL DCL activity has its own, untypical – for developed components – functionality, that had to be implemented there.

The reason for this is that the group of SQL keywords, practised in the module, handle the authorization aspects of data and permits the user to control who has access to see or manipulate data within the database. So, because of the fact that every user in the database is especially allowed to access specific data objects or to perform some of system level activities, there should be a possibility to check whether grantee has really gained privileges (or group of privileges). For example, after creating a new user and granting him chosen object privilege, students should

have the chance to connect to the database as a new user and try to execute an operation on a granted object.

Thus, in DCL module, the *Login* and *Password* text areas appear in students' window, in order to make the verification, whether the effect of written commands is in accordance with student's supposition, possible (Fig. 11).

Fig. 11 The final view of the DCL task realization window

For this purpose the button *Change User* was coded in *view.php* file:

```
<input type="button" name="changeuser"
  onclick="logOn(thisForm.login,thisForm.password)" value="Change User"/>
```

and special function *logOn* was implemented in a script:

```
function logOn(txtarea1,txtarea2) {
  var login=(txtarea1.value);
  var pass=(txtarea2.value);

  <?php echo("firstVar = $cm->id;");?>
  window.location.href = "view.php?id="+firstVar+"&login="+login+"&pass="+pass;
}
```

After re-logging, an empty area that allows entering SQL commands is displayed and, after executing a DCL statement, result or appropriate message is shown (Fig. 12).

```
Error :
UPDATE command denied to user 'emp1'@'localhost' for table 'sales'
Query: :UPDATE sales
      SET sales_person='smith ann'
      WHERE region='north warsaw'
```

Fig. 12 Exemplary error message

3 Transaction Realization

Using the database, it must be remembered that in many cases, commands sent to it must be treated as a transaction, characterized by the following properties: **Atomicity, Consistency, Isolation, Durability (ACID)**. As a part of their training, students are acquainted with knowledge how to define a scope of transaction and how the database server ensures its properties. Research in this field included the solution of two problems.

The first one concerned the evaluation of the possibilities and methods of the SQL commands execution with usage of a transaction. The second was related to a schedule of the transaction realization using the locks and defined levels of isolation.

A. Transaction Atomicity

To meet the demands of the first issue, the environment was prepared to test the ACID properties of transactions. For this purpose a series of experiments was performed.

For example, the command to transfer the fees for studies from student's account (account_ID=100 and bank_ID =1) to the account of the university (account_ID=200 and bank_ID=2) was analyzed. Two different situations were examined:

1. Commitment of the transaction.

```
BEGIN TRANSACTION;
UPDATE account SET stan = stan - 500
    WHERE account_ID = 100 AND
bank_ID = 1;
UPDATE account SET stan = stan+100
    WHERE account_ID = 200 AND
bank_ID = 2;
COMMIT;
```

2. The transaction roll backing.

```
BEGIN TRANSACTION;
UPDATE account SET stan = stan - 500
    WHERE account_ID = 100 AND
bank_ID = 1;
UPDATE account SET stan = stan+100
    WHERE account_ID = 200 AND
bank_ID = 2;
ROLLBACK;
```

Experiments carried out on the MOODLE platform, in the new ACID module, confirm the possibility of explicit definition of the transaction scope by the START/BEGIN TRANSACTION command and the COMMIT or ROLLBACK. This allows demonstrating students the existence of transaction atomicity property.

The second option for defining the transaction scope is to use PHP script commands to start and to end the transactions implicitly. [Swan 2009]. In order to do that, the code of view.php file should be supplemented with appropriate instructions:

```

create a connection variable;
if (connection established){
    Create the variable that holds SQL statements;
    Start a transaction;
    Parse the statements in the context of a database connection;
    Execute the statements;
    Recognize mistake and rollback changes
    while (there are rows in the result set){
        get the number of columns used in SQL statement;
        for ($i is less than number of columns){
            return column value for fetched row;
        }
    }
}
else
    print an error message and exit the program;
Free up the resources used to perform the query;
End the transaction;
Close the database connection;

```

It should be noted that commands realized without explicit definition of the transaction can be treated as an independent one [Hareźlak and Werner 2010]. This is tantamount to a default setting of AUTOCOMMIT option during a connection to the database. This results in an automatic commission of each DML command by the server. Taking the possible models of the database engines activity into consideration, the tests of changing these settings by the command SET AUTOCOMMIT OFF were examined as well.

B. Isolation Levels – Consistency and Isolation

During works on the MOODLE extension, much attention was paid to the issues related to the organization of access rights to resources (command LOCK/UNLOCK) and the subsequent transaction properties – consistency, and isolation.

The need to ensure that commands of two concurrent transactions operating on the same data sets, resulting in equipping the developed module with two windows. Owing to this, the possibility of transactions execution in two concurrent sessions was achieved (Fig. 13).

During tasks realization, students are familiarized with the names of the isolation levels and syntax of commands setting. This solution also allows to observe the effects of the undesirable phenomena accompanying the particular isolation level. Among these there can be found: dirty read or deadlock. The example of such a task is shown in the Fig. 13.

SQL TSQL Task1

Describe database server behaviour in points X, Y in READ UNCOMMITTED isolation level.
Write statements for different transactions A and B in two separate windows.

1A. start transaction;
1B. start transaction;
2A. select * from BOOK;
(X)
2B. update BOOK set
3B. title='Iks' where author='Smith';
(Y)
3A. select * from BOOK;
4A. rollback;
4B. rollback;

Database was created

Enter SQL

```
SET ISOLATION LEVEL READ UNCOMMITTED;
START TRANSACTION;
SELECT * FROM book;
```

```
SET ISOLATION LEVEL READ UNCOMMITTED;
START TRANSACTION;
UPDATE book SET title='Iks' where author='Smith';
```

Fig. 13 Window of a task realization in SQL TSQL module

C. Durability Effect

Another challenge to be faced was to demonstrate and clarify the mechanisms for ensuring durability of transactions. The set of objects used by a server to protect a database from system failure consists of databases copies and transaction logs. Therefore, in the ACID module the functionality of a database backup via an interactive window was implemented. This copy is subsequently used for restoring captured database.

At the present stage of system development, when working with MySQL server, students have two possibilities for database backup. One of them allows remembering the data stored in the table, in an external file, by executing the command `SELECT * INTO OUTFILE 'OutputFileName' FROM TableName`. The second one gives the possibility of backup all the tables with their data, by pressing Create Backup.

Database Type

Choose Database type

MySQL Dump Settings

Path to MySqlDump
 Path to MySql

Fig. 14 Parameters needed for students' database backup

This button reads the needed parameters from the form displayed by *mod_form.php* file (Fig. 14) and calls *mysqldump* program:

```
$rev = system($dml->mysqldump.
'-u '.$CFG->dbuser.
'-p'.$CFG->dbpass.
' '.$dml->dbname.
'>'.$dml->path_for_user_db1.'.temp.sql');
```

In other database system the DDL command, like BACKUP DATABASE, can be used instead.

4 Database Programming

Participants, who learned the rules of SQL queries constructing and acquired the abilities connected with data manipulation are ready to get to know the aim and the rules of database programming. Each database server is equipped with its own programming language, mainly based on SQL commands, enhanced with variables declaration, commands managing the flow or defining the loops. These languages are used to implement database stored procedures and triggers, both of which play a crucial role in databases. Stored procedures allow to place the business logic on the server and they are a very important part of database protection, hiding its structure from the user. The second of the programmed object types, triggers, enable an active database reaction to the changes performed within its records.

Utility of the discussed objects decided about introducing to the platform a new module, facilitating their study. Owing to the fact that database procedures are an object of a particular database, in this module there were used mechanisms, developed beforehand, separating working areas of particular course participants. Additionally, a requirement concerning student's preparation for the tasks he is to face was introduced. This is the result of the fact that to be able to design and implement database procedures properly, one needs to be familiar with SQL, or even DDL, commands. Therefore, if a student starts performing the tasks from the *Database programming* module, he needs to have SQL, DML and DLL modules already passed. Making such a verification is possible owing to system tables extension with table *prefix_promote*. Records of this table consisting of *user_id*, *SQL_points*, *DCL_points*, *DML_points*, *DDL_points* columns, collect points for the tasks solved correctly. Zero appearing in at least one of the columns in this table or lack of a record for a specified student means that he has not yet realized all the needed tasks and is not prepared for training issues from TSQL module. The situation when a course participant has already acquired essential knowledge and has omitted particular course modules is also accepted. For such students there is a test prepared, consisting of interactively realized operations of data selection and

modification, and of creating database objects. In such cases TSQL module form opens with a new button *test yourself* and deactivated *create database* (Fig. 15).

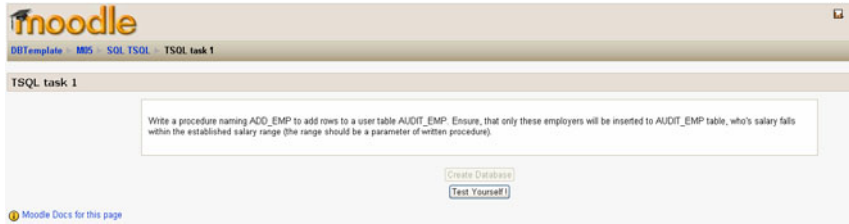


Fig. 15 Form displayed for students with no credit

After choosing the first one, next form containing the test is being opened (Fig. 16).

The test is divided into groups, in accordance with SQL, DCL, DDL, DML modules, represented by adequate buttons.

During the test it is accepted to make one mistake in each of the test tasks. Exceeding this number of mistakes results in being excluded from a particular part of the classes. Performing the previous SQL, DML and DDL tasks can change this situation.

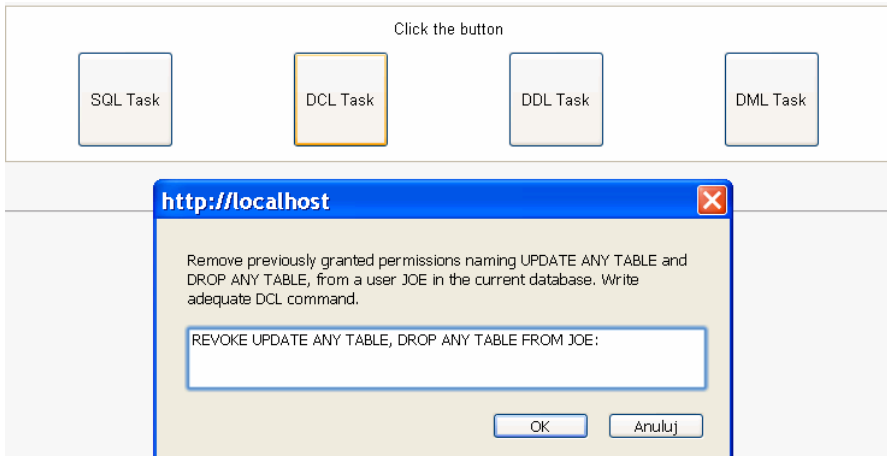


Fig. 16 DCL task complete window

Successfully finished part of the test is signaled by appropriate message and deactivation of a given button (Fig. 17).

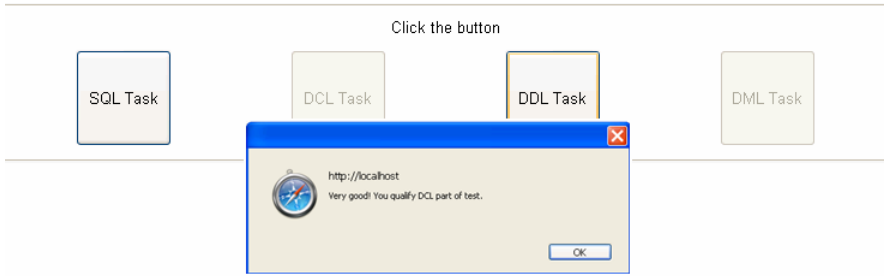


Fig. 17 The result of successfully completed DCL test

The code performing these functions is included in *view_php* file and consists of steps presented below:

```
function SolveDCL(){
    var commandTxt = prompt ("task content", "");
    if (commandTxt is correct){
        //update column DCL_points in prefix_promote table
        <?php
            $answers=mysql_query(
                'UPDATE '.$CFG->prefix.'promote p
                SET DCL_point = DCL_point + 1
                WHERE p.user_id='.$USER->id
            );
            alert ("Very good! You qualify DCL part of test.");
            document.getElementById('btn2').setAttribute('disabled', 'disabled');
        }
    else
        alert("Wrong answer! Try again later.");
}
```

When all missing parts are accepted, *create database* button is activated and student can continue his work by typing his own procedure or by loading it from a file (Fig. 18).

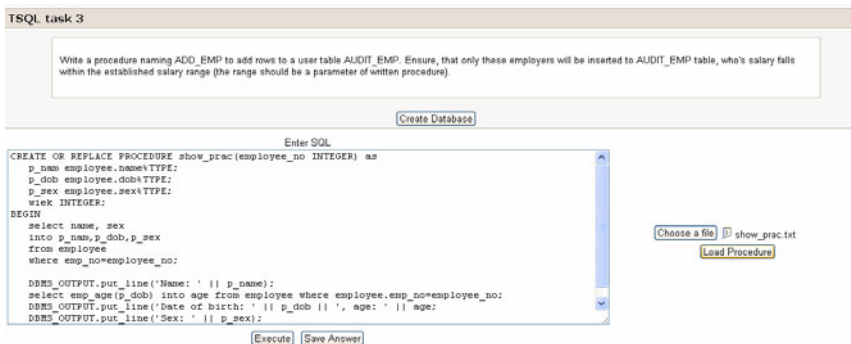


Fig. 18 TSQL window with *Load Procedure* button

5 Conclusions

In the paper the possibility of using e-learning platform as an environment for education of database issues has been analyzed. Among them, basic database objects and users creation, analysis and management of a transaction were considered. For this purpose appropriate modules: DDL, DCL – for the first tasks and DML with ACID for second one, were created. Besides, the database programming has also been discussed.

The architecture of all modules is similar but their functionality enforced introducing of specific solutions suitable for tasks that modules should implement. For example DML, DDL and DCL ones include database creation for every user of the platform to ensure independency in his task realization. The ACID module, on the other hand, possesses complex functionality connected with transaction management.

Proposed functionalities were verified by the group of 12 students of Computer Science Faculty.

During the tests, the correctness and efficiency of queries execution via MODDLE platform were compared with results achieved for the same group of tasks performed on a local and remote database servers.

Results obtained in all of the mentioned environments were comparable. Response time was the shortest for local connections. Time delay achieved in case of both remote connections didn't have any influence on the comfort of work and wasn't significant.

All of the extensions were made using php scripts and were tested in the chosen database servers: MySQL, MS SQL Server and Oracle DBMS.

Results of all tests were satisfactory, confirming preliminary assumptions for using e-learning MOODLE platform in the interactive teaching of database issues.

References

- [Bylina at al. 2008] Bylina, B., Walczyński, T., Bylina, J.: The review of e-learning operation basing upon an activity of students and lecturers. *Advanced problems of Internet Technologies*. Academy of Business in Dąbrowa Górnicza (2008)
- [Garcia-Molina and al. 2003] Garcia-Molina, H., Ullman, J.D., Widom, J.: *Database system implementation*. WNT, Warszawa (2003) (in Polish)
- [Gumińska and Madejski 2007] Gumińska, M., Madejski, J.: Web based e-learning platform as a source of the personalised teaching materials. *Journal of Achievements in Materials and Manufacturing Engineering* 24 (2009), http://www.journalamme.org/papers_vol124_2/24251.pdf (2007)
- [Harezlak and Werner 2010] Harezlak, K., Werner, A.: E-learning database course with usage of interactive database querying. In: *Internet – Technical Development and Applications Series*. AISC, vol. 64 (2004); ISBN: 978-3-642-05018-3
- [Swan 2009] Swan, B.: *Acquiring access to SQL Server databases using PHP* (2009), http://www.microsoft.com/poland/technet/bazawiedzy/centrumrozwiazan/cr360_01.msp (in Polish)
- [Welling and Thomson 2005] Welling, L., Thomson, L.: *PHP and MySQL web development*, 3rd edn. Helion, Poland (2005) (in Polish)

Human Activity Supporting by Deontological Knowledgebases

J.L. Kulikowski

Polish Academy of Sciences, M. Nalecz Institute of Biocybernetics and Biomedical Engineering, 4 Ks. Trojdena Str., 02-109 Warsaw, Poland
juliusz.kulikowski@ibib.waw.pl

Abstract. A concept of computer systems supporting human activity by recommendations for actions leading to desired intermediate and final goals is here presented. The recommendations are given in a standard form of deontological statements specifying the current states, the desired next states and operations transforming the current into the desired states. The quality of recommended actions is by parameters describing various favorable and unfavorable effects assessed. The quality parameters are presented as vectors in semi-ordered linear (Kantorovitsch) space where preferences between vectors is induced by definition of a positive vectors' cone. It is shown how the positive cone by systems of linear inequalities (without constant terms) can be defined so as to establish a compromise between different quality parameters in order to chose the most preferable actions. Two methods of the above-mentioned methods extension on the paths of actions comparative quality assessment are proposed. The presented basic concepts by simple examples are illustrated.

1 Introduction

Human activity can formally be described by networks of logically connected actions. For this purpose various, on graph theory [Baker and Eris 1964], probabilistic theory [Jensen 2000], formal linguistics [Rusinkiewicz and Bregolin 1994], logical [Nowakowska 1979] and other methods based approaches can be used. However, in activity description several additional aspects of action networks like: 1st multi-scaling, 2nd fuzziness, 3rd time-dependence,- should be taken into consideration. *Multi-scaling* means that any activity network can on one hand be considered as a part of a higher-level activity network and, on the other one, its components (actions) can be displayed into some lower-level structures consisting of sequences of the following steps: real situation evaluation, desired situation awareness, specification of possible operations changing the given situation, admissible operations comparative assessment, choosing operation for realization, and finally – execution of the chosen operation. Such a sequence of steps, shown in Fig. 1, will be called below an *action cycle (AC)*.

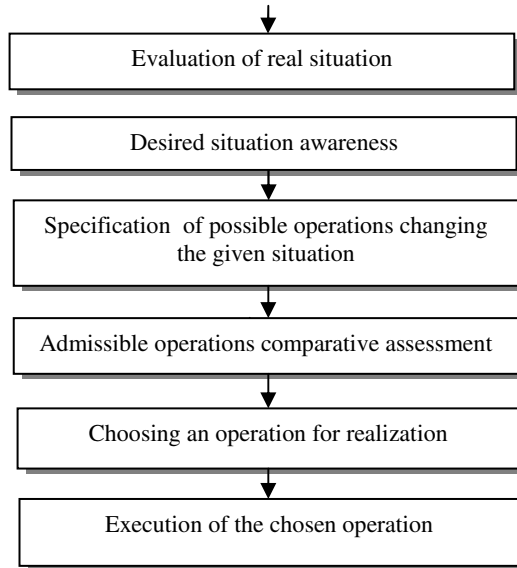


Fig. 1 General scheme of a selected action cycle (AC)

Fuzziness of action networks means that some of the steps of its action cycles (e.g. actual situation evaluation, specification of possible situation changing operations, comparative operations assessment) are based on uncertain, out-of-date or incomplete information. At last, time-dependence of activity networks means that as the time is going on, its planned actions should be adjusted to the changing external situation. In multiple application areas the ACs can be supported by computer systems. Such a support for several reasons may be desirable: 1st knowledgebases may contain more complete and better founded information about the positive, neutral and/or negative effects of the actions, 2nd computer-based relative assessment of various aspects of actions may be more effective than this on human intuition based one, 3rd computer-based decisions are more than the intuitive ones reliable in the sense that if in the same situation repeated they remain the same. And still, the computer systems should play an advising rather than deciding role, because human goals selection criteria between and even within the ACs may be changed and very often they are not exactly defined. Moreover, for the last decades, the attempts to make the computer decision making similar to the human one, including its flexibility and capability to deal with incomplete and/or uncertain information, have been strengthened. For this purpose various approaches, on Bayesian networks [Jensen 2000], possibility theory [Kłopotek 1994], fuzzy sets [Dubois and Prade 1991], rough sets [Pawlak 1991], non-classical logics [Bolc et al. 1998], artificial neural networks [Hecht-Nielsen 1991], etc. based have been proposed and

intensively investigated. On an opposite pole, approaches based on negotiations theory, games theory, multi-agent systems [Barthelemy and Janowitz 1991], [Ferber 1999], etc., assuming decision reaching in groups of many collaborating subjects also should be mentioned. So, the traditional decision making concepts based on strong [19] or on multi-criteria optimization, which since the 40ths up to 90ths years of the 20th century were dominating, excepting some specific application domains, to a second plan have been moved. This was caused by the fact that in practice decisions are mostly in long multi-ACs processes made in which the effects of any faults can iteratively be reduced. Otherwise saying, imperfect but quick decisions at a long are better than the perfect but delayed ones. This should be taken into account in any computer-based decision-aiding systems design.

The aim of this paper is presentation of an approach to decision making aided by information drawn from a specific kind, called *deontological* (*gr. déon = duty, what should be done*) knowledgebases. The idea of using deontological statements in decision making was originally presented in [Kulikowski 2006] while in [Kulikowski 2009] the role of deontological statements in knowledge representation has been described. A typical deontological statement in natural language has the form of a recommendation:

If A then do B in order to reach C,

where *A* is an assertive statements describing a current situation, *C* is an assertive statement describing a desired situation and *B* describes recommendation for undertaking action transforming situation *A* into situation *C*. Apparently, there is no difference between deontological and implicative (*If A and B then C*) statements. In fact, to deontological statements as to recommendations no logical values (e.g. *true – false*) but rather some *utility weights* should be assigned. Moreover, at least three aspects by the utility weight should be expressed: 1st real possibility of undertaking *B* in some given circumstances, 2nd effectiveness of transforming *A* into *C* if *B* has been undertaken, and 3rd additional (side) effects of undertaking *B* in the situation *A*. It also may happen that *C* is not a final but rather an intermediate goal in a long-term sequence of actions. At last, the above-mentioned aspects of actions cannot be evaluated but on an uncertainty level. That is why inference based on deontological statements cannot be realized as a logical implication of typical assertive statements.

We call a *deontological knowledgebase (DKB)* a collection of deontological statements organized for actions planning in a certain application domain. Selection from a *DKB* logically consistent deontological statements, their utility evaluation and realization of the *ACs* recommended by the selected deontological statements is in fact a core of any computer-aided human action. The *DKB*-based decision making systems belong to a large class of case-based decision systems. The quality (effectiveness, etc.) of single decisions made by such systems is usually lower than this of decisions based on accurately a given reality describing

(e.g., by functional equations, probability distributions, etc.) models and on adequate optimization methods. However, the problem is that such accurate models together with the corresponding input data are usually not available. A *DKB* reminds a cookbook providing recipes of dishes preparation without explanation of chemical backgrounds of cooking processes. Unlike the cookbook, a *DKB*-based decision system should also provide information about preferred dishes composition in a meal, costs of their ingredients, dietetic values and/or contraindications, etc. In exploration of such systems three basic tasks thus should be solved: 1st retrieval in the *DKB* deontological statements potentially useful to a given decision problem solution, 2nd composition of *ACs* suitable to decision problem solution on the basis of retrieved deontological statements, 3rd construction of paths of *ACs* leading from initial to final (desired) state, evaluation of their effectiveness and selection of the most preferable ones. The effects of actions undertaken within the selected path after each *AC*'s realization should be assessed and, if necessary, the plan of next actions adequately to the assessment should be modified. So, realization of a multi-step computer-supported human action needs multiple acts of information from *DKB* deriving. Let us remark that till now, no *DKB*-based decision making system organized and working exactly according the below presented concept exists. However, we are familiar with simpler, similar to them subsystems "Help" attached to numerous advanced application programs. A typical "Help" system consists of an alphabetically ordered "Index" of items and of sets of assigned to them "Recommended actions". The items correspond there to the statements *C* describing "desired situations" while the "current situations" *A* are by default established. The "recommended actions" exactly correspond to the actions *B* in the deontological statements.

According to the below-proposed approach, actions are considered as paths consisting of linearly ordered sequences of *ACs*. The *ACs* and paths of actions are characterized and evaluated by real parameters organized as vectors in a semi-ordered linear vector space (Kantorovitch space). This assumption a discrimination among the desired, undesired and conditional side effects of actions makes possible. The paper is organized as follows. In Section 2 computer representation of deontological statements in *DKB* is presented. Section 3 concerns principles of multi-aspect evaluation of *ACs*. Construction of paths of actions and their final comparative evaluation are described in Section 4. In Section 5 final conclusions are summarized.

2 Deontological Statements Representation in *DKB*

Representation of deontological statements in a natural form, as shown above, is, because of high redundancy and of synonymic and homonymic problems, highly to computer implementation inconvenient. The assertive statements *A* and *C* describing the initial states and the desired final states in a decision problem should

be in a more concise form by a formal language characterized. For this purpose, like in the case of textual documents retrieval, key words can be used. The terms used as key words from adequate to the decision problems domain ontologies [Zilli et al. 2009], if any exist, or from the lists of the most frequently used key words can be taken.

A query ordered by an user to the *DKB* is a transposition of a deontological statement and may have a standard natural form:

What (X) should be done if A takes place and C is desired?

The first problem is how the states *A* and *C* in the *DKB* should be represented in order to effectively retrieve the replies to the queries. In an advanced case some corresponding domain ontologies as sources of situations description can be used. A domain ontology is defined as a quadruple consisting of: *a/* a set of concepts describing the given domain, *b/* a taxonomy of the concepts, *c/* a set of relations among the concepts, and *d/* a set of basic assumptions concerning the relations []. So-defined ontology is a formal model of the application domain, sufficient to formulate questions concerning the states that in the given application domain may arise. In simpler cases, the states *A* and *C* by key-words can be characterized as it is illustrated below.

Example 1

Description of states are given in natural form, as below:

Starting situation (statement A):	Desired situation (statement C):
1. John is 5 th February in Warsaw;	John arrives 5 th February to London;
2. Mary caught a cold;	Mary became healed;
3. A square matrix <i>M</i> is given;	A reversed matrix M^{-1} is calculated;
4. My car engine cannot start;	My car engine has started;
5. One should employ several computer programmers;	Some computer programmers have been employed;

etc. As in typical assertive statements, in *A* and *C* subjective, predicative and objective phrases can be distinguished. A subjective phrase consists of a *subject*, i.e. a *noun* denoting an (abstract or real) object, a person, an event, a process, etc., and, possibly, some *attributes* describing quantitative or qualitative *features* of the subject expression by *adjectives* or *numerals*. Construction of an objective phrase is similar to this of a subjective phrase; however, it may occur optionally. A predicative phrase consists of a *predicate* in the form of a *verb* and (eventually) some *adverbials* of time, place, manner, concession, number, degree, etc. or in the form of a mathematical or logical operator. In order to avoid misunderstandings, in deontological statements representation by key words the last should be organized into separated by semicolons groups of key words corresponding to the subjective, predicative and objective phrases. The above-given examples of deontological statements can thus by key words can be expressed as follows:

Statement A:	Statement C:
1. N ; to be in, 5 th February, Warsaw;	N ; to arrive to, 5 th February, London;
2. N ; to catch; cold;	N ; to be, cured of;
3. M ; to be; square matrix;	M^{-1} ; to be, calculated;
4. Engine, car; to be, not starting;	Engine, car; to start up;
5. N ; offering, job; computer programmers;	Computer programmers; to be, employed; N ;

etc. N is used for *somebody* if personalization of a deontological statement is not necessary; otherwise, real proper names should be used, e.g.:

6) Director Smith will be in the office at 10.00 a.m.;	Documents are in time delivered to Director Smith;
which can be expressed as follows:	
6. Director Smith; to arrive, office, 10.00 a.m.;	Documents; to be, delivered to; Director Smith;

The recommendations for actions B may take several different forms:

- a) *Simple direct* form:
Take British Airways flight;
- b) *Composite direct* form:
Rest home *AND* take 1 tablet of aspirin 2 times/day till you feel sick;
- c) *Simple indirect* form:
Use the function *reverse matrix* in *MS Excel* program;
- d) *Composite indirect* form:
Try again *OR* follow the *car manual* *OR* call for a *traffic assistance* •

It follows from the above-given examples that simple recommendations consist of single instructions for actions while the composite ones consist of several instructions joined by logical conjunctions. Direct in instructive form given recommendations are ready to be undertaken while the indirect ones refer to documents containing detailed (usually composite) instructions for action.

3 Multi-aspect Evaluation of Simple Actions

Human actions from various points of view can be evaluated. Even if exact, e.g. physical or financial units are used, direct costs or beneficial effects can more exactly be calculated than the indirect ones. Side effects of actions usually only partially are foreseeable and in computer-aided, on strong mathematical models based, actions planning are very often neglected. The below-presented approach to the *DKB*-recommended actions evaluation is based on the assumptions that *a/* various aspects of actions' quality and/or effects by numerical parameters can be evaluated, *b/* the parameters as components of a multi-dimensional linear vector space

can be considered, $c/$ the vectors can be in Kantorovitch sense semi-ordered [Kantorovich 1959] making comparative evaluation of actions and of their linearly ordered sequences (paths) possible. The assumptions are stronger with respect to those e.g. on purely topological semi-ordering of actions for their relative assessment. Stronger assumptions follow from the fact that usually, more complete and exact meta-information about the ACs, if from a *DKB* taken, is available.

Kantorovitch space (K -space) is a linear vector space U whose elements (vectors) have been semi-ordered due to introduction of a *positive vectors* cone K^+ notion. The K -space linearity means that: 1/ a null-vector θ , 2/ a unity vector $\mathbf{1}$, 3/ sum of vectors $u' + u''$ and 4/ multiplication of vectors by real numbers $a \cdot u$ have been in K -space defined and satisfy the transposition and the distributiveness of multiplication with respect to summation rules. For any vector u the identities:

$$0 \cdot u \equiv \theta \text{ and } u + \theta \equiv u \quad (1)$$

hold. Moreover, a difference of vectors as:

$$u'' - u' \equiv u'' + (-1) \cdot u' \quad (2)$$

is defined. A positive cone K^+ is defined as an infinite convex cone containing the null vector θ as its apex. It follows from the definition that:

- a) if u is a vector such that $u \in K^+$ and a is a non-negative real number then $a u \in K^+$;
- b) if u', u'' are any vectors such that $u', u'' \in K^+$ then $u' + u'' \in K^+$.

The null vector θ is a common apex of K^+ and of the negative $K^- \equiv (-1) \cdot K^+$ cone. The definition of the positive cone does not define its exact geometrical form; its basis in multidimensional space may be spherical, ellipsoidal or even convex polyhedral. The last form will be considered below in details.

If a pair of vectors u', u'' satisfies a relation:

$$u'' - u' \in K^+ \quad (3)$$

then it is called that u' *precedes* u'' (u'' *follows* u') what shortly is denoted by $u' \prec u''$ or by $u'' \succ u'$. Otherwise, the given pair of vectors is called incomparable and denoted by $u' ? u''$. The relation of precedence \prec is reciprocal ($u \prec u$ for any u in the K -space), anti-symmetrical (if u' and u'' are different and $u' \prec u''$ then it is not $u'' \prec u'$) and transitive (if it is $u' \prec u''$ and $u'' \prec u'''$ then it follows that $u' \prec u'''$). Therefore, taking into account that not for all pairs of vectors in the K -space the relation (3) holds, we conclude that \prec is a relation of semi-ordering of vectors in the K -space. On the other hand, the incomparability (?) relation is symmetrical but neither reciprocal nor transitive. The precedence and incomparability of vectors in K -space are illustrated in Fig. 2.

Evaluation of human actions' effects. In multi-aspect human actions evaluation it is assumed that:

- a) to various aspects of ACs evaluating numerical parameters are assigned;
- b) the parameters can be considered as components of vectors in a linear vector spaces;
- c) along the sequences of consecutive ACs the evaluating parameters are cumulated;
- d) relative quality assessment of ACs is based on a semi-ordering relation in a K -space in which the positive cone K^+ adequately to the given actions planning problem has been established.

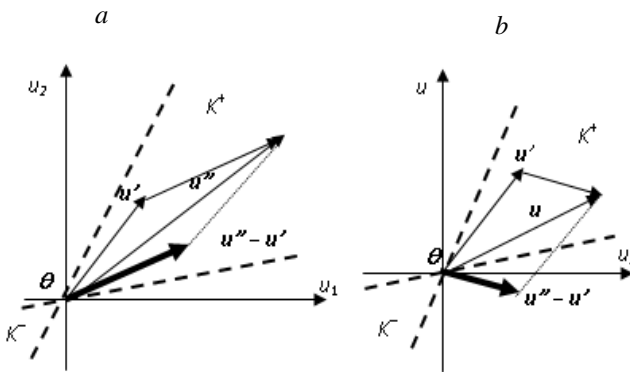


Fig. 2 Relations between vectors in K -space: $a/ u' \prec u''$, $b/ u' ? u''$

A choice of positive cone's K^+ components plays thus in actions planning problem a substantial role. For this purpose, the following groups of ACs' realization effects should be taken into consideration:

I. Favorable effects:

- direct profits of reaching the desired state;
- indirect (side) profits of reaching the desired state;
- expected progress rate in reaching a final goal, etc.

II. Unfavorable effects:

- direct costs of AC realization;
- indirect (side) costs of AC realization;
- expected risk of AC realization failure, etc.

III. Conditional effects:

- additional effects depending on the favorable effects intensities;
- additional effects depending on the unfavorable effects intensities.

The *AC* quality evaluating parameters in the groups of desired (favorable) and undesired (unfavorable) effects can be different. It will be shown below how the general *K*-space notion can be used to chose within a unified model a compromise between the desired and undesired effects of *AC*s. The following example illustrates the above-given notions.

Example 2

A medical treatment process aimed at healing osteoarthritis of the hip in a patient is planned. The following states of the patient can be specified:

I. Advanced osteoarthritis; patient to surgical intervention not ready.	II. Advanced osteoarthritis; patient ready to total heap replacement.	III. Patient after total heap replacement; osteoarthritis effects removed.
---	---	--

The planned action should thus consist of two *AC*s transforming, respectively, the states $I \rightarrow II$ and $II \rightarrow III$. For realization of the $I \rightarrow II$ transforming action several recommendations (deontological statements) from a *DKB* have been obtained:

1. *If I then rest home for a month and watch a strengthening diet in order to reach II;*
2. *If I then apply a series of 20 physiotherapeutic procedures in order to reach II;*

etc. For evaluation of the corresponding *AC*s the following effects can be taken into consideration:

- a) Favorable effects:
 - body weight reduction;
 - LDL/HDL-cholesterol rate reduction;
 - probability of final goal reaching.
- b) Unfavorable effects:
 - financial cost;
 - time delay in final goal reaching;
 - probability of desired state reaching failure.
- c) Conditional effects:
 - diastolic blood-pressure reduction;
 - inconvenience level of realization, etc.

The above-given effects on the basis of the *DKB*-provided data or of human experience can be evaluated. However, some of them should be transformed in order

to be used as K -space components. This in particular concerns the probability measures, whose values should be kept between 0 and 1 and which along the paths L of ACs shouldn't be added but multiplied. For example, the probability of reaching a final goal by realization of a sequence of ACs is given by the formula:

$$P(L) = P_0 \cdot P_{1|0} \cdot P_{2|1} \cdot \dots \cdot P_{k|k-1} \quad (4)$$

where $P_{i+1|i}$, $i = 0, 1, \dots, k-1$, denote conditional probabilities of successful reaching by a simple AC on the path L the desired next $(i+1)^{\text{st}}$ state, assuming that the i -th step has been reached. P_0 denotes the probability of reaching the initial, 0^{th} state of the path L . In this case, for a progress in final goal reaching characterization the probabilities $P_{i+1|i}$ can be replaced by *action progress rates* given by the formula:

$$h_{i+1|i} = \ln \frac{1}{1 - P_{i+1|i}} \quad (5)$$

It can be shown that $h_{i+1|i}$ takes value 0 if $P_{i+1|i} = 0$, it drives at infinity when $P_{i+1|i}$ drives at 1, and it satisfies the above-mentioned addition requirements along the path L . The corresponding reversed formula takes then the form:

$$P_{i+1|i} = 1 - e^{-h_{i+1|i}} \quad (6)$$

However, a serious constraint is connected with calculation of the *action progress rates*: the formula (5) holds only if $P_{i+1|i} < 1$. Therefore, in practice it should be chosen a small enough parameter ε , $0 < \varepsilon \ll 1$, and it should be assumed that P_0 , $P_{i+1|i} < 1 - \varepsilon$ for all $i = 0, 1, \dots, k-1$. In such case the maximum probability of reaching the final goal will be approximately $P(L) \approx 1 - (k+1) \cdot \varepsilon$ and this value should be accepted as a "practically sure" goal reaching measure. Similar approach can be used to any other limited rates using as parameters characterizing the effects accumulated along the path L of actions. E.g. the "inconvenience level of realization" can be defined as a rate:

$$f_T = \frac{T_c}{T} \quad (7)$$

where T_c denotes mean time [hours/day] necessary for realization of the given action, T – total free time {hours/day} being at the patient's disposal. In this case $0 \leq f_T \leq 1$ and it, like a probability measure, should be recalculated using the formula (5) in order to make it additive •

Other types of ACs' parameters transformation in the case of their threshold-type nonlinearities are needed. Two typical examples of such nonlinearities are shown in Fig. 3. A "dead interval" case (Fig. 3a) means that non-zero effects u occur only if original parameter's value v goes beyond a fixed interval. A "saturation" case means that the effect u to the original parameter v is proportional only if the last within a finite interval is kept, otherwise its value remains constant (Fig. 3b).

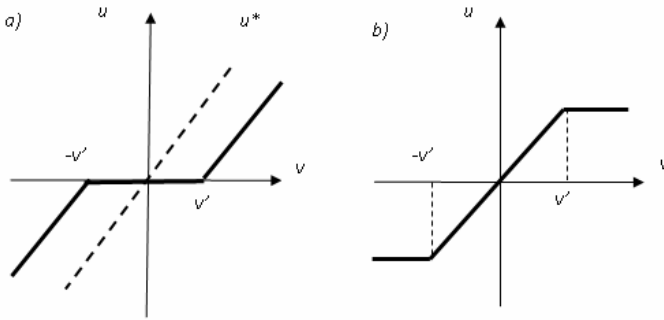


Fig. 3 Examples of threshold-type nonlinearities of ACs' parameters: a) with "dead interval", b) with saturation

$$u^* = \begin{cases} (v - v') & \text{if } v > v', \\ 0 & \text{if } -v' \leq v \leq v', \\ (v + v') & \text{if } v < -v' \end{cases} \quad (8)$$

In the saturation case, instead of the effect u the rate of its top value $u(v')$ reaching:

$$f_v = \frac{v}{v'} \quad (9)$$

(like in the "inconvenience level of realization" case) can be used and according to the formula (5) non-linearly transformed. Finally, a set of real parameters evaluating the AC realization effects and satisfying the K -space assumptions can be established.

Unfavorable effects of actions can be reformulated so as to be represented and replaced by opposite with them favorable effects. The purpose of this reformulation is characterization of ACs in a K -space by non-negative parameters only such that in general their maximization is desired. Unfavorable additive parameters like costs z , delays t etc. if their maximal values (respectively, Z , T etc.) are fixed, by their reserves $Z-z$, $T-t$, etc. can be replaced and as other favorable additive parameters can be maximized.

If $Q_{i+1|i}$ ($= 1 - P_{i+1|i}$) denotes a conditional probability of *failing* in reaching by an AC the desired next $(i+1)^{\text{st}}$ state, assuming that the i -th step has been reached, then the expression:

$$g_{i+1|i} = -\ln Q_{i+1|i} \quad (10)$$

can be used as *goal reaching chance* (a favorable parameter).

Construction of a positive cone for comparison of vectors. For the ACs comparative analysis the primarily specified set of actions' desired and undesired effects should be parameterized and reformulated so as to be presented in an unified form of a vector whose all components as "desired" or at least as "neutral" effects can be interpreted. The positive cones can be defined so as to have an elliptic or a convex-polygonal cross-section. The last case, leading to a cone having in fact the form of a pyramid described by 2-dimensional planar faces intersected at the θ apex, is to ACs assessment more convenient. In such case, the faces of a positive pyramid can be defined by a set of linear inequalities without constant terms describing relative preferences between the pairs of parameterized effects. This is illustrated in Fig. 4 where four typical necessary conditions for 2-component vectors assumed to be positive are presented. Fig. 4a illustrates a situation of positive sign of a vector meaning that at least one of its two components is positive while the other one is non-negative (i.e. its positive- or null-value is admitted). Fig. 4b presents a stronger situation in which positive are assumed vectors whose both components are positive. In Fig. 4c assumed positive are vectors whose component u_p is non-negative and $u_q > u_p$. In Fig. 4d at least one component of positive vectors is non-negative.

Fig. 4. Types of pyramids of "positive" vectors in K -space: a) both components are non-negative, b) both components are positive, c) one non-negative component (u_p) is lower than the other, positive one (u_q). d) at least one of two components is positive.

The corresponding inequalities take the following form:

- In the case 4a: $u_p \geq 0, u_q \geq 0$;
- In the case 4b: $u_p \geq \beta \cdot u_q, u_q \geq \beta \cdot u_p$;
- In the case 4c: $u_p \geq 0, u_q \geq (1+\beta) \cdot u_p$;
- In the case 4d: $u_p \geq -(1+\beta) \cdot u_q, u_q \geq -(1+\beta) \cdot u_p$.

for a small positive β . Replacement of the $>$ -type inequalities by their stronger versions (e.g., $u_p \geq \beta \cdot u_q$ instead of $u_p > 0$) is caused by the necessity of the K^+ pyramid being defined as a closed set.

When two actions $AC^{(a)}$ and $AC^{(b)}$ are to be compared, a pair of vectors $\mathbf{u}^{(a)} = [u^{(a)}_1, \dots, u^{(a)}_q]$, $\mathbf{u}^{(b)} = [u^{(b)}_1, \dots, u^{(b)}_q]$ characterizing their effects and their difference $\Delta(\mathbf{u}^{(a)}, \mathbf{u}^{(b)}) = [u^{(a)}_1 - u^{(b)}_1, u^{(a)}_q - u^{(b)}_q]$ should be taken into consideration. A system of linear inequalities imposed on the variables u_1, \dots, u_q , should thus be chosen according to the preferences assigned to the differences of the corresponding vector components $u^{(a)}_1 - u^{(b)}_1, \dots, u^{(a)}_q - u^{(b)}_q$.

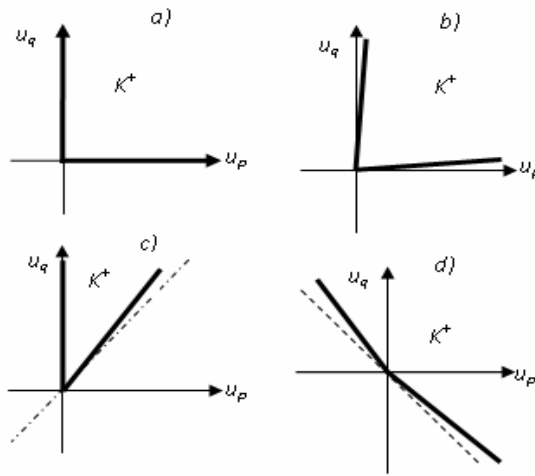


Fig. 4 Types of pyramids of “positive” vectors in K -space: *a*) both components are non-negative, *b*) both components are positive, *c*) one non-negative component (u_p) is lower than the other, positive one (u_q). *d*) at least one of two components is positive

Example 3

For comparative assessment of ACs consisting in spending a time in a physical rehabilitation camp the following desired effects are taken into consideration: *a*/ economical effectiveness (reversed to cost) u_1 , *b*/ offered total time u_2 of physical trainings, *c*/ offered total time u_3 of physiotherapeutic procedures. The following conditions for undesired effects comparison are established:

- The economical effectiveness and the total time of physical trainings are positive;
- The total time of physiotherapeutic procedures of physiotherapeutic procedures is non-negative;
- The total time of physiotherapeutic procedures should not exceed the total time of physical trainings.

This leads to the following inequalities describing the cone K^+ :

$$u_1 > 0; \quad u_2 > 0; \quad u_3 \geq 0; \quad u_3 \leq u_2 \quad (11)$$

However, the positive pyramid K^+ should be closed by its faces and this is why the inequalities $u_1 > 0$, $u_2 > 0$ by their stronger version:

$$u_1 + u_2 \geq \beta \cdot u_3 \quad (12)$$

should be replaced, β being a small positive constant.

Let for four ACs the following data be given:

Table 1 Data for comparative assessment of four ACs

	c [\$]	u_1 [h/\$]	u_2 [h]	u_3 [h]
$AC^{(a)}$	2400	0.0188	25	20
$AC^{(b)}$	3000	0.0120	20	16
$AC^{(c)}$	3600	0.0100	16	30
$AC^{(d)}$	2800	0.0129	16	20

The economical effectiveness has been calculated as a sum of training and procedure hours ($u_1 + u_2$) divided by the corresponding cost c . For a comparison of vectors their differences are calculated:

Table 2. Differences of compared vector's components

	$\mathbf{u}^{(a)} - \mathbf{u}^{(b)}$	$\mathbf{u}^{(a)} - \mathbf{u}^{(c)}$	$\mathbf{u}^{(a)} - \mathbf{u}^{(d)}$	$\mathbf{u}^{(b)} - \mathbf{u}^{(c)}$	$\mathbf{u}^{(b)} - \mathbf{u}^{(d)}$	$\mathbf{u}^{(c)} - \mathbf{u}^{(d)}$
u_1	0.0020	0.0088	0.0059	0.0020	-0.0009	-0.0029
u_2	5	9	9	4	4	0
u_3	4	-10	0	-14	-4	10

Looking at the Table 2 one can observe that the inequalities (13) by the differences $\mathbf{u}^{(a)} - \mathbf{u}^{(b)}$ and $\mathbf{u}^{(a)} - \mathbf{u}^{(d)}$ only are satisfied. In both cases, assuming that $\beta > 0.8$ the inequality (14) is also satisfied. Hence, it can be concluded that in the K -space described by the given inequalities the relationships $\mathbf{u}^{(b)} \prec \mathbf{u}^{(a)}$ and $\mathbf{u}^{(d)} \prec \mathbf{u}^{(a)}$ hold. Moreover, it can be noticed that no reversed difference of vectors satisfies the inequalities (13) and (14). Therefore, the relationships $\mathbf{u}^{(a)} ? \mathbf{u}^{(c)}$, $\mathbf{u}^{(b)} ? \mathbf{u}^{(c)}$, $\mathbf{u}^{(b)} ? \mathbf{u}^{(d)}$ and $\mathbf{u}^{(c)} ? \mathbf{u}^{(d)}$ have been found. Finally, $AC^{(a)}$ as the most preferable solution can be chosen. However, in general, more than one and most preferable mutually incomparable solutions can be found. In such case a selection of the best one by narrowing the positive pyramid K^+ can be reached •

4 Paths of Actions Planning, Evaluation and Selection

Logical structure of recommendations for composite actions can be represented in a canonical form:

$$P = \bigvee_{r=1}^R [\bigwedge_{s=1}^S a_s^{(r)}] \quad (13)$$

where \vee and \wedge are, respectively, the symbols of multiple logical *OR* and *AND* conjunctions; R and S_r are maximal numbers of disjunctive and conjunctive terms, $a^{(r)}_s$ denotes a simple action. This structure can also be presented by a graph shown in Fig. 5.

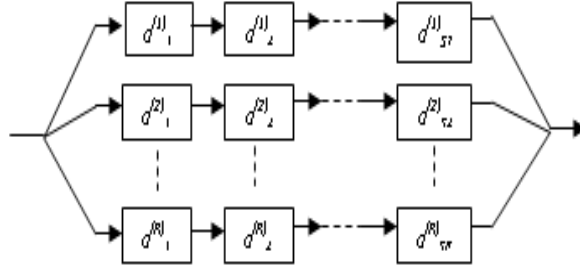


Fig. 5 General structure of a composite action: serial connections correspond to *AND* while parallel to *OR* logical conjunctions between simple actions

When the most preferable simple actions according to the rules described in Sec. 3 have been chosen, it arises a question how the most preferable path of actions (one of the R admissible ones) should be found. The problem would be trivial if all simple actions $a^{(r)}_s$ using the same criteria represented by vectors in a fixed K -space are evaluated. However, in general, applying exactly the same criteria to evaluation of several different ACs (e.g., of three different concepts of spending holiday times: a) by resting at a sea shore, b) by travelling by a hired car, c) by resting at holiday cottage and gardening) may be difficult even if not meaningless. Two methods of overcoming this difficulty are proposed:

1. Accumulation of general profits. The method consists in selection (if possible) of one or more additive parameters (called *general profits*), common for all simple ACs evaluation, calculation of their accumulated values along the paths of ACs and comparison in a corresponding K -space of so-obtained positive vectors. However, the general profits accumulated values calculation in different ways in the case of favorable and reformulated unfavorable parameters should be performed. In the first case, it consists in simple summation of the parameter's values. In the second case, the accumulated value v of a reformulated (favorable) parameter based on the unfavorable (z_s) ones limited by their maximum value Z is given by the formula:

$$v = Z - \sum_{s=1}^r z_s \quad (14)$$

The values v should thus be calculated for each path of ACs. The so-obtained favorable parameters in each path should be collected as components of vectors $U^{(1)}, \dots, U^{(R)}$ and compared as elements of a K -space based on an adequately to the nature of the given composite action constructed positive pyramid K^+ .

2. Credibility of the ACs assessment. This approach can be used if no common general profits to the simple ACs of a given composite action can be assigned. Taking into account that the components of the composite action previously as the most preferable ACs have been selected the credibility of the selection according to the following assumptions can roughly be assessed:
 - a) Credibility of a path of actions assessment is the lower the higher is the number of ACs the given path consists of;
 - b) For any two paths of equal lengths lower credibility is assigned to the path whose minimal credibility of ACs assessment is lower;
 - c) For any two ACs lower assessment credibility is assigned to the one whose assessment is based on lower number of parameters;
 - d) For any two ACs whose assessment is based on the same number of parameters lower assessment credibility is assigned to the one whose maximal preference has been based on higher number of mutually incomparable maximal quality vectors in the corresponding K -space.

In any case of composite actions assessment at least one of the above-formulated proposals can be used. On the other hand, in any case more than one mostly preferred solution may occur.

5 Conclusions

Computer-aided decision systems may be an effective tool supporting human activity in various application areas. Their role not obviously consists in solution of sophisticated logical inference problems; capability to retrieve adequate solution methods in large resources of former solutions or of on human experience based recommendations is non-less important. To this purpose deontological knowledgebases (*DKB*) can be used. They may provide a user with deontological statements containing recommendations for actions leading from a given present state to a desired final state. As such, they make possible construction of sets of alternative paths of actions (*ACs*) from which the most preferable one or ones should be selected. The selection should be based on evaluation of the favorable and unavoidable unfavorable effects of actions and on finding among them a compromise. For this purpose using the concept of semi-ordered linear vector space has been proposed. However, this (known as Kantorovitsch space or K -space) concept, as it has been shown in the paper, should be to the purpose adopted and slightly modified. As a result, a combined theoretical model consisting of *DKB* and K -space notions is proposed as a basis for design of human activity aiding computer systems.

References

- [Baker and Eris 1964] Baker, B.N., Eris, L.R.: An introduction to PERT-CPM. Richard D Irwin Inc, Homewood (1964)
- [Barthelemy and Janowitz 1991] Barthelemy, J.P., Janowitz, M.F.: A formal theory of consensus. *SIAM J. Discrete Math.* 4, 305–322 (1991)

- [Bolc et al. 1998] Bolc, L., Dziewicki, K., Rychlik, P., Szalas, A.: Inference in non-classical logics. In: Automation of Inference. AOW PLJ, Warsaw (1998)
- [Dubois and Prade 1991] Dubois, D., Prade, H.: Fuzzy sets and systems. Academic Press, New York (1980)
- [Ferber 1999] Ferber, J.: Multi-agent systems. Addison-Wesley, New York (1999)
- [Hecht-Nielsen 1991] Hecht-Nielsen, R.: Neurocomputing. Addison-Wesley, New York (1991)
- [Jensen 2000] Jensen, F.V.: Bayesian networks and decision graphs. Springer, Heidelberg (2000)
- [Kantorovich 1959] Kantorovich, L.V., Vulich, B.Z., Pinsker, A.G.: Functional analysis in semi-ordered spaces. GITTL, Moscow (1959) (in Russian)
- [Kłopotek 1994] Kłopotek, M.: Conditional independence concept in probability theory versus that of Dempster-Shafer theory. In: Intelligent Information Systems Proc. of the Work held in Wigry, IPI PAN, Warsaw (1994)
- [Kulikowski 2006] Kulikowski, J.L.: Model of deontological inference in decision system with knowledge base. In: Grzech, A. (ed.) *Żywnieria Wiedzy i Systemy Ekspertowe*, pp. 163–172. OW Politechniki Wrocławskiej, Wrocław (2006)
- [Kulikowski 2009] Kulikowski, J.L.: Problems of knowledge representation in computer-assisted decision making systems. In: Hippe, Z.S., Kulikowski, J.L. (eds.) *Human-Computer Systems Interaction*. AISC, vol. 60, pp. 39–54. Springer, Heidelberg (2009)
- [Nowakowska 1979] Nowakowska, M.: Action theory. PWN, Warsaw (1979) (in Polish)
- [Pawlak 1991] Pawlak, Z.: Rough sets – theoretical aspects of reasoning about data. Kluwer Academic Publishers, Boston (1991)
- [Rusinkiewicz and Bregolin 1994] Rusinkiewicz, M., Bregolin, M.: Transactional workflows in distributed systems. In: Intelligent Information Systems Proc of the Work held in Wigry. IPI PAN, Warsaw (1994)
- [Zilli et al. 2009] Zilli, A., Damiani, E., Ceravolo, P., Corallo, A., Gianluca, E.: Semantic knowledge management. An Ontology-based Framework. Information Science Reference, Hershey (2009)

Multi-aspect Character of the Man-Computer Relationship in a Diagnostic-Advisory System

E. Nawarecki¹, S. Kluska-Nawarecka^{2,3}, and K. Regulski¹

¹ AGH University of Science and Technology, Krakow, Poland
nawar@agh.edu.pl, regulski@metal.agh.edu.pl

² Foundry Research Institute in Cracow, Poland
nawar@iod.krakow.pl

³ The Academy of Information Technology WSInf, Lodz, Poland

Abstract. Chapter reviews the solutions used in diagnostic-advisory system, dedicated to the needs of the foundry industry. Multimedia techniques used in the system were selected to create a comprehensive capabilities of contact between user and computer system. The first part of the considerations apply to work in diagnostic mode, including identification of castings defects and determine the causes of their occurrence, the second part is a look at some of the features of the system implemented in a consultative mode, such as the integration of knowledge or technical and market expertise. Many illustrations was used to illustrate the various forms of contact with the user.

1 Introduction

Rapidly advancing industrial development of diagnostic and advisory systems is associated, on the one hand, with increasing use of advanced formal methods (fuzzy logic, rough sets, descriptive logic, Bayesian networks, ontologies) while, on the other, it means an improvement in computer hardware and software tools. However, there is still a third aspect, determining not so much the possibility to design systems of this class as rather the conditions of their dissemination, i.e. the degree of acceptance from industrial users. This aspect also concerns the means and tools used to ensure proper man (system user) - computer relationship.

Rational determination of this relationship can be sometimes quite difficult, as it requires taking into account not only the considerations of a technical nature, but also (and perhaps especially) the psycho-physical and social factors that arise from the habits and the way of thinking of a technologist (or expert) in a given sector of industry. The point here is, therefore, to make the formalized knowledge contained in a computer system available in the form acceptable by a technologist

and consistent with the patterns of thinking he has been accustomed to use on a regular basis.

Specific solutions in the scope of the forms of knowledge presentation and organisation of user interface depend, of course, on the specific features of the industry sector, to which the information - advisory system is dedicated. Therefore, avoiding too far-reaching generalizations, in the present work it was decided to describe selected solutions regarding a computer - user relationship as applied in an information - decision system tailored to the needs of foundry industry and created in cooperation with the Foundry Research Institute, the Faculty of Industrial Computer Science and Modeling, and the Chair of Computer Science, AGH University of Science and Technology.

The starting part of the paper gives a general description of problems related with the diagnosis and technical advisory activity in foundry. Against this background, some of the solutions used in a multimedia knowledge presentation and dialogue procedures applied in the diagnostic and decision-supporting systems have been described. Finally, the functionality range of the main modules of the system has been outlined.

The discussions have been illustrated with drawings and screenshots of the performed applications.

Regardless of the fact that the presented solutions relate to a limited area of the domain knowledge, it appears that some ideas can be in a natural way transferred to a similar class of systems to assist the production processes in various industries.

2 Diagnosis and Technical Advice in Foundry

High complexity of the problem of diagnosing the defects in castings results from both the number of the recognized types of defects (acc. to PN-85/H-83105) as well as the number of potential causes of their formation. Moreover, several possible causes are attributed to each defect, and each cause can give rise to different types of defects.

As a result, the number of possible, from a technological point of view, pairs of statements [defect, cause] becomes very large and indicating proper pair of these concepts can be difficult, especially with incomplete information concerning the technological process parameters.

Additional difficulties in making proper diagnosis result from the fact that the mere identification of defect type formed in casting is also a task sufficiently complicated, since some symptoms of the occurrence of a number of defects can be similar (identical even sometimes).

The diagnostic process consists of several steps, and each of these steps uses slightly different areas of technical knowledge, requiring appropriately selected formalisms of knowledge representation and forms by which it is rendered

available to users of the system. A schematic diagram of the performed diagnostic procedure with indication of the knowledge components used in subsequent steps is shown in Figure 1.

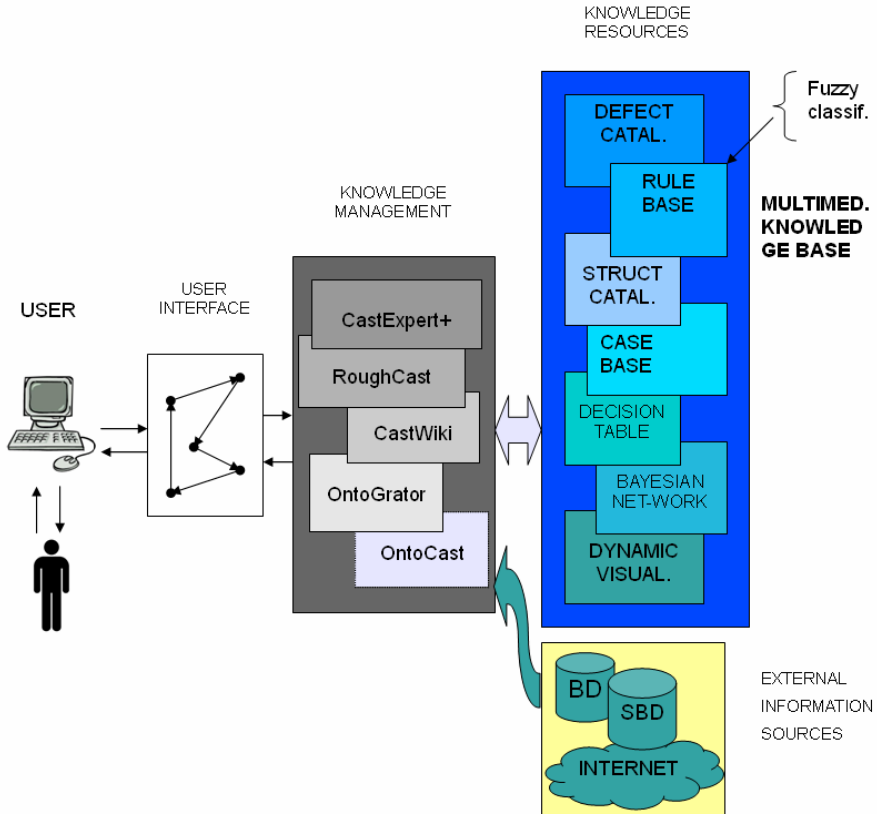


Fig. 1 Schematic diagram of the diagnostic procedure

In step 1 an identification of the defect type is done, using knowledge modules which allow its visual identification assisted with linguistic description.

Step 2, which leads to the determination of causes of the occurrence of defect type identified in step 1, is based on a dialogue with the user, applying a rule-based knowledge (classical logic, fuzzy logic), completed with a graphical representation (catalogue of structures, Bayesian networks).

Relatively simple in implementation is step 3, where to the causes of defects are assigned appropriate corrective actions (preventing the occurrence of causes), using simple rules of classical logic.

The second, besides the diagnostic process outlined above, functionality is that of offering expertise and consultant advice on how to improve the manufacturing process, carry out modernization works and implement new technologies.

Due to a large variety of such tasks, it is difficult to give a universal scheme of dialogue procedures run in parallel with these tasks. Some examples of such procedures will be presented in the discussion of various components of the knowledge and decision-making modules.

Great complexity is also a typical feature of problems associated with technical advice on specific opportunities for improvement of foundry technology, modernization actions, or introducing new technologies. Here, an important role is played by changes in the production process (modernization of equipment, improvement of measurement methods), the appearance of new materials and technologies, and finally, the market dynamics (types of orders, prices of raw materials and products, etc.).

All these factors contribute to the situation in which a computerized diagnostic and advisory system should have the ability to flexibly adapt to the changing needs of its user. This flexibility should apply to both the knowledge area used by the system, as well as the means of processing it and forms of provided access.

To meet these multi-faceted requirements, a concept of the system based on component structure was adopted. Individual components represent the specific areas of knowledge that can be disjoint on some occasions, or similar, identical even, on the other, but which are always represented with different formal methods. This is achieved by the ability to adapt the knowledge area and some means for its processing to the specific characteristics of an ongoing task.

It should be noted that the use of knowledge originating from different components (and sometimes from different sources of information) is invariably associated with the need for its integration, using appropriate methods and tools (e.g. ontologies).

Consequently, from a functional point of view, the system can be represented as a diagram shown in Figure 2, where knowledge resources consisting of a series of components and solutions for knowledge management, i.e. bringing it to the form handed to the user by interface, have been isolated. (It is worth mentioning that this division is for illustration only and may not correspond to the physical structure of the system).

The information resources provided by the system include a multimedia knowledge base (MKB), which forms an integral part of the system, the databases, like SINTE and NORCAST (both installed and operating at the Foundry Research Institute), and an option to retrieve information from the Internet.

A multimedia knowledge base (MKB) includes the following components:

- catalogue of casting defects and photographs of these defects with descriptions, as specified by the standard,
- rule base written in terms of classical logic,

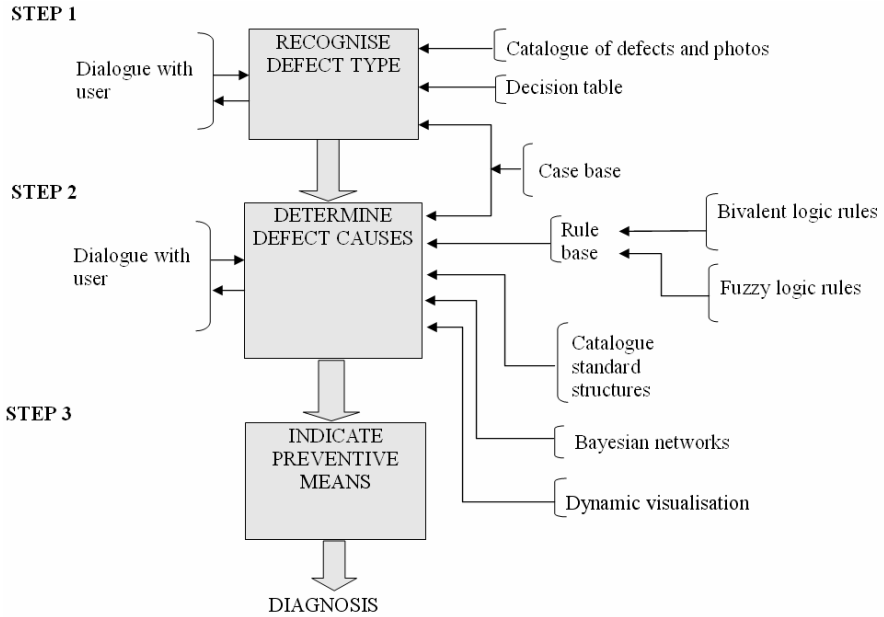


Fig. 2 Functional diagram of the diagnostic-advisory system

- rule base written in terms of fuzzy logic,
- catalogue of standard structures of cast metals and alloys with microscope images of these structures,
- case base containing descriptions and photographs of typical casting defects, stating also causes of these defects,
- attribute tables (which are a specific variety of decision tables), containing linguistic descriptions of defects, based on Polish and foreign standards,
- Bayesian networks, which illustrate causal connections related with the occurrence of certain types of defects,
- dynamic visualisation module, which enables tracing the mould pouring process, thus making the user capable of understanding the mechanism responsible for the formation of some types of defects (e.g. blowholes).

From the above it follows that the knowledge contained in MKB is characterized by a high level of redundancy, as it provides alternative ways to present the same physical process, which is the formation of defects in castings. The idea of this conceived solution was to create a possibility for the selection of components that will best correspond to individual diagnostic situations and can be adapted to the user's mentality and predisposition to make decisions.

The way of formulating the task posed by user (diagnosis, expert opinion, consultation) and the level of information he possesses (knowledge of the parameters of a technological process, predefined nature of responses) determine the choice of

a system module, which will be used in solution of this task. The procedures comprised in this module, implemented further in modules of the knowledge management (KM) define, on the one hand, the components of knowledge that will be used in the course of solving the task while, on the other, they indicate the mode in which the dialogue will be run with user and the form in which the knowledge used in this dialogue will be imparted (linguistic, visual, dynamic).

In the adopted embodiment of knowledge management, the following modules are used:

- CastExpert+ which, being an expanded version of the previously developed CASTEXPERT system [Dobrowolski et al. 2003], performs various diagnostic functions, which in this case include: identifying the type of defect, indicating possible causes of this defect, identifying actions preventing the occurrence of the defect;
- Rough Cast, which uses the knowledge of casting defects expressed in the form of decision-making tables. This module is adapted to work in conditions of incomplete knowledge and is used mainly to identify the type of defect;
- CastWiki, based on the ontological knowledge representation, functioning as a platform for the exchange and saving of foundry knowledge. The task of this module is integration of domain knowledge, originating from various sources and expressed in a linguistic form or in the form of graphic files;
- like CastWiki, OntoGrator is designed to integrate knowledge from heterogeneous sources, with attention focused on the task of eliminating the semantic and syntactic differences, presenting the user with knowledge in a unified form.

A perspective development vision of the existing system is the currently elaborated concept defined as OntoCast, where integration of the above mentioned modules is anticipated, and – as a consequence – further simplification of dialogue procedures and enrichment of the system functionality.

3 Dialogue with the User in CastExpert+ System

Organization of dialogue with user has a decisive influence on the course and outcome of the diagnostic process. The components of knowledge available to the user at the subsequent stages of this process should facilitate responses adequate to the specific character of a diagnostic task.

In determination of defect type (step 1), the dialogue begins with presentation of the successive images comprised in a catalogue of defects. As an example, one of the pages from this catalogue has been shown in Fig.3.

This page contains a symbol, a description, and a series of the images of a specific defect. In case of doubt, one can refer to a case base (Fig. 4) containing pictures and descriptions of both the typical defect occurrence, as well as cases difficult to diagnose (occurring incidentally).

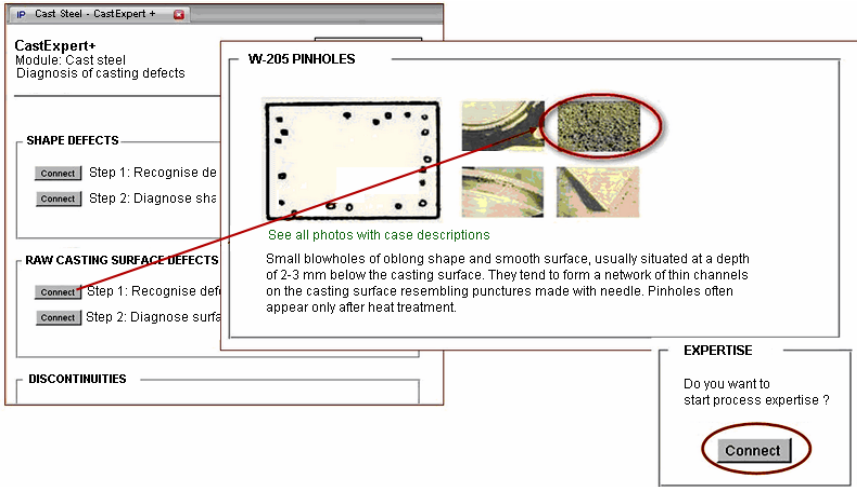


Fig. 3 Catalogue of defects and photos

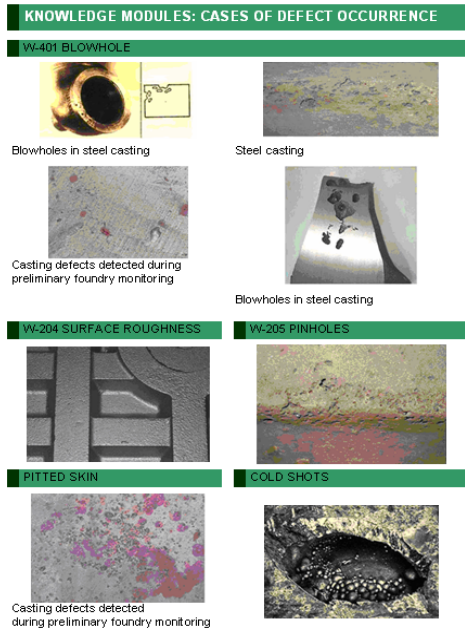


Fig. 4 Knowledge modules: cases of defect occurrence

For better adaptation to the user's predisposition, the CastExpert+ system also introduces a module with description of the defect in the form of decision table, a fragment of which is shown in Fig.5. In this table, individual defects are treated as objects, to which numerous attributes have been ascribed, while their values are a

linguistic description of the defect (object). Here, the defect is identified in a kind of opposite perspective - the user himself indicates the features of the defect (attribute values), and a composition of these features (i.e. an appropriate line in the table) allows indicating the name of the defect.

The decision table is also used in another context, namely as a form of incomplete knowledge representation, and this variant of its application will be described in continuation of this study.

Assuming that the diagnosis of the defect type has already been made, one has to specify reasons for its occurrence (step 2). Here, the dialogue concerns the conditions (parameters) of the technological process performance. An example of such a dialogue is given in Fig. 6.

#	A	B	C	D	E	F
	Defect name	Standard	Symbol	damage type	Visibility	damage size
1						
20	Cold laps	CZ	341	wrinkles	sharply outlined	distinct
25	Crush, Bruising, Push up	CZ	116	cavity	sharply outlined	distinct
30	Dent	CZ	123	dent	sharply outlined	distinct
45	Internal contraction crack	CZ	313	discontinuity	invisible	distinct
48	Mechanical damage	PL	W101	dent	sharply outlined	distinct
53	Misrun	PL	W102	part of casting missing	sharply outlined	distinct
58	Knob	PL	W103	deformation	sharply outlined	distinct
178	Cold crack	D	21	discontinuity	sharply outlined	
179	Crack in core	D	23	buildup		
180	Hot crack	D	48	fissures	visible with naked eye	scattered
184	Mould crack	D	11	buildups	visible with naked eye	local
185				knobs		
186	Cold fracture or cold crack	FR	C111	discontinuity	hardly visible	distinct
194	Cold crack	FR	C211	fissure	sharply outlined	distinct
195	Crack	FR	C221	fissure	visible	distinct
199	Hot crack	FR	C222	fissure	sharply outlined	distinct
201	cold lap, cold shots	FR	C311	discontinuity	sharply outlined	distinct

Fig. 5 Fragment of decision-making table for cast steel

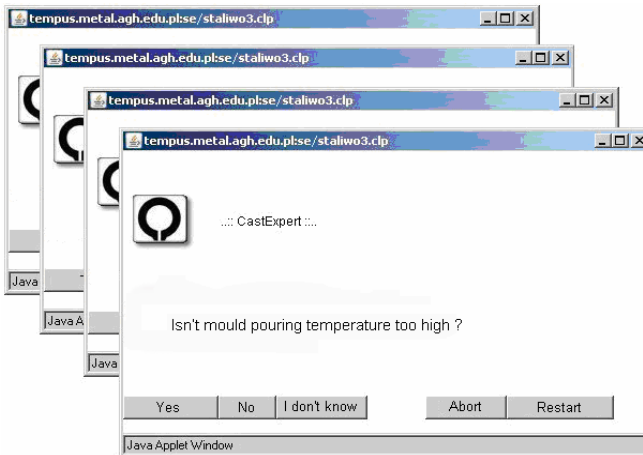


Fig. 6 Dialogue with user

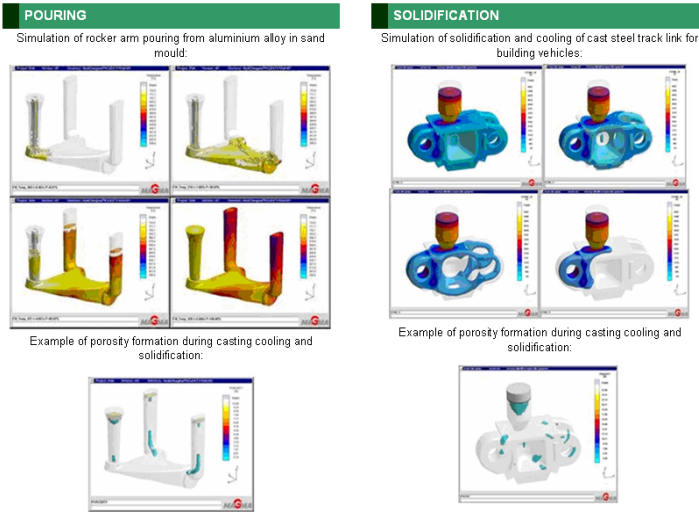


Fig. 8 Dynamic visualisation of pouring and solidification process

Having determined the cause of defects, the next step is making the final diagnosis with an indication of actions that should be taken to prevent formation of this defect (step 3). This stage usually takes place without user's intervention. An example of the final diagnosis is given in Fig. 9.

The above description of the diagnostic process has been given mainly to show various forms of knowledge visualisation, which are expected to make the user getting more easily adapted to the operating mode of an expert system. At the same time, a presentation of this type enables the user to acquire certain routine in working with computer and prompts reflection on the role of technological conditions in a casting process.

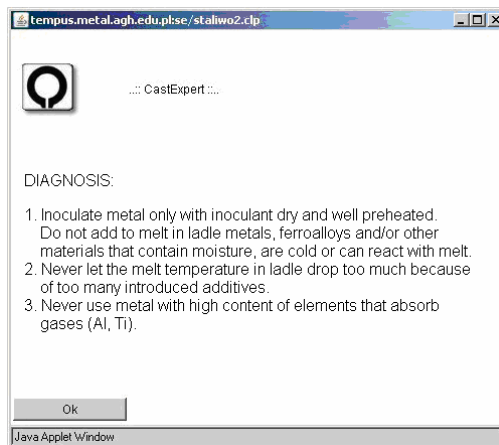


Fig. 9 Example of the final diagnosis

4 Analysis and Integration of Information

In the foregoing discussion, the forms of contact with the user applied in conventional diagnostic procedures were described. Here the solutions for an analysis and integration of information and knowledge addressed not so much to a technologist conducting the manufacturing process, as rather to an expert trying to capture the unknown interrelations between the process parameters, or to a researcher, whose aim is to deepen the knowledge, or detect new relations and regularities describing given class of thermophysical and manufacturing processes, will be given.

4.1 Representation of Incomplete Knowledge in RoughCast Module

In this module, the knowledge about the defects is represented in the form of already mentioned (see item 3) decision-making tables. The concept of decision table can be applied to a model of the information system expressed in the form of an aggregate [Pawlak 1982]:

$$SI = \langle X, A, V, f \rangle \quad (1)$$

where:

$$f: X \times A \rightarrow V \quad (2)$$

is mapping of the Cartesian product $X \times A$ in a space of attribute values $V = UV_j$, which assigns values to each pair (object, attribute), that is:

$$f(x_i, a_j) = v_{ij}, \quad x_i \in X, \quad a_j \in A, \quad i = 1, \dots, n; \quad j = 1, \dots, n \quad (3)$$

The function f is often expressed as a $T^{n \times n}$ table, an example of which is the table representing the knowledge about the defects in castings shown in Figure 5, where X represents the set of defects, a are the qualities that describe these defects, and v is the value of the j^{th} feature for the i^{th} defect (defined for a given case in a linguistic form.)

In describing the defects in castings, one can often encounter the situation when the characteristics of a defect cannot be specified explicitly, or there is no possibility to determine a given characteristic. In this case, formula (3) assumes the following form:

$$f(x_i, a_j) = [v'_{ij}, v''_{ij}] \quad (4)$$

where: v'_{ij}, v''_{ij} determine the range of values of the j^{th} attribute for object x_i , while in the case of a linguistic variable, a subset of values $\{v_{ij}\} V_j$ should rather be referred to.

If the value of certain feature (or of an attribute a) cannot be estimated, then: $f(x_i, a_j) = V_j$ and the situation corresponds to blank spaces in the table in Figure 5.

From a formal point of view, as a consequence of the replacement of expression (3) with (4), the information system (1) becomes a system with incomplete information. It is the fact well known that in the theory of rough sets for this class of systems [Pawlak 1982; Kluska-Nawarecka et al 2009], in determination of a set of objects (Y) of the required properties (attribute values), the concepts of upper (\overline{Y}) and bottom (\underline{Y}) approximations are used, where:

- \underline{Y} is the lower approximation of the searched set Y , if all the elements $x_i \in \underline{Y}$ certainly belong to Y .
- \overline{Y} is the upper approximation of the set Y , if it contains all the elements $x_i \in \overline{Y}$, which may belong to Y .

Implemented in the diagnostic-decision system described here, the RoughCast module allows various operations to be performed on sets with incomplete information, including not only the derivation of approximations ($\underline{Y}, \overline{Y}$), but also the following analyses:

- study of the effect of individual attributes on the accuracy of approximation, defined as:

$$\mu(a, X) = \text{card}(\underline{Y}) / \text{card}(Y) \quad (5)$$

- determination of a relationship between objects (representing defects) through designation of a set (or its approximation) of objects for which the selected attributes have the same values, that is:

$$f(x_i, a_j) = f(x_k, a_j), \quad x_j \neq x_k \quad (6)$$

referring condition (6) to a greater number of attributes, which enables introducing the notion of similarity between defects, defined as a number of the defect attributes x_i, x_k having the same values.

From a technological point of view, the study of changes in the approximation accuracy (μ) can provide an important guidance on the significance of each attribute when determining the type of defect (step 1 in the diagnostic procedure).

The technologist can also draw very interesting conclusions from the similarity of defects, since it indicates a similar nature of the causes of these defects (step 2).

It should also be noted that the decision table (Fig. 5) includes defects defined by standards published in different countries (Polish, Czech, French). In this context, the degree of similarity between the defects shows similarities (or differences) in standards used by different countries.

As an illustration, Figure 10 shows one of the forms used in RoughCast module, while Figure 11 is a screenshot giving details of the upper and lower approximation of a specific set of defects determined by the following formula:

$$t = [\text{damage type} = (\text{discontinuity, crevice})] \quad [\text{distribution} = \text{local}] \quad (7)$$

The screenshot shows the RoughCast 2009 interface. At the top, there is a logo for 'RoughCast 2009' and a navigation bar with links: 'Main Page', 'Start Test', 'Help', 'Database', and 'RoughCast Description'. Below the navigation bar, there are two main panels. The left panel is titled 'defect shape' and contains a list of attributes with checkboxes: 'icing sugar', 'irregular', 'straight', 'regular', 'ramified', 'wide', 'narrow, rounded edges', 'narrow', 'curved walls', 'zigzagged', and 'zigzagged'. The 'regular' and 'narrow, rounded edges' checkboxes are checked. Below this list are 'Next' and 'Reset' buttons, and a 'Help' link. The right panel is titled 'damage type' and contains a list of attributes with checkboxes: 'casting part missing', 'lumps', 'knobs', 'cavity', 'channels', 'cold shots', 'hot tear', 'oxide spots', 'bulldup', 'scabs', 'unevenness', 'spots', 'deformation', 'breaking off', 'swell', 'envelope', 'cold lap', 'pores', and 'discontinuity'. All checkboxes in this panel are unchecked.

Fig. 10 Forms selecting attribute values in the RoughCast module

If the user deems thus obtained approximation insufficient, he may enter the value of another attribute (providing he possesses such information). As a condition for completion of the procedure, the highest accuracy of approximation ($\mu_{max} \mu$), or obtaining the lower approximation in the form of a single element ($card \underline{Y} = 1$) can be adopted.

The screenshot shows a window titled 'Possible casting defects'. It contains two sections: 'Upper approximation:' and 'Lower approximation:'. Each section lists seven defects with their corresponding codes and languages. Below the lists are 'Next' and 'End' buttons, and a 'Help' link.

Possible casting defects

Upper approximation:

- 1) 341 COLD LAPS (Czech)
- 2) C211 COLD CRACK (French)
- 3) C221 CRACK (French)
- 4) C331 COLD LAP NEAR CORE OR OTHER METALLIC PART (French)
- 5) C411 CONCHOIDAL FRACTURE, ICING SUGAR (French)
- 6) W301 HOT CRACK (Polish)
- 7) W303 CONTRACTION CRACK (Polish)

Lower approximation:

- 1) 341 COLD LAPS (Czech)
- 2) W301 HOT CRACK (Polish)
- 3) W303 CONTRACTION CRACK (Polish)
- 4) C221 CRACK (French)
- 5) C331 COLD LAP NEAR CORE OR OTHER METALLIC PART (French)
- 6) C411 CONCHOIDAL FRACTURE, ICING SUGAR (French)

Next End Help

Fig. 11 Upper and lower approximation of casting defects set

4.2 Ontology and Knowledge Integration

As already mentioned, one of the features of the system is sharing the knowledge originating from the distributed and often heterogeneous sources [Kluska-Nawarecka et al 2010; Kluska-Nawarecka et al. 2007].

In the offered solution, integration and, as a consequence, knowledge dissemination, are carried out in the two, alternatively used, modules identified as Cast-Wiki and Ontogrator. These modules differ in the way of using formal methods and procedures for the knowledge processing, and finally in a form in which the knowledge sharing is done. The user can decide on the choice of one of these tools - according to the specific tasks he intends to solve.

In both these solutions, the first stage was the creation of a domain ontology, which in this case concerned the knowledge of the manufacture of castings. To accomplish this task, an OWL language was used together with the corresponding Protégé environment and methodology 101 [Cullot et al 2007; Konstantinou et al 2006]. The corresponding example of a hierarchical ordering of the knowledge about cast materials is shown in Figure 12.

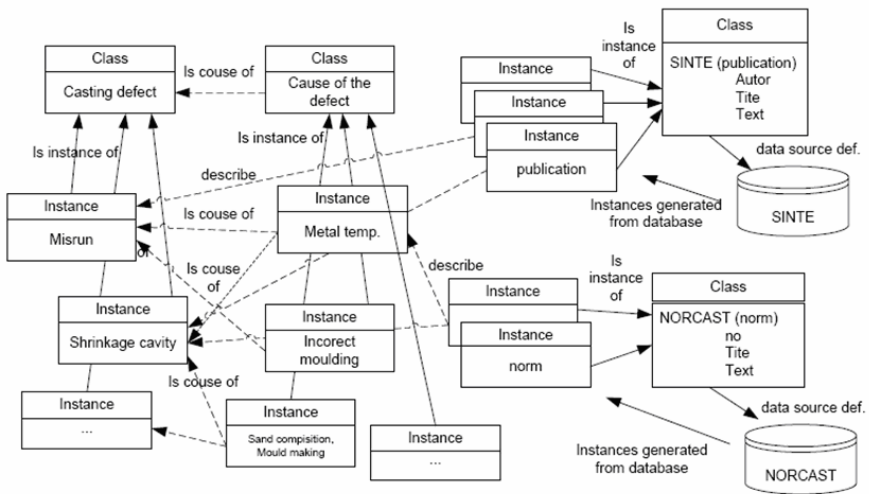


Fig. 12 Upper and lower approximation of casting defects set

The CastWiki module was implemented basing on a Wikipedia concept. It was decided to adopt this type of approach taking into consideration the following characteristics:

- the Wiki-type tools are a popular source of information and knowledge, well known to most of the Internet users,

- the Wiki system allows maintaining permanently updated knowledge resources, including comments and discussions,
- the Wiki technology is relatively simple and requires minimal skills in introducing and editing new content,
- the Wiki-type structure allows description of terms and concepts in natural language, ensuring simultaneously their effective identification through a unique URL identifier.

So, CastWiki is a theme-oriented platform for storing and sharing the knowledge on castings expressed in the form of:

- descriptions in natural language,
- image files,
- hypertext links to other entries in CastWiki,
- links to all resources shared on the network, having their own URL (catalogues, photos, animations, databases).

These opportunities are illustrated by an example given in Figure 13.

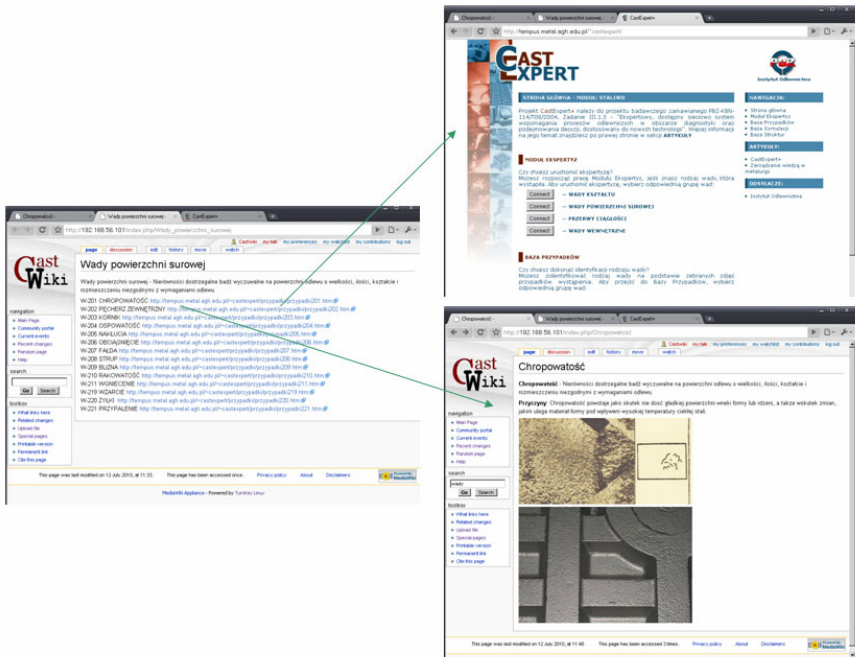


Fig. 13 Illustration of CastWiki proceeding

The structure of the knowledge resources in a CastWiki module is organised in such a way that it provides the user with a number of possible to introduce modifications, including, among others, the following:

- introduction of new classes / instances directly to the ontology,
- creating new definitions or modification of the already existing ones,
- creation of new classes in ontology with the possibility of gradual additional defining.

The possibility of co-creation of the knowledge resources by users of CastWiki promotes deeper understanding of the role and functionality of this module, and consequently contributes to its effective use.

5 Final Remarks

The main aim of this study was to present the diversity of the forms of knowledge representation, used in the currently applied diagnostic and advisory system, enabling flexible and multi-threaded contact with the, comprised in it, information resources. The scenarios of dialogue and methods for visual representation of knowledge elements are adapted to the specific tasks and user predispositions, providing him with clear and inspiring idea of how to best work with the system.

Focusing on the above aspects, the work abandons formal considerations and implementation issues, as they could enlarge too much the volume of the study. This area of the conducted studies is related in various publications [Kluska-Nawarecka et al 2007; Górný et al 2010].

The authors believe that the most interesting of the presented solutions, having no counterparts in any of the known embodiments of systems of this class, are the following ones:

- use of the dynamic presentation of knowledge in the form of the results of computer simulation that provides inspiration for deeper understanding of the essence of thermophysical processes determining the properties of product (casting),
- use of decision tables (also called attribute tables) as a form of knowledge representation which, on the one hand, enable its direct use (in a dialogue mode) while, on the other, allow an in-depth analysis of the consequences of uncertainty of some elements of knowledge,
- using procedures based on the rough sets theory, designing of modules that, by using ontological methods, provide opportunities for automatic integration of knowledge, which allows the use of distributed information sources, such as standards and catalogues published in different countries, distributed databases, network, Internet, etc.

It seems that the system can serve as an important support tool for technologists and specialists in foundry industry, providing inspiration in designing particular

References

- [Cullot et al. 2007] Cullot, N., Ghawi, R., Yétongnon, K.: DB2OWL: A Tool for automatic database-to-ontology mapping. In: Proc. of the 15th Italian Symp. on Advanced Database Systems, Italy, Torre Canne di Fasano (BR), pp. 491–494 (2007)
- [Dobrowolski et al. 2003] Dobrowolski, G., Marcjan, R., Nawarecki, E., Kluska-Nawarecka, S., Dziaduś, J.: Development of INFOCAST: Information system for foundry industry. *TASK Quarterly* 7(2) (2003)
- [Górny et al. 2010] Górny, Z., Kluska-Nawarecka, S., Wilk-Kołodziejczyk, D., Regulski, K.: Diagnosis of casting defects using uncertain and incomplete knowledge. *Archives of Metallurgy and Materials* (2010)
- [Kluska-Nawarecka et al. 2007] Kluska-Nawarecka, S., Smolarek-Grzyb, A., Wilk-Kołodziejczyk, D., Adrian, A.: Knowledge representation of casting metal defects by means of ontology, archives of foundry engineering. *Polish Academy of Sciences* 7(3), 15/3, 75–78 (2007)
- [Kluska-Nawarecka et al. 2009] Kluska-Nawarecka, S., Wilk-Kołodziejczyk, D., Dobrowolski, G., Nawarecki, E.: Structuralization of knowledge about casting defects diagnosis based on rough sets theory. *Computer Methods In Materials Science* 9(2) (2009)
- [Kluska-Nawarecka et al. 2010] Kluska-Nawarecka, S., Mrzygłód, B., Durak, J., Regulski, K.: Analysis of the applicability of domain ontology in copper alloys processing. *Archives of Foundry, Polish Academy of Sciences* (2010)
- [Konstantinou et al. 2006] Konstantinou, N., Spanos, D.E., Chalas, M., Solidakis, E., Mitrou, N.: VisAVis: An approach to an intermediate layer between ontologies and relational database contents. *Web Information Systems Modeling* (2006)
- [Nawarecki et al. 2007] Nawarecki, E., Kluska-Nawarecka, S., Dobrowolski, G., Marcjan, R.: OntoGRator – an intelligent access to heterogenous knowledge sources about casting technology. *Computer Methods in Material Science* 7 (2007)
- [Pawlak 1982] Pawlak, Z.: Rough sets. *Int. J. of Information and Computer Science* 11(5), 341–356 (1982)

Services Merging, Splitting and Execution in Systems Based on Service Oriented Architecture Paradigm

A.Grzech, A. Prusiewicz, and M. Zięba

Institute of Informatics, Faculty of Computer Science and Management,
Wroclaw University of Technology, Poland

{adam.grzech,agnieszka.prusiewicz,maciej.zieba}@pwr.wroc.pl

Abstract. The aim of the paper is to discuss some selected issues related to services merging, partitioning and execution in systems based on service oriented paradigm. The main feature of such systems is that the required services may be efficiently and flexibly composed of available atomic (elementary) services providing certain and well-defined functionalities. It is rather obvious that the flexibility of such a services delivering system may be limited by the amount and cost of communication necessary to support increasing atomic services granularity. It is assumed that the cost of complex service delivery is composed of exchanged data flows processing and communication costs and the services quality depends on delays introduced by available resources for data flows characterizing services requests followed by specified requirements.

1 Introduction

Systems based on SOA (Service Oriented Architecture) paradigm offer complex services, which are delivered as composed of atomic services [Johnson et al. 1995; Milanovic and Malek 2004]. The main feature of such an attempt is that the required complex services may be efficiently and flexibly composed of available atomic services providing certain, well defined, required and personalized functionalities. Requested complex services are characterized by set of various parameters specifying both functional and nonfunctional requirements. The former define exact data processing procedures, while the latter describe various aspects of required service quality. The set of parameters describing requested complex service form SLA (Service Level Agreement) [Anderson et al. 2005; Narayanan and McIlraith 2003].

Functionality of the requested complex service is available as a sum of atomic services functionalities. In order to deliver complex service with requested functional and non-functional properties appropriate atomic services must be chosen in

the process of complex service composition [Jaeger et al. 2005]. Required functionality, uniquely defined in the SLA, determines set of required atomic services as well as a plan according to which the atomic services are performed in distributed environment. Non-functionality of the requested complex service, which is mainly related to QoS (*Quality of Service*) issues, in most cases, i.e., in distributed environment, may be assured or obtained by proper resources (processing and communication) and tasks (atomic services) allocation [Grzech 2004; Grzech and Swiatek 2008; Grzech and Swiatek 2009].

Discussed complex services delivery approach is available only at the distributed environment; possible parallel execution of distinguishable atomic services requires allocation of proper amount of processing and communication resources in parallel manner. The distributed environment may be obtained both by allocation of separated or virtualized resources.

In order to obtain various required QoS in distributed environment well-known QoS strategies, i.e., best-effort, integrated services and differentiated services concepts may be applied. Usefulness of the mentioned concepts strongly depends on formulation of the non-functional SLA's parts. Application of the best-effort concept, based on common resources sharing leads to solution, where the same, higher enough, average quality of service is delivered to all performed requirements. The next two, previously mentioned, concepts offer differentiated quality of service for requests (also guarantees) and are mainly based on resources reservation for individual requests (integrated services concept) or for classes of requests (differentiated services concept) [Grzech 2002].

One of the most important and commonly used non-functional properties of the requested complex services, reflecting services' QoS, is service response time. Value of the latter is mainly influenced by three factors: execution time of atomic services, load of the system (number of requests performed in parallel in the system) and communication delays introduced by communication channels among servers executing the atomic services. In order to guarantee requested response time, given in the SLA, all these factors must be taken into account in the process of service composition [Garey et al. 1976; Graham et al. 1979].

The paper is devoted to discuss some selected issues concerning merging and partitioning of (atomic) services available in services repository and applied to obtain complex services functionalities. The discussed issue has the following motivation. In some distributed processing and communication resources environment some set of complex services, composed of atomic services from well-defined set, may be obtained. It is rather obvious that higher level of growing atomic services granularity (followed by decreasing functionality) increases obtained complex services flexibility which is paid by increasing amount of data exchanged among data processing unites. On the other hand, merging of atomic services leads to services with broader spectrum of functionalities limiting both flexibility of the services delivering system and amount of required communication resources. In such circumstances it is worth to ask about optimal atomic services granulation level

assuring services flexibility and communication costs trade-off. It is assumed that collected knowledge about required complex services, given by the complex services structures as composed of particular atomic services, allows to discover atomic services substructures, which are performed more or less frequently. High and increasing frequency of performance of some substructures composed of atomic services may lead to question about possible merging of the atomic services into one, greater (in sense of delivered functionalities) atomic service. The services' merging both decreases services flexibility and saves communication resources. The atomic services, requested in sequence, may be worth to be merged if it leads to limit amount of required communication resources. On the other hand some services, assuring some functionality, are worth to be divided into separate parts, if the separation decreases services delivery costs.

The other question is about the optimal number of services that may be executed in parallel according to the resources limitations. The total number of services executed in parallel may be reduced to the optimal one by applying services merging process.

2 Services Merging Problem Formulation

Complex services may be represented by set of directed paths along which flow data required to complete the performed services. The paths may be represented by amount of data flow exchanged among adjacent nodes at the paths (atomic services) as well as by processing and communication resources required to assure the required atomic services functionalities. It is simply illustrated at the Figure 1, where width of the edges represents occurrence frequencies of the paths, and – in consequence – amount of data flows over the edges connecting nodes (atomic services) distinguished in the discussed environment.

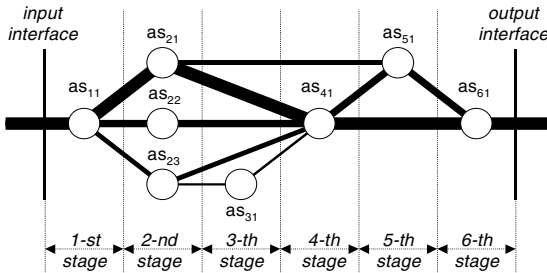


Fig. 1 Different level of paths utilization at the available complex services graphs

The services merging problem is at first limited to the simplest case, i.e., where two atomic services are performed in sequence and the amount of data traffic,

transferred through two processing units delivering appropriate distinguished services (functionalities), is a measure of the services merging capabilities. The discussed attempt is based on assumption that if the two distinguished atomic services are used frequently in gain to assure frequently required complex services, it is worth to combine the two services together. The obtained service, when performed, may limit required amount of required communication and processing resources.

2.1 Merging of Sequential Services

Let us consider a distinguished path at the complex services delivery environment. The path is a sequence of n atomic services (nodes of the complex service structures graph) connected by graph edges. Path's nodes (processing units delivering required services) are described by processing time, which is proportional to the amount of in-flowing data, while the edges, belonging to the paths, are characterized by introduced delay, being a function of amount of transferred data.

Let us consider two adjacent atomic services: $(k - 1)$ -th (as_{k-1}) and k -th (as_k), which are performed one-by-one. The amount of data, processed at the $(k - 1)$ -th node is a sum of data incoming from the $(k - 1)$ -th service (as_{k-1}) and originated at all other possible graph nodes (atomic services) – it is denoted by $f_{in,k-1} + f_{k-2,k-1}$. Some part of the traffic, generated after completing the atomic service as_{k-1} is sent to adjacent – at the path - k -th node ($f_{k-1,k}$), while the rest of the traffic is sent to other atomic services ($f_{k-1,out}$). The amount of data, processed at the k -th unit is a sum of flows from $(k - 1)$ -th service and from the rest of the graph nodes (services) – it is denoted by $f_{in,k}$ (Figure 6).

For purpose of analysis of the considered services, as well as various versions of services merging, it is assumed that the cost of processing of data amount

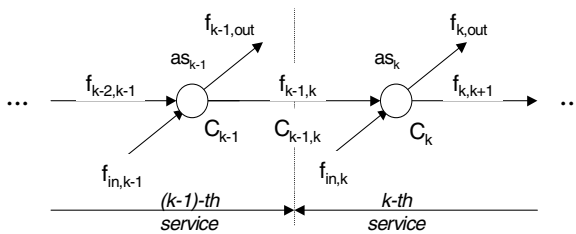


Fig. 2 Data flows through the two separate services

in-coming to the two considered $(k-1)$ -th and k -th nodes (services) and cost of communication between them are measured by processing and communication delays denoted by d_{k-1} , d_k and $d_{k-1,k}$, respectively.

The average delay introduced by the considered tandem of separated services for given amount of data flows (Figure 2), i.e., the cost of delivering $(k-1)$ -th and k -th functionalities (denoted by $\bar{d}_{k-1,k}$) is given by:

$$\begin{aligned} \bar{d}_{k-1,k}(f_{k-2,k-1}, f_{in,k-1}, f_{k-1,k}, f_{in,k}) = \\ \frac{1}{\Lambda_{k-1,k}} \left[(f_{k-2,k-1} + f_{in,k-1}) d_{k-1} (f_{k-2,k-1} + f_{in,k-1}) + \right. \\ \left. f_{k-1,k} d_{k-1,k} (f_{k-1,k}) + (f_{k-1,k} + f_{in,k}) d_k (f_{k-1,k} + f_{in,k}) \right] \end{aligned} \quad (1)$$

where:

$$\begin{aligned} - d_{k-1}(f_{k-2,k-1} + f_{in,k-1}) &= \frac{1}{C_{k-1} - (f_{k-2,k-1} + f_{in,k-1})}, \\ - d_{k-1,k}(f_{k-1,k}) &= \frac{1}{C_{k-1,k} - f_{k-1,k}}, \quad d_k(f_{k-1,k} + f_{in,k}) = \frac{1}{C_k - (f_{k-1,k} + f_{in,k})} \end{aligned}$$

are, respectively, delays introduced by: the $(k-1)$ -th processing unit, the communication channel and the k -th processing unit,

- $\Lambda_{k-1,k} = f_{k-2,k-1} + f_{in,k-1} + f_{k-1,k} + f_{in,k}$ is a total data flows incoming to and processed by the two processing ($(k-1)$ -th and k -th) units as well as flowing through the communication channel connecting the two processing units,
- C_{k-1} and C_k are capacities of the $(k-1)$ -th and k -th processing units assuring as_{k-1} and as_k atomic services functionalities.
- $C_{k-1,k}$ is a capacity of the communication channel between the two units (as_{k-1} and as_k services).

Combining the two separate services as_{k-1} and as_k into one new service $as_{(k-1,k)}$ (Figure 3) means, first of all, that no communication utilities are required to transfer data between the previously mentioned two separate processing units delivering requested services.

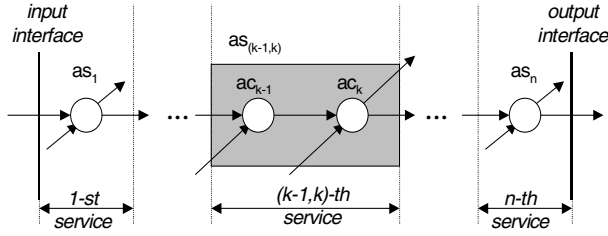


Fig. 3 New $as_{(k-1,k)}$ service obtained by combining ac_{k-1} and ac_k services

Combining the above mentioned services (as_{k-1} and as_k into one $as_{(k-1,k)}$) means also that the amount of data flows – compared with amount of flows to and from previously considered separated services – are changed; the total flow outcoming from the $(k-1)$ -th processing unit, are also processed by the k -th processing unit; so the data flows processed by the k -th processing unit is higher then in previously discussed case by $f_{k-1,out}$ (Figure 4).

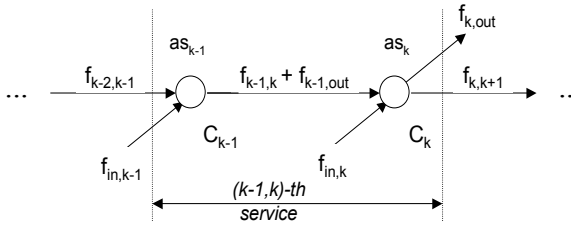


Fig. 4 Data flows through the combine services

The average delay introduced by the combined services for given amount of data flows (Figure 4), i.e., the cost of delivering combined $(k-1)$ -th and k -th functionalities (denoted by $\bar{d}_{(k-1,k)}$) is given by:

$$\begin{aligned} \bar{d}_{(k-1,k)}(f_{k-2,k-1}, f_{in,k-1}, f_{k-1,out}, f_{in,k}) = \\ \frac{1}{\Lambda_{(k-1,k)}} \left[(f_{k-2,k-1} + f_{in,k-1}) d_{k-1}(f_{k-2,k-1} + f_{in,k-1}) + \right. \\ \left. (f_{k-1,out} + f_{in,k}) d_k(f_{k-1,out} + f_{in,k}) \right] \end{aligned} \quad (2)$$

where:

- $\Lambda_{(k-1,k)} = f_{k-2,k-1} + f_{in,k-1} + f_{k-1,out} + f_{in,k}$ is a total data flows incoming to and processed by the two combined processing units, and
- $d_{k-1}(f_{k-2,k-1} + f_{in,k-1}) = \frac{1}{C_{k-1} - (f_{k-1,k-1} + f_{in,k-1})}$,
- $d_k(f_{k-1,out} + f_{in,k}) = \frac{1}{C_k - (f_{k-1,out} + f_{in,k})}$

are delays introduced by the combined $(k-1)$ -th and k -th processing units.

Based on the above expressions, the discussed approaches (separated or combined services) may be compared or selected as dependable on delays difference:

$$\begin{aligned} \bar{\Delta}_{k-1,k} &= \bar{d}_{k-1,k}(f_{k-2,k-1}, f_{in,k-1}, f_{k-1,k}, f_{in,k}) - \\ &\bar{d}_{(k-1,k)}(f_{k-2,k-1}, f_{in,k-1}, f_{k-1,out}, f_{in,k}) \end{aligned} \quad (3)$$

It is rather easy to observe, that the delays difference $\bar{\Delta}_{k-1,k}$ strongly depends on the amount of data exchanged between the two merged services. If the two services are used frequently one after another in required complex services and if most of the data flow outgoing from as_{k-1} incomes to as_k , it is worth to merge the two distinctive services. If the amount of exchanged data are relatively small, it is better to keep the two services separated in gain to obtain as higher as possible resources utilization and to support systems flexibility.

The above discussed delays difference ($\bar{\Delta}_{k-1,k}$) may be applied to decide about atomic services merging or decomposition for known services requirements. The delays difference, treated as a cost of services delivery, may be also extended by owning cost of processing and communication resources.

2.2 Merging of Parallel Services

Let us consider two atomic services ($l1$ -th and $l2$ -th), which are performed in parallel manner (Figure 5).

The amount of data traffic processed at the $l1$ -th processing unit (in gain to assure the $l1$ -th service) is a sum of two data flows: from $(l-1)$ -th service ($f_{l-1,l1}$) and from another services ($f_{in,l1}$). Similarly, the amount of data traffic arriving to the $l2$ -th processing unit is a sum of two data flows: from $(l-1)$ -th service ($f_{l-1,l2}$) and from another services ($f_{in,l2}$).

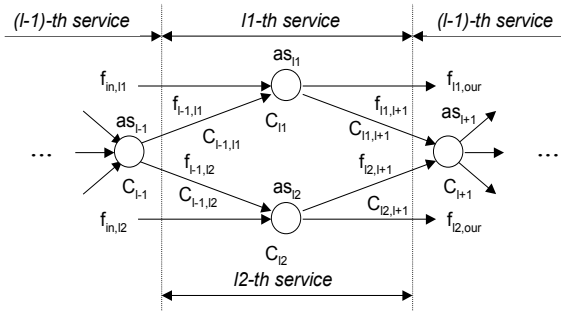


Fig. 5 Data flows for two parallel services

It is assumed, similarly as for previously discussed merging of services performed sequentially, that the cost of processing of data amount in-coming to the considered services and cost of communication among distinguished atomic services are measured by processing and communication delays. The performance of the discussed two services (ac_{l1} and ac_{l2}) are described by the following processing and communication delays: transfer of data between $(l-1)$ -th and $l1$ -th units ($d_{l-1,l1}$), transfer of data between $(l-1)$ -th and $l2$ -th units ($d_{l-1,l2}$), processing at the $l1$ -th units (d_{l1}), processing at the $l2$ -th units (d_{l2}), transfer of data between $l1$ -th and $(l+1)$ -th and units ($d_{l1,l+1}$) and transfer of data between $l2$ -th and $(l+1)$ -th and units ($d_{l2,l+1}$).

The average delay introduced by the considered services, for assumed amount of data flows and resources (Figure 5), i.e., the cost of delivering $l1$ -th and $l2$ -th functionalities (denoted by $\tilde{d}_{l1,l2}$), is given by:

$$\begin{aligned} \tilde{d}_{l1,l2}(f_{l-1,l1}, f_{l-1,l2}, f_{in,l1}, f_{in,l2}, f_{l1,l+1}, f_{l2,l+1}) = \\ \frac{1}{\Lambda_{l1,l2}} [f_{l-1,l1}d_{l-1,l1}(f_{l-1,l1}) + f_{l-1,l2}d_{l-1,l2}(f_{l-1,l2}) + \\ (f_{in,l1} + f_{l-1,l1})d_{l1}(f_{in,l1} + f_{l-1,l1}) + (f_{in,l2} + f_{l-1,l2})d_{l2}(f_{in,l2} + f_{l-1,l2}) + \\ f_{l1,l+1}d(f_{l1,l+1}) + f_{l2,l+1}d(f_{l2,l+1})]. \end{aligned}$$

where:

- $d_{l-1,l1}(f_{l-1,l1}) = \frac{1}{C_{l-1,l1} - f_{l-1,l1}}$, $d_{l-1,l2}(f_{l-1,l2}) = \frac{1}{C_{l-1,l2} - f_{l-1,l2}}$,
- $d_{l1}(f_{l-1,l1} + f_{in,l1}) = \frac{1}{C_{l1} - (f_{l-1,l1} + f_{in,l1})}$,

- $d_{l2}(f_{l-1,l2} + f_{in,l2}) = \frac{1}{C_{l2} - (f_{l-1,l2} + f_{in,l2})}$,
- $d_{l1,l+1}(f_{l1,l+1}) = \frac{1}{C_{l1,l+1} - f_{l1,l+1}}$, $d_{l2,l+1}(f_{l2,l+1}) = \frac{1}{C_{l2,l+1} - f_{l2,l+1}}$,
- $\Lambda_{l1,l2}$ is a sum of data flows incoming to $l1$ -th and $l2$ -th units and outgoing from these units to $(l+1)$ -th unit.

The above expressions may be reduced if it is assumed that the data flows transferred from the $(l-1)$ -th service to $l1$ -th and $l2$ -th services are equal, i.e., $f_{l-1,l1} = f_{l-1,l2}$.

Merging of $l1$ -th and $l2$ -th units, assuring $l1$ -th and $l2$ -th into one l -th service, produces changes in data flows. The possible reduction of the average delay $\tilde{d}_{l1,l2}$ is due to decreasing an amount of data transferred from $(l-1)$ -th unit (Figure 6).

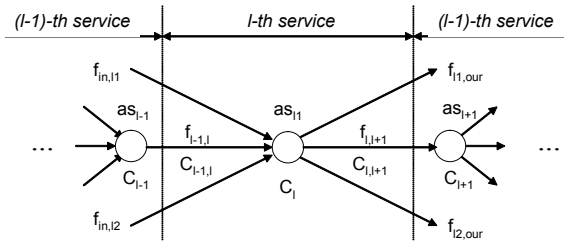


Fig. 6 Data flows for combined parallel services

The performance (communication and processing and communication delay) of the new service as_l , obtained by described above combining as_{l1} and as_{l2} services, is analyzed based on the following assumptions:

- the amount of data generated by the $(l-1)$ -th service satisfies condition: $f_{l-1,l1} = f_{l-1,l2} = f_{l-1,l}$,
- capacities of communication channels linking the $(l-1)$ -th and adjacent services satisfy condition: $C_{l-1,l1} = C_{l-1,l2} = C_{l-1,l}$,
- amount of data generated by as_l service is equal to sum of amounts produced by as_{l1} and as_{l2} services, i.e., $f_{l,l+1} = f_{l1,l+1} + f_{l2,l+1}$,
- capacity of l -th processing unit is equal to sum of $l1$ -th and $l2$ -th units capacities, i.e., $C_l = C_{l1} + C_{l2}$,

- capacities of communication channels linking the $(l+1)$ -th and antecedent services satisfy condition: $C_{l1,l+1} = C_{l2,l+1} = C_{l,l+1}$.
- amount of data incoming to processing unit delivering l -th services, from services other than $(l-1)$ -th service, satisfies equality: $f_{in,l} = f_{in,l} + f_{in,l}$.

The average delay introduced by the considered service ac_l , for given amount of data flows and resources (Figure 6), i.e., the cost of delivering l -th service (denoted by \tilde{d}_l) is given by:

$$\tilde{d}_l(f_{l-1,l}, f_{in,l}, f_{l,l+1}) = \frac{1}{\Lambda_l} [f_{l-1,l}d(f_{l-1,l}) + (f_{l-1,l} + f_{in,l})d_l(f_{l-1,l} + f_{in,l}) + f_{l,l+1}d(f_{l,l+1})] \quad (4)$$

where:

- $\Lambda_l = f_{l-1,l} + f_{in,l} + f_{l,l+1}$, $d_{l-1,l}(f_{l-1,l}) = \frac{1}{C_{l-1,l} - f_{l-1,l}}$,
- $d_l(f_{l-1,l} + f_{in,l}) = \frac{1}{C_l - (f_{l-1,l} + f_{in,l})}$, $d_{l,l+1}(f_{l,l+1}) = \frac{1}{C_{l,l+1} - f_{l,l+1}}$.

Comparing the two delays, i.e., $\tilde{d}_{l1,l2}$ and \tilde{d}_l delays:

$$\tilde{\Delta}_l = \tilde{d}_{l1,l2}(f_{l-1,l1}, f_{l-1,l2}, f_{in,l1}, f_{in,l2}, f_{l1,l+1}, f_{l2,l+1}) - \tilde{d}_l(f_{l-1,l}, f_{in,l}, f_{l,l+1}) \quad (5)$$

it is rather easy to observe, that the considered delays difference strongly depends on the amount of incoming data flows. If the amount of data flow, generated by as_{l-1} service is high enough and it is sent to both as_{l1} and as_{l2} services, merging of the two services into one service may reduce average communication costs.

Value of the delays difference $\tilde{\Delta}_l$ - for known data flows as well as for given processing and communication resources - may be applied select the proper granulation (merged or split) of atomic services for known required set of complex services.

The above presented delays comparison is limited to the delays introduced by assumed service structure for given amount of resources. The comparison is limited to delays, but can be easily extended to comparison where both delays and resources owning cost are taken into account. The performed, simplified comparison of the atomic services performance, shows also that the effective atomic services aggregation (merging) or granulation (slitting) depends on analyzed service system load, i.e. number and structure of complex services performed in the assumed environment uniquely defined by the set of available atomic services.

3 Parallel Services Execution

Consider K atomic services (as_{l1}, \dots, as_{lK}), which can be executed in parallel (Figure 7). Total data flow process by lk -th processing unit is a sum of two data flows: data flow $f_{l-1,lk}$ from $l-1$ -th unit and $f_{in,lk}$ from other units. The capacity of lk -th processing unit is denoted as C_{lk} . Total data flow generated by lk -th unit is a sum of data flow, which is delivered to $l+1$ -th processing unit ($f_{lk,l+1}$) and to other units ($f_{lk,out}$). The capacities of communication channels linking $l-1$ -th service and lk -th services are denoted as $C_{l-1,lk}$. The capacities between lk -th services and $l+1$ -th service are represented by $C_{lk,l+1}$.

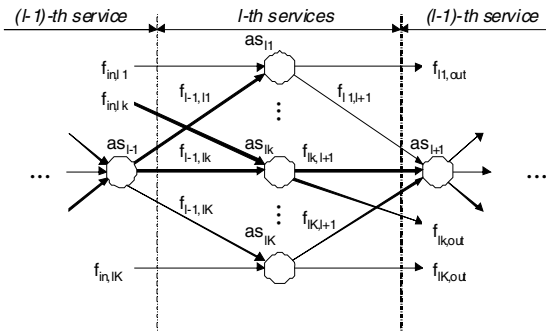


Fig. 7 Data flows for K parallel services

The main problem of parallel service execution is to find optimal number of services, which can be executed in parallel according to chosen criterion. The initial number of services executed in parallel (K) can be reduced to optimal number (k^*) in services merging process. As a criterion we consider average delay, which is in this case composed of: communication delays between $l-1$ -th service and lk -th services, processing delays on lk -th units and communication delays between lk -th services and $l+1$ -th service.

Assume following conditions to be satisfied:

- the amount of data generated by the $(l-1)$ -th service satisfies condition: $f_{l-1,l1} = \dots = f_{l-1,lk} = \dots = f_{l-1,lK} = f_{l-1,l}$
- the amount of additional data on the input of each lk -th services satisfies condition: $f_{in,l1} = \dots = f_{in,lk} = \dots = f_{in,lK} = f_{in,l}$

- the amount of data returned by each of lk -th services is equal:

$$f_{l1,l+1} = \dots = f_{lk,l+1} = \dots = f_{lK,l+1} = f_{l,l+1}$$

Additionally, the assumptions about capacities in parallel merging services given in section 2.2 are also present in this case. The average delay of presented on Figure 7 services configuration is equal:

$$\tilde{d}_{lK}(f_{l-1,l}, f_{in,l}, f_{l,l+1}) = \frac{I}{A_{lK}} \left[K \frac{f_{l-1,l}}{C_{l-1,l} - f_{l-1,l}} + K \frac{(f_{l-1,l} + f_{in,l})}{C_l - (f_{l-1,l} + f_{in,l})} + K \frac{f_{l,l+1}}{C_{l,l+1} - f_{l,l+1}} \right] \quad (6)$$

where:

- $A_{lK} = Kf_{l-1,l} + Kf_{in,l} + Kf_{l,l+1}$

The average delay can be simplified to the following formula representing average delay on single lk -th unit:

$$\tilde{d}_{lK}(f_{l-1,l}, f_{in,l}, f_{l,l+1}) = \frac{I}{f_{l-1,l} + f_{in,l} + f_{l,l+1}} \left[\frac{f_{l-1,l}}{C_{l-1,l} - f_{l-1,l}} + \frac{(f_{l-1,l} + f_{in,l})}{C_l - (f_{l-1,l} + f_{in,l})} + \frac{f_{l,l+1}}{C_{l,l+1} - f_{l,l+1}} \right] \quad (7)$$

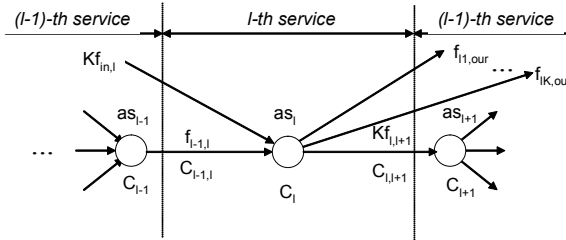


Fig. 8 Data flows after merging K parallel services

The K services in parallel services can be merged to one service (see Figure 8). The average delay introduced by the considered combined service as_l is as follows:

$$\tilde{d}_{l1}(f_{l-1,l}, f_{in,l}, f_{l,l+1}) = \frac{I}{f_{l-1,l} + Kf_{in,l} + Kf_{l,l+1}} \left[\frac{f_{l-1,l}}{C_{l-1,l} - f_{l-1,l}} + \frac{(f_{l-1,l} + Kf_{in,l})}{KC_l - (f_{l-1,l} + Kf_{in,l})} + \frac{Kf_{l,l+1}}{C_{l,l+1} - Kf_{l,l+1}} \right] \quad (8)$$

The decision about merging K services into one service in most cases is pointless due to increasing amount of data transfer using communication channel between services as_l and as_{l+1} (The constant value of communication capacity $CS_{l,l+1}$ was assumed in this case and the data flow sent in this channel increases K times). It is more accurate to merge two parallel services in each step and monitor the values of average delay. The problem of finding optimal number of parallel services can be presented as simple optimization task:

$$k^* = \arg \min_k \tilde{d}_{lk}(K, k, f_{l-1,l}, f_{in,l}, f_{l,l+1}, \Gamma) \tag{9}$$

where:

- K is initial number of services executed in parallel,
- Γ is a merging policy,
- k is the number of services executed in parallel after applying merging process. For instance, if an initial number of services executed in parallel K is equal 5 and merging process were made two times as a result three services are executed in parallel ($k = 3$).

The value of average delay strongly depends on merging policy (Γ). Different values of average delays can be obtained for the same number of services if different merging policies are applied in optimization process. Two marginal policies of merging parallel services can be distinguished for taken assumptions: merging two services (two merging candidates) with the lowest values of sum data flows incoming to considered services and outgoing from them to $l + 1$ -th service (balanced merging, Figure 9, a) and merging two services with highest incoming and outgoing data flows (incremental merging, Figure 9 b)).

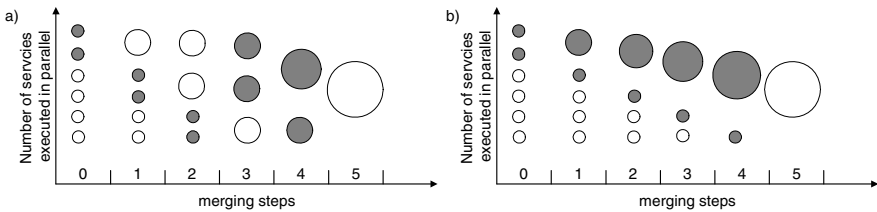


Fig. 9 Two merging polices in parallel services processing: a) Balanced merging and b) incremental merging

For stated assumptions in first policy (Figure 9, a)), in each iteration those two services are merged, which were merged rarely in previous iterations. Considering second policy, in each iteration two services are merged, which were merged most often in previous iterations. It seems natural, that it is better to use first policy for merging parallel services, because in each step those two services are merged, for

which sum of incoming and outgoing data flow for newly created service is minimal and as a consequence, the data flows incoming to $l + 1$ -th service are more balanced.

Assume, that initial number of services executed in parallel K is equal 29 and incoming and outgoing data flows are known ($f_{l-l} = 1, f_{in,l} = 3, f_{l,l+1} = 0.5$). Communication and processing capabilities are also given ($C_{l-l} = 5, C_l = 15, C_{l,l+1} = 15$). We are interested in finding optimal number of services for which the average delay is the lowest. We consider two different merging policies described in previous section.

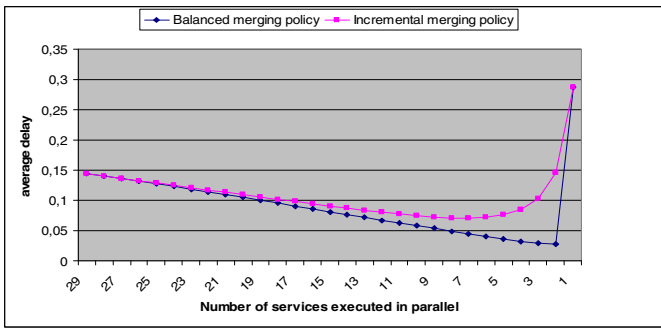


Fig. 10 Average delays for different numbers of services executed in parallel considering two politics of merging

Figure 10 illustrates the average delays for two merging polices. It can be observed that the first, balanced merging policy gives lower average delays for all numbers of services executed in parallel. If the data flows in channels linking l -th services and $l + 1$ service are very close to capacities of that channels the average delay increases rapidly. The lowest average delay value were achieved for 2 services executed in parallel ($k^* = 2$).

In some telecommunication systems the loss of some capacity of processing units may occur while merging two services. Consider two services, for which capacities of processing units are denoted: C_{lk1} and C_{lk2} . Then the capacity of merged services $C_{l(k1,k2)}$ is equal:

$$C_{l(k1,k2)} = \alpha(C_{lk1} + C_{lk2}) \tag{10}$$

where $1 - \alpha$ is percentage loss of total capacity of processing units in merging process.

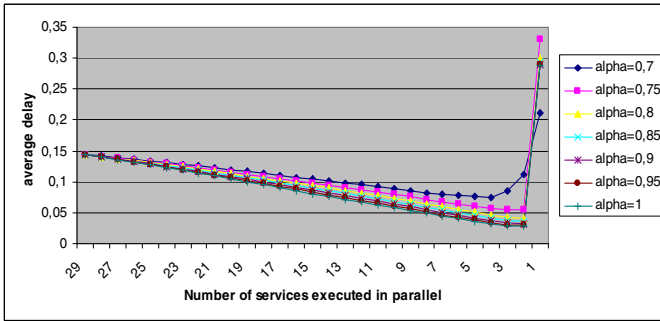


Fig. 11 Average delays for different numbers of services executed in parallel for different alpha values (for balanced policy of merging)

In Figure 11 average delays for different processing capacity loss scenarios are presented (assuming balanced policy of merging). The lowest value of average delay is observed if there is no loss of capacity ($\alpha = 1$). The optimal number of parallel executed services for 30 % percent loss of processing capacity ($\alpha = 0,7$) is equal 4 ($k^* = 4$).

4 Conclusions

The primary goal of the paper was to discuss circumstances under which it is worth to perform merged services instead of applying them as a set of separated services in distributed environment. Merging and splitting of atomic (elementary) services is a natural question; the gain is to find equilibrium between services flexibility and resources utilization efficiency. It is rather obvious that increasing services granulation (splitting) level leads to increasing flexibility, which is paid by increasing communication costs, mostly determined by the amount of data flows, which have to be exchanged among processing units delivering requested and required functionalities. In most cases, the required functionality should be delivered satisfying some assumed non-functionality features (delay, security, reliability, etc.). Presented approach is based on assumption that the services quality depends on delays introduced by available resources for data flows characterizing services requests followed by specified requirements. Performed comparison of services qualities, offered in merged and split services modes, shows that the quality of complex services, assured by set of atomic services available in distributed environment, strongly depends on structures and functionalities of the requested services. The presented, simplified analysis of services quality and resources utilization, can be easily extended to cases, where other, more sophisticated and realistic services quality measures are applied. Further research is devoted to apply different services quality as well as resources utilization measures.

Acknowledgment

The research presented in this paper has been partially supported by the European Union within the European Regional Development Fund program no. POIG.01.03.01-00-008/08.

References

- [Anderson et al. 2005] Anderson, S., Grau, A., Hughes, C.: Specification and satisfaction of SLAs in service oriented architectures. In: 5th Annual DIRC Research Conf., pp. 141–150 (2005)
- [Garey et al. 1976] Garey, M., Johnson, D., Sethi, R.: The complexity of flowshop and job-shop scheduling. *Mathematics of Operations Research* 1, 117–129 (1976)
- [Graham et al. 1979] Graham, R.L., Lawler, E.L., Lenstra, J.K., Rinnooy Kan, A.H.G.: Optimization and approximation in deterministic sequencing and scheduling: a survey. *Annals of Discrete Mathematics* 3, 287–326 (1979)
- [Grzech 2002] Grzech, A.: *Teletraffic control in the computer communication networks*. Wrocław University of Technology Publishing House (2002) (in Polish)
- [Grzech and Świątek 2008] Grzech, A., Świątek, P.: Parallel processing of connection streams in nodes of packet-switched computer communication networks. *Cybernetics and Systems* 39(2), 155–170 (2008)
- [Grzech and Świątek 2009] Grzech, A., Świątek, P.: The influence of load prediction methods on the quality of service of connections in the multiprocessor environment. *Systems Science* 35(3) (2009) (in press)
- [Jaeger et al. 2005] Jaeger, M.C., Rojec-Goldmann, G., Muhl, G.: QoS aggregation in web service compositions. In: *IEEE Int Conf. on e-Technology, e-Commerce and e-Service*, pp. 181–185 (2005)
- [Johnson et al. 1995] Johnson, R., Gamma, E., Helm, R., Vlisides, J.: *Design patterns; elements of reusable object-oriented software*. Addison-Wesley, Reading (1995)
- [Milanovic and Malek 2004] Milanovic, N., Malek, M.: Current solutions for web service composition. *IEEE Internet Computing* 8(6), 51–59 (2004)
- [Narayanan and McIlraith 2003] Narayanan, S., McIlraith, S.: Analysis and simulation of web services. *Computer Networks: The International Journal of Computer and Telecommunications Networking* 42(5), 675–693 (2003)

Importance Measures in Reliability Analysis of Healthcare System

E. Zaitseva

Department of Informatics, University of Zilina, Zilina, Slovakia
Elena.Zaitseva@fri.uniza.sk

Abstract. Healthcare system is modern complex system that includes four principal components in point of view of reliability engineering. They are hardware, software, human factor and organization component. There are different methods in reliability engineering for analysis and quantification of every of these components. But new tendency in reliability analysis needs methods that evaluate the system as a single whole. In accordance with this tendency one aspect of reliability engineering (importance analysis) is considered in the paper. The importance reliability analysis allows to estimate influence of every healthcare system component to the system reliability and functioning. New algorithms for importance analysis of healthcare system are proposed in this paper.

1 Introduction

Reliability, availability and performance are importance properties of any modern technological system (manufacturing, telecommunication, pattern recognition, power etc.). These properties depend on the combinational of number of interrelated processes of component degradation of failure and repair, of diagnostics and maintenance, which result from the interaction of different part including not only the hardware but also the software, the human and the organizational system parts. Reliability engineering allows to analyse and to estimate system properties as reliability, availability and performance. Reliability engineering methods aim at the quantification of the probability of failure of the system and its working.

Modern system is complex and includes different parts that demands special knowledge and methods in reliability engineering [Pham 2003]. For example, specified methods allow to examine hardware, other methods are used for quantification of software. There are special methods and algorithms for analysis of human and organization factors, and maintenance influence to system performance. E.Zio in paper [Zio 2009] has been defined four principal parts (components) for modern real-world system: hardware, software, organizational and human. Reliability engineering was originally developed to handle rationally the failures of the components of the first and second types. But the experience accumulated on

occurred industrial accidents in the last few decades has clearly shown that the organizational and human factors play a significant role in the risk of system failures and accidents. This is due also to the fact that the reliability of the hardware components utilized in technological systems has significantly improved in recent years [Zio 2009]. As a consequence, the relative importance of the errors of the organizations managing the systems and of the human operators running them on the risks associated to the operation of these systems has significantly increased. This explains the significant focus on Organizational and Human Reliability Analysis (HRA) and on its full integration within systematic risk analysis and reliability assessment procedures [Zio 2009; Lyons et al. 2004]. All these aspects in here in healthcare system too.

Initial reliability engineering methods for healthcare system have been considered more 30 years ago in paper [Taylor 1972]. E.F.Taylor declared principal items of reliability engineering in a healthcare system as reliability analysis of medical equipments and devices. Reliability quantification of equipment and devices has been principal tendency in medicine until recently. Information technology development causes application new type of healthcare systems that consist of two interdependent components as hardware and software [Cohen 2004]. Therefore new methods for shared analysis of hardware and software part of a healthcare system have been developed. One of these methods has been presented in paper [Taleb-Bendiab et al. 2006]. But human errors problem in a healthcare system has been considered as independent problem of reliability analysis [Lyons et al. 2004]. There are some investigations that propose to examine a health care system as system of two principal components: the first is technical component that include equipments and devices with hardware and software, and the second is human factor. For example, B.S.Dhillon in [Pham 2003] has been considered this type of healthcare system interpretation. But a healthcare system as complex system of four components (hardware, software, organizational and human) is not considered in reliability engineering until now.

In this paper one of possible ways of healthcare system interpretation as complex system under [Zio 2009] is considered. Reliability quantification of such system is implemented based on the reliability importance analysis. Importance analysis allows to estimate an influence of different system component functioning or failure to the system reliability (availability, performance) change [Wang et al. 2004]. New method for importance analysis of healthcare system is proposed in the paper. According to this method the investigated system includes technical components (hardware and software), human factor and organization component, and is interpreted as *Multi-State System* (MSS). This interpretation of system allows to investigate different performance levels of system functioning, that didn't include level of functioning and fault only. This method has been developed based on the results presented in papers [Zaitseva 2009]. New method for quantification of MSS reliability has been proposed in [Zaitseva 2009]. In paper [Zaitseva 2010]

this methods has been adapted for analysis of human factor. Basic aspects of application this result to healthcare system reliability analysis has been considered in [Zaitseva 2010]. New variant of mathematical model of healthcare system is considered in this paper and is analysed below. Evaluation of different variants of mathematical model of healthcare system allows to obtain real and exact measures of reliability. Some new measures for reliability analysis of healthcare system are proposed in this paper too.

The paper is organized as follows. In the following Section 2, the basic conception of healthcare system reliability analysis is introduced. Then Section 3 presents mathematical model of healthcare system based on MSS and new method for importance analysis of this system. Numerical example is provided in Section 4.

2 Reliability Analysis of Healthcare System

Healthcare system (for example, magnetic resonance imaging scanners, remote patient monitoring systems, surgical robotics, telemedicine systems, etc.) is complex system that includes four principal components (Fig.1): hardware, software, human factor and organization components. There are many methods and algorithms for independent analysis and quantification of each of these components in reliability engineering. The hardware component and software component conform to hardware and software parts of medical devices and equipments. The human component of the system models a physician work and conforms to probability of medical errors. The organization component of the system unites management and maintainability aspects of the healthcare system. But there are some difficulties for presentation of a real-word healthcare system based on four principal components only. Some time definition of hardware and software components as original components is difficult problem because both have to be part of single device or equipment. In this case they can be united in one component that is named as “technical component”. For example, technical component consists of special devices and standards-based devices in the structure of healthcare system according to [Pham 2003; Zaitseva 2010]. The first type of these devices is a medical device based on personal computers. For example, it is the medical decision support system, system for integration electronic medical records or picture archiving communication systems. The second type is a special medical device that can be used for special operation only (as magnetic resonance imaging scanners, for example). It is possible to define other interpretation of basic structure of a healthcare system, but principal condition of successful reliability analysis of such system is united evaluation of (a) system technical part (hardware, software, special and standards-based devices) and (b) social part (human factor, organization process).

Definition of a healthcare system typical structure is the first part of reliability analysis problem in medicine. The second part of this problem is caused by diversity of mathematical conceptions that are used for reliability analysis of each

system component. A lot of methods and algorithms for analysis basic components of healthcare system allow to evaluate they in detail. But these methods and algorithms based on different methodology and mathematical conceptions for different system components and therefore in some situation there are problems for united quantification of healthcare system performance. For example, analyzable system or its component can be presented as Markovian model or as structure function that defines system performance for every combination of system components states. Therefore the actual problem of reliability analysis in medicine is development new methods for estimation every system component based on common mathematical methodology. Decision of this problem (evaluation and quantification of system components reliability, availability and dependability) includes four steps (Fig. 2):

- Formalization of analyzable system;
- Definition of mathematical model for system description;
- Choice methodology conception for performance analysis;
- Definition of analysis aspects (definition of reliability indexes and measures for system evaluation).

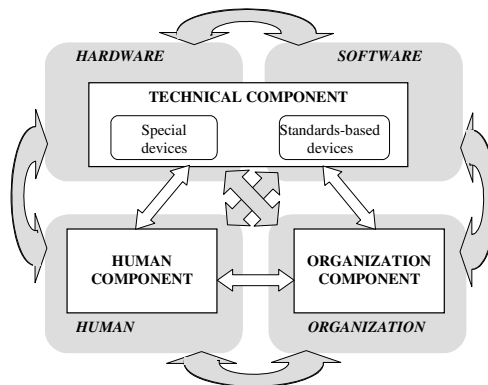


Fig. 1 Typical structure of a healthcare system for reliability analysis

The first step declares a healthcare system structure and allows to define basic components of the system for performance or reliability analysis.

According to the second step the mathematical model for a healthcare system representation is determine. There are two principal models in reliability engineering for investigated of object description. It is Binary-State System or *Multi-State system* (MSS). The system and its components are allowed to have only two possible states (completely failed and perfect functioning) in a Binary-State System. This approach is well known in Reliability Analysis, but can prevent the examination of many situations where the system can have more than two distinct states [Zio 2009]. MSS reliability analysis is a more flexible approach to evaluate

system reliability, as it can be used when both the system and its components may experience more than two states, to include, completely failed, partially failed, partially functioning and perfect functioning. The MSS scientific achievement has been documented in [Lisnianski and Levitin 2003; Pham 2003; Zio 2009].

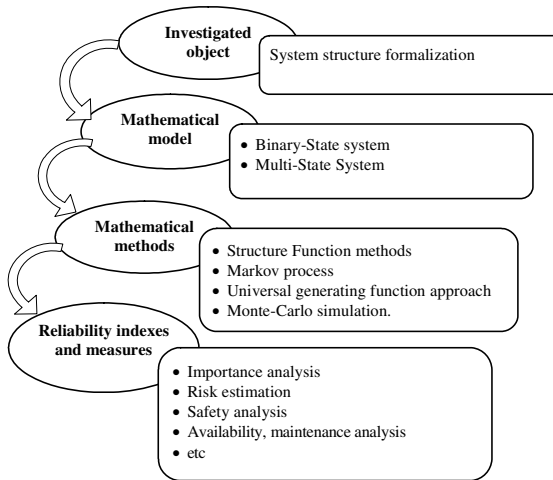


Fig. 2 Scheme of the healthcare system performance analysis

The third step determines basic mathematical methods for system analysis. There are four mathematical conceptions for elaboration of methods in reliability engineering (Fig.2). For example, Markov processes are used to analyze the system state transition process and the structure function approach is used to investigate the system topology. The Universal generating function approach allows to evaluate system of large dimension. Monte-Carlo simulation is effective mathematical tools but need high computation resources. Combination of different mathematical conception causes elaboration of proficient methods for system reliability analysis [Zio 2009; Pham 2003].

The last step in performance analysis of the healthcare system permits to define some necessary indices and measures system evaluation. One type of possible measures is importance measures that are probabilities of system performance state changes depending on change of system component state [Lisnianski and Levitin 2003].

The importance analysis is one of reliability engineering problems. An important problem in reliability engineering is to evaluate the relative importance of the various elements and components comprising the system. Indeed, the identification of which element mostly influence the overall system performance allows one to trace technical bottlenecks and provides guidelines for effective actions of system improvement. In this sense, importance measures are used to quantify the contribution

of individual elements to the system performance (that means by reliability, availability, risk). Importance measures quantify the criticality of a particular component within a system design. They have been widely used as tools for identifying system weaknesses, and to prioritise reliability improvement activities.

There are many methods and algorithms for calculation of importance measures. Some time different mathematical conceptions in these methods and algorithms don't compare calculated importance measures. Therefore effective and successful analysis of the system performance needs elaboration of algorithms for calculation of importance measures based on united mathematical conception.

3 Importance Analysis

Importance analysis is part of reliability analysis. The importance concept and the first importance measure were introduced by L.W.Birnbaum. The concept presumed orderly arrangement of components in a system, that some of the components are more important than others in providing certain system characteristics. Besides assisting system design and optimisation, importance analyses are useful for diagnosing failures and generating repair checklists. Some examples of importance analysis effective application have been presented in papers [Aven and Nokland 2010; Marseguerra and Zio 2004; Pham 2003]. First of all importance analysis has to answer to questions: How does a change in one component affect system or How can system reliability be best improved (which components should be upgraded firstly)?

To address such question, there are definition of importance and *Importance Measure* (IM). IM quantifies the criticality of a particular component within a system. They have been widely used as tools for identifying system weaknesses, and to prioritise reliability improvement activities. According to papers [Fricks and Trivedi 2003; Zaitseva 2009] the most-used IM is *Structural Importance* (SI), *Criticality Importance* (CI), *Birnbaum importance* (BI), *Fussell-Vesely importance* (FVI), *Component Dynamic Reliability Indices* (CDRI) and *Dynamic Integrated Reliability Indices* (DIRI). These measures allow to investigate different aspect of influence of component functioning to system reliability. There are some techniques for calculation of IMs based on different mathematical conceptions (Fig. 2), for example, as Markov Chains [Fricks and Trivedi 2003], Monte Carlo simulation [Marseguerra and Zio 2004], Logical Differential Calculus [Zaitseva 2009] or other [Aven and Nokland 2010]. The mathematical model of analysed system is one of principal conditions for definition of IM calculation technique. Consider the mathematical model and measures for importance analyses of Healthcare systems below.

3.1 Mathematical Model

A healthcare system is complex, whose overall performance can settle on different levels (e.g. 100%, 80%, 50% of the nominal capacity), depending on the operative conditions of their constitutive components. Therefore MSS is more preferable model for mathematical representation and quantification of healthcare system reliability.

There are different methods for MSS importance analysis. One of them is based on the system behaviour description by structure function. In this case system is interpreted as MSS of n component, and MSS and each of n elements can be in one of m possible states: from the complete failure (it is 0) to the perfect functioning (it is $m-1$). Every system component x_i is characterized by probability of the performance rate (component state):

$$p_{i,s} = \Pr\{x_i = s\}, \quad s = 0, \dots, m-1 \tag{1}$$

The structure function defines correlations between MSS performance level and different components states:

$$\phi(x_1, \dots, x_n) = \phi(\mathbf{x}): \{0, \dots, m-1\}^n \rightarrow \{0, \dots, m-1\}. \tag{2}$$

The following assumptions are used for structure functions (2) [Lisnianski and Levitin 2003; Zaitseva 2009]: (a) it is the *Multiple-Valued Logic* (MVL) function; (b) the structure function is monotone and $\phi(s) = s$ ($s \in \{0, \dots, m-1\}$); (c) all components are s-independent and are relevant to the system.

The assumption (a) is important to exploit the mathematical tools of MVL for the reliability analysis. For example, Direct Partial Logic Derivatives are used in importance analysis of MSS [Zaitseva 2009]. Direct Partial Logic Derivatives are part of Logic Differential Calculus and are used for analysis of dynamic properties of MVL function. Basic conception of MSS reliability analysis by Logic Differential Calculus has been considered in [Zaitseva 2009] and Direct Partial Logic Derivatives have been proposed for this analysis, because these derivatives reflect the change in the value of the underlying function when the values of variables change and can be applied for analysis of dynamic behaviour of MSS that is presented as the structure function (2) according to the assumption (a).

Direct Partial Logic Derivative with respect to variable x_i for a MSS structure function permits to analyse the system reliability change from j to h when variable value changes from a to b [Zaitseva 2009]:

$$\partial\phi(j \rightarrow h) / \partial x_i(a \rightarrow b) = \begin{cases} m-1, & \text{if } \phi(a_i, \mathbf{x}) = j \text{ and } \phi(b_i, \mathbf{x}) = h \\ 0, & \text{in the other case} \end{cases}, \tag{3}$$

where $\phi(a_i, \mathbf{x}) = \phi(x_1, \dots, x_{i-1}, a, x_{i+1}, \dots, x_n)$; $\phi(b_i, \mathbf{x}) = \phi(x_1, \dots, x_{i-1}, b, x_{i+1}, \dots, x_n)$; $a, b \in \{0, \dots, m-1\}$.

In paper [Zaitseva 2009; Zaitseva 2010] different types of MSS component changes and their analysis by Direct Partial Logic Derivative (3) has been considered in details. Proposed method for MSS importance analysis based on Direct Partial Logic Derivative technique allows to compute different IMs including known measures [Zaitseva 2009]. And this method can be used for different investigated objects and human factor too [Zaitseva 2010].

3.2 MSS Reliability Function

Reliability function for Binary-State System is defined as probability of system function without failure during given period of time. But for MSS the reliability function has some interpretation.

In paper [Lisnianski and Levitin 2003] MSS reliability function $R(t)$ is the probability of the system being operational throughout the interval $[0, t)$:

$$R(t) = \Pr\{T \geq t, \phi(\mathbf{x}) > 0\}.$$

In some fixed time t the reliability function $R(j)$ can be interpreted as:

$$R = \Pr\{\phi(\mathbf{x}) > 0\}.$$

But for MSS there are some level of system performance and reliability analysis of this system needs to include estimation of probability of system to be in every of these performance state. Therefore some definitions of reliability function for MSS have been proposed. One of them allows to presented probability of MSS to be in state, that isn't less than performance level j ($m-1 \geq j \geq 0$):

$$R(j) = \Pr\{\phi(\mathbf{x}) \geq j\}.$$

There are one more interpretation of MSS reliability function, when MSS reliability function is defined as probability of system reliability that is equal to the performance level j :

$$R(j) = \Pr\{\phi(\mathbf{x}) = j\}, \quad j = 1, \dots, m-1. \quad (4)$$

Note, MSS unreliability is

$$F = R(0) = 1 - \sum_{j=1}^{m-1} R(j). \quad (5)$$

3.3 Importance Measures of MSS

Importance analysis allows to estimate influence of every system component state changes to MSS reliability (performance level). The possibility of system failure is investigated previously to obtain information about system component with maximal and minimal influence for MSS unavailability. Consider principal types of IMs for MSS quantification.

SI is one of the simplest measures of component importance and it concentrates on the topological structure of the system. This measure determines the proportion of working states of system in which the working of the i -th component makes the difference between system failure and its working [Zaitseva 2009]:

$$I_s(x_i) = \frac{\rho_i}{m^{n-1}} \quad (6)$$

where ρ_i is number of system states when the breakdown of the i -th system components results the system failure and this number is calculated as numbers of nonzero values of Direct Partial Logic Derivative (3) $\partial\phi(1 \rightarrow 0)/\partial x_i(1 \rightarrow 0)$.

There is one more definition of SI [Zaitseva 2009]. It is modified SI that determines influence of the i -th system component breakdown to MSS failure:

$$I_{s_m}(x_i) = \frac{\rho_i}{\rho_i^{(1,1)}} \quad (7)$$

where $\rho_i^{(1,1)}$ is number of system states when $\phi(1_i, \mathbf{x}) = 1$ (it is computed by structure function of MSS).

Therefore SI $I_s(x_i)$ conforms to probability of MSS failure among all possibility system state caused by breakdown of the i -th system component. Modified SI $I_{s_m}(x_i)$ is probability of MSS if the i -th component fails. A system component with maximal value of the SI measure ($I_s(x_i)$ and $I_{s_m}(x_i)$) has most influence to MSS failure or this component failure causes high possibility of MSS failure.

SI measures aren't dependent on components state probability (1) and characterize only topological aspects of MSS performance. These measures are used for prevention system analysis or reliability analysis in step of a system design [Fricks and Trivedi 2003; Zaitseva 2010].

BI of a given component is defined as the probability that such component is critical to MSS functioning [Fricks and Trivedi 2003]. This measure represents loss in the MSS when the i -th component switches to state below a . Consider BI for estimation of MSS failure (system reliability level changes from "1" to "0"):

$$I_B(x_i) = |\Pr\{\phi(1_i, \mathbf{x}) = 1\} - \Pr\{\phi(0_i, \mathbf{x}) = 1\}| \quad (8)$$

BI can be used to evaluate the effect of an improvement in component state on system functioning. BI larger value has system component that is more critical for MSS performance level (reliability or availability).

Note that BI measure (8) of the i -th component only depends on the structure of the system and states of the other components, but is independent of the actual state of the i -th component. But CI allows to remove this imperfection of BI (8). According to [Fricks and Trivedi 2003] CI is the probability that the i -th system component is relevant to MSS functioning at time t and has failed before time t :

$$I_C(x_i) = I_B(x_i) \cdot \frac{P_{i,0}}{F}, \quad (9)$$

where $I_B(x_i)$ is the i -th system component BI measure (8); $p_{i,0}$ is probability of the i -th system component failure (1) and F is probability of system failure (system unreliability) (5).

FVI quantifies the maximum decrement in MSS reliability caused by the i -th system component state deterioration [Fricks and Trivedi 2003]. This measure for MSS failure can be defined as:

$$I_{FV}(x_i = s_i) = 1 - \frac{\Pr\{\phi(s_i, \mathbf{x}) = 1\}}{R(1)} \quad (10)$$

where $s_i = \{0, 1, \dots, m-2\}$ and if $s_i = 0$ the measure (10) allows to estimate system performance level decrease for full unreliability of the i -th system component; $R(1)$ is probability of system work with performance level "1" and is defined in (4).

There is one more type of IMs for MSS that are DRIs. These measures have been defined in paper [Zaitseva 2009]. DRIs allow to estimate a system component relevant to MSS and to quantify the influence of this component state change to the MSS performance. There are two groups of DRIs: Component Dynamic Reliability Indices (CDRIs) and Dynamic Integrated Reliability Indices (DIRIs).

CDRI estimates the influence of the i -th component state change to MSS and is probability of MSS performance change depending on the i -th component state change. Consider CDRIs for MSS failure. In this case the CDRI is probability of MSS failure that is caused by the i -th system component state decrease. But for a coherent MSS the i -th system component breakdown cause MSS failure [Zaitseva 2009]. Therefore CDRIs for MSS failure take into consideration the probability of MSS failure provided change of the i -th component state form working to failed and the probability of inoperative component state:

$$I_{CDRI}(x_i) = I_{S_m}(x_i) \cdot p_{i,0} \quad (11)$$

where $I_{S_m}(x_i)$ is the modified SI (7) and $p_{i,0}$ is probability of component breakdown that is declared in (1).

DIRI is the probability of MSS failure that caused by the one of system components state breakdown. DIRIs allow to estimate probability of MSS failure caused by some system component (one of n):

$$I_{DIRI} = \sum_{i=1}^n I_{CDRI}(x_i) \prod_{\substack{q=1 \\ q \neq i}}^n (1 - I_{CDRI}(x_q)) \quad (12)$$

where $I_{CDRI}(x_i)$ is calculated by (11).

4 Example of Importance Analysis of Healthcare System

Consider example of some healthcare system as of the Decision Support System for Preliminary Diagnostics in Oncology [Zaitseva 2010]. One of possible interpretation of this system according to the conception of complex system reliability analysis in [Zio 2009] is in Fig.3. This system structure includes three principal

components (Fig. 3): technical component, human component and organization component. The technical component consists of three elements as the special devices, hardware and software. Each of these elements needs more detail reliability analysis by special methods. Human component joins human factors that have to influence to the decision about patient diagnosis and treatment (for example, it can be medical errors, patient mistake in description of his state etc.). Organization component presents different management decision for the system function, safety and maintenance. Need to say that these management decisions have not to influence to the technical part only but to human factor too. There are other interpretations of this system structure that has been presented in [Zaitseva 2010]. These differences in the system structure interpretation are caused by influence of personal expert knowledge to the system formalization in the first step of reliability analysis (Fig. 2). The system structure formalization depends on experiences specialists in reliability engineering, as well as medics. And there are not strict rules for initial system representation therefore reliability analysis provides for some variants of system structure so that to increase analysis precision.

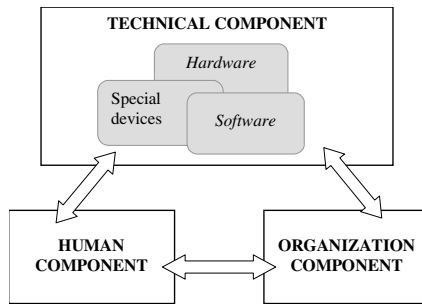


Fig. 3 Typical structure of a healthcare system for reliability analysis

Use the MSS mathematical model for description and quantification of the system in Fig.3. The system consists of three components ($n = 3$) and has three levels of performance ($m = 3$). The component probabilities in Table 1 have been determined for this system by the expertise. In Table 1 the system component state “0” considers to the component failure; the component state “1” is component functioning with some unimportant restriction; the component state “2” is perfect functioning. Implement the importance analysis of this system by (4) – (12).

Implementation of importance analysis needs structure function of the system. There are some definitions of the structure function for this system because this function is constructed based on the expert knowledge [Zaitseva 2010]. One of variants that has been proposed by experts for reliability quantification of the system is in Table 2. Consider importance analysis for the Decision Support System for Preliminary Diagnostics in Oncology based on the structure function in Table 2, where x_1 is performance state of the technical component; x_2 and x_3 is performance state of the human and organization components of the system.

Table 1 Component probabilities

m	$p_{1,s}$	$p_{2,s}$	$p_{3,s}$
0	0.1	0.2	0.3
1	0.3	0.3	0.4
2	0.6	0.5	0.3

According to the structure function in Table 2 and component state probabilities in Table 1 the reliability function of this system is defined by (4) and for two working performance levels is: $R(1) = 0.503$ and $R(2) = 0.309$. The system unreliability is calculated based on (5) and is $F = 0.188$. So the total probability of the Decision Support System for Preliminary Diagnostics in Oncology failure is 0.188 and probability of the perfect working of this system is 0.503. The probability 0.309 conforms to the system functioning with insignificant complications.

Table 2 Structure function of the healthcare system in Fig.2

$x_1x_2x_3$	$\phi(x)$	$x_1x_2x_3$	$\phi(x)$	$x_1x_2x_3$	$\phi(x)$
0 0 0	0	1 0 0	0	2 0 0	0
0 0 1	0	1 0 1	0	2 0 1	1
0 0 2	0	1 0 2	0	2 0 2	1
0 1 0	0	1 1 0	0	2 1 0	1
0 1 1	0	1 1 1	1	2 1 1	1
0 1 2	0	1 1 2	1	2 1 2	2
0 2 0	0	1 2 0	1	2 2 0	1
0 2 1	1	1 2 1	1	2 2 1	2
0 2 2	1	1 2 2	2	2 2 2	2

Importance measures (6)-(12) allow to estimate different aspects of particular system component influence to the system performance level (reliability). These measures the Decision Support System for Preliminary Diagnostics in Oncology (Fig.3) are in Table 3.

Table 3 Importance measures for the system in Fig.2

i	Importance measures							
	$I_s(x_i)$	$I_{s_m}(x_i)$	$I_B(x_i)$	$I_C(x_i)$	$I_{FV}(x_i=1)$	$I_{FV}(x_i=0)$	$I_{CDRI}(x_i)$	I_{DIRI}
1	0.330	1.000	0.133	0.071	0.573	0.930	0.100	0.490
2	0.330	1.000	0.105	0.112	0.519	0.833	0.200	
3	0.330	0.600	0.047	0.075	0.412	0.624	0.180	

For this system according to SI measures in Table 3 the all components have equal influence to the MSS in point of view of the system structure. But modification SI makes more exactly impact of each component failure to the system reliability. So the third component (organization component) failure has lesser influence to the system reliability and its correct work. And the system isn't working in case if the first or the second components fail ($I_{s_m} = 1$). Therefore there is some chance for the system functioning if the organization component fails, but the system has not to work if the technical or human component is failure.

BI measure equals the probability that the MSS is in a state in which the functioning of the i -th component is critical. Therefore working states of the 1-st system component ensures the maximal probability of the MSS functioning ($I_B(x_1) = 0.133$). Therefore the system has maximal probability to fail if the technical component is not functioning.

CI of the 2-nd system component (human component) has the largest value. Therefore this component has maximal importance for this MSS taking into account the probability of this component failure.

The 1-st and the 2-nd FVI measures have larger values. FVI of the 3-rd component is minimal and in other words the organization component failure has minimal influence to the system failure.

CDRI is probability of the MSS failure caused of the i -th component break down. DIRI is similar measure, but allows to analyse the MSS failure depending on unavailability of the some system component. These measures take into consideration both the system topology and component performance (component state probability). According to CDRI the human component has maximal importance for the system. The 1-st component causes the system failure with minimal probability because this component has high importance in point of view of the structure and functioning (see SI) but its probability of failure is minimal. So the system failure caused by break down of the 1-st component is not high. DIRI is probability of the system failure if one of system component (this component is not fixed) is not functioning. In other words failure of the 1-st or 2-nd or 3-rd component causes with probability 0.49 that the system fails.

Therefore importance analysis allows to investigate all influences of the system component functioning to the system reliability changes. This analysis reveals unreliable components in the system structure and functioning. Advantage of the analysis is possibility to use it in system design and SI and modification SI are useful in this stage first of all.

5 Discussion and Conclusions

The analysis on occurred medical mistake in the last decades has clearly shown that the organization and human factor play signification role in the risk of diagnosis and treatment [Lyons et al 2004]. This is due also to the fact that the reliability of the technical components has significantly improved in recent years. As a consequence, the influence of the errors of the organizations managing and of the human operators to systems operation has significantly increased. Therefore

correct reliability analysis of healthcare system needs consideration of three basic component of the system at the least: technical (hardware and software), human and organization. But some recommendation and rules for representation of investigated system based on this structure is not proposed now. This problem needs decision in reliability engineering. In this paper one of possible way for decision of this problem has been considered by example for the Decision Support System for Preliminary Diagnostics in Oncology. It allows to implement reliability analysis of this system. Influence of possibility human and organization errors, and faults in the technical system component has been quantified by importance analysis.

New importance analysis method has been proposed for healthcare system investigation based common mathematical conception in the paper. It allows including to the unified analysis heterogeneous component as technical component, as well as human and organization components. This method estimates influence of each system component failure to the system reliability behaviour. And next investigation of this problem in reliability analysis of healthcare system needs generalization of importance analysis for every system performance levels from failure to perfect working.

Acknowledgment

This research was partially supported by NATO grant CBP.EAP.CLG 984228 “Intelligent assistance systems: multisensor processing and reliability analysis”.

References

- [Aven and Nokland 2010] Aven, T., Nokland, T.E.: On the use of uncertainty importance measures in reliability and risk analysis. *Reliability Engineering and System Safety* 95(2), 127–133 (2010)
- [Cohen 2004] Cohen, T.: Medical and information technologies converge. *IEEE Engineering in Medicine and Biology Magazine* 23(3), 59–65 (2004)
- [Fricks and Trivedi 2003] Fricks, R.M., Trivedi, K.S.: Importance analysis with Markov Chains. In: *Proc. IEEE the 49th Annual Reliability & Maintainability Symposium*, Tampa, USA, pp. 89–95 (2003)
- [Lisnianski and Levitin 2003] Lisnianski, A., Levitin, G.: Multi-state system reliability. Assessment, optimization and applications, p. 358. World Scientific, Singapore (2003)
- [Lyons et al. 2004] Lyons, M., Adams, S., Woloshynowych, M., Vincent, C.: Human reliability analysis in healthcare: A review of techniques. *Int. Journal of Risk & Safety in Medicine* 16(4), 223–237 (2004)
- [Marseguerra and Zio 2004] Marseguerra, M., Zio, E.: Monte Carlo estimation of the differential importance measure: application to the protection system of a nuclear reactor. *Reliability Engineering and System Safety* 86(1), 11–24 (2004)
- [Pham 2003] Pham, H. (ed.): *Handbook of Reliability Engineering*, p. 659. Springer, London (2003)

- [Taleb-Bendiab et al. 2006] Taleb-Bendiab, A., England, D., Randles, M., et al.: A principled approach to the design of healthcare systems: Autonomy vs. governance. *Reliability Engineering and System Safety* 91(12), 1576–1585 (2006)
- [Taylor 1972] Taylor, E.F.: The reliability engineer in the health care system. In: *Proc. IEEE the 18th Annual Reliability & Maintainability Symposium, USA*, pp. 245–248 (1972)
- [Zaitseva 2009] Zaitseva, E.: Importance analysis of Multi-State System by tools of Differential Logical Calculus. In: Bris, R., et al. (eds.) *Reliability, Risk and Safety. Theory and Applications*, vol. 3, pp. 1579–1584. CRC Press, Boca Raton (2009)
- [Zaitseva2010] Zaitseva, E.: Reliability Analysis Methods for Healthcare system. In: *Proc. IEEE the 3rd Int Conf. on Human System Interaction, Rzeszow, Poland*, pp. 212–216 (2010)
- [Zio 2009] Zio, E.: Reliability engineering: Old problems and new challenges. *Reliability Engineering and System Safety* 94(2), 125–141 (2009)

Towards a Formal Model of Visual Design Aided by Computer

E.J. Grabska

Faculty of Computer Science, Jagiellonian University, Kraków, Poland

ewa.grabska@uj.edu.pl

Abstract. This paper aims at contributing to a better understanding of essential concepts of visual design aided by computer. Towards this end, we first present an ontology of conceptual visual design. Then, we define particular components of the formal model paying attention to the role of different kinds of classification during the design process. Moreover, we present two types of logic models used in computer tools supporting the design process. Finally, a formal model of computer-aided visual design is defined. The model is illustrated on examples of designing floor layouts based on a graph-based data structure gathering information for design knowledge and describing two categories of thinking during design process.

1 Introduction

During the conceptual phase of design process designers operate at various levels of abstraction. Then their dialogue with some medium (a sheet of paper, a monitor screen) is absolutely essential. Usually sketching is one of the best ways to absorb early design ideas. But in the Internet age the designer can also create early drawings on the monitor screen with the use of appropriate CAD tools. Recent frameworks for conceptual design emphasizes a role of creative visual thinking during design process. Almost half the brain is devoted to the visual sense and the visual brain is capable of interpreting visual objects in many different ways [Ware 2008]. Visualization of main ideas and concepts during conceptual designing characterizes so called *visual design* [Grabska 2007].

This paper proposes a new model of visual design aided by computer. The multidisciplinary nature of visual design process, non formalized knowledge and the semantic inconsistency of information coming from different design contexts make computer methods supported the conceptual design very complex and effort consuming. Therefore before the presentation of the proposed model an ontology of the conceptual visual phase of design process will be proposed [Yurchyshyna 2009]. The objective of an ontology is to represent non-ambiguous computational semantics. In computer science an ontology is formally defined as a “specification of a conceptualization”.

Another aim of this paper is to propose a formal coherent framework for visual designing. The framework enables one to handle different types of thinking in a formal way, and as a consequence reveals more detailed characteristics of design creativity.

2 An Ontology of Visual Design Aided by Computer

In this section the following main concepts of the ontology will be described:

- (i) *classes*: concepts and sets;
- (ii) *attributes* defined on classes and depend on them;
- (iii) *types* classifying elements of classes;
- (iv) *reasoning mechanisms*: statements in the form of sentences or logical formulas describing the logical inferences that can be drawn from an assertion in a particular form;
- (v) *relations*: specifying how objects and types classifying particular classes can be related to one another;
- (vi) *functions* assigning one class to another.

The considered ontology in the domain D of visual design aided by computer takes into consideration that the designer has an internal world being a mental model of a design task that is build up of concepts and visual perceptions stored in his mind, and an external world composed of representations outside the designer [Gero et al. 2002]. Both drawings created by the designer and their internal representations can be treated as situations in the external world build up outside him. The designer takes decisions about design actions in his internal world and then executes them in the external world. In this paper we assume that the designer's decision making process is supported by the computer-aided design system.

2.1 Classes and Attributes

First, four basic concepts are introduced:

1. a *design task* s – descriptions of what must be true in order for some assertion to be accepted as a design solution expected,
2. a *visualization site* v – a drawings along with a surface on which it is drawn; different surfaces can be used for drawing, e.g., a sheet of paper or a monitor screen,
3. a *data structure* h – a specialized format for organizing and storing data, and
4. a *physical design action* a – one of the following activities: drawing, copying and erasing elements of graphical outputs.

These four concepts allows one to determine four classes being sets of the following objects:

1. S – a set of *design tasks*,
2. V – a set of *visualization sites*,

3. H – a set of *data structures*, and
4. A – a set of *actions*.

Attributes defined on the sets assign properties, features and parameters to particular elements of these sets.

2.2 Types and Reasoning Mechanism

The types of the ontology are connected with sets of objects. For sets of objects we define the way of classification in the following way:

Definition 1. A **classification** is a triple $C_O = (O, T_O, I^O)$, where

- O is a set of objects to be classified,
- T_O is a set of types used to classify objects of O , and
- I^O is a binary relation between O and T_O that specifies which objects are classified as being of which types.

Objects being design solutions are classified by design requirements in the form of expressions of the *propositional logic*. Physical actions are classified using either structure-less or structural objects. For graph based data structures being design drawing representations the *first-order logic* is used as a reasoning mechanism. Information stored in the data structures corresponding to design visualization is translated to sentences of the first-order logic.

In the first order logic we start with defining a specified vocabulary for this logic.

Definition 2. A **vocabulary** of the first-order logic is a triple $W = \{B, F, R\}$, where

- B is a set of constant symbols,
- F is a set of multi-argument function symbols, and
- R is a set of multi-argument relation symbols.

We assume that we have a set of *variables*, which we usually write as x and y , possibly along with subscripts. The set of *terms* is formed starting from constant symbols of B and variables and closing off under function application, i.e., if l_1, \dots, l_n are terms and $f \in F$ is an n -ary function symbol, then $f(l_1, \dots, l_n)$ is also a term. An *atomic formula* is either of the form $r(l_1, \dots, l_k)$, where $r \in R$ is an k -ary relation symbol and l_1, \dots, l_k are terms, or of the form $l_1 = l_2$, where l_1 and l_2 are terms. The set of general logical formulas is built over atomic formulas using logical connectives and quantifiers, and closed under the consequence relation. The formulas contain variables universally quantified over appropriate component types. Formulas which do not have free variables are called *sentences* [Fagin et al.1995].

The semantics of first-order formulas uses *relational structures*. A relational structure consists of a domain of individuals and a way of associating with each of the elements of the vocabulary corresponding entities over the domain. Thus, a constant symbol is associated with an element of the domain, a function symbol and a relation symbol are associated with a function and a relation, respectively. Both function and relation are defined on the domain.

The definition of a relational structure for the proposed ontology is as follows:

Definition 3. A relational W -structure L consists of:

- a domain D of conceptual visual design aided by computer,
- an assignment of a k -ary relation $r^L \subseteq D^k$ to each k -ary relation symbol $r \in R$,
- an assignment of a n -ary function $f^L: D^n \rightarrow D$ to n -ary function symbol $f \in F$,
- an assignment of a $b^L \in D$ to each constant symbol $b \in B$.

The next step to define the formal semantics of first-order formulas is specification of an interpretation of variables. A valuation u on a structure L is a function from variables to elements of D . Given a structure L and a valuation u on L , u is inductively extended to a function that maps terms to elements of D . Let $u(b) = b^L$ for each constant symbol b and then the definition of u is extended by induction on the structure of terms by taking $u(f(l_1, \dots, l_n)) = f^L(u(l_1), \dots, u(l_n))$.

Given a relational structure L with a valuation u on L , $(L, u) \models \Phi$ denotes that a formula Φ is true in L under the valuation u . The truth of the basic formulas is defined as follows:

- $(L, u) \models r(l_1, \dots, l_k)$, where $r \in R$ is a k -ary relation symbol and l_1, \dots, l_k are terms, iff $(u(l_1), \dots, u(l_k)) \in r^L$,
- $(L, u) \models l_1 = l_2$, where l_1 and l_2 are terms, iff $u(l_1) = u(l_2)$,
- $(L, u) \models \neg \Phi$ iff $(L, u) \not\models \Phi$,
- $(L, u) \models \Phi_1 \wedge \Phi_2$ iff $(L, u) \models \Phi_1$ and $(L, u) \models \Phi_2$,
- $(L, u) \models \exists x \Phi$ iff $(L, u[x/a]) \models \Phi$ for some $a \in D$, where $u[x/a]$ denotes the valuation with $u(x) = a$.

2.2 Relations and Functions

The basic relations between the three design classes: design tasks, computer visualizations and actions are defined between components of their classifications. Therefore in this subsection we start with definitions of classifications of the design classes.

Design tasks being elements of the first design class describe situations which are classified by types expressing designer's requirements.

Definition 4. The classification of design tasks is a triple $C_S = (S, T_S, \vdash^S)$, where

- S is a set of objects to be classified, called *design situations*,
- T_S is a set of types called *design requirements* used to classify objects of S , and
- \vdash^S is a binary relation between S and T_S that specifies which objects are classified as being of which types.

If a design situation $s \in S$ is classified as being of type $t \in T_S$, we write $s \vdash^S t$. Types of T_S being design requirements are formulated in the form of expressions or sentences of the *propositional logic*.

Designer's external world is associated with many drawings which allow the designer to interweave analysis with synthesis. The next design class contains visualization sites, i.e., arbitrary surfaces on which drawings are made along with these drawings. Two different drawings on the same surface, e.g., on the sheet of paper or on the monitor screen determine two different visualization sites. Visualization sites can be treated as situations in the external world built up outside the designer, and as such they belong to appropriate types of their own classification, just as design task situations do.

Definition 5. By the classification of visualization sites we mean a triple $C_V = (V, T_V, \vdash^V)$ consisting of:

- V – a set of objects to be classified, called *visualization sites*,
- T_V – a set of types used to classify the visualization sites of V , and
- \vdash^V is a binary relation between V and T_V that specifies which objects are classified as being of which types.

Visual perception plays an essential role for the classification of visualization sites. The types of T_V associated with visualizations sites are related to geometrical properties of drawings: appropriate geometrical objects and their transformations which allow for obtaining admissible components of design objects. This kind of classification is specified by means of types in the form of expressions of the propositional logic. Every visualization site belongs to a set of visualization sites that a collection of types classifies.

The two classifications C_S and C_V makes possible to define two basic relations in the domain D of visual design, namely *signalling* and a *semantic convention*.

If a visualization site $v \in V$ is used to find a design situation $s \in S$ then v *signals* s and we write $v \rightarrow s$. *Signalling*, denoted by \rightarrow is a binary relation from the set V of visualization sites to the set S of design situations.

Designer's requirements can be treated as constraints on expected design solution. Visualization site drawings are different from forms wherein the designer expresses requirements related to the design tasks. When taking physical actions the designer encodes information about the object being designed in the fictional

depicted world. He/she also deals with visual organization of the drawing, which includes form, proportion, line, shape and so on. The correspondence between constraints on drawings being types of C_V and constraints on designer's requirements formulated as types of C_S determines a *semantic convention* defined by a binary relation from T_V to T_S denoted by \Rightarrow .

The two relations: signalling and semantic convention form a mapping from the Cartesian product $V \times T_V$ of the classification C_V to the Cartesian product $V \times T_S$ of the classification C_S .

At the present time designer's classification of visual sites can be supported by the computer aided design system. In the domain D of visual design, where monitor screen with design drawings are considered as visualization sites, drawings are automatically transformed into appropriate data structures and then information stored in the data structure is translated to formulas of the first-order logic, which constitute design knowledge. The relational structure of the first-order logic is in the form of the data structure corresponding to the design drawing created by the designer on the monitor screen and assigned by the internal representation function.

Let H be a set of data structures h . The basic function of D is an *internal representation* $\tau: V \rightarrow H$, i.e., the function τ assigns data structures $h \in H$ to design drawings of visualization sites $v \in V$.

Let Ψ_H be a set of first-order logic formulas supporting the classification of the visualization sites on the basis of data structures of H . If a visualization site $v \in V$ is classified as being of type $t \in T$ and $\psi \in \Psi_H$ is a logic formula supporting the classification of v on the basis of data structure $\tau(v)$ then $t \wedge \psi$ classifies also v and we say that v belongs to $t \wedge \psi$.

Recently, a design process is often treated as a sequence of actions, which changes the external world. These actions called *physical design actions* consist in drawing, copying and erasing elements of graphical outputs (Suwa et al. 2000). Physical design actions, which result in modifications of drawings, automatically impose changes both in the data structures and design knowledge. Thus logic sentences form dynamic design knowledge, i.e., physical actions performed by the designer on drawings simultaneously modify this knowledge.

The last design class is related to physical design actions treated as a certain kind of events in the external world that start with an initial situation and result in another situation.

Definition 6. The classification of physical design actions is a triple $C_A = (A, T_A, \vdash^A)$ where:

- A is a set of objects to be classified, called *physical design actions*,
- T_A is a set of types used to classify the actions, and

- l^A is a binary relation between A and T_A that specifies which objects are classified as being of which types.

We define the next relation in the domain D , namely a tertiary *input and output* relation $V \times A \times V$ between a set V of visualization sites and a set A of actions. The relation is denoted by \rightsquigarrow and $v_i \rightsquigarrow^a v_o$ means action a has v_i as an *input* visualization site and v_o as an *output* visualization site. We assume that each action has unique input site and output site.

An output visualization site presenting the final design solution is a result of sequences of actions. We introduce an operation \bullet , called *composition* on the set A of actions, which is closed under \bullet . The operation is partial and associative.

For all actions a, a' in A and for all visualization sites v, v' in V the composition $a \bullet a'$ is defined by means of an extension relation \rightsquigarrow with the input visualization site v and with the output visualization site v' defined in the following way:

Definition 7. $v \rightsquigarrow^{a \bullet a'} v'$ iff there is a visualization site $v^* \in V$ such that $v \rightsquigarrow^a v^*$ and $v^* \rightsquigarrow^{a'} v'$.

In other words, an output visualization site v' is a result of the composition $a \bullet a'$ defined on an input visualization site v if and only if there exists a direct visualization site v^* being an output visualization site for the action a and an input visualization site for the action a' .

3 A Formal Model

Summing up the discussion on the four design classes and relations between them, a Computer Visual Design model (CVD – model) is defined.

Definition 8. A Computer Visual Design Model (CVD-model) is a 9-tuple

$P = (C_S, C_V, C_A, H, \tau, \Psi_H, \longrightarrow, \Rightarrow, \rightsquigarrow)$, where:

- $C_S = (S, T_S, l^S)$ is a classification of design tasks,
- $C_V = (V, T_V, l^V)$ is a classification of visualization sites,
- $C_A = (A, T_A, l^A)$ is a classification of physical design actions,
- H is a set of data structures,
- $\tau: V \rightarrow H$ is an internal representation function,
- Ψ_H is a set of first-order logic formulas defined on the basis of H ,
- $\longrightarrow \subseteq V \times S$ is a signalling relation,
- $\Rightarrow \subseteq T_V \times T_S$ is a semantic convention relation, and
- $\rightsquigarrow \subseteq V \times A \times V$ is an input and output relation.

It is worth noticing that the proposed model can also be used to describe the design process without the support of a computer tool. In this case visual sites can be

for instance in the form of sketches on a sheet of paper and they do not have internal data structures, i.e., the sets H and \mathcal{Y}_H are empty and the function τ is not defined. Consequently, a *Visual Design Model* (VD-model) is a 6-tuple $\mathbf{P}' = (C_S, C_V, C_A, \rightarrow, \Rightarrow, \rightsquigarrow)$.

4 An Example

To illustrate *CVD*-model we sketch the prototype system *HGSDR* (Hypergraph Generator Supporting Design and Reasoning) (Grabska et al. 2009). The system allows the designer to edit and automatically applies operations on hyper-graphs being internal representations of drawings.

Let us consider a simplified specialized CAD editor of the *HGSDR* system for designing floor layout composed of polygons which are placed in an orthogonal grid. These polygons represent functional areas or rooms. Mutual location of polygons is determined by the designer. Lines with small squares on them represent the accessibility relation among components, while continuous lines shared by polygons denote the adjacency relations between them (see: Figure 1). The sides of each polygon are ordered clock-wise starting from the top left-most one. In a design drawing only qualitative coordinates are used i.e., only relations among graphical elements (walls) are essential.

In *CVD*-model related to the example each design task $s \in S$ is designing floor layout. An example of designer's requirement being a type t of the classification C_S can be as follows: t - a bathroom is accessible from a sleeping area.

Fig. 1 presents one of the steps of creating the layout of a one storey house with a garage. The example is very simple but its role is only to show the way of generating the hyper-graph corresponding to the design drawing edited by the designer on the monitor screen. Let assume that v is a visualization site corresponding to this drawing along with the monitor screen. Let the function τ assigns an internal representation in the form of a hyper-graph h , i.e., $\tau(v) = h$. This representation h is shown in Figure 2.

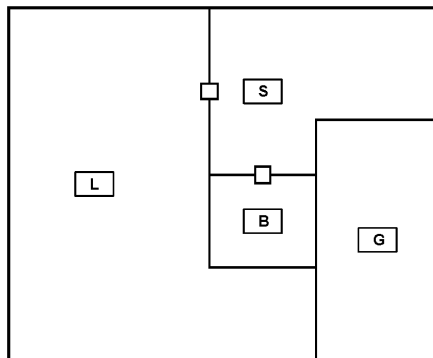


Fig. 1 The design drawing of v

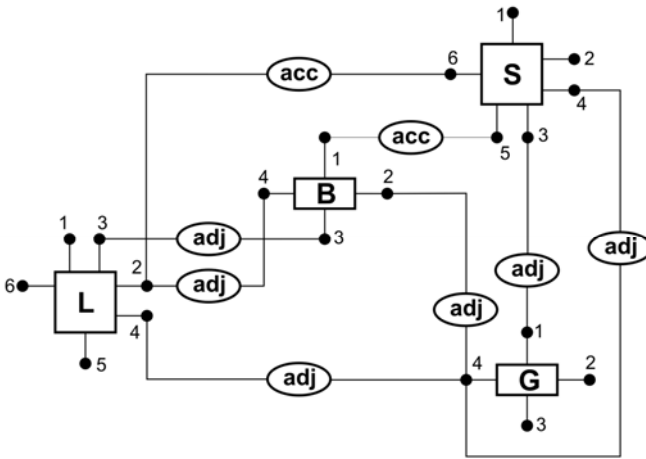


Fig. 2 The hyper-graph $h = \tau(v)$

A hyper-graph has two types of hyper-edges, called *component* hyper-edges and *relational* hyper-edges. Hyper-edges of the first type correspond to drawing components and are labelled by component names. Hyper-edges of the second type represent relations among fragments of components and can be either directed or non-directed in the case of symmetric relations. Relational hyper-edges of the hyper-graph are labelled by names of relations. Component hyper-edges are connected with relational hyper-edges by means of nodes.

The considered hyper-graph h is composed of twelve hyper-edges: four component hyper-edges that correspond to the four polygons of the design drawing and eight relational hyper-edges. Two of these relational hyper-edges represent the accessibility relation; first – between rooms corresponding to polygon L and polygon S and next – between polygon S and polygon B. The remaining six hyper-edges represent the adjacency relation.

Let us come back to create the design drawing presented in Figure 1. Before it is obtained, the first drawing edited by the designer represents the area of whole apartment. The initial hyper-graph representing this first drawing generated automatically is composed of one hyper-edge connected with four external nodes representing sides of the area and placed in the drawing according to the geographical location of the sides they correspond to. In the next steps, the designer divides the whole apartment area into four parts representing a living area (L), sleeping area (S), bathroom (B) and a garage (G). Then he/she draws small squares on the lines representing the common walls between the living area and the sleeping area, and between the sleeping area and the bathroom. As a consequence, the hyper-edge operations on the initial hyper-graphs are invoked automatically. As the result of the operations the hyper-graph representing the four areas and accessibility and adjacency relations between them is generated. Each areas is considered as component hyper-edges together with the nodes assigned to them.

Design actions $a \in A$ on visual sites cause changes in hyper-graphs serving as a base for reasoning about design. During the design process knowledge corresponding to design drawings created by the designer is first translated to an appropriate hyper-graph and then to sentences of the first-order logic. In this process a problem-oriented relational structure, which assigns elements of hyper-graphs to entities of the specified first-order logic alphabet is used.

In the example related to design floor layout

1. constant symbols of B represent walls of rooms of a layout,
2. F contains one single argument function symbol, which determines a room to which a given wall belongs, and
3. $P = \{adj, acc\}$ is a set of relation symbols, where adj and acc are two binary relation symbols representing adjacency and accessibility between rooms.

The relational structure is in the form of a hyper-graph $\tau(v)$, $v \in V$. The domain of this structure includes:

- a set of component hyper-edges, and
- a set of hyper-graph nodes.

Relations between design components presented in the drawing of v are specified between fragments of these components, which correspond to hyper-graph nodes. The interpretation of each relation is the hyper-edge relation of the hyper-graph such that there is a relational hyper-edge coming from a sequence of nodes of at least one component hyper-edge and coming into a sequence of nodes of other component hyper-edges.

Let us come back to the hyper-graph shown in Figure 2. Denote by $n_{B,1}$ and $n_{S,5}$ the first node attached to hyper-edge B and the fifth node attached to hyper-edge S . The atomic formula $acc(n_{B,1}, n_{S,5})$ - a bathroom is accessible from a sleeping area belongs to the syntactic knowledge about the designed drawing presented in Fig.1. This syntactic knowledge, obtained from the relational structure being a hyper-graph corresponding to this drawing, is in the form of atomic formulas which facilitate reasoning about features of designs. Moreover, changes in design knowledge resulting from drawing modifications consist in changes atomic formulas and they can be traced by the designer.

5 Categories of Creative Thinking

A manner in which the designer thinks about design problems is one of essential aspects for creative design. There are two major categories of thinking: *divergent* and *convergent* [Lawson 2001]. Divergent thinking is imaginative and intuitive, whereas the convergent one is logical and rational. There exists the hypothesis that “creativity involves the capacity to spontaneously shift back and forth between analytic and associative modes of thought according to the situation” [Gabora 2010]. Therefore “thought is triggered by stimuli, which activate memory in

specific patterns that support divergence or convergence” [Goldshmidt 2010]. Taken as a whole, design is a divergent task. However, during the process of creative design good designers are able to develop and maintain several lines of thought, both convergent and divergent.

The *VD* model allows one to define divergent and convergent categories of thinking more formal. The former type of thinking is based on abduction and it is typical in an inventive design when a number of unknown design concepts is sought. This ability has been associated with skill in the arts and it has been interpreted as an open-ended approach seeking alternative. In the presented model the definition of divergence is as follows:

Definition 9. Let *VD* be a model of visual design and t_1, \dots, t_n be a sequence of types of T_S classifying design task of S , and \Rightarrow be a semantic convention. The system *VD* **imposes divergence** on the types t_1, \dots, t_n iff there exist an input visualization site $v \in V$ and a sequence of actions $a_1, \dots, a_m \in A$ such that:

1. The types t_1, \dots, t_n allow the composition action $a_1 \bullet \dots \bullet a_m$ on the visualization site v .
2. The output visualization site v' for the action $a_1 \bullet \dots \bullet a_m$ allows new types $t_1^*, \dots, t_k^* \in T_V$ for $k > 1$ such that $v' \models^V t_i^*$ for all $i = 1, \dots, k$.
3. On the semantic convention \Rightarrow , each type t_i^* indicates at least one new type t of T_S .

In other words, the composition of actions a_1, \dots, a_m leads to the output visualization site which allows the designer to discover new facts different from types classifying visual sites signalling the considered design solutions. Each of these fact can inspire the designer to formulate a devised requirement.

During design process the designer develops also convergent lines of thought, which require deductive and interpolative skills. Occurring emergence of new shapes is an example of the convergent thought. The designer discovers a new shape called *emergent* (which had not been consciously constructed).

Example: Let us consider the scribble shown in Fig. 3a. An example of emergent shape drawn white line is presented in Fig. 3b. The perceptual action allows the designer to notice this shape and associates it with shapes of a lamp.



Fig. 3 An emergent shape

This association becomes a new inspiration in creating a form of the designed lamp (Figure 4) and enables the designer to formulate a devised requirement t (a new type of T_S) and it is an example of the convergent thought.



Fig. 4 Designing a lamp

Definition 10. Let VD be a model of visual design and t_1, \dots, t_n be a sequence of types of T_S classifying design task of S , and \Rightarrow be a semantic convention. The model VD **imposes convergence** on the types t_1, \dots, t_n iff there exist an input visualization site v in V and a sequence of actions a_1, \dots, a_m in A such that:

1. The types t_1, \dots, t_n allow the composition action $a_1 \bullet \dots \bullet a_m$ on the visualization site v .
2. The output visualization site v' for the action $a_1 \bullet \dots \bullet a_m$ realizes a type t^* in S_V such that $v' \models^V t^*$.
3. On the semantic convention \Rightarrow , the type t^* indicates a new type t of T_S .

Drawing and extraction of visual information during the process of visual designing needs both convergent and divergent thought. It imposes using entire brain.

6 Conclusions

Nowadays, visual designer environment plays an essential role in designing. The proposed system of Computer Visual Design System presents significant aspects of visual design process. To develop this system, which is necessary, in the one hand for deeply understanding the fundamentals of conceptual design and in the second hand to devise new visual tools, a higher level of abstraction had to be used. The notions of classifications of design tasks, of visualization sites and physical design actions have been introduced. These notions enable us to formulate design constraints in a natural way.

The new framework for conceptual design allows one to hold concepts from different disciplines (engineering and psychology) in a formal way and shows influence of different perspectives on the design theory.

References

- [Fagin et al.1995] Fagin, R., Halpern, J.Y., Moses, Y.: Vardi MY Reasoning about knowledge. MIT Press, Cambridge (1995)
- [Gabora 2010] Gabora, L.: Revenge of the neurds: Characterizing creative thought in terms of the structure and dynamics of memory. Creativity Research J. 22(1), 1–13 (2010)

- [Gero et al. 2002] Gero, J.S., Kannengiesser, U.: The situated function-behaviour-structure framework. In: Gero, J.S. (ed.) *Artificial Intelligence in Design 2002*, pp. 89–104. Kluwer, Dordrecht (2002)
- [Goldschmidt 2010] Goldschmidt, G.: Ubiquitous serendipity: Potential visual design stimuli are everywhere. In: *NSF Workshop: Studying Visual and Spatial Reasoning for Design Creativity*. Pre-Workshop Paper (2010)
- [Grabska 2007] Grabska, E.: *Computer-aided visual design*. EXIT, Warszawa (2007) (in Polish)
- [Grabska et al. 2009] Grabska, E., Borkowski, A., Palacz, W., Gajek, S.: Hypergraph System Supporting Design and Reasoning. In: Huhnt, W. (ed.) *Computing in Engineering EG-ICE Conference*, pp. 134–141 (2009)
- [Grabska 2010] Grabska, E.: The theoretical framework for visual thinking in design creativity. Invited paper of the NSF International Workshop on Studying Visual and Spatial Reasoning for Design Creativity (2010) (in printing)
- [Lawson 2001] Lawson, B.: *How designers think: the design process demystified*. Butterworth Architecture, Oxford (2001)
- [Suwa et al. 2000] Suwa, M., Gero, J.S., Purcell, T.: Unexpected discoveries and s-invention of design requirements: Important vehicles for a design process. *Design Studies* 21(6), 539–567 (2000)
- [Ware 2008] Ware, C.: *Visual thinking for design*. Elsevier, Amsterdam (2008)
- [Yurchyshyna 2009] Yurchyshyna, A.: *Modeling of conformity checking in construction: An ontological approach*. PhD thesis, Nice-Sophia Antipolis University (2009)

Agricultural Products' Storage Control System

W. Sieklicki, M. Kościuk, and S. Sieklicki

Gdańsk University of Technology, Gdańsk, Poland
wiktorsieklicki@pg.gda.pl, mk84@interia.pl,
sieklick@eti.pg.gda.pl

Abstract. The paper discusses the idea and the design of a remote control system for storage management of agricultural products which temperature may rise as the result of biological processes during the storage. An actual potatoes storehouse is discussed as an application for the proposed automation system. Because of existing buildings and infrastructure at the farm, wireless data transfer system has been proposed for communication between sensors, data acquisition module (Storehouse Local Controller) and human-machine interface (HMI) module. Special rod-like temperature and humidity sensors have been designed and built to measure the temperature and humidity inside the potatoes stack at the storehouse. A complete system allows to monitor the storage process and control it on the run maintaining the desired temperature of the stored products. The study is an elaboration of previously proposed control system, including ventilation system working algorithm, further analysis of working range of the wireless data transfer modules and agricultural products temperature changes characteristics.

1 Introduction

In order to provide a high quality of agricultural products it is important to maintain a correct and precise temperature and humidity in a storehouse while storage. Biological processes during a storage of e.g. potatoes, carrots, grain evoke spontaneously rising temperature which is highly undesired [Schippers 1977; Bohl et al. 2002; Czerko 2003]. Automation system which helps to maintain a desired temperature is thereby considered highly valuable for a proper storage. Records of key aspects of a storage e.g. temperature variation or velocity of temperature changes during the storage, together with monitoring the humidity may certify that product had been stored in a proper manner, being in this way its quality certificate. Recently used wire based measuring systems of the temperature and humidity have a suitable measurements accuracy and reliable data storage. Their disadvantages though are that the connection wires laid in the potato stack often disturb every work done in the storage facility (e.g. emptying the storehouse), wires in the storehouse often breaks resulting in loss of a signal and finally, data transmission

to the farm management person residing in a control building is threatened by heavy machines working in the farm. This problem exists especially in old farms which are modified and modernized in order to improve the quality of a storage. Moreover, existing farms owners may not have possibility to lay down transmitting cables from the storage houses to the control building because of the difficult terrain or distance. Installing a wireless storage management system in existing farms gives the possibility to the farm owners to compete with newly built farms with highly automated storage management systems [Czerko 2007].

Aim of this work is to propose a wireless system for temperature and humidity measurements with control of airflow flaps and fans in an existing storage facility. A system composes of a PC (to monitor and control the storage process by managing person), from 10 to 15 potatoes stack temperature sensors, from 2 to 8 potatoes stack humidity sensors, one external and one interior temperature sensor, 2 airflow mass sensors, 1 airflow temperature sensor, 3 airflow flaps actuators, 2 fans, a Storehouse Local Controller (SLC) acquiring data from sensors placed in the potatoes stack and controlling the airflow flaps and fans by wires (Fig.1).

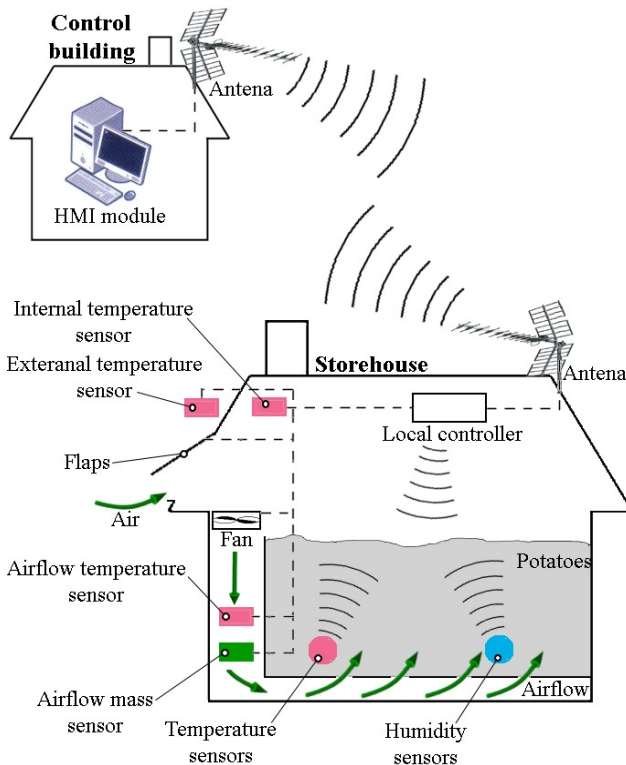


Fig. 1 An initial fragment of a cognitive map with false transitivity

2 Control System Components

A complete system is divided onto two parts: a control house – where human-machine interface module is placed, including a PC; and a storehouse – where SLC, a set of sensors and an airflow control system are placed (Fig.2.).

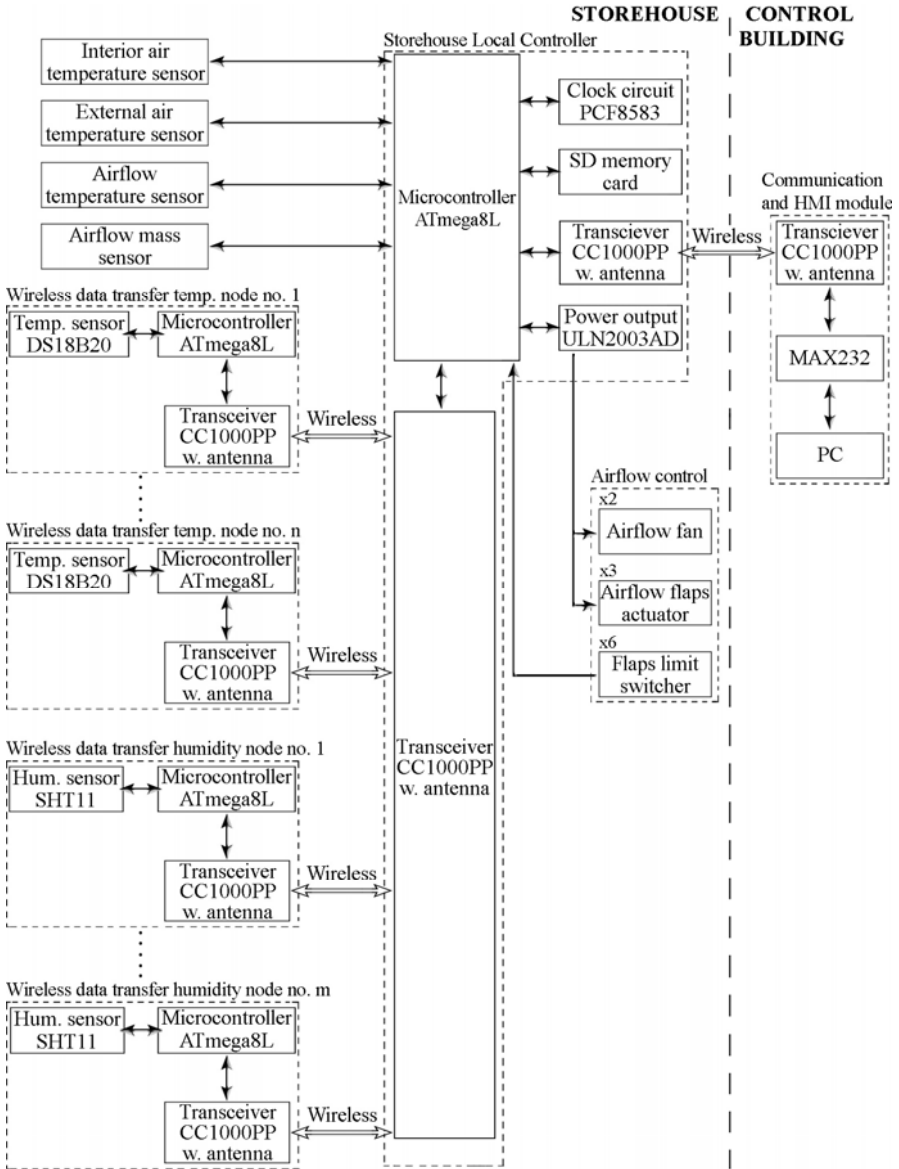


Fig. 2 Storage control system scheme

2.1 Measurement Nodes

The temperature and the humidity measurements of the potatoes stack have to be made frequently in order to maintain the proper storage conditions. For this reason rod-like measuring nodes for temperature and humidity measurement have been developed. A complete system is designed to operate with n temperature measurement nodes ($1 < n < 15$) and m humidity measurement nodes ($1 < m < 8$). Each node is a 1.5 m length rod equipped with digital temperature sensor DS18B20 or digital humidity sensor SHT11 at one of the rod's end [WWW-1 2010; WWW-2 2010]. At the other end of the rod-like measurement node a sealed box is mounted consisting an AtMega8 controller with batteries and a CC1000PP transceiver [WWW-3 2010]. The AtMega8 controller controls the power supply for the temperature sensor and the transceiver. It also measure batteries' discharge level and inform the system managing person through HMI about potential necessary maintenance.

By default, every one hour sensors DS18B20 and SHT11 are powered up and the current temperature and humidity readouts are acquired by AtMega8. The data is then send by a CC1000PP to the SLC. After receiving information from sensors, controller turns all the electronic parts of the measuring node a to low-power consumption state in order to save the batteries' energy. Every AtMega8 can measure the batteries discharge level of a particular node and inform the system user through the HMI about necessary maintenance of the node. Every measurement node has got an individual address used by CC1000PP transceiver to communicate between each other. Nodes are capable of transmitting data about the temperature and humidity to the local controller placed in radius of approximately 50 m from nodes. All components of measuring nodes, including the AtMega8 controllers and CC1000PP transceivers, are low-price electronic devices, thus multiple usage may be considered in the proposed system.

2.2 Storehouse Local Controller (SLC)

The Storehouse Local Controller's task is to control the storage process having the current temperature and humidity of the potatoes stack data feedback. Executive elements of the system are fans and flaps, which may provide an additional air-flow through the potatoes stack if the controller defines it as necessary. The SLC consists of ATmega32 controller, two CC1000PP transceivers, a PCF8583 Clock Circuit, an interchangeable Secure Digital (SD) card reader and an ULN2003AD power output integrated circuit. SLC is mounted on the wall of the building and is powered from the local mains. Using one of the transceivers, it communicates with temperature and humidity measurement nodes issuing orders to send data and acquiring the current readouts. A circular antenna is used to communicate with temperature and humidity nodes. The antenna is located inside the storehouse in close proximity of a potatoes stack. The second transceiver is used to

communicate with the communication module placed in a control building. With this transceiver the SLC may send information to a PC, which is a part of an integrated HMI module. A directional antenna, which is located outside the storehouse on the building wall, is used for this reason.

Three airflow temperature sensors are connected to the controller by wires – they are located nearby the air intake (outside and inside the building) and in the airflow tunnel. The data gathered with use of those sensors is used to define a possibility to change the current temperature and humidity of a stack. Fans and flaps may be then utilized in order to force the correct airflow.

The SLC can work independently from the HMI module and in that case, it operates without human attendance. The data acquired by the controller working in this state is stored on the SD card in order to send it to the HMI when connection become possible. The data can be also transferred with use of an SD card to any computer equipped with SD card reader.

2.3 Communication Module

Superior to the SLC is a communication module placed in the control building. Communication module consists of a ATmega32 microcontroller, a CC1000PP transceiver and a MAX232 integrated circuit. The transceiver sends and receives data from the SLC, whereas the MAX232 enables communication with the PC.

2.4 Human-Machine Interface Module

The visualization software application has been developed for the PC. While in automatic mode, application's task is to inform the managing person about the control process state. Actualization of the process parameters is made on the screen every hour and one can verify the data on the run or refer to data written in the history log file. The manual mode makes it possible to manually control the airflow in the storehouse. The history log, may be provided by the application as a chart, presenting changes of a storage parameters within time to be certain about products' quality.

3 Airflow Control System Algorithm

Potatoes stack temperature variations may indicate dangerous biological processes taking place in the stack. To avoid those processes temperature of the stack has to be kept in proper limits and fresh air let through the stack. Meanwhile the stack temperature cannot be less than $\sim 2^{\circ}\text{C}$ and cannot exceed $\sim 22^{\circ}\text{C}$ since those results in drop of the stored potatoes quality. Humidity monitoring is important not to let the potatoes dehydrate themselves if the humidity is too low, whereas high

humidity is an indicator, that biological processes may be taking place in the stack. Proper temperature and humidity of potatoes stack during the potatoes storage depends from their designation and the phase of the storage (Table 1).

Table 1 Temperature and relative humidity requirements in the storage facility for four stages of potatoes storage [Czerko 2008; Wachowicz 1998]

Storage stage	Potatoes designation	Temp. [°C] $T^1 - T^2$	Relative humidity [%] $H^1 - H^2$	Time [weeks]
<i>Healing and epidermis suberisation</i> – I stage –	Seed potatoes	15 – 18	90 - 95	1.5
	Consumption	12 – 15		2
	Processing	10 – 12		4
<i>Cooling</i> – II stage – the temperature decrease of 0.2 – 0.3°C per day from temperature A down to temperature B	Seed potatoes	from 15 – 18 down to 2 – 4	90 - 95	2 – 3
	Consumption	from 12 – 15 down to 4 – 6		
	Processing: Chips Crisps	from 10 – 12 down to 6 – 10 down to 5 – 8		
	Dried, Starch	down to 6 – 7		
<i>Long term storage</i> – III stage –	Seed potatoes	2 – 4	90 - 98	Depends upon necessities
	Consumption	4 – 6		
	Processing: Chips	6 – 10		
	Crisps	5 – 8		
	Dried, Starch	6 – 7		
<i>Reconditioning</i> – IV stage –	Seed potatoes	12 - 15	75 - 80	3 – 5
	Consumption	10 - 11	85 - 90	1.5
	Processing			

Four stages of the storage are determined: 1st stage – healing and epidermis suberisation – duration: 1.5 up to 4 weeks – essential at this stage is fast drying the potatoes to limit pathogenic fungi growth; 2nd stage – cooling – 2 up to 3 weeks – aim at this stage is to decrease the temperature of potatoes down to the temperature necessary for long term storage; 3rd stage – long term storage – allow to store the potatoes during the winter, relative high humidity is appropriate but the temperature cannot vary of more than 0.5°C per day; 4th stage – reconditioning – 1.5 up to 5 weeks – preparation to empty the storehouse.

Based on Table 1. rules for flaps and fan operation in respect to potatoes temperatures (T^1 and T^2), humidity (H^2), airflow temperatures is proposed in Table 2.

Table 2 Rules for flaps and fan in respect to potatoes and airflow temperatures and humidity

External temperature T^3	Airflow temperature T^4	Potatoes temperature T^p	Potatoes humidity H^p	Flaps [on+/off-]	Fan [on+/off-]	
$T^3 < T^p - 3^\circ\text{C}$	$T^4 < T^1$	$T^p \leq T^2$	Any	-	-	
		$T^p > T^2$	Any	-	+	
	$T^1 \leq T^4 \leq T^2$	$T^1 \leq T^p \leq T^2$	$T^p < T^1$	Any	-	-
			$H^p < H^2$	-	-	
			$H^p > H^2$	-	+	
		$T^p > T^2$	$H^p < H^2$	-	+	
			$H^p > H^2$	+	+	
			Any	+	+	
	$T^4 > T^2$	$T^p < T^1$	$H^p < H^2$	-	-	
			$H^p > H^2$	-	+	
		$T^1 \leq T^p \leq T^2$	$H^p < H^2$	-	-	
			$H^p > H^2$	-	+	
Any			+	+		
Any			+	+		
$T^p - 3^\circ\text{C} \leq T^3$ or $T^3 \leq T^p + 5^\circ\text{C}$	$T^4 < T^1$	$T^p < T^1$	Any	-	-	
		$T^1 \leq T^p \leq T^2$	$H^p < H^2$	-	-	
			$H^p > H^2$	-	+	
	$T^1 \leq T^4 \leq T^2$	$T^p < T^1$	$H^p < H^2$	-	-	
			$H^p > H^2$	-	+	
			Any	-	+	
		$T^1 \leq T^p \leq T^2$	$H^p < H^2$	-	-	
			$H^p > H^2$	+	+	
			Any	+	+	
	$T^4 > T^2$	$T^3 < T^1$	$H^p < H^2$	-	-	
			$H^p > H^2$	+	+	
		$T^1 \leq T^p \leq T^2$	Any	+	+	
$T^p > T^2$	Any	+	+			
$T^3 > T^p + 5^\circ\text{C}$	$T^4 < T^1$	$T^3 < T^1$	Any	-	-	
		$T^1 \leq T^p \leq T^2$	$H^p < H^2$	-	-	
			$H^p > H^2$	-	+	
		$T^p > T^2$	$H^p < H^2$	-	-	
	$H^p > H^2$		-	+		
	$T^1 \leq T^4 \leq T^2$	$T^p < T^1$	$H^p < H^2$	-	-	
			$H^p > H^2$	-	+	
			Any	-	+	
		$T^1 \leq T^p \leq T^2$	$H^p < H^2$	-	-	
			$H^p > H^2$	-	+	
			Any	-	+	
	$T^4 > T^2$	$T^p > T^2$	$H^p < H^2$	-	+	
$H^p > H^2$			+	+		
$T^p < T^1$		$H^p < H^2$	-	-		
		$H^p > H^2$	-	+		
$T^1 \leq T^p \leq T^2$	$T^p > T^2$	$H^p < H^2$	-	+		
		$H^p > H^2$	+	+		
	$T^p < T^1$	$H^p < H^2$	-	-		
		$H^p > H^2$	-	+		
$T^4 > T^2$	$T^1 \leq T^p \leq T^2$	$H^p < H^2$	-	+		
		$H^p > H^2$	+	+		
	$T^p > T^2$	$H^p < H^2$	-	+		
		Any	+	+		

Decreasing the potatoes stack humidity below the lower value of the potatoes stack optimal humidity H^1 (Table 1.) is affecting product's quality, but it is not a factor influencing occurrence of unwilled biological processes in the stack. Therefore only the higher value of the potatoes stack optimal humidity H^2 is taken into the consideration in presented algorithm.

The stage of the storage is taken into account to alter the airflow control algorithm. During the first stage of the storage flaps are opened and fans are turned on more often. While during the third stage of the storage fans are turned on and flaps are opened at least once a day for at least half an hour but not more than it takes to lower the temperature of the stack more than 0.5°C.

The system is not supposed to prevent from extensive cooling nor heating since the systems is utilizing external or internal air circulation only. If the potatoes stack is exposed to lower temperature than 2°C or higher than 22°C system is informing managing person about the situation and continue working in accordance with the algorithm. Managing person though is alarmed and may take the command in any moment.

4 Wireless Data Transfer

Wireless data transfer between sensors and SLC as well as between SLC and human-machine interface (HMI) module is possible thanks due to the CC1000PP device. It consists of a CC1000 integrated circuit and several passive elements which allow to tune the CC1000 to preferred frequency. The CC1000PP transceivers used in this system are tuned to 870 MHz – the frequency that belongs to the Industrial Scientific Medical (ISM) band [Tanenbaum 2004]. In this set the device has the output power of max. 5 dBm. The output power is programmable. The CC1000PP may be set to power-down mode, what makes it perfect for a battery-powered devices.

The CC1000PP is configured by an external controller using 3-wire serial communication interface. The bi-directional, synchronous data flow needs another two wires – one for data and one for clock signal. The device can be configured for three different data formats: Synchronous Non Return to Zero (NRZ) mode, Synchronous Manchester encoded mode and Transparent Asynchronous UART mode. In the first two modes CC1000 provides the clock signal at clock wire and the data wire is used to send or receive bits at the rising edge of clock signal. The difference between these modes is that NRZ mode does not use encoding, whereas Synchronous Manchester encoded mode does the Manchester encoding. In the third mode the data wire is used to send data but no synchronization clock is provided – the synchronization should be done by the controller.

The CC1000PP is in this application equipped with circular or directional antenna. The Received Signal Strength Indicator (RSSI) is considered to check the

signal strength. With this feature it is possible to verify whether two devices are able to communicate with each other what helps to define if directional antennas are pointed in the proper way. Wireless data transfer tests presented in Section 6. were carried out with usage of this feature.

5 Communication Algorithm

Measurements algorithm include gathering data from the measuring nodes (Fig.3). Each temperature and humidity measurement nodes begin their work cycle with measuring the temperature/humidity. In this manner node checks whether the sensor is running correctly. If there is any error with the sensor the node is able to send information about it through the AtMega8 and CC1000PP transceiver, further to the AtMega32 and finally to the visualization software in the PC or write it onto SD memory card. If there are no errors and the sensor is working properly the node awaits to be called by the SLC. When this occurs the node read data from the sensor and sends the sensor's state and the temperature/humidity readout together with the batteries' state value to the SLC and further to the PC or SD memory card. In return the node receives data which states for how long it should be powered-down. The power-down period depends on the system work type (automatic or manual) and whether the situation needs more or less frequent measurements.

The SLC starts its work by checking whether all it's external circuits run correctly. After that it checks the actual time using the integrated clock circuit and initialize system gathering data from measurement nodes one, by one by sending them orders to response. If the automatic mode is chosen, the SLC checks fans and flaps state according to the control algorithm. Next, it saves the data on the SD memory card and listens whether the HMI is calling. If so, it exchanges data with it. If the HMI has not been calling for some time, the SLC may start a new read cycle, as it gets clock circuit call out signals. In this situation, after more than one reading cycle, the information exchanged with HMI consists of multiple readings data.

The visualization screen shows all the measured values on the run as well as fans and flaps states when running in automatic mode. If the personnel wants to take over the control, HMI makes it possible. With the HMI the personnel is able to switch to manual mode and control the state of the flaps, turn the fans on and off and set the desired measuring intervals. It also gives the possibility to show the history data logs – one can check all the data previously saved by the system and make sure the storage process was maintained properly. As the HMI doesn't take part in automatic control process, it may remain turned off if the visualization of the process is not needed.



Fig. 4 A farm at which wireless data transfer tests were carried out

The CC1000PP transceiver offers up to 10 dBm of an output power when tuned for 433 MHz frequency and up to 5 dBm when tuned for 868 MHz frequency. The first test was to define the difference between maximum communication distance between CC1000PP transceivers tuned for 433 MHz and CC1000PP transceivers tuned for 868 MHz while using the same type of antenna. Communication modules used in this tests were equipped with copper wire omnidirectional antennas connected with H-155 cable to the transceiver. Data transmission tests involved sending a set of data between communication modules and verifying their correctness and speed of acquiring the data. HMI module was placed inside of the control building whilst its antenna was placed next to the window at the ground floor of the building. In order to define the maximum communication range for both frequencies in respect to the terrain shape and obstacles SLC was mounted on a mobile platform. Communication module was placed approximately 0.5 m above the ground level. Results of this test are shown in Fig. 5, where the maximum communication distance for communication modules tuned for 868 MHz is defined by the yellow area and the maximum communication distance for communication

modules tuned for 433 MHz is defined by the orange area. In this test maximum communication distance for CC1000PP transceivers tuned for 433 MHz and 868 MHz were found to be 490 m and 146 m respectively. In both cases the possible communication was strongly influenced by the buildings in the way of the signal. Terrain shape was found to be an insignificant factor, although bushy trees decreased the signal strength.

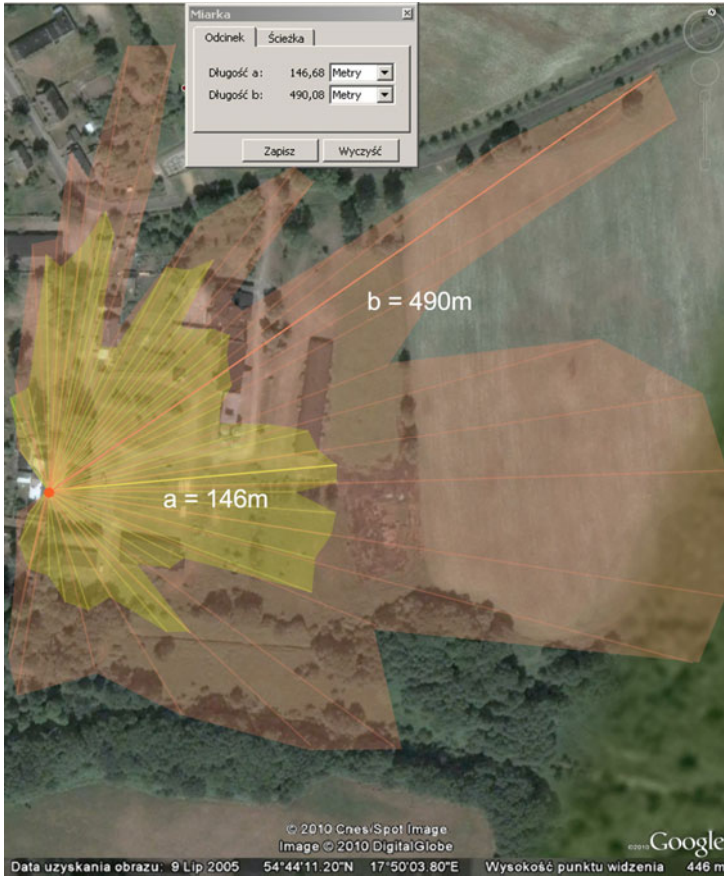


Fig. 5 The range of a possible wireless data transfer with use of copper wire omnidirectional antennas tuned for 433 MHz (orange) and 868 MHz (yellow) in a mockup of a farm satellite photo

Similar test was made to investigate the relationship between the maximum communication distance and type of the antenna used. CC1000PP transceivers in this test were tuned for 868 MHz and either copper wire omnidirectional antennas or directional 6-elements antennas having a 7dBi of forward gain were used (Fig.6). Directional antenna situated in the control building though was directed

towards the SLC communication module each time the one had moved. This test revealed that as long as the communication was made on an open ground, the system equipped with directional antennas had much greater maximum communication distance. Moving behind or entering any building though made those two systems work equally with almost the same maximum communication distance.



Fig. 6 Exemplary directional antenna AK 7/405-435

Further investigation of a communication possibility in respect to the position of the SLC communication module included placing the antenna inside of the storehouse buildings and varying the height and the position at which the antenna was placed. For this test copper wire omnidirectional antennas were used. Communication module at the control building was situated in the same position as in the previous tests. At this particular farm storehouses are situated 74 m and 140 m away from the control building and communication modules tuned for either of the investigated frequencies were able to communicate at that distance on an open ground with no problems. It was found, that the communication is possible as long as the antennas are placed higher than 0.3 m above the level of a ground in both buildings if no thicker obstacles than wooden walls are in between the communication modules. Communication possibility was found highly affected if any brick or metal walls are in between the communication modules. In the case of a storehouse located 140m away from the control building antenna situated behind any brick wall (approx. 40 cm thick) resulted in loss of a signal for either of the frequencies. In the case of a storehouse located 74 m away from the control building communication was possible if the antenna was placed in most of the places inside the storehouse. Although the further to the back of the storehouse the more places occurred problematic for the communication. More than one brick wall in the later case was also too much for the system to communicate.

In the last test the communication between two communication modules situated in the same storehouse building was investigated. In this case CC1000PP transceivers were tuned for 868 MHz and copper wire omnidirectional antennas were utilized. Communication modules were able to communicate with no problems in the storehouse of approx. 400 m², where the greatest distance in the straight line between communication modules was 20 m. Most of the inside walls and floors are wooden though and if one of the communication modules was placed outside the storehouse, signal was not always strong enough to communicate.

It is considered that in this particular farm copper wire omnidirectional antennas may be used for communication between control building and the SLC with one of the communication modules placed outside of the building. Directional

antennas were found to be the best choice for the communication on a distance greater than 350 m. In the case of installing the system in any other farm, where the distances between the buildings are similar, directional antennas are therefore preferable. The communication between measurement nodes and the SLC communication module is based on the copper wire omnidirectional antennas. Thus, one of the SLC communication modules has to be placed inside of the building. Moreover the distance between SLC communication module and measurement nodes should not exceed 40 m in a straight line if no brick walls are inside of the building and should not exceed 20 m in a straight line if storehouse has brick walls inside. Even though, not more than one brick wall shall be left in between.

7 Temperature Measurement Tests

In order to verify the correctness of measured variables readouts of temperature measuring nodes were gathered and compared to standard thermometer readings in the storehouse filled with the white mustard seeds. Tests were made in a summer season, when air temperature was 21°C. White mustard was stored in 50 cm high clamps. Temperature readouts sent to the HMI module were updated every 9 sec. After placing the temperature measurement node in the clamp temperature readouts stabilized after approximately 7 min and showed 12.4°C. Since that moment the temperature was found to be stable and no clutters were noticed till the node was moved (Fig.7).

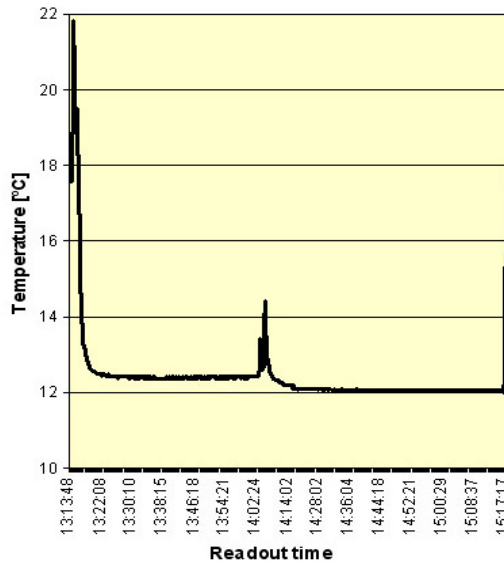


Fig. 7 Exemplary temperature readouts with use of a temperature measuring node

A standard mercurial thermometer placed next to the measuring node showed 12.2°C with no variations as well. After 40 minutes of testing in one spot the measuring node was moved onto another place. This time it indicated 12.1°C what was again in good correlation with readings from mercurial thermometer which showed 12°C. Those tests confirmed a high precision of the temperature measurements.

9 Conclusions

The system was designed and built for the purpose of use in an existing old farm. Five temperature sensors and one humidity sensor, as well as a SLC were built. Dedicated software for HMI module was written.

A proposed system gives the person managing the storage of agricultural products possibility to monitor temperature and humidity inside the stack of stored products. This information, gathered in the form of a history log, may certify that product has been stored in a proper manner, being in this way its quality certificate.

Wireless data transmission for communication between either HMI module to the SLC and the SLC to the measurement nodes is described and its correctness in respect to transmission range was verified. Measurement nodes, thanks to the wireless data transmission, may be placed in any spot in the storehouse what was not possible with any wire based system. This is considered a great improvement comparing to standard, wire based systems.

Tests of wireless data transfer and temperature readouts confirmed high reliability of the system and provided information about possible range of transmission with use of different communication antennas. Assuming the possibility of mounting antennas on the outside walls of the buildings and usage of copper wire, omnidirectional antennas the system is functional with storehouses placed up to approximately 140 m away from the control building. The system was found to work properly if the distance between measurement nodes and the SLC communication module did not exceed 20 m.

Future works are aimed in realizing an airflow control system which would actively influence the climate in the storehouse. The HMI module features and visualizing software will be improved accordingly to information received from management persons using the system already.

The system may also be expanded by a GSM module connected to the SLC. This solution would enable to inform a system user about any alarms and transfer information about the storage process to the management person wherever the one is.

References

- [Bohl et al. 2002] Bohl, W.H., Oberg, N., Kleinkopf, G.: Variable frequency drive fan control for potato storage, *The Spudvine*, pp. 1–2 (November 2002)
- [Czerko 2003] Czerko, Z.: Potatoes – new challenges. *Instytut Hodowli i Aklimatyzacji Roślin, Radzików* (2003) (in Polish)

- [Czerko 2007] Czerko, Z.: Space adaptation for potatoes storage. *Wiadomości Rolnicze* 12(40), 6–7 (2007) (in Polish)
- [Czerko 2008] Czerko, Z., Zgórska, K.: Storage technology of potatoes designated for processing. *Zeszyty problemowe postępów nauk rolniczych* 530, 69–79 (2008) (in Polish)
- [Schippers 1977] Schippers, P.A.: The rate of respiration of potato tubers during storage. *Potato Res.* 20(4), 321–329 (1977)
- [Tanenbaum 2004] Tanenbaum, A.S.: *Computer networks*. Prentice Hall PTR, New Jersey (2004)
- [Wachowicz 1998] Wachowicz, E.: Modeling of the selected processes in potatoes storehouses. In: *Inżynieria Rolnicza, Rozprawy habilitacyjne nr 2 series 4(5)* (1998) (in Polish); ISSN 1429-7264
- [WWW-1 2010] DS18B20 datasheet,
<http://datasheets.maxim-ic.com/en/ds/DS18B20.pdf>
(accessed March 2, 2010)
- [WWW-2 2010] SHT11 datasheet,
<http://www.sensirion.com/images/getFile?id=25> (accessed March 2, 2010)
- [WWW-3 2010] CC1000 datasheet,
<http://focus.ti.com/lit/ds/symlink/cc1000.pdf> (accessed March 2, 2010)

A Dynamic Programming Approach for Ambient Intelligence Platforms in Running Sports Based on Markov Decision Processes

J. Vales-Alonso¹, P. López-Matencio¹, J.J. Alcaraz¹, J.L. Sieiro-Lomba¹, E. Costa-Montenegro², and F.J. González-Castaño²

¹ Department of Information Technology and Communications, Polytechnic University of Cartagena, Spain
{javier.vales,pablo.lopez,juan.alcaraz,josel.sieiro}@upct.es

² Department of Telematic Engineering, University of Vigo, Spain
{kike,javier}@det.uvigo.es

Abstract. Outdoor sport practitioners can improve greatly their performance if they train at the right intensity. Nevertheless, in common training systems, performance is only evaluated at the end of the training session, and sensed data are incomplete because only human biometrics are analyzed. These systems do not consider environmental conditions, which may influence athletes' performance directly during instruction. In this paper, we introduce a decision making method for a multi-step training scenario based on dynamic program optimization and formulated as a Markov Decision Process, which allow athletes to complete heterogeneous training programs with several levels of exercise intensity. This methodology is applied in a pilot experiment of cross-country running. Environment and athletes are monitored by means of a wireless sensor network deployed over the running circuit, and by mobile elements carried by the users themselves, which monitor their heart rate (HR). The goal is to select, for a given user, a running track that optimizes heart rate according to a predefined training program. Results show that the proposal is of practical interest. It achieves a notable success in heart rate control over non-optimal track selection policies. The importance of environmental data is shown as well, since heart rate control improves when those data are taken into account.

1 Introduction

Wireless Sensor Networks (WSN) processing capabilities are growing rapidly, and they are getting increasingly embedded and seamlessly integrated in the physical world. WSN-assisted environments will be sensitive and responsive to the presence of people, allowing the development of context-aware personalized services.

Among the applications fields for these contextual services, sports may be one of the most benefited. For instance, in outdoors sports such as cross country

running and jogging, practitioners commonly use wearable computing devices, which provide useful telemetry about runner biometrics and practice-related events. Among other parameters, they measure heart rate, track routes, speeds and distances, and so forth. Often, the complexity of the data collected requires subsequent analysis by a human coach, sometimes with the aid of specific software. In this scenario the athlete is left alone during the training session, and she/he can only take decisions *a posteriori*.

Nowadays, corrective feedback on real-time runner performance and environment-awareness are limited in training systems. The evolution of computing and communications technologies may provide adaptive coordination between the user and the environment, modifying the system behavior according to changes in the environment or the user conditions (*e.g.* air temperature or runner location).

This work introduces an ambient intelligence (AmI) system prototype for running sports in open areas. Fig. 1 shows our approach. Athletes train in a field with track alternatives. Each track has a different *hardness* degree and the weather conditions may vary (*temperature*). The goal is to select at each path junction the most suitable track for the runner, in order to fulfill an overall training program. In our system we aim at controlling the HR of the user. Different training HR ranges are possible (*e.g.* cardio-training, fat-burn, etc.).

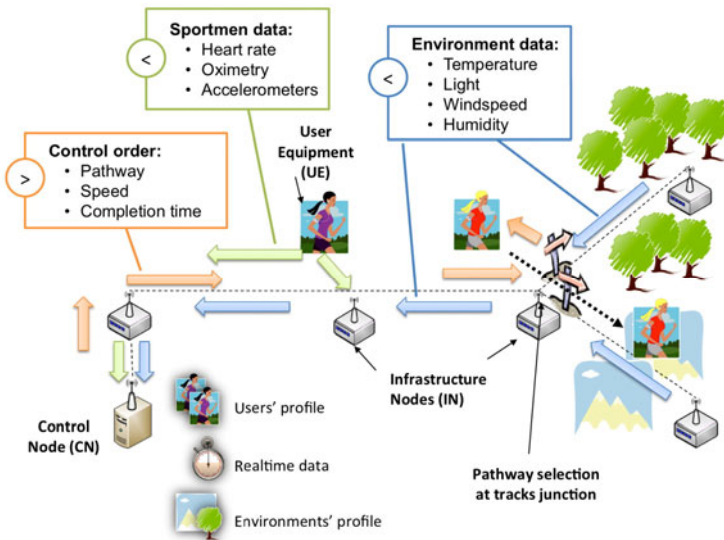


Fig. 1 System architecture and data flow of the sensed data and the command flow to/from the Control Node (CN)

Before training starts, athletes or coaches select the desired HR profiles during exercise, *e.g.* 50% tracks in cardio-training HR range and 50% tracks in fat-burn range. We denote this kind of training as *heterogeneous*, as it involves different HR ranges. Then, the criterion for path selection is to choose the track that maximizes the correct distribution of the HR intensity levels during training. This is a multi-step decision making process, since decision at step n will affect HR at steps $n+1$, $n+2$, and so on. In addition, as the conditions of our problem change over time, we require the system to be able to record athletes' performance along the session, as well as the environmental conditions.

In this work we improve the system previously presented in [Vales-Alonso et al 2010; López-Matencio et al 2010]. In those previous works the goal was to keep the athlete in static *-i.e.* homogeneous- training range (either cardio-training or fat-burn regime), introducing the simplification of selecting a track regardless of the future decisions (single step decision). In the present chapter, however, we address the problem of multi-step training decisions.

The technique we have selected for the multi-step decision process is a mathematical optimization method based on Dynamic Programming (DP) [Bellman 2003]. Specifically, a Markov Decision Process (MDP) is used due to the stochastic nature of system evolution. DP allows breaking a multi-step planning problem into simpler steps at singular moments in time (in our case, at each track selection instant). The Bellman equation recursively determines a correspondence between the increment of the overall benefit (or cost) in two consecutive periods. It evaluates how the decision makes the system and its benefit (or cost) evolves over time. Section 3 describes the formulation of our problem as a MDP dynamic program.

Besides, as training conditions can rapidly vary, the system is expected to take real-time decisions to meet the user needs. Therefore, constant user and environmental monitoring are necessary. Although the present work mainly focuses on developing a suitable decision engine, the development of an architecture that enables real-time data harvesting was also emphasized. Thus, efficient communications and location protocols are also two key areas of this work. To perform these tasks our system relies on a WSN deployed over a rough outdoor area. This networked system (see Fig. 1) has two main elements: Infrastructure Nodes (IN) and User Equipment (UE). The former are static and measure environmental variables and track the runner position along the course. The latter are carried by users and, thus, are mobile devices. INs include sensing capabilities to acquire ambient temperature, and mobile devices can sample the pulse rate of the runner. The "intelligence" in our approach (that is, the track selection procedure) resides in a particular fixed node (Control Node or CN) that runs the decision engine with the aforementioned dynamic programming optimizer. This special node receives the data from all nodes. The information gathered maps temperature to the training tracks and keeps runners located. In addition, the runners' HR is continuously delivered to the CN via the INs. This allows feed the dynamic programming core with real-time feature measurements.

The prototype system has been implemented using standard WSN hardware (MICAz and IMOTE2 motes from Crossbow Technology Inc.). The INs and the CN have been implemented using MICAz devices that provide processing and communication capabilities. In addition, both types of nodes measure the environmental parameters. Although in our prototype only temperature monitoring is activated, the MTS400 sensor board may measure light, temperature, humidity and barometric pressure. In addition, each MICAz is equipped with two panel antennas (2.4 GHz Stella Doradus 24-8080 planar antenna). The main reason to select these antennas is to expand communication range between the stations, allowing a lower density of IN nodes, and, therefore, lower installation and operation costs. Fig. 2(a) shows two deployed INs.

The UE node was designed to measure human biometrics (HR in our implementation). The UE acts as an interface between the system and the user, delivering training orders to the athletes. To perform these operations, several modules, as shown in Fig. 2(b) and 2(c), compose the UE: an IMOTE2 IPR2400 serves as an expandable wireless sensor platform; an IMB400 plays speech messages, *e.g.* “select track #2” (there is one message recorded for each command in the IMOTE2 memory); and an integrated iPOD 3211 pulse oximeter device from Nonin Medical company measures the HR. We chose the iPOD for its lightness, size, and easy integration with the UE (via a RS-232 interface).

Since network nodes must be autonomous, power saving is an important requisite for protocol design/selection. Thus communication protocols have been designed to keep signaling messages at a minimum. For routing, a simple tree topology was selected, since information flows from the IN nodes to the CN or vice versa. This suggests a tree-form network topology, with the CN as a root. This topology reduces signaling burden, and simplifies routing.

The previous work [Vales-Alonso et al. 2010] describes the system prototype and the communication protocols in depth. Finally, our approach can be generalized to other outdoor sports (*e.g.* cycling or biathlon), where terrain slopes and atmospheric conditions play an important role in practice [Vihma 2009].

The rest of this chapter is organized as follows: Section 2 discusses related work. Section 3 explains the analytical method we have developed to build the decision engine. A pilot experiment is presented in Section 4. Section 5 analyzes the results and, finally, Section 6 concludes the chapter.

2 Related Work

Wearable computing is nowadays common in athletes’ training. These devices are no longer exclusive for elite sports, but available to the general public (*e.g.* heart rate monitors). In fact, the widespread use of commodity hardware (*e.g.* the Apple iPhone) has also lowered the barrier to create mobile training applications [Saponas et al. 2008].

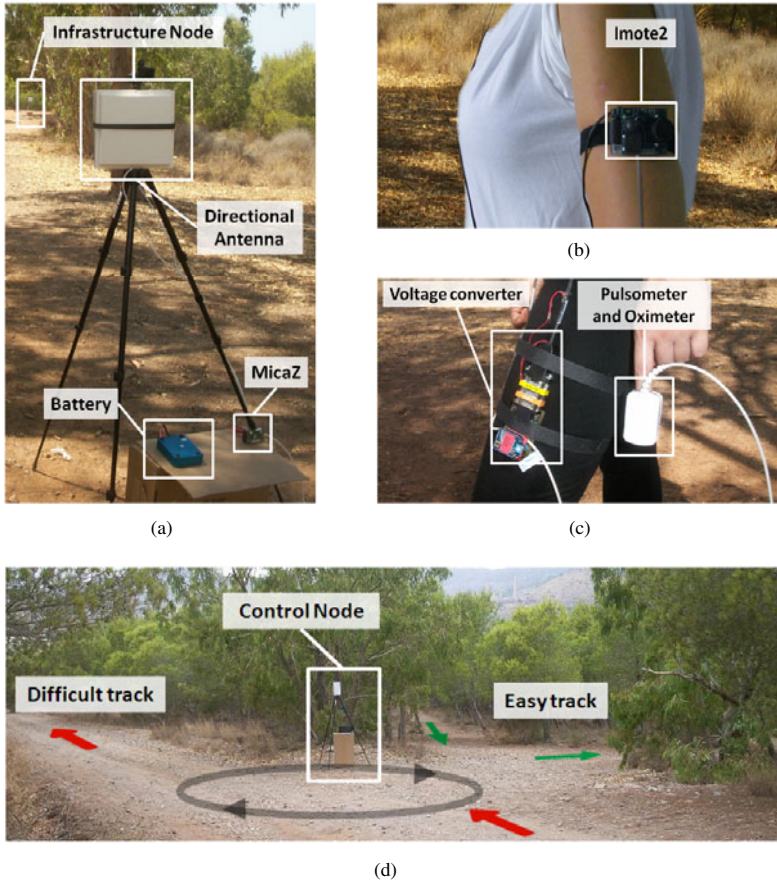


Fig. 2 Deployed hardware. (a) Infrastructure node. (b) Upper User Equipment. (c) Lower User Equipment. (d) Crossroad

To some extent, the availability of these devices is simply the first step towards the advent of true contextual services. Future developments will aim at expanding the range of monitored data (environmental conditions, detailed user data) and producing useful actions and information based on them. WSNs represent one of the enabling technologies for that evolution.

Several context-aware applications for athletes' training have already been introduced. In previous work [Vales-Alonso et al. 2010; López-Matencio et al. 2010] we show the usefulness of the ambient intelligence paradigm for sports practice. Our goal was selecting, for a given user, suitable tracks where HR would typically lie in the desired HR range. Nevertheless, we did not consider the

optimal fulfillment of a heterogeneous multi-step training program, but a simple homogeneous single-step problem instead. In [Vales-Alonso et al. 2010] decision making was based in (m, s) -splines interpolation of the HR signal, whereas in [López-Matencio et al. 2010] a k-NN classification engine was developed. Besides, [Vales-Alonso et al. 2010] demonstrated the feasibility of the communication and location system architecture described in the introduction, by means of tests in real outdoor training scenarios.

In MarathonNet [Pfisterer et al. 2006], a WSN monitors runners in marathon events. Sensors on runners collect HR, time and location data. These data are sent to a central database via base stations along the track, where they are subsequently analyzed. Base stations can communicate with the central database by means of GPRS, WLAN, or a wired network link. The sensor nodes in our work have similar functionality, but they also act as information routers.

More intelligent systems have been developed for sports scoring purposes. For example, in [Spelmezan and Borchers 2008], a sensor system intends to provide immediate feedback to alert users of incorrect movements and body positions. The prototype employs sensors attached to the human body and embedded in the boots that detect bad snowboarding practices. In [Ghasemzadeh and Jafari 2009], the authors describe a feedback system based on sensors in a golf club, which capture the golf swing. Swing motion is preprocessed locally, and then sent to a control station for further analysis. The quality of the swing motion is expressed as the degree of deviation from the target line and is computed by a linear discriminant analysis technique.

To sum up, most previous systems do not sense environmental data, but only information from the athlete, and they only provide limited real-time feedback in some cases. Our proposal aims to overcome such limitations, providing useful automated feedback to runners based on an ambient intelligent system, capable of monitoring training performance selecting correct tracks according to a predefined heterogeneous training program.

Dynamic programming techniques are applied in many areas such as economics, medicine or artificial intelligence, which included some developments in sports. In [Clarke 1988] the authors calculate at any stage of the innings the optimal scoring rate, including an estimate of the total number of runs to be scored and the chance of winning in an one-day game. The results obtained help to study optimal batting tactics (*e.g.* the best run rate at any stage of the innings).

Optimization problems in outdoor sports like competition glider flying and sailboat racing are challenging due to uncertain environmental conditions. The pilot must take a continuous series of strategic decisions, with imperfect information about the conditions he will encounter along the course and later in the day. In a competition, these decisions seek to maximize cross-country speed, and hence the final score in the contest; even in noncompeting flying the pilot must fly fast enough to complete the chosen course before the end of the day. The work in [Almgren and Tourin 2004] addresses the problem of uncertain future atmospheric

conditions by constructing a nonlinear Hamilton-Jacobi-Bellman equation for the optimal flight speed, with a free boundary describing the climb/cruise decision.

All these mentioned works illustrate the mere usefulness of dynamic programming techniques to provide good decisions along a sport game. Nevertheless, none of them are implemented in physical systems. In this work we apply dynamic programming optimization to provide athletes with actual real-time feedback.

3 Decision Engine: Markov Decision Process

The decision engine runs in the Control Node (CN), and is in charge of the system's intelligence. Its goal is to maintain the runners' HR within a given target intensity. The relationship between HR and sport activity is affected by many factors, such as the effect of temperature and exercise intensity, as studied in [Backx et al. 2003]. Thus, the runner's HR and, in turn, his/her performance, is difficult to characterize. As an example of this difficulty, Fig. 3 depicts the HR samples of a runner training session from the validation of our pilot (see Section 4).

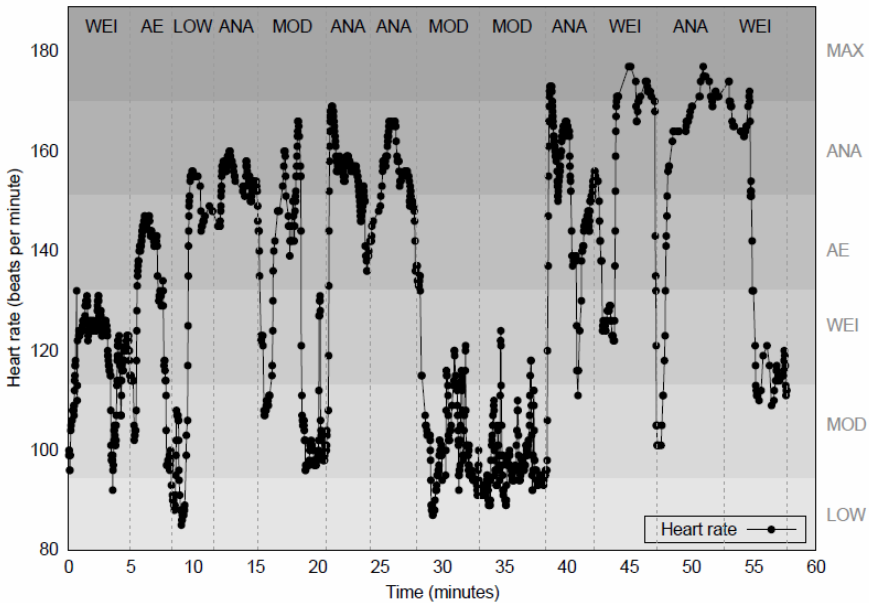


Fig. 3 HR samples of a training session

Heart rate can be divided in different intensity levels or classes according to the type of training [Haskell et al 2007]:

VO₂ MAX (Maximum effort) if HR is between 90% and 100% of the maximal recommended HR.

Anaerobic (Hardcore training) if HR is between 80% and 90% of the maximal recommended HR.

Aerobic (Cardio training/Endurance) if HR is between 70% and 80% of the maximal recommended HR.

Weight Control (Fitness/Fat burn) if HR is between 60% and 70% of the maximal recommended HR.

Moderate activity (Maintenance/Warm up) if HR is between 50% and 60% of the maximal recommended HR.

In addition, we considered two additional ranges in this work, for the correct characterization on the HR evolution process:

High if HR is beyond the maximal recommended HR.

Low if HR is below 50% of the maximal recommended HR.

An accurate formula to compute the maximal HR [Tanaka et al 2001] is:

$$HR_{max} = 208 - 0.7 \times age \quad (1)$$

Fig. 3 indicates the HR range of each training period. Note that HR samples may belong to different HR ranges, even for the same stage. Throughout this paper we consider that a stage belongs to a given HR regime if it contains a majority of samples in its range.

For our formulation, let us assume a partition of HR ranges consisting of m non-overlapping levels, each one comprising heart rate in the set $HR^i = [hr_i, hr_{i+1}]$ for $i = 1, \dots, m$, where hr_i denotes the lower bound of the heart rate associated to range i .

As stated in the introduction, in our scenario, the runners train on a circuit containing several tracks (t tracks), and each track is associated to different difficulty level. At each stage, a track must be selected. Therefore, a training session consists in a sequence of N tracks. The goal of the session may be either to perform all tracks within a selected HR range, or to fulfill a multi-goal training, for instance performing $N-4$ tracks in range HR^2 and the remaining four in HR^3 . The HR will depend, among others, on the following aspects: athlete's condition, previous stages, the hardness of the track, and environmental conditions. From these variables (features in decision making terminology), the system must decide which track to follow for the upcoming stage.

Thus, the CN must select, at each decision epoch, the best track in order to accomplish the overall training goal. Previous approaches (see Section 2) consider only the problem of selecting the best decision for one stage, independently of the future evolution. That is, optimization is done for a single-step scenario.

In this work, we consider the multi-stage case. That is, track selection in order to fulfill training goals, considering the future evolution of the athletes. This problem can be formulated as a dynamic program. This formulation comprises the following elements:

- The stage where the system is at. In a finite horizon problem, as the one described (the training session consists of N stages), the stage is denoted by $n = 0, \dots, N$.

- The state of the system at stage n , x_n , containing all the required information for the CN to take a decision at each stage (e.g. the HR). Let X denote the set of possible states.
- The control applied at each stage, u_n , is the track selected (each one with a different difficulty level).
- The reward obtained at each stage, $r(x_n, u_n)$, is a function of the state and the control selected. Its formulation depends on how the system's goal is defined.
- The value function, $J^n(x_n)$, is the maximum expected reward that can be gathered from stage n to stage N , i.e. if the optimal decisions u_n are taken from the current stage to the last one.

In the general multi-goal case, the target is to perform R^1 tracks in HR^1 , R^2 in HR^2 and so forth. Hence $N = R^1 + \dots + R^m$. Whenever a track is traversed in HR^i there must be a positive reward added to the value function if the number of remaining loops in HR^i is positive. Otherwise, there is no reward. Therefore in the general problem, x_n comprises three elements: (1) the athlete's HR, which we denote as hr_{x_n} , (2) the selected track and (3) the number of loops remaining of each HR class to accomplish the objective.

Therefore, the reward function for the problem is defined as:

$$r_n(x_n, u_n) = \begin{cases} 1, & R^{hr_{x_n}} > 0 \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

Note that, in this case, the reward does not depend either on the selected control or the stage, but only on the state.

The dynamic program can now be formulated. It basically consists computing the value function $J^n(x_n)$ recursively as the sum of the reward expected in state x_n plus the value function for the next stage, $n+1$. An optimal policy has the property that whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy regarding the state resulting from the first decision (Bellman's principle of optimality).

$$J^n(x_n) = r(x_n) + \max_{u_n} \left\{ E \left\{ J^{n+1}(f_n(x_n, u_n)) \right\} \right\}, \forall x_n \in X, n = 0 \dots N \quad (3)$$

To solve this equation, the transition mapping $f(\cdot)$ must be known. However, the transition from one state to another is not deterministic, due to the random nature of HR in outdoor sports. To overcome this limitation, we can model our system as a Markov Decision Process (MDP). That is, the decisions must be made in a Markov context where the system stochastically evolves from one state to the next

one. In a MDP it is assumed that the probabilities of going from state i to state j when control u_n is applied at stage n are known, yielding to transition probability matrices $\mathbf{P}^{(n)}(u_n)$. Therefore the recursive equation can be written in the following matricial form:

$$\mathbf{J}^n = \mathbf{r} + \max_{u_n} \left\{ \mathbf{P}^{(n)}(u_n) \mathbf{J}^{n+1} \right\} \quad (4)$$

In order to compute the transition matrices of the system we have considered a static transition scheme. That is, the transition probabilities do not depend on stage n . In our experiments, we have computed the probability transition matrices by evaluating the frequencies of the transition events experimentally (see Section 4). Therefore, we only need to compute the transition probability by estimating the relative frequencies of going from going from each combination of track and HR (track_i, HR^i) to each subsequent combination of track and HR (track_j, HR^j).

Since N stages are considered, the dynamic program has a finite horizon and the value function for the last stage is $\mathbf{J}^N = \mathbf{r}$. Besides, since in real training the athletes start with a warm-up (WU) period, no reward is considered for the first stage, that is, $\mathbf{r}_0 = \mathbf{0}$.

If the initial state x_0 is known, the value function is evaluated just as $J^0(x_0)$. With the reward function as previously defined, this value can be interpreted as the average number of tracks that the athlete will perform in the correct HR range if the optimal control is applied. The closer the value to N , the better the training. In the next section we evaluate this model with a simple experiment, and compare the results with some simple non-optimal policies.

4 Pilot Experiments

A pilot experiment was deployed in a cross-country training circuit using the architecture depicted in Fig. 1. As described previously, the goal was to keep the planned activity levels for athletes.

Fig. 4 shows the operation of the decision engine. The system requires information from previous training sessions to build the transition matrices. Then, those data are used as an input to the MDP block, which selects the optimal policy using the table that corresponds to current environmental conditions. Note that the optimal policy is in fact a mapping between all the system states and their associate best controls (the tracks). This representation of the optimal policy is often denoted as *lookup table* in MDP-related literature. At each decision epoch, the Track Selection block retrieves from the lookup table the track that the runner should follow according to the athlete's HR, the last track selected, and the remaining tracks in each HR class (*i.e.* the parameters determining the current state).

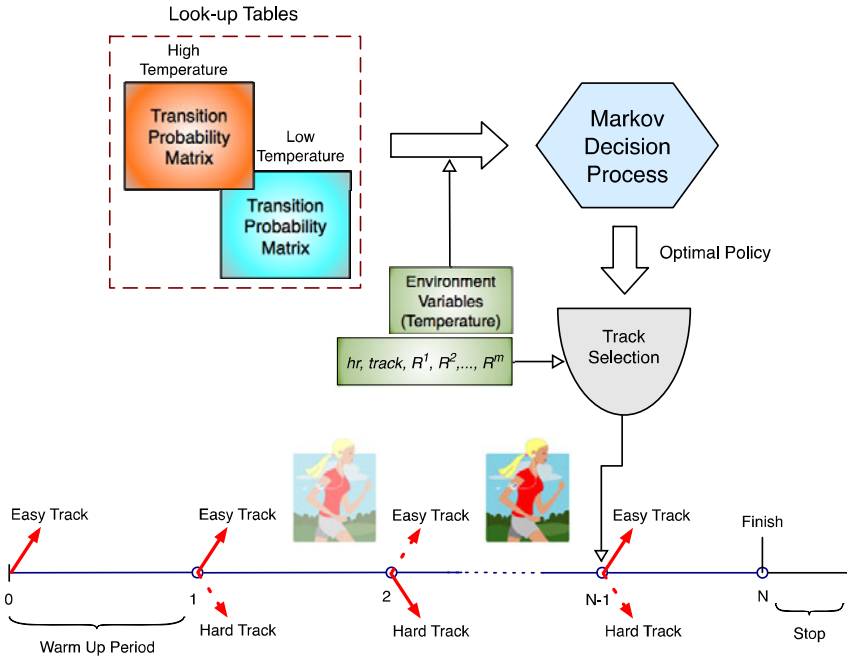


Fig. 4 Decision engine functional scheme

4.1 Deployment

The prototype has been validated in a cross-country circuit near Cartagena (Spain) (see Fig. 5). It consists of two interconnected loops (red and blue) with different hardness and environmental conditions due to:

- Closeness to the coast in some areas of the circuit, in which wind is constant.
- Different terrain slopes, since part of the circuit is on a hill whose height is 97 meters over sea level, with some slopes of 14%.
- Shadow, depending on training hours and the trees along the circuit (as can be observed in the red circuit in Fig. 5(a)). This has influence on temperature.
- Different lengths: 1.1 and 0.9 kilometers for the red and blue tracks, respectively.

The red track is considered hard due to its length and steeper slope. The blue track is considered easy. We deployed ten INs in the hard sector of the training area and eight in the easy one, as shown in Figure 5(b). The CN was placed in the junction of both loops. Thus, regarding the model described in Section 3, the number of control decisions is $t = 2$, that is, $u_n = \{Hard, Easy\}$, independently of the stage n . The optimal policy must decide between the hard or the easy track at each stage (see Fig. 5(a)).

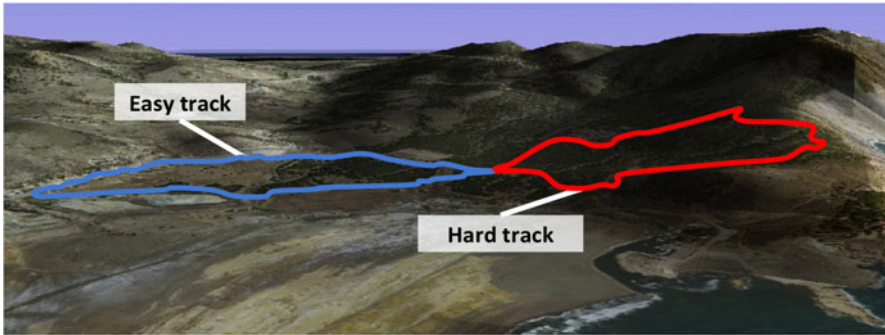
4.2 Computation of the Transition Matrix

To compute the transition matrix an athlete performed a series of training tests:

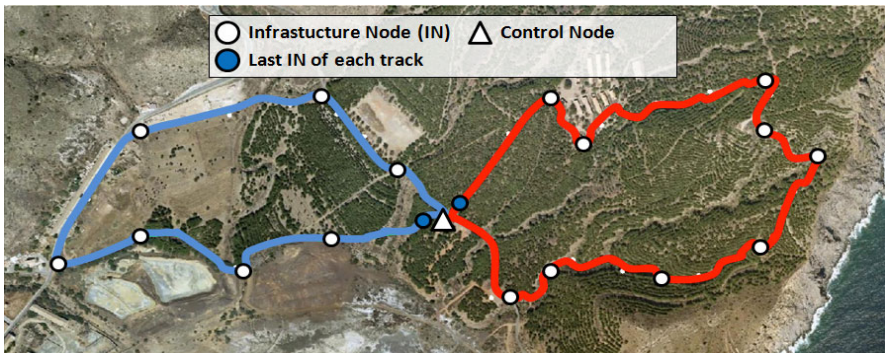
1. There was an initial warm-up (WU) (the first in Fig. 3). During which the athlete trains in the easy sector of the circuit. The data from this sector is discarded (there is no reward associated to this stage in our mathematical model).
2. Each training session consists of 12 loops along the circuit (after the warm up sector). At each loop the athlete ran either in the hard track or in the easy one. The order was the same for all sessions (Table 1). Ambient temperature was monitored as well.

Table 1 Course profile of a training session. WU- warm up period; E- easy track; H- hard track

Loop	1	2	3	4	5	6	7	8	9	10	11	12	13
Hardness	WU	E	E	E	H	E	E	H	H	E	H	H	H



(a)



(b)

Fig. 5 Aerial sights of the training circuit. 5(a) Cross training circuit. 5(b) Deployed infrastructure

From these experiments, the ratio of transitions from one state to any other state was computed, thereby obtaining the transition matrices. Indeed, three matrices were obtained:

- Low Temperature matrix, which only accounts for transitions in low temperature (*i.e.* average track temperature below 25 Celsius degrees).
- High Temperature matrix, which only accounts for transitions in high temperature (*i.e.* average track temperature higher than 25 Celsius degrees).
- Absolute matrix, computed using information from all transitions, regardless of the temperature.

Besides, to reduce the overall number of states of the system, which may render the algorithm computationally infeasible, only three HR classes have been used:

- HR^1 , comprising Low, Moderate and Weight Control HR classes.
- HR^2 , which corresponds to Aerobic HR class.
- HR^3 , comprising Anaerobic, VO2 MAX and High HR classes.

5 Results

The performance of the dynamic program can be computed applying the transition matrices of the experiments in the method described in Section 3. Recall from Section 3 that the value function represents the number of tracks that are successfully performed in some of the requested HR classes. The corresponding results are described in this section. Moreover, the optimal policy is compared to the following ones:

- Easy policy, *i.e.* selecting always the “easy” track of the circuit.
- Hard policy, *i.e.* selecting always the “hard” track of the circuit.
- Worst policy, easily computed substituting the max operator for the min operator in Eq. (4).

The initial state is always $x_0 = (HR^1, Easy, R^1, R^2, R^3)$, where R^1, R^2, R^3 denote the training configuration. Hence, the value function of the whole test is computed from Eq. (4) as $J^0(x_0)$. Let us remark that this performance metric is scalar, since all athletes depart from the same initial state. In the next sections, different training programs are evaluated. In addition, note that in all tests the temperature is considered constant during all the training.

5.1 Single-Goal Optimization

In this case the training program to fulfill is $(R^1 = 0, R^2 = i, R^3 = 0)$ that is, performing all tracks in the aerobic (cardio-training) regime. Figure 6 shows the

results for $i = 1, \dots, 10$ using the absolute matrix. As expected, the optimal policy achieves better performance than non-optimal ones. In addition, Fig. 6 depicts a comparison of the optimal policy using either a specific matrix (high and low temperature) or the absolute one. If the specific ones are used, the value function improves significantly, especially in high temperature conditions.

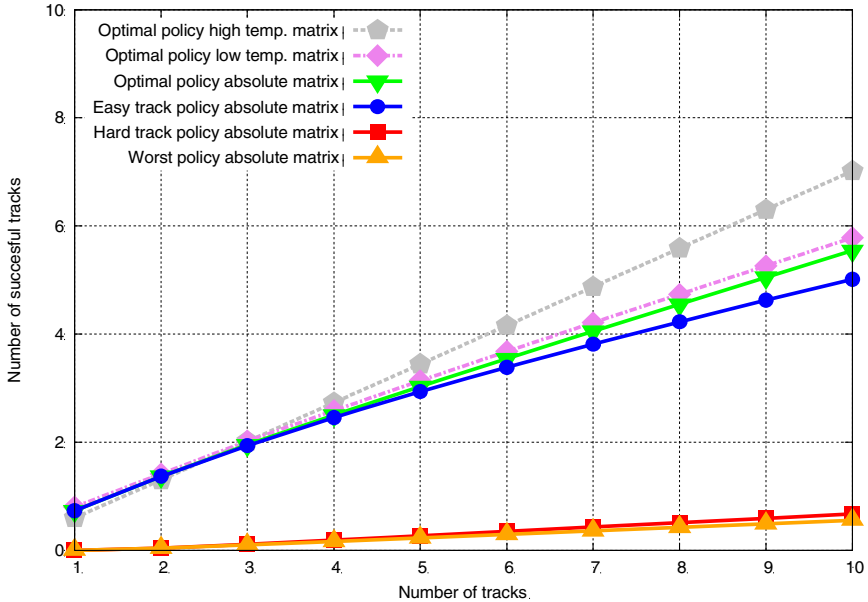


Fig. 6 Single-goal (0,i,0) training. Comparison of policies. Optimal policy results with specific and absolute matrices

This experiment clearly demonstrates that using environmental information is highly advisable.

5.2 Multi-goal Optimization

In addition, two multi-goal trainings have been studied:

- $(R^1 = 10 - i, R^2 = i, R^3 = 0)$. In this case, 10 loops must be performed with a different distribution in two HR classes. Figure 7 shows the results. The trends are similar to those of the single-goal experiment, yet with a larger difference between the optimal policy and non-optimal ones.
- $(R^1 = 7 - i, R^2 = i, R^3 = 3)$. In this case, 7 loops are always performed with a HR distribution across the three classes. Figure 8 shows the results. Again, similar trends as in previous experiments have been obtained, confirming the suitability of our proposal.

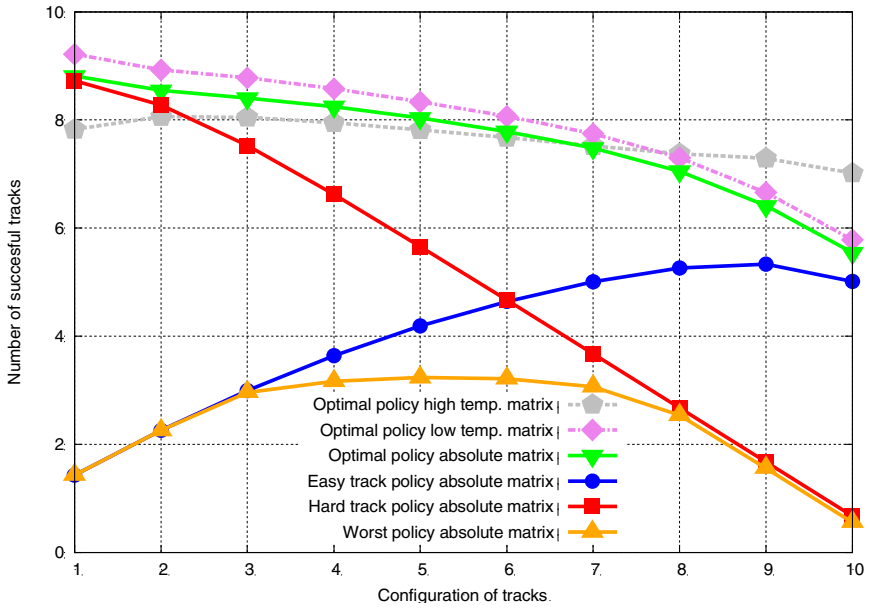


Fig. 7 Multi-goal (10-i,i,0) training. Comparison of policies. Optimal policy results with specific and absolute matrices

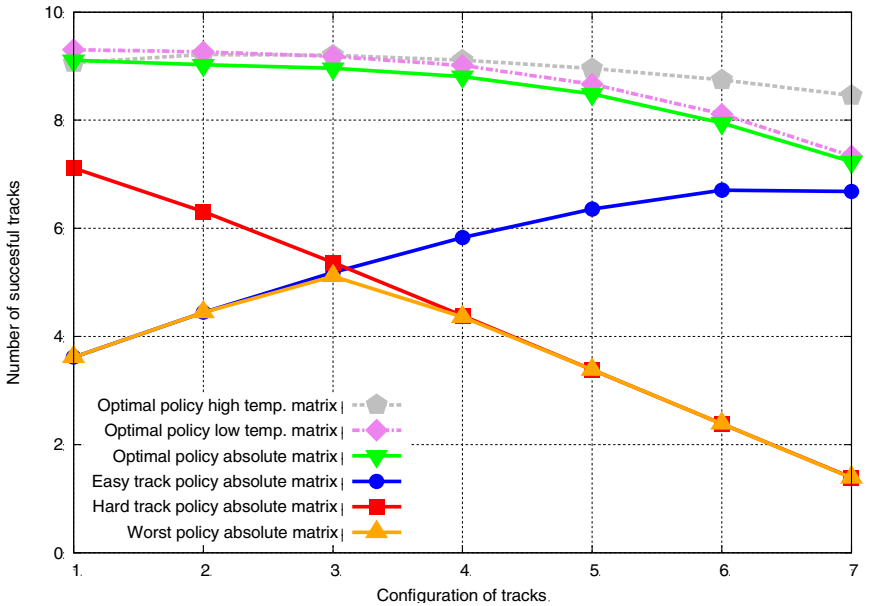


Fig. 8 Multi-goal (7-i,i,3) training. Comparison of policies. Optimal policy results with specific and absolute matrices

6 Conclusions

This paper presents a theoretical framework for applying dynamic programming to outdoor cross-country running sport. The computation of optimal policies requires the transition matrices of Markov decision processes. These matrices have been obtained experimentally in a pilot test-bed that allowed us to evaluate our methodology for different training goals. The results show the benefits of the optimal policies over trivial ones, as well as the necessity of ambient monitoring to improve system performance. A wireless sensor network to monitor athletes' parameters also provides that environmental monitoring. The whole system (WSN + decision making process) represents a first stage towards an ambient intelligence environment for outdoor sports.

Acknowledgment

This work has been supported by grants TEC2010-21405-C02-02 CALM (Ministerio de Ciencia e Innovación, Spain), TSI-020301-2008-2 PIRAmIDE (Ministerio de Industria, Turismo y Comercio, Spain). It has also been developed within the framework of "Programa de Ayudas a Grupos de Excelencia de la Region de Murcia", funded by Fundacion Seneca, Agencia de Ciencia y Tecnologia de la Region de Murcia (Plan Regional de Ciencia y Tecnologia 2007/2010).

Referentes

- [Almgren and Tourin 2004] Almgren, R., Tourin, A.: Optimal Soaring with Hamilton-Jacobi-Bellman Equations (2004), <http://www.cims.nyu.edu/~almgren/optsoar/optsoar.pdf> (accessed October 25, 2010)
- [Backx et al. 2003] Backx, K., Someren, K.V., Nevill, A., et al.: Mathematical prediction of one hour cycle time trial performance under different ambient temperatures. *Medicine & Science in Sports & Exercise* 35(5), S30 (2003)
- [Bellman 2003] Bellman, R.E.: Dynamic programming. Princeton University Press, Princeton (2003) (republished)
- [Clarke 1988] Clarke, S.R.: Dynamic programming in one-day cricket-optimal scoring rates. *J. of the Operational Research Society* 39(4), 331–337 (1988)
- [Ghasemzadeh and Jafari 2009] Ghasemzadeh, H., Jafari, R.: Sport training using body sensor networks: A statistical approach to measure wrist rotation for golf swing. In: *The 4th Int. Conf. on Body Area Networks*, Los Angeles, CA (2009)
- [Haskell et al. 2007] Haskell, W., Lee, I., et al.: Physical activity and public health: updated recommendation for adults from the American College of Sports Medicine and the American Heart Association. *Circulation* 116(9), 1081 (2007)
- [López-Matencio et al. 2010] López-Matencio, P., Vales-Alonso, J., González-Castaño, F.J., et al.: Ambient intelligence assistant for running sports based on k-NN classifiers. In: *Proc. of 3rd Int. Conf. on Human System Interactions*, Rzeszow, Poland, pp. 605–611 (2010)

- [Pfisterer et al. 2006] Pfisterer, D., Lipphardt, M., Buschmann, C., et al.: Marathonnet: adding value to large scale sport events—a connectivity analysis. In: Proc. of the First Int. Conf. on Integrated Internet Ad Hoc and Sensor Networks, p. 12. ACM, New York (2006)
- [Saponas et al. 2008] Saponas, T., Lester, J., Froehlich, J., et al.: Ilearn on the iphone: Real-time human activity classification on commodity mobile phones. University of Washington CSE Tech. Report UW-CSE-08-04-02 (2008)
- [Spelmezan and Borchers 2008] Spelmezan, D., Borchers, J.: Real-time snowboard training system. In: CHI 2008 Extended Abstracts on Human Factors in Computing Systems. ACM, New York (2008)
- [Tanaka et al. 2001] Tanaka, H., Monahan, K., Seals, D.: Age-predicted maximal heart rate revisited. *J. of the American College of Cardiology* 37(1), 153 (2001)
- [Vales-Alonso et al. 2010] Vales-Alonso, J., López-Matencio, P., González-Castaño, F.J., et al.: Ambient intelligence systems for personalized sport training. *Sensors* 10(3), 2359–2385 (2010)
- [Vihma 2009] Vihma, T.: Effects of weather on the performance of marathon runners. *Int. J. of Biometeorology*, 1–10 (2009)

Prometheus Framework for Fuzzy Information Retrieval in Semantic Spaces

A. Andrushevich¹, M. Fercu¹, J. Hopf¹, E. Portmann², and A. Klapproth¹

¹ CEESAR – iHomeLab, Lucerne University of Applied Sciences and Arts,
Horw, Switzerland
{aliaksei.andrushevich,michael.fercu,joern.hopf,
alexander.klapproth}@hslu.ch

² Information Systems Research Group, University of Fribourg, Fribourg, Switzerland
edy.portmann@unifr.ch

Abstract. This paper introduces a novel vision for further enhanced Internet of Things services. Based on a variety of data (such as location data, ontology-backed search queries, in- and outdoor conditions) the Prometheus framework is intended to support users with helpful recommendations and information preceding a search for context-aware data. Adapted from artificial intelligence concepts, Prometheus proposes user-readjusted answers on umpteen conditions. A number of potential Prometheus framework applications are illustrated. Added value and possible future studies are discussed in the conclusion.

1 Introduction and Related Work

In the epigrammatic triumphant history of the Internet, first the World Wide Web was created as a CERN-project initiated by Timothy Berners-Lee. In the early Web, retrospectively referred to as Web 1.0, a small number of so-called information producers published their insights as a collection of static HTML pages and a great mass of consumers was opposed to these insights.

In the late 90s DiNucci first mentioned the term Web 2.0 and thus caused the advent of a new slogan. Afterwards O'Reilly declared that Web 2.0 technically did not differ from the earlier Web 1.0. In contrast to static expert-generated content, interactive elements are crucial in Web 2.0. Ever since the first use of the term Web 2.0 Berners-Lee deployed it as a marketing buzzword. He tried instead to advertise his future visions of the WWW with his ideas about the Semantic Web. The Semantic Web is an emerging development of the Internet in which not only the meaning or semantics of information is defined but also services on the Web, making it possible for machines to understand and satisfy the requests of both

people and machines. Because of the enhancement to a machine-understandable Internet, the Semantic Web is sometimes called Web 3.0. Berners-Lee specified the Semantic Web as a component of Web 3.0.

In fall 2009, O'Reilly and Battele went further deeper by defining another upcoming buzzword "Web Squared" where the Web no longer is a collection of static pages that describe something in the world. Instead they outline in [O'Reilly and Battelle 2009] the Internet of Things. The Internet of Things exemplifies ubiquitous computing and "things that think". It describes a form of physical computing and is a non-deterministic, open network in which self-organized or intelligent entities will be interoperable and able to act independently – pursuing their own or shared objectives – depending on the context, circumstances or environment as described in [IoT Summary 2005].

These networks are delineated as ubiquitous computing models, which are post-desktop models of human-computer interaction, considered as an advancement from the desktop paradigm. However, when Web meets the world the vast data produced are mostly stored or provided in an unstructured way distributed on different systems; globally considered. An important case in ubiquitous computing for this reason is to find relevant information.

To find context relevant information in connection to real-world human-settled-environment services, processes and systems become more crucial. A viable way to improve existing searches and to approach a universal Semantic Web (that is virtual Internet together with the real Internet of Things) is to teach the Web based on an automatically built ontology the meaning of real world parameter values. Additionally machine learning, a scientific discipline that is concerned with the design and development of algorithms, can be used to learn based on sensor and/or Internet data as Bishop elucidates in [Bishop 2008]. After Kasabov approaches to ML are expert systems whereby "an expert system is a program that can provide expertise for solving problems in a defined application area in the way the experts do" as explained in [Kasabov 1996].

To present expert-system-based real-time information in a clever way and to force the users to interact with the information, real and virtual worlds can melt into a new augmented reality. Thus augmented reality can be considered as an event of ubiquitous computing where virtual computer-generated symbolisms are superimposed into physical real-world environments, creating a mixed reality as Azuma et al. explain in [Azuma et al. 2001].

In Greek mythology, Prometheus (Ancient Greek for forethought) was a champion of humankind known for his intellect. He is said to be the benefactor of culture and the great instructor of all human beings. The ambitious project's goal, named after this transcendent ideal, is to offer the human race further techniques to master their human duties and responsibilities in an easier way by pooling virtual and real world aspects.

2 Applications of Fuzzy Sets Theory

This section aims to introduce some concepts of fuzziness which deals with vague reasoning. To emphasize the benefits that fuzziness brings to artificial intelligence the first section 2.1 brings the affinity of human thinking and fuzziness in. Section 2.2 introduces fuzzy set theory and classification and section 2.3 fuzzy expert systems.

2.1 Fuzziness and the Human Factor

Mentioned in [Zadeh 1965] inter alia, fuzzy logic – a particular type of multi-valued logic emerged as a corollary of Zadeh’s proposition of fuzzy set theory— follows the way humans think and helps to better handle real world facts, since human reasoning is undichotomic, contrasting computers, where all is either true (1) or false (0). It deals with haziness and the conceptions are polysemous in terms of that they cannot be sharply defined. Fuzzy logic brings imprecise human facts over to accurate mathematical models.

While variables in mathematics usually take numerical results, in fuzzy logic, the non-numeric linguistic variables are often used to cultivate the locution of rules and facts. A linguistic variable such as “size” can have a value just like ‘tall’ or its antonym ‘short’. However, the great utility of linguistic variables is that they can be modified through linguistic transformation which can be associated with given functions. The question whether a person is ‘tall’ cannot be unmistakably answered, because it is not possible to clearly state if a person is ‘tall’. An answer may depend on individual cognition and further for the individual itself it may even not be feasible to give a strict answer for the simple reason that belonging to a set (e.g. size) is often not sharp but fuzzy, involving a partial matching expressed in the natural language by the expressions ‘quite’, ‘slightly’, ‘more or less’, etc.

Figure 1 shows a tender varying curved line that passes gently from ‘not-tall’ to ‘tall’. Therefore this line stipulates the transition of the linguistic variable “size”. Both people are ‘tall’ to some degree (as both people are ‘short’ to some degree) but the female is significantly ‘less tall’ than the male. The vertical axis is an index reputed with the membership value between 0 and 1; the curve is noted as membership function.

2.2 Introduction to Fuzzy Set Theory and Classification

Fuzzy sets are an extension of the classical sets and incorporate special membership levels. In classical set theory, the membership of elements in a set is assessed in binary terms according to a two-valued condition; an element either belongs or doesn’t belong to the set. By contrast, fuzzy set theory permits the gradual

assessment of the membership of elements in a set; this is described by dint of a membership function valued in the real unit interval $[0..1]$. Therefore fuzzy sets generalize classical crisp sets, since the indicator functions of classical sets are special cases of the membership functions of fuzzy sets.

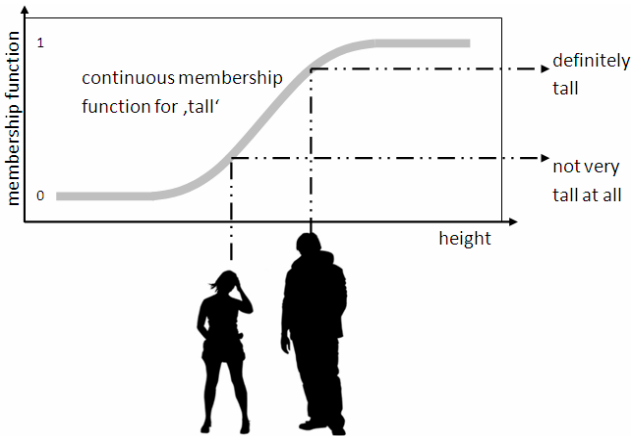


Fig. 1 The fuzzy height set illustrate the continuous membership function for the linguistic variable 'tall'

Fuzzy classification is an upgrading of traditional classification; equally fuzzy sets extend classical sets. The term classification describes the way of clumping elements into clusters, so that elements in the same cluster are as identical as possible, and elements in different clusters are as diverse as possible. In sharp classification each element is associated with just one cluster; as a result the belonging of the elements to clusters are reciprocal and exclusive. On the other hand fuzzy classification allows elements to belong to several clusters at the same time; and again like fuzzy sets, each element has a membership degree which reveals how far it belongs to the various clusters. Thereto fuzzy clustering algorithms allow the modeling of uncertainty associated with vagueness and imprecision and putting this into mathematical equations as described in [Portmann and Meier 2010]. In general fuzzy clustering algorithms a fuzzy cluster is represented by a representative element (typically the cluster centre) and the membership degree of an element to the cluster is decreasing with increasing distance to the cluster centre.

To minimize elements with a small distance to the cluster it should be assigned a high membership level whereas elements with larger distances should have low membership levels. A clustering algorithm begins with a random initialization and updates the membership levels and the prototype in an iterative procedure.

2.3 Fuzzy Expert Systems

Expert systems (introduced by Feigenbaum) are prolific examples within the wide scope of artificial intelligence as in [Russell and Norvig 2003] explained. Expert systems are knowledge-driven systems that can form conclusions based on knowledge on a particular field. The knowledge is represented by 'if-then' rules. By applying consequences on the stated rules, expert systems may deduce optimal decisions.

The major challenge is to commute the knowledge of subject matter experts into 'if-then' rules which are as exact as possible even given that the human representation of the knowledge cannot be well-defined determined. This downside is hurdled by usher fuzzy rules as exemplified in [Grekovs 2002].

Fuzzy rules are a collection of linguistic statements that describe how to make a decision regarding classifying an input or controlling an output:

if $(p_1 \text{ is } \mu_{p_{1_i}}) \wedge (p_2 \text{ is } \mu_{p_{2_j}}) \wedge (\dots)$

then $(c \text{ is } \mu_{c_k})$;

where μ is a membership function

$i, j, k \in \mathbb{N}$

p is a preposition

c is a conclusion

Or more exemplarily:

if $(\text{input}_1 \text{ is membership function}_1)$

and/or $(\text{input}_2 \text{ is membership function}_2)$

and/or (\dots)

then $(\text{output}_n \text{ is output membership function}_n)$

Consider the following rule for instance:

if person is short

and weight is high

then person is overweight

There would have to be membership functions that define what we mean by 'short persons' (input1), 'high weight' (input2) and 'overweight' (output1).

The process of taking an input such as "size" and processing it through a membership function to decide what 'short' means is stated fuzzification. The principle at that is once more to map the inputs from a set to values [0..1] using a set of input membership functions.

Thus fuzzy expert systems are usually involved when processes cannot be described by exact algorithms or when these processes are difficult to model with conventional mathematical models.

3 Challenges and Related Components

This section intends to introduce the faced challenges and their related components. Section 3.1 clarifies fundamental challenges to be outgrown in the Internet of Things applications. Section 3.2 highlights the semantic homes and environments. Section 3.3 discusses necessary sensing and data provision infrastructure.

3.1 Fundamental Challenges

Coming to a new unknown public place like a railway or underground station, airport, hospital, mall, industrial facility, corporate office, university campus all of us can remember trying to quickly find the right location, information desk, directional hint and other service availability information. Environments with quickly changing geophysical parameters like hospitals, care houses, logistic units, military and production line facilities are other problematic control domains.

Static information assistance systems related to mass produced products became a reality by the introduction of bar codes. Using modern data gathering technologies like RFID, NFC and WSN, it is possible and rational to associate various real-world entities with personalized dynamic content from weblogs, social networks, folksonomies, news feeds, location and condition tracking systems and so forth.

Semantically relevant suggestions to recognized intentions can be delivered to the user in many ways, starting from simple visualization hints, to enhanced human-machine interface influences directly on human action as in [Stapelkamp 2007; Sears and Jacko 2007] illustrated. Embedded devices for implementation of such systems will vary from the everyday smart phone to newly developed augmented reality systems.

3.2 Semantic Spaces: Home and Environment

According to scientific studies, the average urban human spends about 80 to 90% of his time indoors. Buildings, houses, public places, industrial and military facilities, and even private and/or public transportation allow mankind to penetrate into all of our planet's places despite a variety of external conditions. The main indoor environment metagoals of security, safety, comfort and energy-efficiency have been implied since the very beginning of civilization. However, the approaches to pursue them are mainly restricted by the technological level available. The Internet technologies are increasingly and widely distributed today, giving us an opportunity to consider, evaluate, process and optimize the semantics behind previously developed indoor service data and information. The semantic home and environment is the concept of "thinking ambient intelligence" that is aware of its

inhabitants (i.e. humans, animals, robots and smart objects). The extension of previously developed services to include semantically defined raw data and processed information lets emerging semantic-aware systems form (better) deductions about occurring events that affect the inhabitants. The use of multi-modal analysis of real-world sensor data and information from local and remote sites of interest is one important but challenging requirement to deducing the meaning behind these events. The deductive process is not trivial. Human actions, behaviors, habits, thoughts, intentions, emotions and health conditions are the key factors that have to be considered while tackling semantic home and environment systems, since they define event context. Context awareness is a fundamental component for informational assistance and intelligent environment behavior systems, on one hand requiring constant intelligence rule updates as in [Driver et al. 2007] illustrated. On the other hand the replayed scenarios are usually accompanied by a certain precondition sets such as: time, location, nearby resources, nearby inhabitants, action sequence, general sensory conditions such as weather, network status, service status, and others; that help in the deduction of context.

Looking from the conceptual point of view, true informational assistance is hardly imaginable without bijectional correspondence between real world and the informational model of reality. Hence, bidirectional communication methods between chip-enabled real physical objects and their informational copy are another necessary component of the ambient intelligence needed here. Following the requirements, the concepts of the Internet of Things have recently made an impressive step towards implementing the one-to-one world-to-model paradigm.

3.3 Sensing Infrastructure and Data Provision

The sensing infrastructure plays an important role in the provision of input data and processed information. Sensing infrastructures have variable complexity depending on the building or environment that they are installed in. For example, within a large building the data travels from disparate sensor platforms within the individual or linked buildings of different domains and authorities that were installed at different times by different vendors, with several grades of access rights and network isolation challenges to be considered along the way, via parallel sub-networks within the same building or environment. Sensing devices across the different networks vary in design sophistication in terms of connectivity, memory, mobility, energy budget and processing power; the extraction of data symbols from devices with various capabilities often incurs practical difficulties. In terms of sophistication, there are many simple battery-powered wireless devices that can be placed around a location to broadcast the current environmental condition at a periodic rate without local storage, but only few devices are designed with costly but powerful mechanisms for processing data locally, adapting their filters for noisy raw data, or supporting some query processing system for requesting data from a declarative database that is on the sensor node. Examples of sensor data spans

across multiple domains: classical physics measurements and area measurements for ambient light, presence, temperature, infrared, absolute and relative humidity, strain; building automation data, current and voltage measurement, state of a switch or dimmer for light or blinds, smoke and fire detection system status, alarm and perimeter security status, audio-video feeds, thermostat value and setting, current weather conditions including barometric pressure and wind speed, location in localization system; industrial applications in machine health and resource plant monitoring; network availability and load of server and infrastructure systems, network security status and important changes; active and passive radio-frequency identification nodes and tag data; and many other specialized sensors and also actuators built for specific tasks.

For less sophisticated sensor devices, the data is not stored locally, but is instead transmitted across the network to a central collection point, whereas more sophisticated devices can store a considerable number of data points locally. The tradeoff of sophistication balances the cost of low-processing easy-to-move battery-powered devices capable to run for several sequential years against those of more sophisticated processing nodes that can be directly queried. Less sophisticated sensors can also use simpler, more optimized and occasionally less standardized protocols by comparison to more widespread and complex protocols such as IPv6 that allow higher interactivity in more sophisticated nodes; herein there is a challenge mitigated by an intermediary that translates the traffic to a more sophisticated protocol or stores it directly at a collection point of the particular sensing platform. These intermediate processing layers may make fuller use of individual sensor functionality than an abstraction layer does when trying to gather data through various application-level and lower protocols. There are two challenges here: to provide robustness for lost and noisy raw data and to expose the data in a way useful to information consumers. Services can expose the data in a sensing infrastructure such that a client may quickly and directly make a query with respect to a variety of contextual conditions (with heavy weight placed on location) provided by the client on what appears as one contiguous data unit, without the challenge associated with slowly querying and aggregating data from disparate sensors across the location.

Data collected into one or more query-able stores or a distributed database must be represented as data views for both public guests and elevated-privilege security-maintenance end-user clients in a standardized way that the client expects, such as using XML and other Web 2.0 and Semantic Web technologies that can expose this data using widespread standards. Here, a layer of middleware is useful in translating the data to a common standard, e.g. SPARQL, understood by a querying client newly connected to the building's middleware. Beyond basic services, additional middleware components can make available more complex services such as services that include heavily pre-processed data that includes data beyond current conditions but also forecasts trends based on data mining historical data against currently developing conditions. Between basic and complex services, a variety of data is expected to be available to the client.

Sensor data and related services need not be restricted to a local area, even though post-processed values are cached or stored locally; the data from these distributed databases can be exposed onto the Internet of Things to be made available to a sensing infrastructure of a wider area for remote data queries of clients not yet arrived at the physical destination. This model is similar to that of the information provided by the growing number of Semantic Web services for weather forecasts, traffic forecasts, train schedules, product ratings, flight costs, outdoor and indoor directions and maps are available today and the shop sales, and other information of tomorrow; and data extracted from known and well-structured older Web 2.0 documents also into a formal knowledge concept representation. An important consideration for a resource-constrained mobile device is to consider its bandwidth when providing it a remote view of the data.

When the data is queried, an important but overlooked challenge in providing a result is to define a meaningful data structure that acts as a transport container: a structure that can define the semantic of the data being transported. Here it is obviously preferable that semantic-tagged data uses an exportable data structure model format that is componentized, standardized and commonly accepted in order to avoid redundancy in re-interpreting, or translating between other data models representing data with the same underlying semantics. Thus, whether it is simple raw data in a known standard format, e.g. SensorML, or information processed or extracted from data by some high-level transformation, any service that shares this data for export also exposes a semantic model with commonly-understood components when the response to a query is intended to be exchangeable.

4 System Concept Description

This section characterizes the system concept. Section 4.1 reveals an innovative information retrieval approach for future searches. Section 4.2 illustrates building blocks of the Prometheus framework. Based on ontologies, section 4.3 shows in brief, how to train the intelligent environment itself the semantic of human-used terms. Section 4.4 presents retrieved semantic data. And finally, section 4.5 discusses human-environment and human-building interaction.

4.1 Towards an Innovative Information Retrieval Method

The data is no longer produced by humans alone but more and more by sensors as well. As in [O'Reilly and Battelle 2009] conceptualized, today's cameras and phones are mutated into artificial ears and eyes like a sort of "sixth sense" applications. Sensors for motion and location provide continual detail where someone is, what one is looking at and how fast and in which direction one is moving. Data

can be gathered and offered on a real time basis. In order to arrange these loose data, they need to be collected from adequate and trustworthy sources.

For this purpose information retrieval (IR), which establishes the retrieval of information from an object such as a document as outlined in [Baeza-Yates and Ribeiro-Neto 2010], comes in. In view of this, information retrieval represents the entire searching science for documents, for information within documents and for metadata about documents, as well as that of searching databases, the WWW, and other sources. At first software agents of a certain type (WWW, HTTP protocol) are instructed to collect documents in preparation for classification; in the simple example of a Web-page with natural language and images a web crawler begins at a trusted starting point and crawls along hyperlinks for each hyperlink determining the credibility and link value by use of a page ranking algorithm like HITS; and this data is reused to order search query results later. By data clustering documents can be automatically grouped into classes. Fuzzy representations are useful to handle the imprecision resulting from the automatic interpretation of extracted content meanings. The resulting interpretation is storable according to a defined domain ontology. The semantics extracted from HTML content are imprecise since they are roughly guessed when extracted and not well-defined; and the same for other text-based and multimedia objects. Semi-structured XML data is better machine readable since it is syntactically defined, but tag definitions provided by user groups are application specific and mostly ambiguous. Not so with the last example: documents tagged in representation languages designed for the Semantic Web, such as emerging RDF/S or DAML+ OIL, include formal semantics, defined rules, and an ontology vocabulary (for various types of ontologies, including service ontology) that create a well-structured machine-readable document; a significant improvement in classification. The second significant part of information retrieval is in the decision making expert systems that first contextually search for queried terms and second provide a ranked ordered list of suggested related terms of nearby clusters and other conditionals.

Generally, information retrieval systems are used to diminish what is called information overload. On the basis of a fully automated ontology with the aid of different sensors for context the IR-located information becomes arranged in a helpful manner for the user in his preferences and circumstances (e.g. location, connection speed, etc.) expanding the query beyond only the limited set of terms that describes the user need.

Paraphrased in [Marsland 2009], AI is the intelligence of machines that perceives its environment and takes actions which maximize its chances of success. A major focus there is to train the system to recognize patterns and make intelligent decisions based on this data without human intervention. However, the term of AI goes beyond the human intelligence limits and contains computational, memory and solving abilities and properties that humans do not possess.

4.2 The Prometheus Framework

Prometheus is a software/hardware information retrieval data processing system for the provision of the most relevant context aware information. The system uses fuzzy logic to construct the term ontology based on sensor and network distributed data. The core of Prometheus is a distributed cognitive and decision making software framework meant to be flexibly usable by humans and other software/hardware services and systems.

Functional components of Prometheus include:

- Data input subsystem gathering sensor data and moving it from sensors to between middleware and databases
- Cognitive ability subsystem implemented using several approaches from cognitive sciences (i.e. symbolic, static, behavioral, emotional) using multiple sensors data analysis capable of hierarchical activities recognition for living with predictive profiles
- Context-aware decision-making subsystem based on ontology representation and fuzzy expert systems
- Interaction subsystem with multiple machine-to-machine and human-to-machine interfaces yet unified with a backend
- Adaptable to environment communication framework allowing automatic transparent data commutation between (upcoming) WBAN, WPAN, WLAN, and Internet

Several significant impacts of Prometheus include but are not limited to:

- Increased ability for context prediction and accountability
- Diminished information overload by the context-relevant data filtering
- Workflow optimization in team and innovative project environments
- Increased energy efficiency, comfort and security of human spaces

4.3 Ontologies and Information Retrieval for Semantic Homes and Environments

Looking at implementation, the information retrieval service is primary based on fuzzy expert systems with help of weak term ontology. Ontology is, following Gruber, a "formal, explicit specification of a shared conceptualization", that is a formal notation of a concept set within a domain and the relationships between those concepts as in [Gruber 1993] exemplified. An ontology is often needed to reason about the properties of a domain and can be used to define a field.

In [Portmann and Meier 2010] it is shown how folksonomies' subjacent tags can be harvested. To this Portmann and Meier apply particular metrics such as the

Jaccard coefficient in order to meter the proximity between tags. The ontology can then be compiled without human intervention on basis of fuzzy clustering algorithms. On the basis of this ontology a meaning, or semantic, can be deduced.

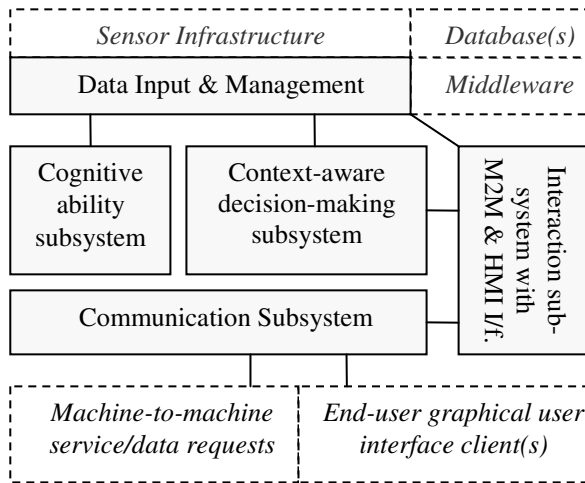


Fig. 2 Components of Prometheus framework

The Prometheus object and service search concept is based on a previously fuzzy-built and constantly automatically updated ontology, enriched with inferences from fuzzy expert systems, in order to arrive to satisfactory solutions. Fuzzy expert systems are fitting instruments for this kind of reasoning.

The interconnection between terms in the ontology allows the intelligent house to associate the user behavior and intentions with physical, economical and social parameters of the environment.

Depending on past user behavior and current context-relevant information, the Prometheus framework adaptively learns from each individual user. Behavior analysis requires storage of goals, actions, conditions and results made previously by the user and achieved as a user profile. Synthesized from this semantic description of the user prediction-based behavior assistance can be provided on the basis of fuzzy expert systems by the recognition of personal patterns from frequently repeated operations at newly visited unknown environments.

The goal of such assistance is to keep the confidence and comfort at an expected level for people dealing with unknown environments. However, every new environment will most likely not contain an exact copy of previously used objects, services and processes. In this case the most logical solution is to find objects with similar functionality and to inform the user about differences and optimal ways to access them. Moreover, informational assistance can be given directly at the “thinking event” moment, not later when the focus has already been switched to another topic.

4.4 Retrieved Semantic Data

Enhancing the system with external data can not only improve usability but also bring an added value in a form of new location-based services as refer to in [Tsetsos et al. 2006]. Adapted from an ontology to help an end-user to get along in an unknown environment, the Prometheus framework draws on, in different dimensions available, input data such as Internet data (e.g. train connections, ratings, directions, indoor plots, taxonomies) and diverse surrounding sensor data (e.g. precise position, weather conditions, traffic jams, local accidents, door or elevator malfunctions, power losses, water and heating shut offs) to present intelligent suggestions (e.g. fastest directions to a certain destination, the next train station, connections, prices, ratings, pictures, descriptions of a specific product).

Additionally, for the moment “not relevant” data like changing weather conditions while shopping is frequently useful right at the next moment because it can influence the user’s (buying) preferences and decisions.

Data to be shown has to be chosen according to the user role, current semantic context and current goal set. Moreover, the group of people has to be also considered while designing the communication dataset. Requirements like people group-oriented data availability, people collaboration encouragement, environment adaptability, accountability, security and privacy shape the dataset.

4.5 Environment and Building Interaction

To reveal relevant data, the Prometheus framework decides under given circumstances how to interact optimally. Using human-computer interaction the building can adapt by changing visual, audio, thermal, humidity, pressure or other physical environment parameters depending on the mood of a tenant.

The search query itself is usually considered as an interface between the user and a search engine. Prometheus is extending the understanding of search queries towards user goals, intentions and behavior. The system does not simply provide the string search box with historically based associative suggestions, but also keeps in mind the semantic context of the user or client system.

The way the semantic environment interacts with a human strongly overlaps with studies from the human-computer-interaction field, as for example the basic interface simplicity requirement. Also, display design principles remain the same.

However, different semantic system parameters define specific functional constraints leading to the terms of human-building-interaction and human-environment-interaction. The relevant data will be presented in a for the user appropriate way adapted for the subjacent hardware.

This work associates buildings with being indoors while environment with being outdoors. The main difference between these two types of human interaction is characterized by different impact factors of the common life-value parameters.

This parameter list is not hard-fixed for every person. It is mainly provided for a demonstrative purpose. Table 1 contains the parameter impact factors belonging to the range [1...5]. The highest value corresponds to the highest impact factor.

Table 1 Simplified Example of Parameters Impact

Parameter	Impact factor on an action	
	Indoors	Outdoors
<i>Weather conditions</i>	2	5
<i>Customization</i>	4	1
<i>Location</i>	3	4
<i>Security</i>	4	4
<i>Comfort</i>	5	2
<i>Safety</i>	4	4
<i>Time</i>	2	5
<i>Cost</i>	4	3

We observe that weather conditions and time of day have much stronger influence on human actions outdoor than indoor. The human action impact of customization and comfort is in opposite relatively little outdoor and significantly high inside of the buildings. Security, safety, location and cost are equally important for human interaction independent of the environment.

5 Practical Applications

The previous more theoretically oriented sections showed stimuli for this section. As an example, a personal digital assistant concept for embedded devices like smart phones and house appliances is presented.

For improved support a future system, like a smart phone, should help people in the same way that present personal digital assistants do. Based on the proposed ontology the digital assistant could educe feasible suggestions for the user. To come up with such suggestions, a future smart phone's built-in personal digital assistant, here referenced as a part of Prometheus system, has to learn first from the elapsed user's environmental data. Then it would be possible to pick (for the user) the best solution based on the fuzzy expert systems method.

A major challenge hereby is to teach the digital assistant appropriate solutions. Therefore the user has sometimes to interact with the personal digital assistant. As known from supervised learning, the user should correct potential personal digital assistants misbehavior. Learning inputs can also come from environmental sensor data in which no user interaction is required.

The simplest example is based on the current (indoor or outdoor) location data and needs of the user. For example, the optimal solution for a query on “shoes” will include costs, time, warranty conditions, the (indoor or outdoor) way, and connection for public transportation. Figure 3 shows the particular possible use case while utilizing the Prometheus framework components.

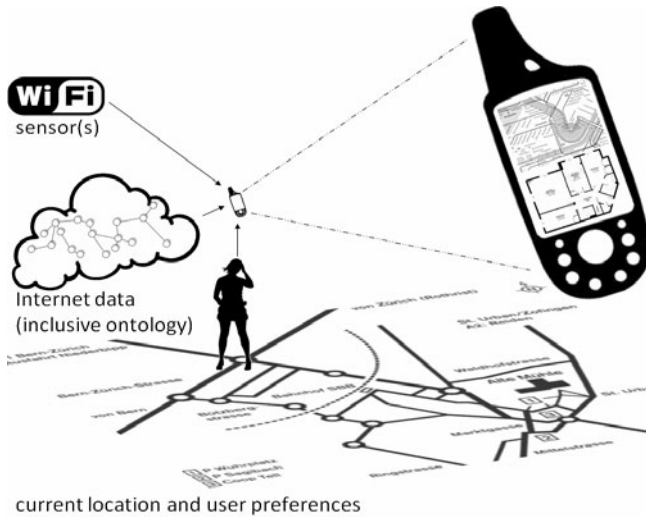


Fig. 3 The Internet- sensor- and user-preference-data are expected come up with bright solutions for the user

6 Conclusions and Outlook

After a short introduction into the fuzzy set theory and its applications, this paper presented the main foreseen challenges to be faced by shaping semantics behind the existing living infrastructure. The presented approach of the Prometheus framework depicts a vision towards a solution of these challenges with the novel approach based on fuzzy set theory in the rapidly developing area of the Internet of Things. Furthermore towards the dataset considerations for semantics implementation particular attention was given to the new terms of human-building and human-environment interaction.

Further subjacent studies on the issue of appropriate outcome-rules for the fuzzy expert system will be needed. It might be that the proposed approach could be improved by taking into account other machine-learning strategies. At the moment Prometheus is thought to be based on fuzzy rules, but in the future it is not limited to only this.

Another point is a possible enhancement of the fuzzy-built ontology towards integrating further sensor data as for example intelligent clothes. Wearable computing is a vigorous research topic, containing user interface design-, augmented

reality-, pattern recognition-, use of wearable's for specific applications or disabilities, electronic textiles and fashion design studies.

The CEESAR-iHomeLab is working on one hand on forming a scientific community for cutting-edge international research projects in the area of ambient intelligence, human-building interaction, user behavior analysis, and assisted living and on the other hand; the Research Center FMSquare implements the ideas of fuzzy methods to various scope of applications, and for this reason both research centers appreciate cooperation with researchers and practitioners.

Acknowledgment

We sincerely thank our colleagues from CEESAR and its iHomeLab (<http://www.iHomeLab.ch/>) project at the Lucerne University of Applied Sciences, who always helped with word and deed. Furthermore, a special thank goes to the Fuzzy Marketing Methods Research Center (www.FMSquare.org) of the University of Fribourg for contributing valuable thoughts.

References

- [Azuma et al. 2001] Azuma, R., Baillot, Y., Behringer, R., Feiner, S., MacIntyre, B.: Recent advances in augmented reality. *IEEE Computer Graphics and Applications* 21(6), 34–47 (2001)
- [Baeza-Yates and Ribeiro-Neto 2010] Baeza-Yates, R., Ribeiro-Neto, B.: *Modern information retrieval*. ACM Press, Essex (1999)
- [Bishop 2008] Bishop, C.M.: *Pattern recognition and machine learning*. Springer, Berlin (2008)
- [Driver et al. 2007] Driver, C., Linehan, E., Spence, M., Tsang, S.L., Chan, L., Clarke, S.: Facilitating dynamic schedules for healthcare professionals. In: *Pervasive Health Conference and Workshop*, pp. 1–10 (2007)
- [Grekovs 2002] Grekovs, R.: *Methods of fuzzy pattern recognition*. Proceedings of Riga Technical University (2002)
- [Gruber 1993] Gruber, T.R.: A translation approach to portable ontology specifications. Knowledge Systems Laboratory, Technical Report KSL 92-71, pp. 199–220 (1993)
- [IoT: Summary] International Telecommunication Union, *The internet of things*. Executive Summary (2005)
- [Kasabov 1996] Kasabov, N.K.: *Foundations of neural networks, fuzzy systems, and knowledge engineering*. MIT Press, Cambridge (1996)
- [Marsland 2009] Marsland, S.: *Machine learning. An algorithmic perspective*. CRC Press, Boca Raton (2009)
- [O'Reilly and Battelle 2009] O'Reilly, T., Battelle, J.: *Web squared: Web 2.0 five years on*. Web 2.0 summit, pp. 1–13. O'Reilly Media Inc, New York (2009)
- [Portmann and Meier 2010] Portmann, E., Meier, A.: A fuzzy grassroots ontology for improving weblog extraction. *J. of Digital Information Management*, 276–284 (2010)
- [Russell and Norvig 2003] Russell, S.J., Norvig, P.: *Artificial intelligence: A modern approach*, 2nd edn. Prentice-Hall, Englewood Cliffs (2003)

- [Stapelkamp 2007] Stapelkamp, T.: Screen- und interfacdesign. Springer Science, Berlin (2007)
- [Sears and Jacko 2007] Sears, A., Jacko, J.A.: Handbook for human computer interaction. CRC Press, Boca Raton (2007)
- [Tsetsos et al. 2006] Tsetsos, V., Anagnostopoulos, C., Kikiras, P., Hadjiefthymiades, S.: Semantically enriched navigation for indoor environments. *Int J. of Web/Grid Services* 2, 453–478 (2006)
- [Zadeh 1965] Zadeh, L.A.: Fuzzy sets. *Information and control* 8, 338–353 (1965)

Evaluating Overheads Introduced by OPC UA Specifications

S. Cavalieri

Department of Electric Electronic and Computer Engineering, University of Catania, Italy
salvatore.cavalieri@diit.unict.it

Abstract. The widespread use of the standard, worldwide and vendor-independent OPC UA specifications in industrial environment introduces many benefits as they allow to keep open the market of the industrial applications. On the other hand, OPC UA adopts a very complex software infrastructure to realise the communication between industrial applications; this complexity may impact on the overall performance of their data exchanges. The aim of this paper is to deal with the performance evaluation of OPC UA. The main features which may influence performance in the client/server exchange of information, will be pointed out; then, the main results of the evaluation of the overhead introduced by these mechanisms onto the OPC UA overall performance will be presented and discussed.

1 Introduction

OPC Unified Architecture (OPC UA) is the current OPC Foundation's technology for secure, reliable and interoperable transport of raw data and pre-processed information from the shop floor into production planning systems [WWW-1].

Definition of OPC specifications started ten years ago to simplify and to standardise data exchange between software applications in industrial environment. The rapid success of OPC specifications was due to the rapid increase in the use of Windows PCs and the choice of Microsoft's DCOM as the technological basis for the OPC specifications, as DCOM was found on every Windows PC. Exactly this point, however, at the same time raised the majority of criticism regarding OPC; OPC technology was too focused on Microsoft, platform-dependent and not firewall-capable, and thus not suitable for use in cross-domain scenarios and for the Internet. When XML and Web Services technologies have been available, the OPC Foundation adopted them as an opportunity to eliminate the shortcomings of DCOM. Since 2003 the OPC XML Data Access (DA) specification has offered a first service-oriented architectural approach besides the "classic" DCOM-based

OPC technology; this Web services-based concept enabled applications to communicate independently of the manufacturer and platform.

Today, the OPC Foundation has introduced the OPC UA standard which is based on a service-oriented and platform-independent approach, creating new and easy possibilities of communicating with Linux/Unix systems or embedded controls on other platforms and for implementing OPC connections over the Internet. The new possibilities of using OPC components on non-Windows platforms, embedding them in devices or implementing a standardised OPC communication across firewall boundaries allow speaking of a change of paradigms in OPC technology. OPC UA servers can be varied and scaled in their scope of functions, size, performance and the platforms they support. For embedded systems with limited memory capacities, slim OPC UA servers with a small set of UA services can be implemented; at the company level, in contrast, where memory resources are not that important, very powerful OPC UA servers can be used with the full functionality. A key feature of the OPC UA specification is the definition of the UA security model, which wasn't available in the previous versions of OPC specifications; the OPC UA Security governs the authentication of clients and servers and ensures data integrity, trustworthiness and authorisation within OPC communication relationships.

Due to the current features of the OPC UA specification, SCADA, PLC/PC-based controls and MES systems are unthinkable today without an OPC UA interface. Nowadays OPC UA plays a very dominant role in industrial applications as the most part of the data exchanges between client/server industrial applications are currently based on these specifications. This introduces benefits as allows to keep open the market of the industrial applications, due to the presence of standard, worldwide and vendor-independent specifications. On the other hand, the OPC UA specifications introduce a very complex software infrastructure between industrial applications, which may impact the overall performance of their communications.

Current literature presents few papers dealing with performance evaluation of OPC UA; most of them focus only on particular services and/or aspects of the OPC UA specification. For example in [Braune et al. 2008; Post et al. 2009] performance evaluation is carried on considering only the security mechanisms and services provided by the OPC UA specifications.

The aim of this paper is to deal with the performance evaluation of OPC UA, pointing out all the main features which could influence performance in the client/server exchange of information. On the basis of the considerations about performance of OPC UA specifications, outlined in the paper as said before, the most meaningful results of performance evaluation will be presented and discussed.

2 OPC UA Overview

The OPC UA specifications are made up by 13 parts [Mahnke et al. 2009; OPC Foundation 2009]. The OPC UA architecture models OPC UA Client and Server as interacting partners; each Client may interact concurrently with one or more Servers, and each Server may interact concurrently with one or more Clients. An application may combine Server and Client components to allow interaction with other Servers and Clients.

Client and Server applications use OPC UA Client and Server Application Programming Interface (API) to exchange data, respectively. OPC UA Client/Server API is an internal interface that isolates the Client/Server application code from an OPC UA Communication Stack. The OPC UA Communication Stack converts OPC UA Client/Server API calls into Messages and sends them through the underlying communications entity; on the other hand, each Message received from the underlying communications entity is delivered to the Client/Server application by the OPC UA Communication Stack.

Implementation of the communication Stack is not linked to a specific technology; this allows OPC UA to be mapped to future technologies as necessary, without negating the basic design. Two data encodings are currently defined: XML/text and UA Binary. In addition, two transport mappings are available: UA TCP and SOAP Web services over HTTP. Clients and Servers that support multiple transports and encodings will allow the end users to make decisions about tradeoffs between performance and XML Web service compatibility at the time of deployment, rather than having these tradeoffs determined by the OPC vendor at the time of product definition.

Figure 1 shows the OPC UA Client architecture; a Client Application uses the OPC UA Client API to send OPC UA Service requests to OPC UA Server, and to receive OPC UA Service responses and Notifications from the OPC UA Server (see subsection 2.1 for a definition of Notifications). The OPC UA Communication Stack converts OPC UA Client API calls into Messages and sends them through the underlying communications entity to the Server; the OPC UA Communication Stack also receives Response and Notification Messages (Notifications will be defined in subsection 2.1, as said before) from the underlying communications entity and delivers them to the Client application through the OPC UA Client API.

Figure 2 shows the OPC UA Server architecture. The Server Application is the code that implements the function of the Server. Real objects are physical or software objects that are accessible by the OPC UA Server or that it maintains internally; examples include physical devices and diagnostics counters. Particular objects, called Nodes, are used by the OPC UA Server to represent real objects, their definitions and their References; the set of Nodes is called AddressSpace. Nodes are accessible by Clients using OPC UA Services (interfaces and methods). Figure 2 shows the Nodes in the AddressSpace, the references between them (drawn by arcs connecting the Nodes) and their relationships with the real objects.

Particular objects, called Monitored Items, can be created inside an OPC UA Server, as shown in Figure 2; they are entities created by the OPC UA Client that monitor AddressSpace Nodes and their real-world counterparts, as described in subsection 2.1. Monitor Items are created inside Subscriptions, shown in Figure 2, which are the contexts of the data exchange between server and client, as described in subsection 2.1.

To promote interoperability of Clients and Servers, the OPC UA AddressSpace is structured hierarchically with the top levels the same for all Servers. OPC UA Servers may subset the AddressSpace into Views to simplify Client access.

Like the OPC UA Client, the OPC UA Server uses the OPC UA Server API to send Response and Notification Messages to OPC UA Clients. The OPC UA Communication Stack receives Request and Publish Messages (see subsection 2.1 for a definition of Publish Message) from the OPC UA Client and delivers them to the OPC UA Server Application through the OPC UA Server API.

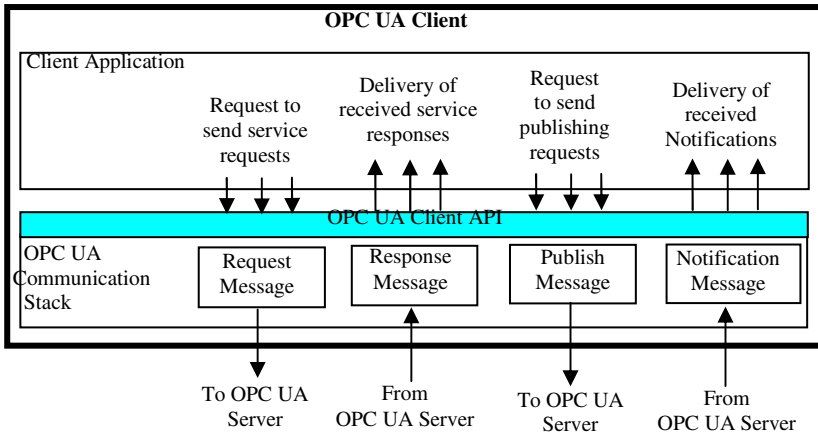


Fig. 1 OPC UA Client

2.1 Client/Server Communication

Communication between client and server is realised through a three-level communication stack: Monitored Item, Subscription and Session. All these three levels are put on the top of a Secure Channel level, described in the subsection 2.2.

The Session is a logical connection between an OPC UA Client and an OPC UA Server created on the top of and in the context of a Secure Channel. The lifetime of a Session is independent of the Secure Channel and another Secure Channel can be assigned to the Session. Servers may limit the number of concurrent Sessions based on resource availability, licensing restrictions or other constraints. Each Session is independent of the underlying communications protocols; failures of these protocols do not automatically cause the Session to terminate. Sessions terminate based on Client or Server request, or based on inactivity of the Client.

A Subscription is the context to exchange values on data changes, aggregates of data and events between server and client. A Subscription requires a Session to transport the data to the client. Subscription lifetime is independent of the Session lifetime and a Subscription has a timeout that gets reset every time data or keep-alive messages get sent to the client.

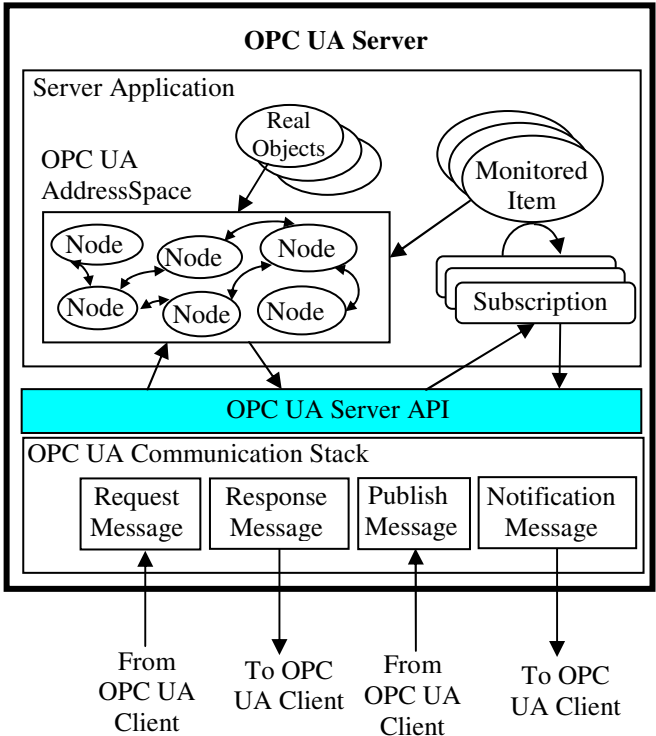


Fig. 2 OPC UA Server

Monitored Items can be created in a Subscription; they are entities in the OPC UA Server created by the OPC UA Client that monitor AddressSpace Nodes and their real-world counterparts. Three types of Monitored Items can be created. The first is used to subscribe for data changes of Variable Values; the second type of Monitored Item is used to subscribe for Events by defining an EventNotifier and a filter for the Event to be monitored. The third type of Monitored Item is used to subscribe for aggregated Values calculated based on current Variable Values in client-defined time intervals. Figure 2 shows Monitored Items and Subscriptions; as can be seen each Monitored Item is related to Nodes into the AddressSpace and is bounded to a specific Subscription.

All Monitored Items have common settings, among which there are the sampling interval and the queue size. The sampling interval defines the rate at which the server checks Variable Values for changes or defines the time the aggregate get calculated. The sample rate defined for a Monitored Item may be faster than the publishing rate of the relevant Subscription; for this reason, the Monitored Item may be configured to queue a certain number of data produced by the Monitored Items.

The simplest way for a Client and Server to exchange data is using the Read and Write services, which allow an OPC UA Client to read and write one or more attributes of Nodes, maintained by the AddressSpace of the OPC UA Server; like most other services, the Read and Write services are optimised for bulk read/write operations and not for reading/writing single values.

A different and more sophisticated way to access data is based on the Monitored Items and Subscriptions; this is the preferred method for clients needing cyclic updates of variable values. In order to explain this kind of data exchange, please refer to Figure 3, which shows a subscription containing, just for example, the three kinds of Monitored Item described before. The figure points out the Publish Interval, which is one of the Subscription settings; the Publish Interval defines when the server clears the Monitored Item queues and conveys their contents into a Notification to be sent to the Client. Notifications will be really sent to the Client by issuing the Publish service, as explained in the following.

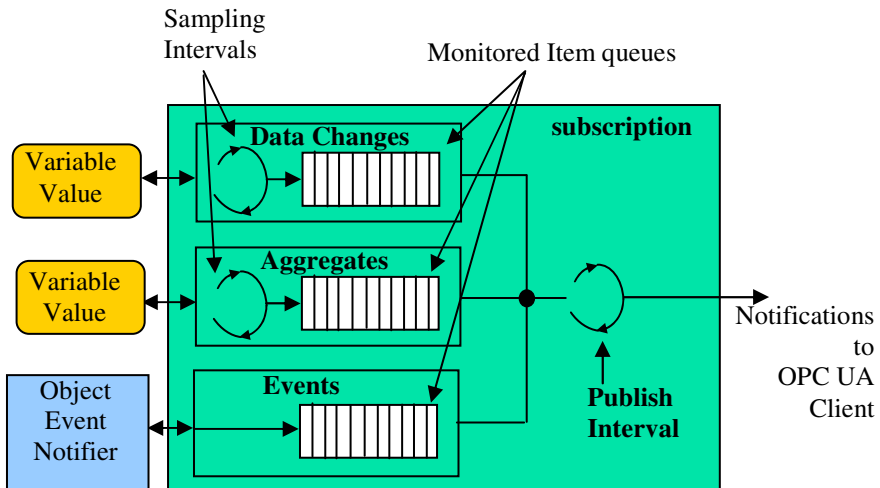


Fig. 3 Subscription

Transmission of Notifications by OPC UA Server is triggered by Publish Requests sent by client. According to OPC UA specifications, a client must send a list of Publish Requests without expecting an immediate response; the server

queues the Publish Requests until a Notification is ready for sending to the client (according to the Publish Interval, as said before). When this occurs, the Notification is sent back to the client through a Publish Request response. The Publish Request is not bound to a specific Subscription and can be used by the server for all the Subscriptions running in the same Session context. To make sure that all Subscriptions can send a notification message at the same time, the client should make sure that there are more outstanding Publish Requests than active Subscriptions.

Figure 4 depicts the exchange of Publish Request and Response between OPC UA Client and Server; as can be seen, data exchange is realised within a Session, previously opened by client and server. The Session may contain several Subscriptions, for each of which Notifications produced on the basis of the Publish Interval are waiting to be transmitted. For each Publish Request sent by the OPC UA Client, exactly one Notification is transmitted; it may belong to one of the current Subscriptions inside the Session.

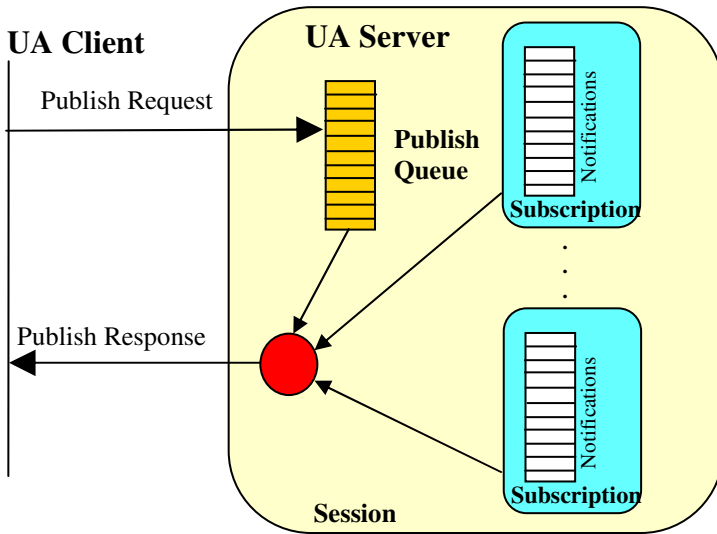


Fig. 4 Transmission of Notifications by Publish Request

2.2 Security Model

OPC UA provides a security model, which includes security mechanisms allowing the authentication of Clients and Servers, the authentication of users, the integrity and confidentiality of their communications, and the verifiability of claims of functionality. Several parameters may be selected and may be set to personalise the security mechanisms to meet the security needs of a given installation. Furthermore, a minimum set of security Profiles that all OPC UA Servers support (even though they may not be used in all installations) has been defined.

Security relies on a Secure Communication channel that is active for the duration of the application Session and ensures the integrity of all Messages that are exchanged.

When a Session is established, the Client and Server applications negotiate a secure communications channel and exchange software Certificates that identify the Client and Server and the capabilities that they provide. Authority-generated software Certificates indicate the OPC UA Profiles that the applications implement and the OPC UA certification level reached for each Profile. Certificates issued by other organisations may also be exchanged during Session establishment.

The Server further authenticates the user and authorises subsequent requests to access Objects in the Server. Authorisation mechanisms, such as access control lists, are not specified by the OPC UA specification; they are application or system-specific.

OPC UA security allows to encrypt and sign Messages; encryption and signatures protect against disclosure of information and protect the integrity of Messages, respectively. OPC UA uses symmetric and asymmetric encryption to protect confidentiality as a security objective; asymmetric encryption is used for key agreement and symmetric encryption for securing all other messages sent between OPC UA applications. OPC UA uses symmetric and asymmetric signatures to address integrity as a security objective. The asymmetric signatures are used in the key agreement phase during the Secure Channel establishment; the symmetric signatures are applied to all other messages.

It's very important to point out that OPC UA specification doesn't make mandatory the use of certificates, digital signatures and data encryption. It's care of the final user of the OPC UA to evaluate when the choice of one or more of the previous security mechanisms is more appropriate; choice must be taken on the basis of the best trade-off between security requirements and the overall performance of the system, which may be influenced by certain security mechanisms as pointed out in this paper.

3 Performance

One of the main requirements for OPC UA is performance; OPC UA must scale from small embedded systems up to enterprise systems with different requirements regarding speed and type of transferred data. In embedded systems, where smaller pieces of data must be transferred in short time intervals, the speed of the data transfer and minimal system load is the most important requirement. In enterprise systems, where structured data must be processed in a transaction- and event-based manner, the efficient handling of structured data is more important than the absolute speed of data transfer [Mahnke et al. 2009].

OPC UA features a very complex architectures made up by a very huge number of mechanisms, each of which can be enabled/disabled and can be personalised by

setting particular parameters. Choice of activation/deactivation of certain mechanisms and choice of the values of the relevant parameters is under the responsibility of the final OPC UA user during the system configuration; each choice must be taken aiming to reach a compromise between the requirements of the industrial application and the overall performance of the system.

For this reason, assessment of performance of OPC UA specifications seems very important in order to verify if and when the requirements of industrial applications (including performance) are met by OPC UA architecture; the impact of each mechanism of the OPC UA specifications on the overall performance of the system should be analysed very carefully. Results of this analysis may help the final user to evaluate when the choice of one or more of the foreseen mechanisms and the choice of the value for each of the relevant foreseen parameters is more appropriate to fulfil all the requirements of the industrial application.

Due to the huge number of mechanisms in OPC UA specifications, it's clear that performance evaluation should be preceded by an analysis aimed to point out the mechanisms of the OPC UA specifications which, more than others, could influence the behaviour of industrial applications using OPC UA to exchange information.

For this reason, the aim of the following subsections is to point out the main features of the OPC UA specification candidate to influence the relevant performance.

3.1 Security

The main question about performance evaluation of security aspects of OPC UA specifications is whether the OPC UA security model is efficient in data transfer.

OPC UA is used at different levels of the automation pyramid for different applications within the same environment. At the plant floor level, an OPC UA server may run in a controller providing data from field devices to OPC UA clients (e.g. HMIs, SCADA). On top of the plant floor at operation level, an OPC UA application may be a client collecting data from the server at the lower level, performing special calculations and generating alarms; an example is represented by an OPC UA client integrated in an ERP system, obtaining information about used devices in the plant floor (e.g. working hours) and creating a maintenance request.

For each application involving OPC UA, the trade-off between security and performance must be reached; at the very top level, security might be more important than performance since the corporate network is connected to the Internet. At the very bottom level, performance could be more important than security when data has to be acquired in very fast and efficient way in order to control a production process.

Performance evaluation seems to play a very strategic role in order to reach the above-mentioned trade-off between performance and security. In particular, at

least two different aspects of the security OPC UA model needs to be investigated during performance evaluation.

The first is related to the use of the certificates and their verification operated by local or remote Certification Authorities (CAs) while opening a secure channel. In e-commerce environment a waiting time of 5-10 seconds until the Web server hosting a Web shop has validated the certificate of the customer, is very common and doesn't represent a very long time to purchase confirmation. However, 5-10 seconds can be a very long time interval for industrial applications, especially for devices located at the field level of the automation pyramid (e.g. applications in chemical or pharmaceutical industries, where very little waiting time could lead to serious problem). It's clear that many sessions in industrial applications may be characterised by very long duration, as they remain open for long period of time; for example, an operator workplace supervising a special area of a power plant can be connected to a server for 10 years without termination. Considering the data exchange inside sessions which remain open for very long periods, waiting times of tens of seconds to validate a certificate when the session is opened (at the start-up) are negligible. But, industrial applications are featured by a lot of applications which must be connected to a server for short period of time; furthermore these applications must access the server when needed without any delays. The most typical example of such applications are supervising and monitoring applications to manage faults or emergencies; these applications create a connection to the server to properly manage fault or emergency, only for the time period needed to resolve the problem. In those cases, any delay in the connection should be avoided.

The other aspect of OPC UA security which seems very important to investigate is relevant to the impact of data encryption/decryption and the digital signature of each message exchanged between OPC UA Client and Server. As known, encryption allows to achieve confidentiality in the data exchange, but in many applications at field device level, confidentiality isn't a strong requirement; for this reason, an analysis of the overhead introduced by encryption during data transfer should be performed in order to highlight if and when data encryption may not be used (as it isn't mandatory according to the OPC UA specifications). The same considerations must be extended to the digital signatures foreseen in the OPC UA specifications (but not compulsory) and aimed to maintain integrity; also in this case a study of the overhead introduced by the signature of each message exchanged seems very important.

3.2 Transport Protocols and Encoding Rules

As said, OPC UA may use the two different transport technologies: UA TCP and SOAP protocol; furthermore, both binary and XML encoding are currently available. Performance evaluation should take into account the different transport protocols and encoding rules, comparing their impact on the data exchange.

3.3 Subscription

Subscription is the mechanism able to deliver information produced in a cyclic fashion.

The previous section pointed out that the subscription is based on several parameters; among them, the Publish Interval seems to play an important role in the overall performance. As said, this parameter determines the time instants at which the Monitored Item queues (containing values coming from changes of variable values or from aggregates of variable values or from events), are emptied and a Notification is prepared to be sent to the Client; these Notifications are then pulled by the Client issuing Publish Requests. It's clear that a small value of the Publish Interval allows the client to receive fresh values, as the Monitored Item queues are emptied very soon, but requires a large amount of Publish Requests to be sent (reducing the available bandwidth in the underlying communication entity). On the other hand, greater values of Publish Interval lead to Notifications containing huge number of data (i.e. variable values, aggregates and events), some of which may be obsolete for the client; in this case Client is compelled to send low numbers of Publish Request.

On the basis of what said, the number of outstanding Publish Requests a client should maintain is another very critic parameter and could influence the overall performance of the system. Frequency of transmission of Publish Requests depends on the Publish Interval value, as said before, but it may be linked also to other events; for example, additional Publish Requests may be required if the latency of the network connection is very high. In any case, the number of outstanding Publish Requests maintained by each Client may strongly impact on the bandwidth utilisation, and could led to bottlenecks in the client/server data exchange.

4 Performance Evaluation Results

The previous section highlighted the main features of the OPC UA specifications which seem, more than others, able to influence the relevant performance. This section will present the performance evaluation and the relevant results, realised in order to investigate their impact on OPC UA data exchange.

4.1 Main Hypotheses about Performance Evaluation

The main hypothesis assumed for the performance evaluation was that to realise an ad-hoc model of the OPC UA specifications and client/server data exchange, instead of fully implementing them using one of the languages, stacks and SDKs supported by the OPC Foundation [WWW-1].

This choice was due to advantages offered by the use of a software model of OPC UA instead of its real implementation on a PC platform; the main advantage is that use of a model avoids the need to detail the internal mechanisms of the OPC UA specifications not directly involved in the aim of the performance evaluation, focusing only on certain mechanisms and their impact on performances. Furthermore, the use of a model allows achieving performance evaluation results not linked to a specific operating system, to a particular software development environment (i.e. to particular libraries), and to specific hardware architecture.

The model of the OPC UA specifications and client/server data exchange has been realised inside OMNeT++ framework [WWW-2]; performance evaluation has been realised through simulation of the model. A great effort has been put to verify that the model behaved exactly as stated by the OPC UA specifications. The model has been defined guaranteeing that the behaviour of the OPC UA specifications and client/server data exchange modelled was close to the real one as much as possible; this has been done executing each OPC UA service modelled and verifying its behaviour inside the OMNeT++ framework using the available tools. Other frameworks have been used to support the OPC UA model: the OpenSSL [WWW-3] and the INET framework [WWW-4]. OpenSSL libraries have been used in the model to realise the main security mechanisms foreseen by OPC UA specifications; in particular, the encryption and decryption mechanisms have been realised using the Basic128RSA15. The INET Framework has been used to realise the Point-to-Point Internet Protocol (PPP) at data link layer, through which the Client and Server data exchange was realised.

Other main hypotheses assumed in the OPC UA model are listed in the following. Both UA TCP and SOAP/HTTP have been used at transport layer, using only the UA binary encoding. Verification of Certificates by local and remote CAs has been modelled using real examples of CAs and deriving from them the average delays needed to perform verification; the values of these average delays have been used in the OPC UA model to represent the time spent for the verification of certificates by local or remote CAs, when requested during the simulation.

4.2 Parameters Used for the Performance Measurements

The success of a performance evaluation depends on the right choice of the most meaningful parameters able to achieve significant performance measurements. Choice of the parameters to be evaluated must be generally done taking into consideration the aim of the performance evaluation itself.

In the case here presented, the performance evaluation aims to analyse the impact of the OPC UA mechanisms pointed out in the previous section, onto the data exchange between OPC UA-based client and server applications. So, the choice of the parameters to be measured must be realised analysing the typical data exchanges featured by client/server industrial applications.

The most common data exchange may be called asynchronous, and occurs when a client (e.g. SCADA) requires to a server (e.g. acquisition board) the (read and/or write) access to one or more variables at unforeseeable time instants. Another kind of data exchange is cyclic or periodic and occurs when a client accesses to values of variables in a cyclic or periodic fashion; an example is a server which produces values of a variable (e.g. temperature, pressure) in a periodic fashion and a client which needs to read these values according to a periodic polling algorithm, with the same production period. In these two different scenarios the single piece of information to be transmitted could be simple (e.g. an integer, a float) or complex (e.g. a data structure made up by several bytes or a large set of variables).

Considering the first kind of data exchange (the asynchronous one), *round-trip time* seems a suitable parameter for the performance measurements; considering the read service issued by a client, round-trip time may be defined as the total response time between the instant at which a request to read one or a set variables is issued by a client and the instant at which the relevant values are delivered to the client. A similar definition may be given for the write service; in this case the time interval is between the instant at which the client delivers the value or the values to the server and the instant at which the server confirms the update of the value/values. Round-trip time seems able to measure the efficiency of the asynchronous data exchange, as it points out the response time of the underlying communication system (including OPC UA stack) for each request issued by a client.

Considering the periodic data exchange, in this analysis, only the data exchange of variables whose values are periodically produced on the server-side have been considered; furthermore, it has been assumed that a client needs to consume these values according to a periodic polling algorithm, with the same production period. Figure 5 shows a server producing values with a certain period T ; a client application should receive each value produced within a deadline. In this case, the *delay* between the instant at which each value has been produced and the instant at which the client receives that value, seems to be of interest during performance evaluation. In fact, it's clear that the average delay for each variable gives a measurement of the efficiency of the data exchange, pointing out the capability to deliver to the client each information periodically produced within the deadline.

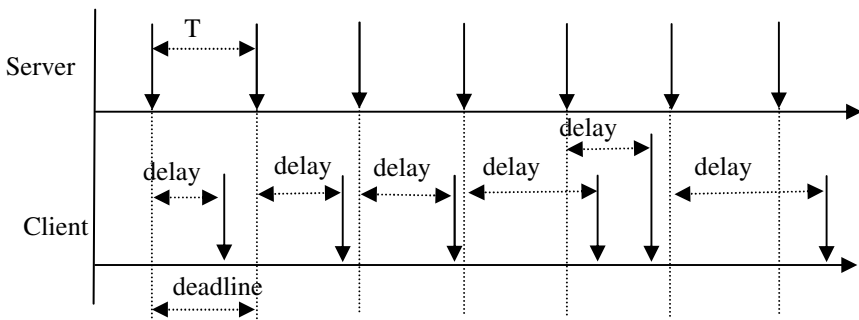


Fig. 5 Delay between each information produced by a server and delivered to a client

4.3 Performance Evaluation about OPC UA Security Mechanisms

The security aspects of OPC UA specifications which have been investigated, are the verification of certificates and the data encryption and signature of each message exchanged between OPC UA client and server. As said before, OPC UA specification doesn't make mandatory the use of certificates, digital signatures and data encryption; this improves the strategic role of a performance evaluation aimed to highlight the relevant overload introduced. Results of the performance evaluation may help the final user of the OPC UA to evaluate when the choice of one or more of the previous security items is more appropriate.

The following main scenarios have been considered during performance evaluation of the OPC UA security: (1) data exchange with no security mechanisms, (2) use of the Secure Channel with no use of certificate (e.g. using passwords or other credentials), (3) use of the Secure Channel with local verification of the certificates (i.e. operated by a local Certification Authority-CA) and (4) use of the Secure Channel with remote validation of the certificates (i.e. operated by a remote hierarchy of CAs).

For the scenarios (2), (3) and (4), the following sub-cases have been considered: (a) no other security option used, (b) use of only digital signature for each message exchanged, and (c) use of both signature and data encryption for each message exchanged.

Combining scenarios (2), (3) and (4) with the previous three sub-cases (a), (b) and (c), and including the scenario (1) alone, 10 different scenarios have been achieved. As said before, the scenario (1) (data exchange with no security mechanisms) is not featured by any sub-cases, so it will simply named scenario number 1 in the following; the other scenarios will be indicated in the following using a number (ranging from 2 to 4) and a letter (a, b and c). The number refers to one of the scenarios (2), (3) and (4) and the letter refers to one of the 3 sub-cases; for example scenario 4.a means use of the Secure Channel with remote validation of the certificates, and no other security option used.

For each scenario, a couple of OPC UA client and server exchanging data has been considered, assuming different available bandwidths: 2, 5 and 10 Mbps; only results related to a 2Mbps bandwidth will be shown in the following.

Data exchange has been assumed to be realised through Read services, i.e. based on asynchronous information flow of bulk of variables generated and transmitted at unforeseeable instants.

A first set of simulations has been carried on in order to highlight the influence of the security mechanisms on the times needed to open and activate a secure session. Only the four main scenarios (1), (2), (3) and (4), without the sub-cases (as they have no influence on the activation of a session), have been considered. Both the two transport mechanisms (UA TCP and SOAP), has been considered. In the case of secure session with the use of remote certification authorities, the times needed to activate a session are very huge (14.0 s for UA TCP and 15.9 s for the SOAP); use of UA TCP leads to a very little save in time (less than 2 seconds). For the other scenarios, times are less than 0.4 s; in the case of lack of security (scenario 1), influence of session activation is absolutely negligible.

Performance evaluation has been carried on also to investigate the influence of digital signature and data encryption on the round-trip times, defined in the previous section. Evaluation of the round-trip times has been achieved considering different sizes (in bytes) of the set of variables read from the server. Figure 6 compares the round-trip times achieved considering the scenarios 1 and 2.c and the two different transport mechanisms; abscissa of the figure is relevant to the size (in bytes) of the set of variables read from the server, as said before.

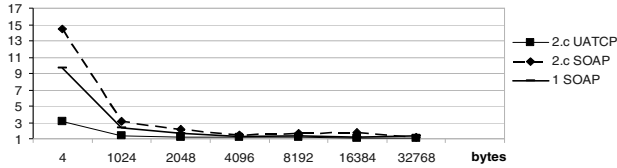


Fig. 6 Round-trip time, Scenarios 1 and 2.c

Each curve shown in Figure 6 is obtained normalising the values of the round-trip time to those achieved considering scenario 1 (no security) and UA TCP. So curve labelled with “2.c UA TCP” refers to the round-trip values concerning scenario 2.c and UA TCP, normalised to the values achieved considering scenario 1 and UA TCP; curve “2.c SOAP” refers to the round-trip values considering scenario 2.c and SOAP normalised to the values achieved considering scenario 1 and UATCP. Finally curve “1 SOAP” refers to the round-trip times relevant to scenario 1 and SOAP normalised to the scenario 1 and UA TCP. Table 1 gives the values of the Round-trip time related to the scenario 1 and UA TCP.

Table 1 Round-trip times related to scenario 1 and UA TCP

Bytes	4	1024	2048	4096	8192	16384	32768
Round-trip Delay (s)	0.003	0.019	0.036	0.068	0.1	0.19	0.39

Figures 7 and 8 compare the round-trip times considering the scenarios (1, 3.c) and (1, 4.c), respectively. Again the round-trip values are normalised as explained before and the abscissa refers to the size of data read from the server.

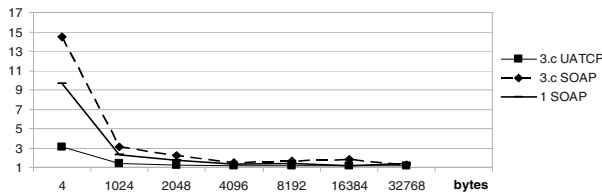


Fig. 7 Round-trip time, Scenarios 1 and 3.c

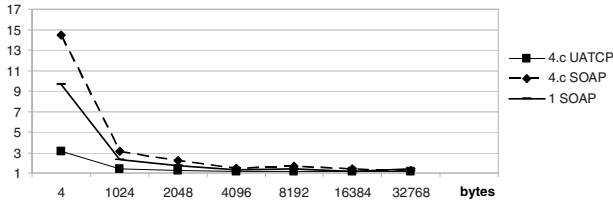


Fig. 8 Round-trip time, Scenarios 1 and 4.c

Figures 6, 7 and 8 point out that use of UA TCP leads to lower round-trip times, also in presence of the security mechanisms. This mainly occurs with small size of data exchanged; when the size of variables exchanged increases, the performance of the UA TCP and SOAP tends to converge, also in presence of security mechanisms.

The last set of measures about security here shown, are relevant to the interest to investigate the influence on the overall performance of the different configurations it's possible to choose inside a secure channel; these configurations are relevant to the sub-cases (a), (b) and (c). The round-trip times have been measured again, considering only the UA TCP transport mechanism and the three scenarios (2), (3) and (4) which include the secure channel. The results achieved for the three scenarios are quite similar, so only those relevant to one of them (the fourth) will be presented in the following.

Figure 9 points out the round-trip times considering only the UA TCP mechanism and scenarios 4.a, 4.b and 4.c; as done before, these values have been normalised to those achieved considering scenario 1 (no security) and UA TCP (shown in Table 1, as said before). As can be seen from the figure, the influence of the digital signature and data encryption mechanisms is very huge for small size of variables exchanged; performances of the different scenarios converge when variables increase in size.

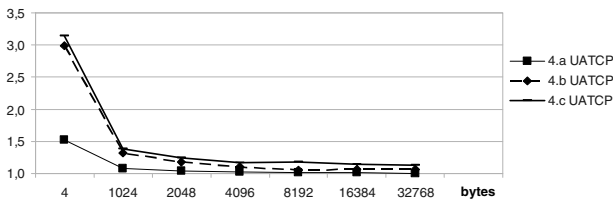


Fig. 9 Round-trip time, Scenarios 4.a, b, c

4.4 Performance Evaluation about OPC UA Subscription Mechanism

Influence of settings related to subscription mechanisms on the overall OPC UA performance has been investigated. This subsection will present the main results achieved.

Two sessions on the top of a secure channel have been considered; one session has been reserved for the exchange of information produced by server in a periodic fashion and linked to a subscription. The other session conveys asynchronous information (i.e. produced at instant not foreseeable in advance), linked to read requests by client; this second session has been considered only to reproduce a realistic scenario featured by a bandwidth occupation made by both periodic and asynchronous data exchanges. It has been considered that the asynchronous traffic occupies the 30% of the available bandwidth; as said before, the following values of bandwidth have been considered during the performance valuation: 2, 5 and 10 Mbps.

The periodic traffic is produced by 3 sets of variables featuring the following production periods: 5, 10 and 15 ms; 5 variables for each set have been considered. The size of each variable is 4 bytes.

For each variable, a Monitored Item has been associated in the OPC UA Server and configured to subscribe for data changes of the same variable; the sampling interval (see Figure 3) has been set equal to the relevant production period. For all the Monitored Items, the queue size has been fixed in order to avoid overflow, i.e. loss of values; this size has been determined running several simulations, analysing the queue occupation and choosing the suitable value able to avoid overflow.

Only one subscription has been considered to convey all the Monitored Items. Different values of Publish Interval have been considered during the performance evaluation, ranging from 5ms to 250 ms. The time interval at which client issues a Publish Request has been fixed to the 80% of the Publish Interval.

Only UA TCP with binary encoding has been considered and the scenario 4.c seen before has been assumed for the secure channel.

For each value produced relevant to a Monitored Item, the delay has been measured; according to what said in section 4.2, delay has been defined as the time interval between the instant at which each value is enqueued in the Monitored Item queue (due to a data change) and the instant at which the client receives the Notification containing this value.

As said in 2.1, a Notification delivered to a client contains all the values contained in a same Monitored Item queue; these values will be featured by different delays as their arrival times in the queue are different, off course. For this reason the lowest delay (relevant to the last value enqueued), the highest delay (relevant to the first value enqueued) and the average delay have been evaluated and updated for each Notification delivered to the client. Results presented in this section will point out these three values.

Figure 10 shows the delay versus the Publish Interval, considering the Monitored Item featuring a sampling interval of 5 ms and an available bandwidth of 2 Mbps. The delay values have been normalised to the sampling interval; this means

that a value of 1 corresponds to a delay equal to the sampling interval. Figures 11 and 12 have a similar content, but they refer to Monitored Item with sampling interval of 10 and 15 ms, respectively; available bandwidth is always 2 Mbps. As can be seen the trends of the delay are quite similar for the three scenarios. Limited values of delay are guaranteed by low values of Publish Interval; when Publish Interval increases, delay may assume too huge values.

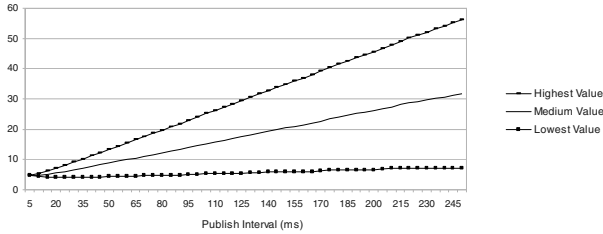


Fig. 10 Delay/Sampling Interval (5 ms), at 2 Mbps

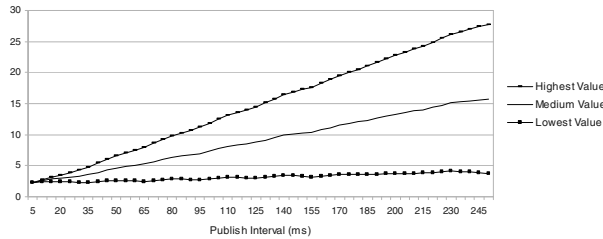


Fig. 11 Delay/Sampling Interval (10 ms), at 2 Mbps

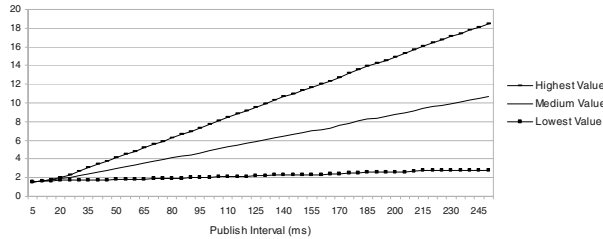


Fig. 12 Delay/Sampling Interval (15 ms), at 2 Mbps

Figures 13, 14 and 15 shows the delay versus the Publish Interval, considering the Monitored Item featuring a sampling interval of 5, 10 and 15 ms, respectively; in this case the available bandwidth is 10 Mbps. Again, the delay values have been

normalised to the sampling interval. Also in this case, the trends of the delay are quite similar for the three scenarios; limited values of delay are guaranteed only by low values of Publish Interval.

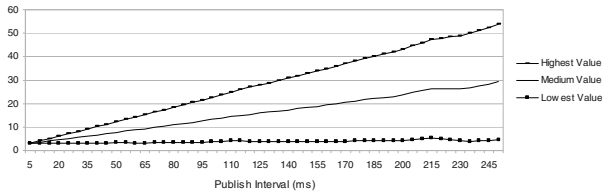


Fig. 13 Delay/Sampling Interval (5 ms), at 10 Mbps

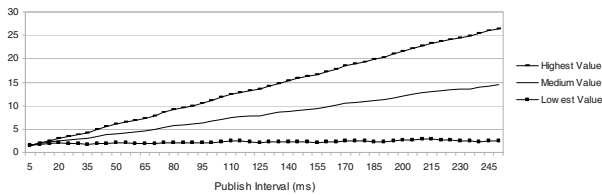


Fig. 14 Delay/Sampling Interval (10 ms), at 10 Mbps

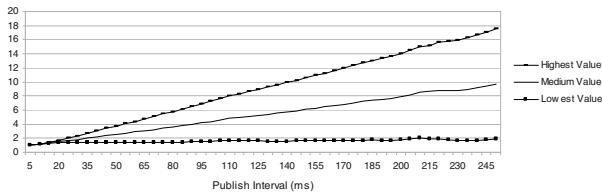


Fig. 15 Delay/Sampling Interval (15 ms), at 10 Mbps

Analysis of these curves is not enough to reach to some conclusions, as low values of Publish Interval means a high frequency transmission of Publish Requests by client, i.e. overload on the transmission medium (as pointed out in Section 3.3). In order to achieve a complete view, the impact of low values of Publish Interval on the bandwidth utilisation must be evaluated.

Figures 16 and 17 show bandwidth utilisation versus the Publish Interval, considering a total bandwidth of 2 and 10 Mbps, respectively. All the values of bandwidth utilisation are above the 30% value, as it has been assumed that the asynchronous traffic occupies the 30% of the available bandwidth, as said before. As can be seen from the figures, low values of Publish Interval lead to a very high bandwidth occupation; this is more evident considering Figure 16, due to the low

value of total available bandwidth. This means that choice of Publish Interval is very critical and difficult as a trade off between bandwidth utilisation and delay must be achieved; for example, comparing Figure 16 with Figures 10, 11 and 12, it's clear that values of Publish Interval close to 35 ms, represents a good trade off between the need to receive (on the client side) fresh values of variables and the need to maintain suitable percentage of available bandwidth.

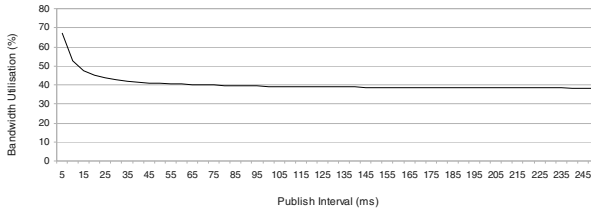


Fig. 16 Bandwidth utilisation, at 2 Mbps

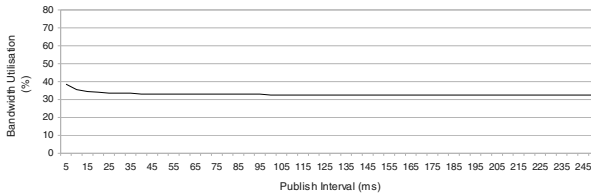


Fig. 17 Bandwidth utilisation, at 10 Mbps

5 Conclusions

The paper has presented the OPC UA specifications, pointing out some features which may have an impact on the overall performance of the client/server data exchange. The paper has then proposed some measurement parameters to be considered in the performance evaluation. Finally, the main results achieved during performance evaluation have been presented and discussed. Results pointed out that some OPC UA parameters (e.g. Publish Interval) are very critical and their setting is very difficult, based to deep and technical considerations and requiring high level of expertise.

References

- [Braune et al. 2008] Braune, A., Henning, S., Hegler, S.: Evaluation of OPC UA secure communication in web browser applications. In: Proc. IEEE International Conference on Industrial Informatics, Daejeon, Korea, pp. 1660–1665 (2008)

- [Mahnke et al. 2009] Mahnke, W., Leitner, S.H., Damm, M.: OPC unified architecture. Springer, Heidelberg (2009); ISBN: 978-3-540-68898-3
- [OPC Foundation 2009] OPC Foundation OPC UA Specification: Parts 1–13
- [Post et al. 2009] Post, O., Deppala, J., Koivisto, H.: The performance of OPC UA Security model at field device level. In: Proc. ICINCO, pp. 337–341 (2009)
- [WWW-1] <http://www.opcfoundation.org>. (accessed March 29, 2011)
- [WWW-2] <http://www.omnetpp.org>. (accessed March 29, 2011)
- [WWW-3] <http://www.openssl.org>. (accessed March 29, 2011)
- [WWW-4] <http://inet.omnetpp.org/> (accessed March 29, 2011)

Part II

Distributed Knowledgebase's and Web Systems

A Document-Centric Processing Paradigm for Collaborative Computing

B. Wiszniewski

Faculty of Electronics, Telecommunications and Informatics,
Gdansk University of Technology, Gdansk, Poland
bowisz@eti.pg.gda.pl

Abstract. Classic models of distributed processing assume documents to be passive objects, sent to remote recipients as messages, or downloaded from remote sites as files. The paper introduces a concept of documents being implemented as *active* objects that can migrate from the originating host to remote sites, and interact there with local users. Upon completing their mission, documents return to the originating host with a resulting content to be archived or processed further. Such a *document-centric* computing paradigm, involving documents implemented as mobile and interactive agents with embedded functionality, is more flexible and usable for knowledge based organizations than *data-centric* computing, with functionality hard-wired in individual processing nodes. This is particularly important for collaborative computing systems, where human and artificial agents render and use services interchangeably. Moreover, it stimulates mobility of users, as it reduces the need for them to stay on-line during the entire computation process and allows for using less sophisticated personal devices.

1 Introduction

Distributed *Mobile Interactive Document (MIND)* architecture [Godlewska and Wiszniewski, 2010] introduces new mechanisms for effective and flexible implementation of collaborative computing systems based on interactive document exchange. So far such systems, e.g., decision support or crisis management systems, virtual collaboration frameworks and alike, have been implementing documents as static units of information, being sent as messages (email attachments or SMSes) to remote users, uploaded to or downloaded from remote servers by interested users. MIND assumes a document to consist of dynamic components, capable of migrating and interacting on their own with cooperating users, who may not only read, edit, fill, or expand document content but also annotate it dynamically or add new components. Document components are implemented as autonomous agents and may use their embedded functionality, services available locally at the current user's

computation device, as well as external services of other hosts to extend local services when necessary. Novelty of the MIND concept with regard to implementing business processes is the possibility to process a document content by a virtual collaboration team in a flexible, effective and creative way, and providing business process design patterns for virtual collaboration that are natural to humans. Actionable content makes documents intelligent and self-manageable units of interface.

1.1 Knowledge Based Organizations

Business process design patterns involves usually two types of actors: artificial (system) agents and human users, who cooperate to solve non-algorithmic problems, such as court trials, integrative bargaining, medical consultations, crash investigation, forensic work, etc. They concern situations when knowledge constitutes a basic resource necessary to succeed in making a *proper* rather than *optimal* decision. A proper decision makes sense in knowledge organizations when no optimization criterion for a decision exists – for example, choice of a therapy for a patient may be proper, i.e., conformant to the best medical practices, but not necessarily optimal, especially when a patient dies. Organizations bound to make such decisions are termed *knowledge based*. Their business processes are *continuous* and *cyclic*, and enable for systematic building of organizational knowledge resources in a form of document repositories. An inherent feature of these processes is intensive interaction of actors using some dedicated communication platform, e.g., a company intranet. Owing to interaction between collaborating users, knowledge may not only be acquired from existing resources, but also expanded, disseminated and explored to discover new knowledge. For this reason participants of such a process are often called *knowledge workers*.

Modern management theory indicates emerging trends in organizations (government agencies, universities, companies) towards the collaborative model indicated above, as a condition to survive on the market [Kahaner 1997]. Of particular importance today are communication techniques and mechanisms, as means for effective knowledge sharing and exchange. Success stories of large scale collaborative organizations are Linux development communities [Stallman 2008] and the human genome project [Human_Genome]. Document-centric paradigm proposed by MIND can certainly leverage comfort of knowledge workers and quality of their collaboration.

1.2 Non-algorithmic Decision Problems

Complexity of interaction between knowledge workers, who collaborate to work out a proper decision goes beyond the classic algorithmic model of a Turing machine. A reason for that is that interaction of independent actors in a collaborative

system involves data resources which are beyond control of the system. This phenomenon has been indicated by Wegner over a decade ago [Wegner 1997], when interaction of people and systems became more sophisticated as architectures and Web technologies advanced. Wegner has proposed to represent a space of implementable computations as shown in Fig.1: P represents a spectrum of parallel computations, D a spectrum of distributed computations, and I a spectrum of interactive computations. Parallel computations are performed in parts executed in the same time, distributed computations are performed in parts executed at separate geographical locations, while interactive computations involve parts reacting to various external events delivering to the system data associated with these events. Each point in space $P \times D \times I$ represents an implementable computation, involving components (processes) performed in parallel, in distribution and in reaction to some external events. Wegner has noted that classic sequential algorithms correspond to the point of origin of space $P \times D \times I$, and compared them to a *business contract*, according to which a user delivers some data to a service and receives in return a specific result, computed in a series of steps. In order to speed up result delivery, or because of data distribution, or both, algorithms are often executed in parallel and/or with distribution. However, for a commonly agreed set of input data their computations are always represented by points in plane $P \times D$.

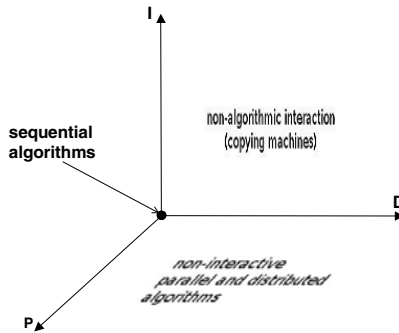


Fig. 1 Space of implementable computations

In opposite to parallel/distributed algorithms, computations corresponding to points in planes $D \times I$ and $P \times I$ are not algorithmic, owing to a potentially infinite set of external events, to which a system should react. When modeled by a Turing machine, computations belonging to planes $D \times I$ or $P \times I$ would require an infinite set of tape symbols, thus going beyond the original Turing machine definition. Wegner gives a simple copying machine as an example of such a non-algorithmic computation: user A types input to a system, which copies it as output to user B, who in return types input copied by the system as output to A, and so on. This process, like writing a paper by two authors for example, may be continued forever. Because users A and B are independent and uncontrollable sources of data, their cooperation cannot be modeled by any point in plane $P \times D$, i.e., any

algorithmic computation. In other words, no algorithm exists that could generate automatically any sensible content of the paper mentioned before.

In collaborative computing systems one should therefore consider computations on data related to new and yet unknown events, as well computations performed in parallel and distribution.

According to Wegner's model, collaborative computations may be defined as those that involve events and data (as represented by axis *I*) generated by human actors. Consequently, if these data are stored and exchanged as documents, collaborative computing becomes document centric.

1.3 Document-Centric Processing Scenarios

Owing to the advance of mark-up languages, making document content readable simultaneously to humans, e.g. with Web browsers, and computer systems, e.g., with XSLT processor, electronic documents provide a natural mean for information exchange in knowledge organizations using collaborative computing. From a technical point of view collaboration of actors implies exchange and collaborative edition of a set of documents circulating in the system, with just a few simple operations performed on texts or images: *insert/paste*, *cut*, *copy* and *delete*. In general there are three basic scenarios for collaborative editing of such documents [Godlewska and Wiszniewski, 2010]:

1. *Pessimistic*, when edited document is a shared resource, and only one copy is available for reading and writing by remote users on a dedicated server. Owing to the shared lock mechanism just one user at a time may edit it. Disadvantages of this scenario are a need to stay on line by all interested users, no provision for concurrency and additional overhead incurred by the shared lock mechanism when the number of users is high.
2. *Optimistic*, when prior to editing a document is copied a certain number of times. Copies are sent to the respective users for editing. Messages concerning content changes of the respective copies are broadcasted in a system, so they may update themselves with each received message. After some period of time each copy will eventually assume a common final state. This scenario is useful when remote users edit disjoint document fragments, so document integration to the final state will require exchange of messages in just one step. A serious disadvantage is a need to implement a sophisticated mechanism for correcting editing actions already performed on overlapping document fragments of remote copies, on the basis of message exchange. Unfortunately previously published algorithms for that turned out later to be flawed and do not guarantee content consistency of distributed copies edited in parallel [Oster et. al. 2005].
3. *Realistic*, when a document is split in logical components, each having its own memory to store its state. Advantage of this scenario is a possibility to use standard object oriented processing mechanisms, especially those of open agent systems, providing document components with mobility and functionality.

A realistic scenario proposed in this paper is advantageous to the previous two ones in several ways. Firstly, decomposition of a large document into a set of logical components performed once is less expensive than its replication. Secondly, upon decomposition of a document, its logical components (usually small) are sent only to the corresponding users, rather than to all of them. Finally, logical components of a document are processed in parallel on distributed user machines, and owing to the disjoint content of the components, messaging between users is limited to a reasonable minimum.

2 The MIND Architecture

The first prototype of MIND was implemented using Java Agent Development Environment (JADE) and XML related notations [Godlewska and Wiszniewski 2009]. Selection of XML was motivated by its simplicity, a reach set of processing tools, and perspectives to survive as a dominant Web standard for years to come – ensuring forward compatibility of MIND documents with formats and technologies that will emerge in the future.

2.1 Mobile Document Life-Cycle

Architecture of MIND is based on a simple combination of two popular Web concepts: automatic *XML data binding* with Java objects [Bourret 2010], and *mobile agents* [FIPA]. Data binding allows for converting units of information contained in document components into functional objects in a computer memory, augmented next with mobility to make them autonomous objects, capable of migrating in an open distributed system. Owing to this, a static representation of an electronic document (in an initial form an empty template), is transformed into a set of dynamic objects that can migrate to remote locations, and perform actions at these locations by interacting locally with their users and services (see Fig. 2).

Lifecycle of a MIND document is initiated by an *originator*, a user who is responsible for designing a logical structure of a *hub document*, using a repository of templates. Its logical structure includes in particular two parts: one specifying a document *migration path*, and another specifying *services* for migrating document components with users at remote locations. Services specified by a document component may be *embedded*, *local*, or *remote*. Embedded services are implemented with scripts carried by the migrating components, what makes the latter relatively independent of local environments at remote sites. Local services specified by a component are requested by it from the local host environment upon arrival. In a case when specialized services are provided by someone else (most often by the originating server), they may be specified as *external* ones. Knowledge workers interact with dynamic components currently residing at their hosts and contribute to their content. Upon completion of the respective final activities specified by the

migration path, dynamic components return to their originator to be integrated into a final document, marshaled back to the static XML form and archived. Archived documents are stored in a repository for further use, in particular for extracting knowledge gathered during their migration.

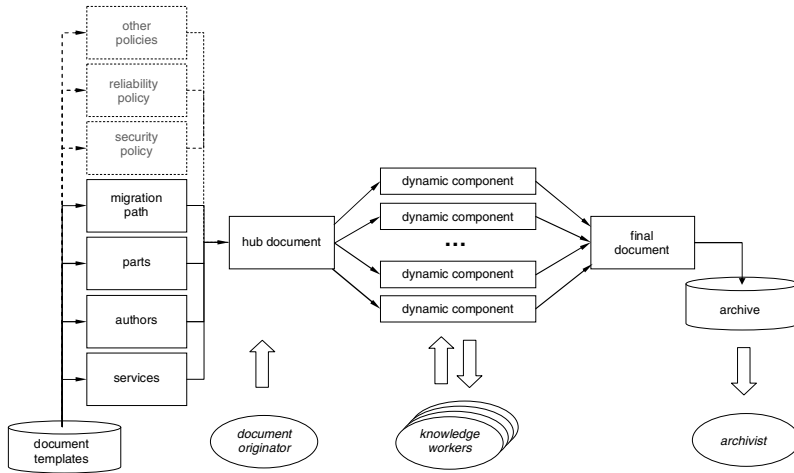


Fig. 2 A conceptual view of a MIND lifecycle

MIND is an open architecture, allowing for functionality extensions of dynamic document components to meet specific quality requirements in a form of policies, in particular *reliability* and *security*. The former will involve mechanisms for document integrity checking and state recovery, while the latter content encryption, digital signatures and self-diagnosability.

2.2 Component Objects

A MIND document is an instance of class <Document> and consists of dynamic objects, as outlined schematically in Fig. 3.

Central components are objects of class <Activity>, which associate responsible authors (objects of class <Author>) and document parts (objects of class <Part>) with a document migration path (an object of class <WorkflowProcess>). Each author may or may not be allowed to modify an associated document part (attribute `active="yes"`|"no"), which according to its history may be new, under processing or done (attribute `state="new"`|"modified"|"verified-positive"|"verified-negative"|"done").

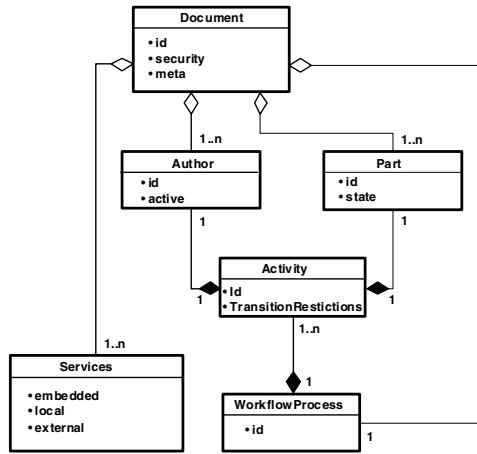


Fig. 3 Dynamic MIND objects

Types of a specific split or join of transitions associated with the activity are specified by attribute TransitionRestrictions, according to the XPDL syntax [XPDL]. Another important MIND document components are instances of class <Services>, which determine functionality of a migrating document. In the first prototype implementation of MIND each activity has only one responsible author, who may perform actions manually. Extension of functionality of <Activity> objects is planned in future versions of the MIND prototype to enable monitoring an activity deadline, and to allow for automatic execution of specific activities by a local system when necessary.

3 Document Mobility

Throughout the rest of this section three key components of the novel MIND architecture are presented: <parts>, that are augmented to mobile objects (as outlined schematically in Fig. 3), <servicesTypes> that provide three kinds of functionality for documents converted to objects, and <WorkflowProcesses> that specify migration paths for the unmarshalled document components. Their logical structure is explained briefly throughout the rest of this section. A complete specification of all MIND components may be found in [Godlewska and Wiszniewski 2009].

3.1 Document Parts

Content of a MIND document is distributed over a set of <part> objects, of the internal structure outlined in Fig. 4.

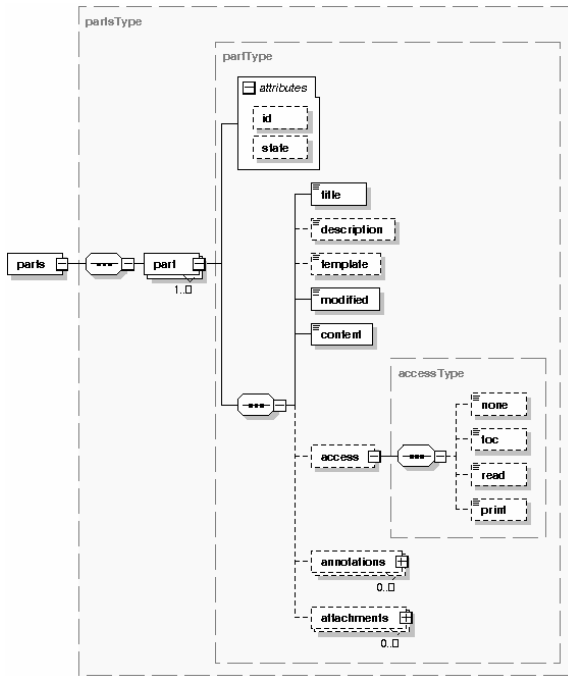


Fig. 4 Specification of a document component content

Elements `<title>`, `<modified>` and `<content>` specify respectively a part title, last date of its modification and the part contents encoded in Base64 format. Two elements `<description>` and `<template>` provide respectively an optional full-text description of the part content, and its relevant template id. The latter may be useful for selecting which application to use at the local station of a knowledge worker based on a particular MIME type specification provided by the template. Besides any content, each part may have a list specifying access rights for each respective author (element `<access>`), a collection of annotations of the content (element `<annotations>`), and a set of other documents attached to the given part (element `<attachments>`).

Elements of `<access>` specify up to four list of authors: without access (element `<none>`), allowed to read table of content (element `<toc>`) or part content (element `<read>`), and to print part content (element `<print>`). Note that element `<write>` is missing in this schema, since authors authorized to modify the content are specified already in respective `<activities>` elements of the `<WorkflowProcesses>` elements. By default, authors not specified in elements `<access>` and `<activities>` of `<part>` have no access rights.

3.2 Document Services

Services of a document component are specified in Fig. 5

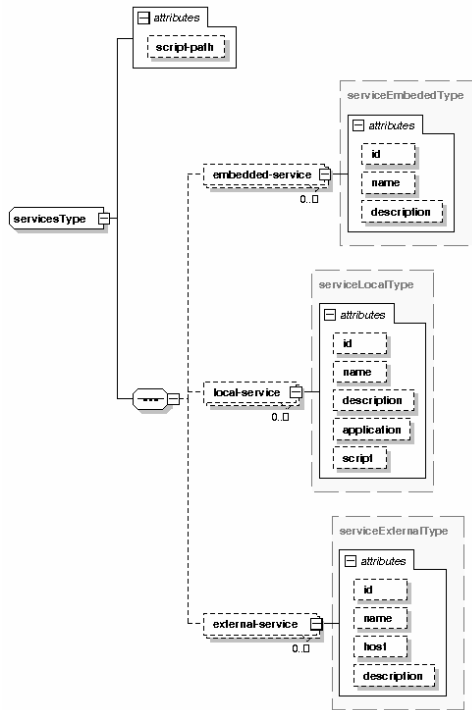


Fig. 5 Specification of document component services

Autonomy of a document component implies its independence of a local computer environment, where it currently resides when accessed or processed by a user specified in its `<access>` or `<activities>` elements. Three types of services can support that: `<embedded-service>`, `<local-service>` and `<external-service>`; they fully control execution context of a component at any node it may reach. This idea is outlined in Fig. 6.

Upon arrival to its destination host, an agent carrying a content of document `<part>`, cooperates with a lightweight MIND browser installed there. Embedded services of a document component enable it to interact directly with services implemented by the browser, in particular the user (knowledge worker) interface.

Local services are provided by other applications installed at the local host, access to which is controlled by the MIND browser. Contents of `<local-service>` elements specify services and tools required by each respective component, which

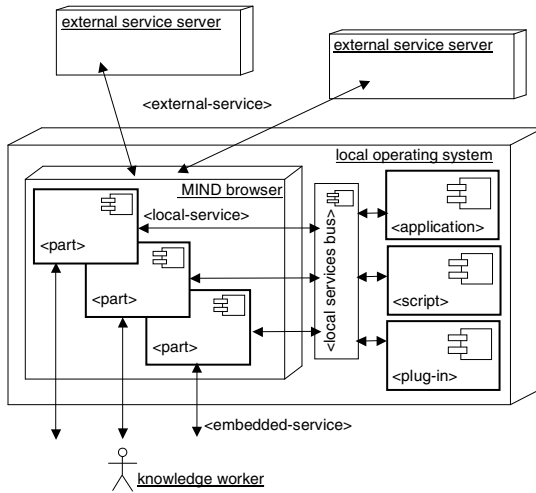


Fig. 6 Execution context of a dynamic MIND component

most often are common and standard. When the required application is different than the one locally installed, but of a similar functionality, e.g. a locally available application is Open Office, but the component requires MSOffice, the component may indicate accepted substitutions. If no substitution is specified by the arriving component, or the substitution is not available for some reason, the component may provide a specialized script for adapting its interface to the locally available tools, or convert its content format to the one expected locally. Finally, if such a conversion is not possible either, MIND browser may request specific external services indicated by the component. It may be resolved by downloading and installing a respective plug-in, or just execution of a service at some remote server.

3.3 Document Workflow

Migration path of MIND components is specified with XPDL, a format supported by the Workflow Management Coalition [WfMC 2008] – see specification of the topmost XPDL elements in Fig.7.

Current prototype implementation of MIND uses a relatively small subset of XPDL elements, what given the evolutionary approach to the XPDL standard development adopted by WfMC, guarantees forward compatibility of MIND documents with workflow systems that may be designed in the future.

Main element of a MIND component workflow is <WorkflowProcesses> and may contain arbitrary many <WorkflowProcess> elements, each of which consists of <Activities> and <Transitions> elements. They describe a Petri net, an alternative representation of the process.

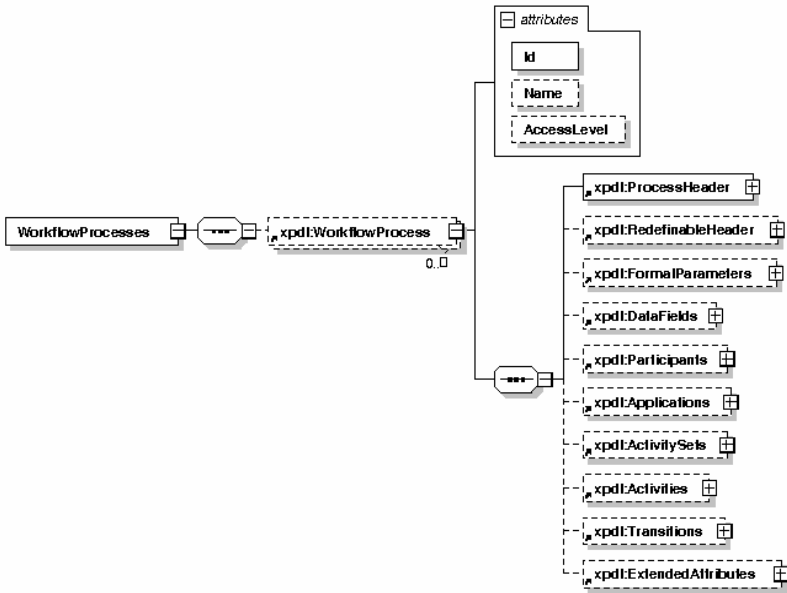


Fig. 7 Specification of a document component workflow

Consider the following piece of XPD code, which describes preparation of a document consisting of two parts, edited in parallel by two independent authors:

```

<xpd:WorkflowProcess>
  <xpd:Activities>
    <xpd:Activity Id="ACT01" Name="Document outline preparation">
      <xpd:Performer>AUT01</xpd:Performer>
      <xpd:TransitionRestrictions>
        <xpd:TransitionRestriction>
          <xpd:Split Type="AND">
            <xpd:TransitionRefs>
              <xpd:TransitionRef Id="TRA01"/>
              <xpd:TransitionRef Id="TRA02"/>
            </xpd:TransitionRefs>
          </xpd:Split>
        </xpd:TransitionRestriction>
      </xpd:TransitionRestrictions>
    </xpd:Activity>
    <xpd:Activity Id="ACT02" Name="Writing part1">
      <xpd:Performer>AUT02</xpd:Performer>
    </xpd:Activity>
    <xpd:Activity Id="ACT03" Name="Writing part2">
      <xpd:Performer>AUT03</xpd:Performer>
    </xpd:Activity>
    <xpd:Activity Id="ACT04" Name="Integration of parts 1 & 2">
  
```

```

<xpdl:Performer>AUT01</xpdl:Performer>
<xpdl:TransitionRestrictions>
  <xpdl:TransitionRestriction>
    <xpdl:Join Type="AND"/>
  </xpdl:TransitionRestriction>
</xpdl:TransitionRestrictions>
</xpdl:Activity>
</xpdl:Activities>
<xpdl:Transitions>
  <xpdl:Transition From="ACT01" Id="TRA01" Name="Send" To="ACT02"/>
  <xpdl:Transition From="ACT01" Id="TRA02" Name="Send" To="ACT03"/>
  <xpdl:Transition From="ACT02" Id="TRA03" Name="Receive" To="ACT04"/>
  <xpdl:Transition From="ACT03" Id="TRA04" Name="Receive" To="ACT04"/>
</xpdl:Transitions>
</xpdl:WorkflowProcess>

```

Author AUT01 prepares a document outline (activity ACT01), which is next multiplied and send (transitions TRA01 and TRA02) to authors AUT02 and AUT03, who write their parts of a document in parallel (activities ACT02 and ACT03). As soon as they finish writing, each part is sent back to author AUT01 (transitions TRA03 and TRA04) for integration. Upon receiving both parts (synchronized merge indicated by `<xpdl:Join Type="AND"/>`) author AUT01 integrates two received parts into a final document during activity ACT04.

Parallel split `<xpdl:Split Type="AND"/>` of transitions implies multiplication of a MIND component, copies of which continue their migration to respective authors independently of one another. Upon parallel join `<xpdl:Join Type="AND"/>` all arriving components are merged into one component again. Another form of split in XPDL (not shown in the example above) is exclusive choice `<xpdl:Split Type="XOR"/>`, which implies selection of one specific path by a MIND component – at random or according to some specified condition. Consequently, exclusive join `<xpdl:Join Type="XOR"/>` implies just a blocking receive operation of one component.

4 Implementability of MIND

Implementability of mobile interactive documents standard languages and platforms has been proved by the prototype version of MIND [Godlewska and Wiszniewski 2009]. All related technologies are freely available and have a long term perspective to survive, what guarantees MIND documents forward compatibility with distributed processing tools and services that may emerge in the future.

A basic notation for specifying a logical structure of hub documents (see Fig. 2 and 3) is XML Schema, which is stable and has good prospects to remain a dominant standard for a long time. Even if XML Schema could be suppressed later by another notation, e.g. Relax NG, general properties of a formal grammar of schema languages will enable automatic conversion between formats [Murata et. al. 2001]. This guarantees persistence of MIND documents archived in the past, and susceptibility for automatic content analysis, in particular data binding to Java objects.

Mobility of document components in the first prototype of MIND has been implemented with JADE, but it also pose no limitation for future use of other agent platforms. This is because all features of the underlying agent platform required by MIND conform to the FIPA standard, which is most likely to survive.

Similarly, XPDL used in the current MIND prototype to specify migration paths conforms to XML syntax – so if in the future new notations emerge (and most likely still conformant to XML), conversions of archived MIND documents will not go beyond a straightforward XSL transformation. Besides, stability of the XPDL format is guaranteed by WfMC.

5 Conclusions

Mobility of document components combined with human interactivity introduce a new dimension in distributed processing. In order to make such a document-centric collaborative computing paradigm fully competitive in the open internet environment two challenges must be further addressed: business process reliability, and document content security. Reliability requires MIND documents to be self-diagnosable, i.e., capable of detecting and recovering from component lost, corruption of its content, migration failures, etc. Security of MIND components involves component contents encryption and digital signatures. Implementing mechanisms similar to those available in PDF documents (classes of users, security levels, etc.), seems to be straightforward.

References

- [Bourret 2010] Bourret, R.: XML data binding resources,
<http://www.rpbouret.com/xml/XMLDataBinding.htm> (last updated March 17, 2010)
- [Godlewska and Wiszniewski, 2009] Godlewska, M., Wiszniewski, B.: Architecture of MIND, a distributed mobile interactive document – feasibility study. Tech. Rep. 18/2009, Faculty of ETI, GUT(2009) (in Polish)
- [Godlewska and Wiszniewski, 2010] Godlewska, M., Wiszniewski, B.: Distributed MIND – a new processing model based on mobile interactive documents. In: Casadio, R., Myers, G. (eds.) WABI 2005. LNCS (LNBI), vol. 3692, pp. 244–249. Springer, Heidelberg (2005)
- [Human_Genome] Human Genome Project Information,
http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml (accessed April 10, 2010)
- [Kahaner 1997] Kahaner, L.: Competitive intelligence. How to gather, analyze and use information to move your business to the top. Touchstone, New York (1997)
- [Murata et al. 2001] Murata, M., Lee, D., Mani, M.: Taxonomy of XML schema languages using formal language theory. In: Proc. Extreme Markup Languages (2001)
- [Oster et al. 2005] Oster, G., et al.: Proving correctness of transformation functions in collaborative editing systems. Tech. Rep. 5795, INRIA, Nancy (2005)

- [Stallman 2008] Stallman, R.: The Free Software Community After 20 Years: With great but incomplete success, what now? (July 24, 2008), <http://www.gnu.org/philosophy/use-free-software.html>
- [Wegner 1997] Wegner, P.: Why interaction is more powerful than algorithms. *Communications of the ACM* 40(5) (1997)
- [WfMC 2008] Workflow management coalition workflow standard: process definition interface – XML process definition language version 2.1a. WfMC-TC-1025 (October 10, 2008), <http://www.wfmc.org>
- [XPDL] XPDL Schema file , http://www.wfmc.org/standards/docs/TC-1025_schema_10_xpdl.xsd (accessed April 10, 2010)

Computing Utilization via Computer Networks

N. Pham¹, B.M. Wilamowski¹, and A. Malinowski²

¹Electrical and Computer Engineering, Auburn University, Alabama, USA
{nguyehu, wilambm}@auburn.edu

²Electrical and Computer Engineering, Bradley University, Peoria Illinois, USA
olekmali@ieee.org

Abstract. The dramatic growth of Internet and network technologies, etc leads to different perspectives of computing methodologies as well as changes of software business model. If the traditional business model for software is one-time payment for a license for one machine with unlimited use, the development of Internet and network technologies, etc makes it possible for users to pay on their consumption as they pay for water, gas and electricity. With advanced technology all computing and storing process can be centralized on the infrastructure of service providers. With this new model, users don't have to concern about deploying their infrastructure, security, etc which will be responsible by service providers. This new trend grows extremely fast in last couple years and attracts a lot of researches from scholars such as Grid Computing model, Client Server model and especially Cloud Computing model with its scalability. In this paper we do not analyze differences between these utility computing models and what model will be the main field in the future. Instead we present how to use computer networks as a mean of computing and simulation and how computer networks are considered as a solution to boost technology development. Two software applications through computer networks were developed and applied successfully in teaching and learning courses in Auburn University and Bradley University are presented in this paper. It is a typical example of enhanced interaction between human and CAD tools while computer networks play a role as a human system interface.

1 Introduction

Since Internet has played an important role in communicating and exchanging information in the world [Wilamowski and Malinowski 2001; Manic et al. 2002], there are many applications developed and deployed on its basis from companies to academic institutions [Wilamowski et al. 1998; Malinowski and Wilamowski 2000; Wilamowski et al. 2000]. Up to now, Internet has been used as an effective means of computing and simulation. Explosion of network technologies and multi-core processor technologies with faster speed makes network computing become an interesting realistic and economic model. Network computing really changes topologies to design software applications [Arano et al. 1996]. Computer networks have been changing not only the engineering view but also the business view as well. Nowadays, many companies mostly rely on computer systems and networks for functions such as order entry, order processing, customer support,

supply chain management, internal communication, and employee administration, etc. In other words, computer networks become a backbone to keep business running. Because of this importance, the reliability and availability of such systems and networks have to be concerned and they are really critical factors of system level management [Juan et al. 2007].

The advance of World Wide Web technologies plus the improvement of network security and network speed, etc make Internet more dynamically interactive with human. It is not just a tool to display and exchange information, it can be used as calculating means for complex problems which can't be solved by a single desktop or laptop. Grid Computing is a clear justification for the power of network computing which was started in the mid 1990s. Grid Computing takes advantage from the existing infrastructure with limited resources of each academic institution and uses computer networks to combine all these institutions together to create a Computing Grid analogous to an electric power grid. This architecture of Grid Computing can be used to solve the biggest and the most complicated computations which may be impossible to be solved or it will takes years and years to finish by a single computer or by a single institution. In other words, a complicated computation can be divided into simpler computations which can be done parallelly by different computers on Grid. Obviously, the technical advances in Computer and Network technologies lead to a new trend of software development in the telecommunication network management industry. The rapid growth in the field of embedded computing as mobile devices creates substantial opportunities for network computing. Over 90% of all processors are sold for embedded use. Mobile devices with limits of computing power, memory capacity or battery capacity are unable to do complex computing. However mobile devices can be connected to networks, then it can be connected to computing utilities through computer networks [Logethran et al. 1998]. Running software through computer networks is a typical software application. Users can use any software via a graphical user interface without installing. This approach has several benefits:

- *Universal user interface on every system:* only one graphical user interface can be accessed to software via Web browsers for all systems.
- *Portability:* all computations are done on the centralized facilities which are often multi-core servers therefore multiple users can remotely access to software at the same time from anywhere by Web browsers.
- *Intellectual protection:* software is a form of intellectual property need to be protected. Copyright violation becomes one of the biggest issues of software companies in recent years. This situation creates a hurdle for development of many software companies. Network computing is one of the choices to solve this problem by allowing users to use software but not own any software version. Therefore, it limits Copyright violation.
- *Legacy software:* old software which can't be compatible with a new platform can be reused by running it in a dedicated environment on the server while its

new graphical user interface is used as a tool to interact with users on new systems for which the particular application is not available.

- *Scalability*: Computing Networks can be deployed and scaled very quickly which is one of the key factor for success of business and development of technology.
- *Computing power*: software applications are located on a server which is a model of super computers, therefore it will improve computing speed. Once a multi-core desktop with a hundred of processors has not become realistic yet, network computing is a good choice to save resources.
- *Compatibility*: software can interact with any platform. It is independent of the operation systems, users do not have to set up or configure software unless it is implemented on the server in the form of stored user profile.

For this particular application in this paper, one of the big issues need to be addressed is how users can control the simulation process and how multiple users can use it at the same time without overloading the system.

2 Overview of Network Programming Technologies

HTML alone does not provide the functionality needed for a dynamic, interactive environment. Additional technologies are used to implement the dynamic behavior both on the client side (inside a Web browser) and on the server side to render a Web page dynamically. Most common network programming tools used for developing dynamic websites on the client side are JavaScript, Ajax extension to JavaScript, and Java or ActiveX applets. Most common tools on the server side are PHP, Sun Microsystems' Java Server Pages, Java Servlets, Microsoft Active Server Pages (ASP) technology, and Common Gateway Interface (CGI) scripts using scripting languages such as PERL, server-side JavaScript, ActiveX and Python, or precompiled binary programs [Malinowski and Wilamowski 2001].

The internet bandwidth is significantly improved with time and already adequate for network computing. However, in order to make applications more robust, the data flow should be designed effectively. The key issue is to solve problems associated with a new way of software development so that software application will be possible through the Internet and Intranet. Therefore task partitioning is very important. Which parts of software should be done on the client machine or the server machine. To get this goal needs to satisfy some requirements as:

- Minimization of the amount of data exchange through computer networks to reduce network traffic
- Task partitioning between the server and the client needs to be effective
- Selection of suitable programming languages used for various tasks
- User interfaces have to be friendly

- Use of multiple servers distributed around the world and job sharing among them to avoid overload for one server
- Security and account have to be safe
- Portability of software used on the servers and the clients
- Other

The next section will discuss two examples using computer networks as an interface to enhance interaction between the client and the server.

3 Examples of Computations through Computer Networks

3.1 Neural Network Trainer

The artificial neural network (ANN) applications are gradually increasing in last couple years. The ANNs are widely applied in the fields of VLSI [Cameron and Murray 2008] [Indiveri et al. 2006], image processing, control systems, prediction, etc. ANNs showed their potential power for a lot of real applications but it is so frustrating to train ANNs successfully. The challenges of this success are how to design a good architecture and a suitable algorithm to train ANNs. Because of this purpose many training algorithms are introduced for ANNs in order to attain faster speed as well as increase success rate. Even though Error Back Propagation (EBP) is considered as the most popular training algorithm of ANNs [Ruhmelhart et al. 1986], it is not an efficient training algorithm because of its slow convergence and inability to handle complicated problems. Many advanced algorithms have been developed lately as gradient descent, conjugate gradient descent, Levenberg Marquardt, Neuron by Neuron (NBN), etc and gave better results. For example, NBN can train ANNs 100 times faster than EBP. With big networks this NBN algorithm has its limitation and its advantages diminish because NBN algorithm requires more computations in each iteration. For all these algorithms storage and computational requirements are different, some are good for this application but not good for the others. It means that it is difficult to find a particular training algorithm that can be best for all applications under all conditions. This paper does not attempt to analyze differences as well as advantages or disadvantages of algorithms. Instead it will introduce a new ANNs training tool which includes in both first order and second order methods and also handles arbitrarily connected neural networks that are not found in the existing trainer as MATLAB Neural Network Toolbox or Stuttgart Neural Network Simulators (SNNS) [WWW 2002].

Neural network trainer NBN 2.0 is developed based on Visual Studio 6.0 using C++ language hosting on the server and communicating with the clients through PHP scripts [Hao and Wilamowski 2009]. Its main interface is shown in Fig. 1.

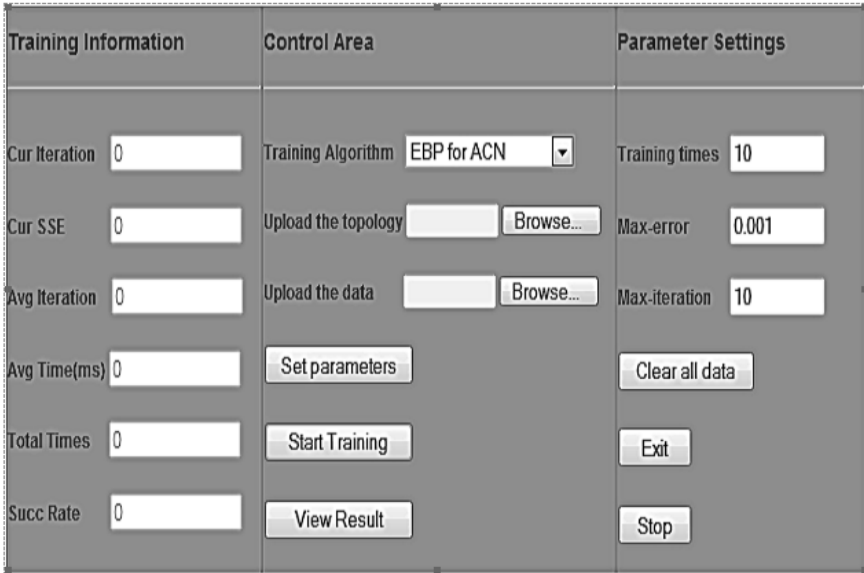


Fig. 1 Neural network trainer interface

NBN 2.0 is developed with four different types of training algorithms:

- Error Back Propagation for arbitrarily connected neuron (EBP for ACN)
- Levenberg Marquardt for multilayer perceptron (LM for MLP)
- Neuron By Neuron (NBN)
- Neuron By Neuron, forward-only (NBN- forward only)

To use this tool users have to upload two files: one topology file and one data file through the neural network interface Fig. 1. As mentioned earlier, this trainer can handle arbitrarily connected networks and it uses the similar solution as Netlist in the SPICE program. These two files have to follow certain syntax so that the training tool can make sense in the correct way.

- Topology file

The topology files are named “*.in”. They are mainly used to construct neural network topologies for training. The topology files consist of four parts: topology design, weight initialization (optional), neuron type instruction and training data specification. The topology design is aimed to create ANN structures. Each line in the topology file is “n [b] [neuron type] [a1 a2 ... an]”, which means the input neurons indexed with a1, a2,..., an are connected to the output neuron b with a specified neural type (bipolar, unipolar or linear). Fig. 2 presents the topology file for the neural network parity-3.

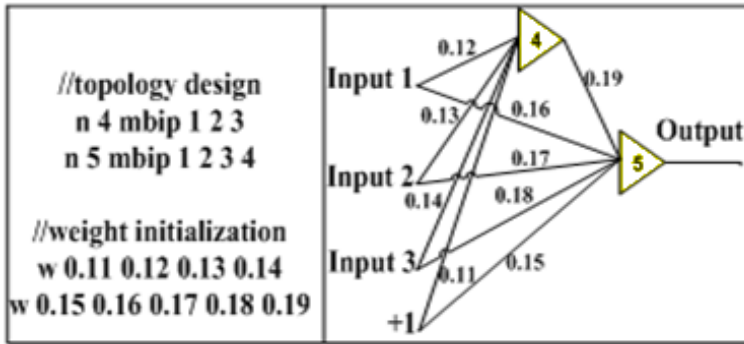


Fig. 2 Weight initialization for parity-3 problem with 2 neurons in FCN network

- Training pattern file

The training pattern files consist of input patterns and related desired outputs. In a training pattern file, the number of rows is equal to the number of patterns, while the number of columns is equal to sum of the number of inputs and outputs for each pattern. However, only with the numerical data in a training pattern file, one can't tell what number of inputs and outputs, so the neural topology should be considered together in order to decide those two parameters (Fig. 3). The training pattern files are specified in the topology files as mentioned above, and they should have the same route as the related topology files.

<i>training data</i>	<i>topology</i>	<i>explanation</i>
-1 -1 -1 -1	//2 inputs and 2 outputs	The first command line of topology shows that there are 2 inputs, since there are 4 columns in training data, so the number of output is 2.
-1 -1 1 1	n3 mbip 12	
-1 1 -1 1	n4 mbip 12	
-1 1 1 -1		
1 -1 -1 1	//3 inputs and 1 output	The first command line of topology shows that there are 3 inputs, since there are 4 columns in training data, so the number of output is 1.
1 -1 1 -1	n4 mbip 12 3	
1 1 -1 -1	n5 mbip 12 3 4	
1 1 1 1		

Fig. 3 Get the number of inputs and the number of outputs from the data file and topology

Besides these two files, there are still couple parameters need to be input from users as *training times* which defines how many times ANNs need to be trained, *max-iteration* which is how many times the same process need to be repeated to update ANNs weights for one training time and *max-error* is an acceptable error limit which is supposed that ANNs will perform well compared with desired outputs. Along with these parameters, there are some other parameters which is

defined as tuning parameters for each algorithm such as “*combination coefficient*”, “*scale constant*”, “*momentum constant*”, “*learning constant*”, “*alpha constant*”, “*beta constant*”, “*gamma constant*”. In order to train ANNs successfully or speed up training process, users are required to tune these parameters to make sure its training process converge and get higher success rate. To reduce data exchange between the client and the server, the neural network interface will check all these parameters on the client side.

During training process the output results will be updated. After training, one result file will be generated which contains all detail information about training algorithm, training pattern file, topology, parameters, initial weights, resultant weights. These results will be saved automatically in database system. Training ANNs usually takes long time, with some big networks it can take thousands of training times or thousands of iterations, therefore this tool is designed in such a way that allows users to stay offline while the training process is running. This approach is very effective when users try to train ANNs by their mobile devices with limited battery capacity.

Another issue of this tool is how to create a friendly graphical user interface with control functions as stop/continue simulation when using software over computer networks. With installed software, these functions are strongly supported by operating systems so it is not a big issue. Using software over Internet is different and is only supported by interactions between the client and server. Because of this reason, software is designed in a different way. Software should have extra control function which has ability to receive requests from the client. JavaScript has certain limitations due to the security model of its implementation by a Web browser. One of those limitations is the inability to retrieve data on demand dynamically from the server. Ajax technology is a solution that allows a JavaScript program embedded inside a Web page to retrieve additional documents or Web pages from the server, store them as local variables, and parse them in order to retrieve data and use it for dynamic alteration of the Web page where the JavaScript is embedded. In this trainer, two buttons “*Stop*” “*Exit*” are used to stop training. The way it works is when a client sends a message to the trainer through the user interface, the trainer will recognize this authoritative message, stop training and send results up to that point back to a client.

In future when Cloud Computing becomes 4th paradigm, software applications can be developed directly on Cloud Operating system as Windows Azure. The approach to design software that is described here may be unnecessary. The Cloud operating system will support all these control functions as any operating systems that are being used.

Simulation Result

(Updated in every second)

Training Information	Control Area	Parameter Settings
Cur Iteration <input type="text" value="7"/>	Training Algorithm <input type="text" value="NBN"/>	Training times <input type="text" value="10"/>
Cur SSE <input type="text" value="0.0018"/>	Upload the topology <input type="text" value="parity3"/>	Max-error <input type="text" value="0.001"/>
Avg Iteration <input type="text" value="6.3333"/>	Upload the data <input type="text" value="parity3"/>	Max-iteration <input type="text" value="10"/>
Avg Time(ms) <input type="text" value="3.4444"/>	<input type="button" value="Set parameters"/>	<input type="button" value="Clear all data"/>
Total Times <input type="text" value="0.0000"/>	<input type="button" value="Start Training"/>	<input type="button" value="Exit"/>
Succ Rate <input type="text" value="1.0000"/>	<input type="button" value="View Result"/>	<input type="button" value="Stop"/>

DONE

Fig. 4 Training result interface

Simulation Curve

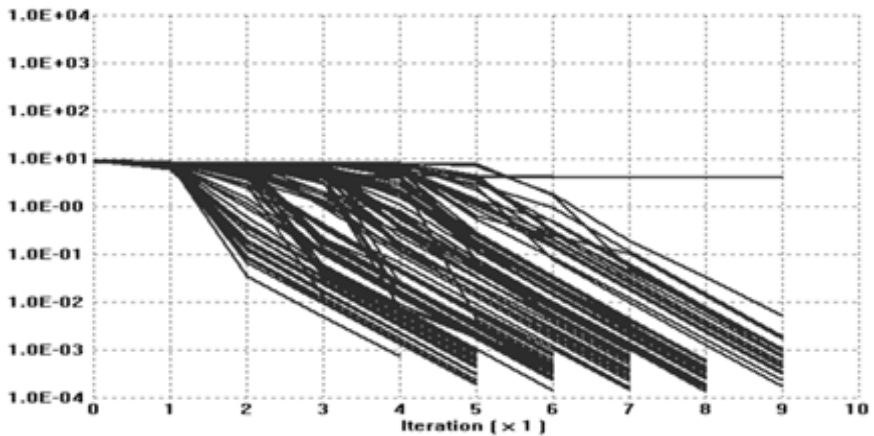


Fig. 5 Output figure

```

Parameters
NBN mu = 0.01000000 scale = 10.00000000
Data File: parity3.in
Topology
3 1 2
4 1 2 3
5 1 2 3 4
Neurons
biplor gain=1.00, der=0.01
biplor gain=1.00, der=0.01
linear gain=1.00, der=0.05
Initial Weights
-0.28000000 0.58000000 0.16000000
0.92000000 0.46000000 -0.08000000 0.28000000
-0.82000000 -0.46000000 0.94000000 -0.62000000 -0.34000000
Results Weights
-0.21352129 0.08990222 0.26889254
-8.09987819 3.40040394 11.61576715 -85.00550976
3.57991511 -1.46565558 -8.29415476 58.78724502 5.95297590
Training Results
Total iteration: 300 Total error: 0.42256835 Training Time: 766

```

Fig. 6 Training result file

3.2 SIP Program

The Spice program implemented through computer network is another example of network computing. The Spice program is the popular software which is widely used to simulate the Integrated Circuits in electronic courses. This is the licensed software which is only available for some machines on campus. It means that students who don't live on campus have to depend on these available machines. Students can have the free version of Spice program but this version can only simulate circuits with limited number of transistors which is often not good enough to run simulation of the Integrated Circuits in electronic courses. In order to make it possible for students to learn electronic courses, the SIP program was developed for this purpose (Fig. 7). The SIP uses Spice3f5 from Berkeley and can simulate the Integrated Circuits with unlimited number of transistors. Users can upload and edit circuit files which have similar formats as capture of the Spice program from MicroSim. After simulation, the SIP program will display results and analysis of simulated data in the form of images or texts [Wilamowski et al. 1998] [Wilamowski et al. 2000]. A unique feature of the SIP versus other Spice simulators is that it is operating system independent. Anyone can access and run a simulation and view results graphically from anywhere with a Web browser via Internet or Intranet.

To start SIP program, users have to select “**View/Edit**” button to enter a circuit file and then save this file by pressing “**SAVE CHANGES**” button. For example after simulating NPN-PNP amplifier circuit Fig.8 the output window will pop up as Fig.9. SIP program can analyze the Integrated Circuits in three different modes: Transient analysis, DC analysis and AC analysis and has some options to display the node analysis as in Spice version of MicroSim.

With the recent technological advances in network speed and network programming, any computation can be done on the server side or the client side only. In order to develop an effective network application, some issues need to be stressed.

- Reduce data exchange between the client side and the server side by deciding which tasks should be done on the client side or the server side.
- Save bandwidth of computer networks to avoid overloading. Because server is still limited by the number of login users at the same time and scalability is still a solving issue, so all computations need to be robust and effective.
- Select the right technology to maintain the ownership of intellectual property when Copyright violation becomes more concerned.

Authorized User: **Anonymous User**
 Accesses: 53
 Last Login: 15:36:04 03/08/2010
 User Level: 1

SPICE Engine	Stored CIR Files	Stored OUT Files
SPICE 3F5 ▾	choose: ▾	choose: ▾
Remote CIR Filename: <input type="text" value="a_dc4.cir"/>	<input type="button" value="View/Edit"/>	<input type="button" value="DELETE"/>
Remote OUT Filename: <input type="text" value="a_dc4.out"/>	<input type="button" value="View/Edit"/>	<input type="button" value="DELETE"/>

Note that everybody can modify and store here files with any name and those could be incorrect. Therefore several exemplary files are write protected. These examples are: A_BP.CIR, A_DC4.CIR, A_DC5.CIR, A_DC6.CIR, A_DIODE.CIR, A_MEM3.CIR, A_SOL3.CIR, and A_TR2.CIR. If you want to modify them you may copy/paste them to a window which is open using a different name.

<input type="button" value="RUN Simulation"/>	<i>simulate CIR file and store output to OUT file [use SPICE .print option]</i>
<input type="button" value="PLOT Data"/>	<i>generate plot using data in OUT file</i>
<input type="button" value="Help"/>	<i>help with the SIP program</i>
<input type="button" value="EXIT"/>	<i>close other open windows</i>

Fig. 7 SIP main interface

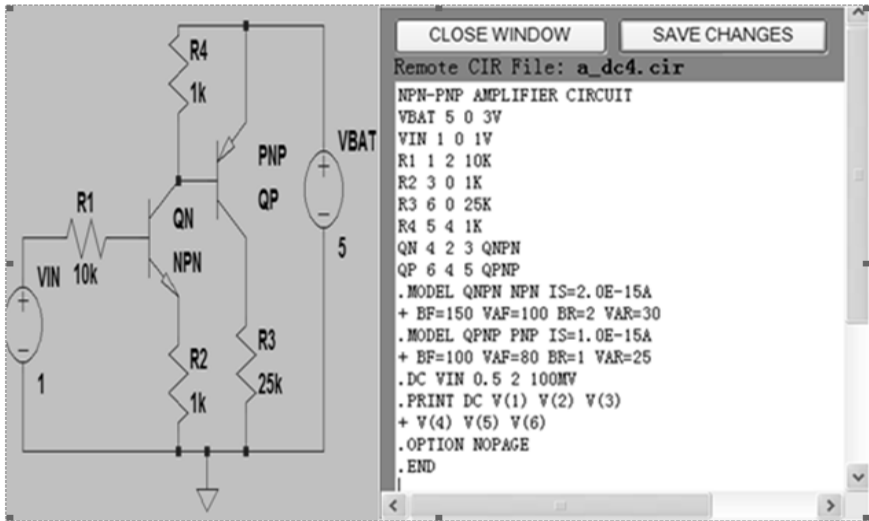


Fig. 8 Schematic and circuit file

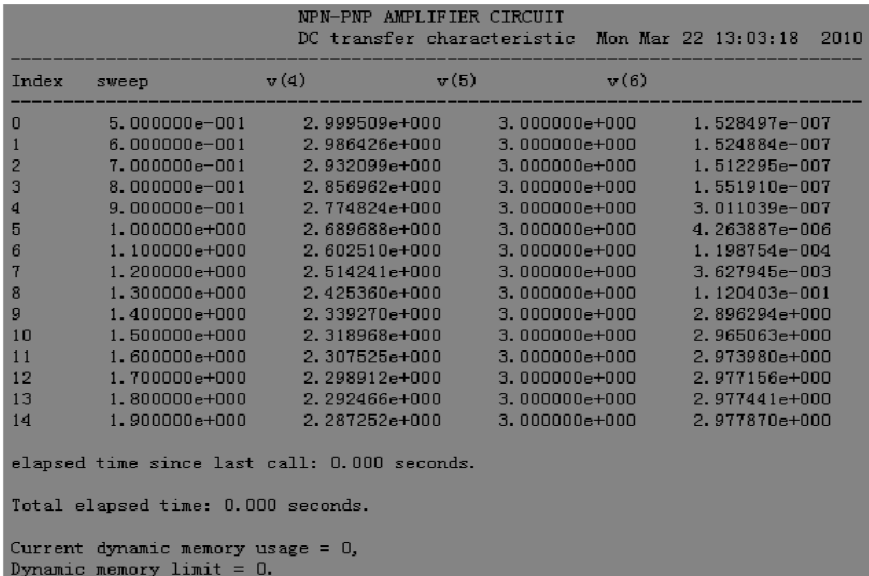


Fig. 9 Simulation output

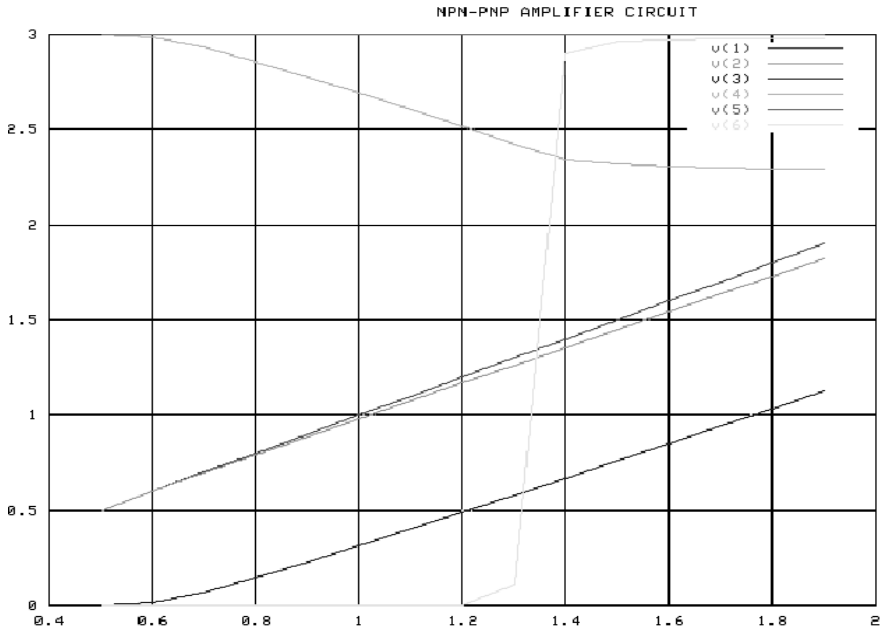


Fig. 10 Output analysis

SIP is a good example to optimize network traffic by task partitioning between the client side and the server side. In case of ANNs trainer users don't have to inspect or analyze data with repeated times, so it is better to generate a graphical image, a result file and send to users. With SIP, users frequently inspect and analyze data many times as changing variables to display voltage nodes or scaling voltage ranges to display, etc Fig.10. In this case there are many requests for different plots of the same data, therefore it could be better to send the data once together with a custom Java applet which could display the same information in many different forms without further communicating with the server.

4 Discussion and Conclusions

This paper shows two examples of how software applications can be implemented through computer networks when network technologies and multi-core processor technologies are advancing. This paper also stresses some reasonable benefits to deploy software applications on computer networks as well as some issues need to be considered to design applications. There are other existing models of computing that are described briefly throughout this paper as Grid Computing, Cloud Computing, etc. They can become the computing models of future because of its computing power, its economic effectiveness and its scalability, etc. Grid Computing models are applied widely in U.S.A Universities to do research about nuclear,

atom and other physics problems. Cloud Computing begins its first steps to deploy Web applications and others.

Two applications described in this paper are very effective in learning and teaching electronic courses and neural network courses. They are not only the useful tools to help students, teachers or scholars to do some simulations but also help them get closer and familiar with technologies in the real world. It means that computer networks can be used as an effective means to boost technology. With these characteristics computer networks shows their power in many real applications in last century and maybe in next century as well. They have been widely deployed in many systems as database management, econ Website, controlling, etc especially in computation.

The NBN 2.0 is available at: <http://131.204.128.91/NNTrainer/index.php>

The SIP is available at: <http://gdansk.bradley.edu/sip/>

References

- [Arano et al. 1996] Arano, T., Aoyama, M.: Emerging technologies for network software development: past, present, future. In: Computer Software and Applications Conf., p. 428 (1996)
- [Cameron and Murray 2008] Cameron, K., Murray, A.: Minimizing the Effect of Process Mismatch in a Neuromorphic System Using Spike-Timing-Dependent Adaptation. *IEEE Trans. on Neural Networks* 19(5), 899–913 (2008)
- [Hao and Wilamowski 2009] Yu, H., Wilamowski, B.M.: Efficient and Reliable Training of Neural Networks. In: Proc of IEEE Human System Int. Conf., Catania, Italy, pp. 109–115 (2009)
- [Indiveri et al. 2006] Indiveri, G., Chicca, E., Douglas, R.: A VLSI array of low-power spiking neurons and bistable synapses with spike-timing dependent plasticity. *IEEE Trans. on Neural Networks* 17(1), 211–221 (2006)
- [Juan et al. 2007] Juan, A.A., Faulin, J., Marques, J.M., Sorroche, M.: -SAEDES: A java-based simulation software to improve reliability and availability of computer systems and networks. In: Simulation Conference, pp. 2285–2292 (2007)
- [Logethran et al. 1998] Logethran, A., Pratiwadi, R., Logenthiran, D., Porebski, A., Thomas, D.W.: Software migration of telecommunication network management systems to the Web using CORBA and Java System Sciences. In: Proc. of the Thirty-First Hawaii Int Conf., Kohala Coast, HI, pp. 637–644 (1998)
- [Malinowski and Wilamowski 2000] Malinowski, A., Wilamowski, B.M.: Web-based C++ compilers. In: ASEE, Annual Conference, St. Louis, MO, CD-ROM session 2532 (2000)
- [Malinowski and Wilamowski 2001] Malinowski, A., Wilamowski, B.M.: Internet technology as a tool for solving engineering problems. In: The 27th Annual Conf. of the IEEE Industrial Electronics Society (tutorial), Denver CO, pp. 1622–1630 (2001)
- [Manic et al. 2002] Manic, M., Wilamowski, B.M., Malinowski, A.: Internet based neural network online simulation tool. In: Proc. of the 28th Annual Conf. of the IEEE Industrial Electronics Society, Sevilla, Spain, pp. 2870–2874 (2002)
- [Ruhmelhart et al. 1986] Ruhmelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by back-propagation errors. *Nature* 323, 533–536 (1986)

- [Wilamowski and Malinowski 2001] Wilamowski, B.M., Malinowski, A.: Paper collection and evaluation through the internet. In: Proc. of the 27th Annual Conf. of the IEEE Industrial Electronics Society, Denver CO, pp. 1868–1873 (2001)
- [Wilamowski et al. 1998] Wilamowski, B.M., Regnier, J., Malinowski, A.: SIP- spice intranet package. In: IEEE Int. Conf. on Industrial Electronics, 192–195 (1998)
- [Wilamowski et al. 2000] Wilamowski, B.M., Malinowski, A., Regnier, J.: SPICE based circuit analysis using web pages. In: ASEE 2000 Annual Conf., St. Louis, MO (CD-ROM session 2520, 2000)
- [WWW 2002] Stuttgart Neural Network Simulator (May 2002), <http://www-ra.informatik.uni-tuebingen.de/SNNS/> (accessed April 10, 2010)

The Method of Communication Quality Improvement in Distributed Systems with Real-Time Services

M. Hajder and T. Bartczak

Department of Distributed Systems, University of Information Technology and Management, Rzeszów, Poland
{mirosław.hajder,tbartczak10}@gmail.com

Abstract. The paper presents a new method of improving the human-machine communication in distributed systems and computer networks, in which a substantial portion of traffic is a multimedia traffic. The method bases on priority processing of short and multimedia services packets. The choice of service time as a parameter evaluating the effectiveness of proposed solutions and an appropriate mathematical model is justified. Simulation studies are used to determine the intensity range of information flow, for which the method is appropriate.

1 Introduction

Since effective methods of transmission have influenced human-machine communication, they can also be done through computer networks. Globalization growing of communication, as well as the dynamic development of data transmission methods mean that providing remote communication between the system and the user is no longer a significant problem.

Regardless of these trends integration of computer and telecommunications networks is progressing [Hajder et al. 2002]. Together with the gradual takeover of services previously reserved for the telecommunication by the computer networks, traffic structure has undergone significant changes. Native support of a network for data packet communication, is used today, also for the implementation of voice, imaging, modeling, distributed, *on-line* control, gaming services and other applications implemented in real time. The IP protocol designed to transfer datagrams, for a long time has been considered as inefficient for handling real time traffic. This stemmed from the fact that each packet of data stream was routed independently, and parameters such as bandwidth, delay and jitter changes

in very wide range during load reevaluation. For this reason, in packet networks, real-time services were more sensitive to congestion than are traditional batch transmission services.

The solution of sensitivity problem was achieved by the application of QoS (*Quality of Service*) [Clark 1998]. In IP networks, primary task of QoS is to minimize the impact of congestion on communication. To perform this task variety of mechanisms and communication protocols such as *IntServ*, *DiffServ*, *RSVP* are developed. The multi-protocol label switching MPLS (*MultiProtocol Label Switching*) which helps to ensure the specified quality of service traffic becomes widely popular.

There are many different methods improving the efficiency of communication (protocols and technologies). One of them is the compression of information, especially the media information, which takes into account the type of channel and type of processed information. Another method minimizes the likelihood of congestion of channels and nodes. In the case of contestations occurrence it minimizes their impact. This method controls the transmission of information by the use of traffic management protocols, and now is very widely implemented.

Effective way of resolving problems caused by high dynamic of traffic is flexible reconfiguration of connections, preferably in the logical level, without modifying the hardware architecture. These connections are dynamically adjusted to the current traffic pattern, thanks of that in a design process incremental [Dutta et al. 2009; Dutta et al. 2004] methods become preferable over static one. Improvement of communication quality can be seen also in the methods based on selective increasing density of information in the frames, packages and messages, the dynamization of priorities, using adaptive algorithms of transmission control, multicast communication methods, and asynchronous transmission modes. The results of empirical modeling and experiments have shown the effectiveness of selected methods based on mentioned concepts [Bartczak et al. 2009].

The method based on dynamic re-prioritization of short frames that contains multimedia information is presented in this work. In further considerations it is assumed that the analyzed communication system is designed to support interactive human-machine interface. Because for the end-users the most important criterion of system evaluation is processing time of the request. The quality of solutions will be estimated on the basis of communication delays.

Due to the heavy use of interactivity in contemporary information systems and their ubiquity in the transmission of multimedia information, the research topic is **important and actual**.

2 The Initial Assumptions and the Research Task

2.1 Environment and Research Tools

How to improve the efficiency of transmission quality depends, above all, from the characteristics of traffic. In particular, exhibit a significant effect: range of protocols used, the distribution of packet length and intensity of occurrence of service requests, the statistical parameters of traffic, customer sensitivity to delays and transmission errors, etc. In view of the considerable changes in the characteristics of communication that have occurred in recent times, known to the authors motion analysis may not be the basis for deciding the future direction of research. Therefore, it became necessary to make your own experiments, which would allow the most important characteristics for a network with extensive use of multimedia services.

For the analysis of network traffic are widely used tools such as *Cisco Works*, *Sniff*, *Mihov IP King*, *Wireshark*, etc. .. However, these tools do not have the features to completing the analysis of motion necessary to focus further research. Therefore, the Department of Distributed Systems of University of Information Technology and Management in Rzeszow developed and implemented Universal Package for Use in Web Traffic Analysis (UPARS). This application, unlike the past, in addition to collecting information about traffic, makes advanced analysis, including analysis of nonlinear time series corresponding to the movement, the impact of various factors on the statistical distribution of the jet, spectral analysis, et al. UPARS was used for detailed studies of traffic on both the starting gate of the University, as well as between the selected organizational units.

Generalized block diagram UPARS was presented in Fig. 1. Each of the packages appear on the measuring device is transmitted to the interface UPARS, and then, after processing is stored in the buffer prior to intercept. At the same time issued a receipt flag packet and received packet is stored in the module collection and if it is an IP packet it is processed in the block analysis. In the course of treatment is excreted in the header, the data are removed, and the processing result is stored in the database. Next, after completion of the process of collecting information on traffic using the internal calculations are performed unit of analysis that will eventually be recorded in the database of results.

The purpose of the tests carried out using the system was to participate UPARS multimedia traffic in the generated traffic, its structure and selected characteristics. Obtaining information on the current structure allowed the motion to direct further research on methods to ensure you get the best performance in a particular network. Note that the distribution of characteristics such as packet length or intensity of their occurrence is strongly dependent on the type of services used, as well

as time of day, day of week or year, the number and range of people using the network. In order to obtain reliable results, recording and analysis of traffic subject to an annual cycle.

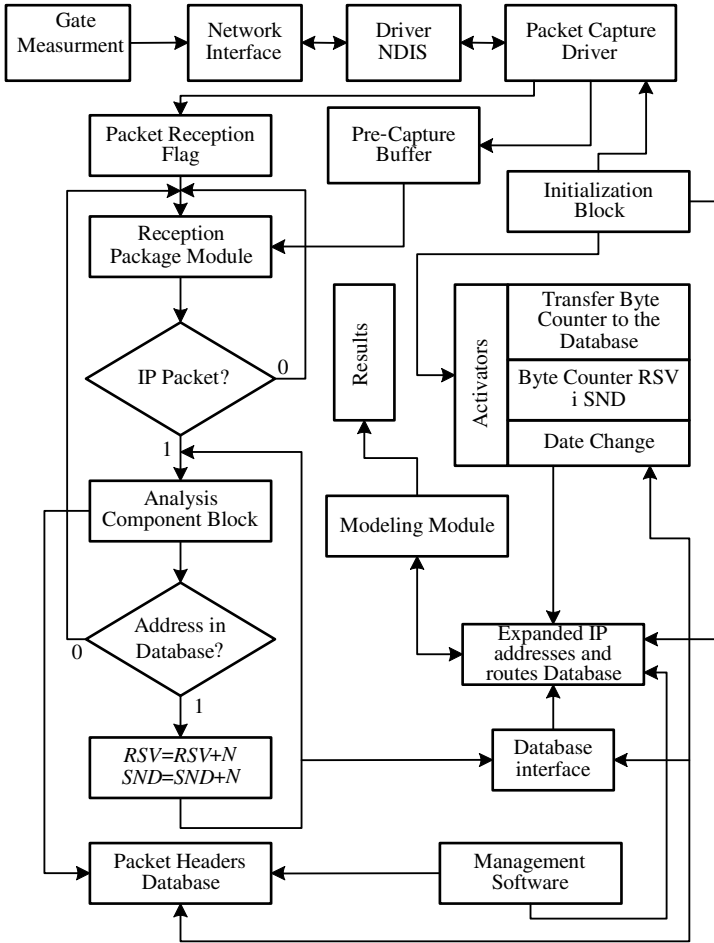


Fig. 1 Functional diagram of the system analysis of network traffic UPARS

Studies have shown that regardless of any characteristics of the functioning of the analyzed network dominated by TCP/IP traffic, and the other types have a slight nature. Although the fluctuations observed minimum degradation of protocols associated with the month, day of week or time of day, in no way alter the above conclusion. On the other hand, studies related to the characteristics of the motion itself shows a clear correlation with time of day, day of week, etc.

In order to determine trends and simplify the subsequent analysis of traffic, in particular its spectral analysis, were measured traffic aggregation. Fig. 2 presents the changes in time for the different aggregation windows. As shown in figure extreme motion values are averaged, but its trends remain unchanged.

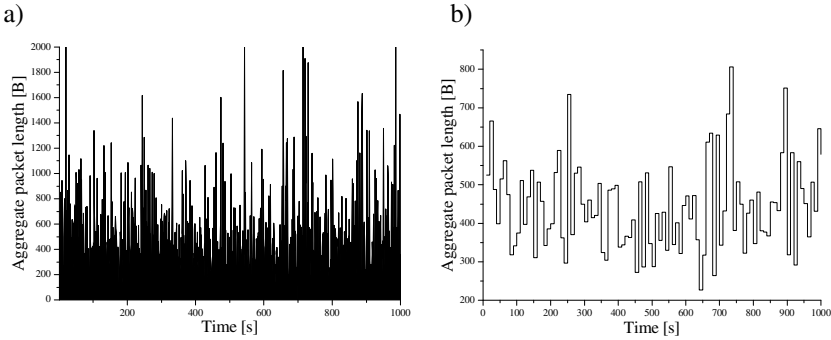


Fig. 2 The results of measurements of traffic on the network interface output, aggregates with different aggregation windows: a) 1 second, b) 10 seconds

Fig. 3 shows the length distribution of packets for real - an important aggregate. In previous studies, been noted that for the minimum (1ms) windows network traffic aggregation there is no clear structure, and this is revealed only at the aggregate level 100ms. The results obtained show that during the gradual change of the window aggregation to less than 1 second observing small changes in the nature of the motion, and after crossing the length of the window, the distribution of stable form which does not change even at further increasing the window of aggregation.

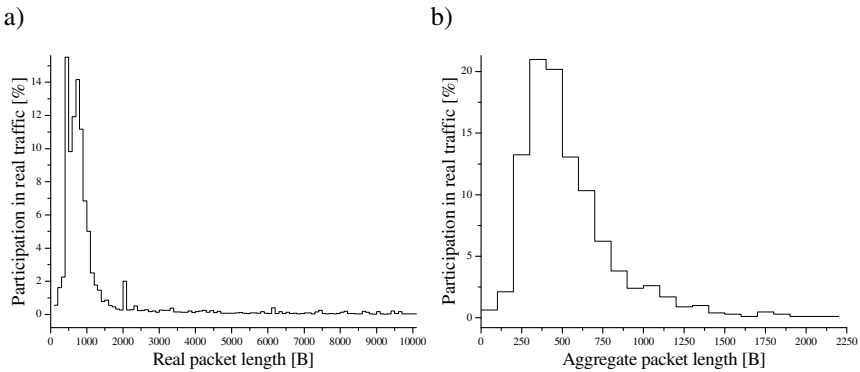


Fig. 3 The distribution of packet length on the output network interface: a) real movement, b) aggregation of traffic from a window of 10 seconds

Approximating curves correspond to the packet length distribution of the log-normal decay function described:

$$f(x) = \left(1/x\sigma_{\ln x}\sqrt{2\pi}\right) \exp\left(-(\ln x - m_{\ln x})^2 / 2\sigma_{\ln x}^2\right), \quad (1)$$

where: \ln – natural logarithm; x – variable; $\sigma_{\ln x}$, $m_{\ln x}$ – distribution parameters, respectively, the expected value and standard deviation. Studies have shown that the movement and its characteristics are dependent static from the time when measurements were made. For example, if you analyzed the daily traffic will be limited, its distribution with very high accuracy corresponds to decomposition (1).

The study results showed that in cases where man-machine interface in a multimedia, network traffic is dominated by a length close to 500 bytes. Furthermore, the dominant movement is particularly sensitive to time delays. These findings became the basis for further research.

2.2 The Choice of Transmission Quality Improvement Method

In this work, interactive client-server systems were reviewed, where communication between nodes is a multimedia and broadcast the same information - Package. ITU-T standards such as H.320 and H.323 and IETF SIP describe in detail the implementation of such services, among others. in an IP network. One of the main indicators of the quality of communication is the level of user acceptance. His determination to apply subjective methods, based on human perception (e.g., MOS - Mean Opinion Score) and objective method based on analysis of real factors (e.g. SNR – Signal to Noise Ratio, MSE – Mean Square Error) [Haverkort 1999].

The quality of communication affects both tackling the application service VVoIP (*Voice and Video over IP*), and the communication network as well. The most important characteristics affecting the perception of communication include: channels of communication bandwidth, packet transmission delay and its variation, the level of transmission errors and distributions: the size of the package and the spacing between them. Assessing the quality of communication was devoted to a series of works. [Haverkort 1999; Calyam et al. 2004; Fische 2004; Hajder and Kielbus 2007] which shows that key factors affecting the acceptance are the transmission delay and the average speed of the stream. The work will focus on the first parameter. Note that the precise packet networks insurmountable delays is virtually impossible. In the event of an overload or damage to both the node and the transmission channel, the packets are routed, which usually involves changing the values of communication delays. These delays can be minimized only through effective management of traffic and maximize network reliability.

Most of the technology supports a variety of lengths packets buffering and thus their dispersion causes communication delays. This dispersion can be mitigated by the priority and defragmenting the longest packet. Another source of delay is just filling a package of information broadcast. In order to make effective use of communication bandwidth, they all must include a situation when an empty package is sent to be regarded as inadmissible. The solution to this problem is to reduce the size of the frame, which will minimize these delays.

In this paper it is proposed to improve the quality of communication between nodes through the use of packet prioritization support multimedia and small size. In order to determine the validity of the use of the method, consider a distributed system of woody switches deployed in the nodes. For the formal description of a switch you can use a mass service system with a multi-dimensional stream of requests [Fische 2004]. Ensuring the correctness of the model requires the input stream with poisson's character have an exponential service request resolution time. The architecture of woody maximum delay occurs between nodes attached to the lowest level of the hierarchy (ie, tree leaves). In homogeneous systems (i.e. those which: for each of the leaves of the tree depth is the same, the parameters are the same nodes of the network, delays in links between the levels are negligibly small) delay of communication between users is equal to the product node and a delay of 1 minus twice the number of levels of the tree.

Assume the parameter that reflects the user acceptance will be the average waiting time $t_{ant,avg}$ to handle the request. For requests without a priority, its value is given by formula:

$$t_{ant,avg} = \sum_{i=1}^m \lambda_i x_i^{(2)} / 2(1 - \rho_1 - \rho_2 - \dots - \rho_m),$$

where: λ_i – intensity of the i input stream; m – number of input streams; $x_i^{(2)}$ – the second initial moment of time demand services i such; $\rho_i = \lambda_i / \mu_i$, μ_i – intensity of the i service request. If we assume the exponential nature of decay times of service requests it: $x_i^{(2)} = 2x_i^2 = 1 / \mu_i^2$. Then $t_{ant,avg}$ is given by formula:

$$t_{ant,avg} = \sum_{i=1}^m \lambda_i x_i^2 / (1 - \rho_1 - \rho_2 - \dots - \rho_m).$$

Consider now a model system with multiple levels of priorities. In general, the node connected to the n input channels through which n_1 requests appear priority and n_2 without it, i.e. $n_1 + n_2 = n$. Analyzed nodes present in the form of multi-service system for the mass of inputs, n of which n_1 supports the claim of priority and n_2 without priority of the twentieth. In this case, you can bring it to a dual-channel system. In order to simplify the process of analysis, a set of input streams with identical priority replace one stream of the sum intensity. The multichannel system with equal priorities, the value $t_{ant,avg}$ is equal to:

$$t_{ant,avg} = \sum_{i=1}^n \lambda_i / (\mu - \sum_{i=1}^n \lambda_i). \text{ On the other hand, for a single channel:}$$

$$t_{ant,avg} = (1 / (\mu - \lambda)) - 1 / \lambda = \lambda / \mu (\mu - \lambda),$$

where: $\lambda = \sum_{i=1}^n \lambda_i$. By setting the two equations right hand sides equal to each other, we obtain: $\sum_{i=1}^n \lambda_i / \mu (\mu - \sum_{i=1}^n \lambda_i) = \lambda / \mu (\mu - \lambda)$. Last equation is

correct, if the condition is met that $\lambda = \sum_{i=1}^n \lambda_i$. Therefore, the value of \bar{t}_{ant} for multidimensional flow demands can replace the importance of \bar{t}_{ant} for the sum of a single stream with intensity. In carrying out similar considerations for the set of priorities, and given that the priority streams are peers, so the system can be regarded as a channel. Requests with identical priorities are set in a common input queue and serviced periodically. In addition, the time readers will ignore requests from different queues, assuming that they all come from one source, the aggregate intensity.

From the standpoint of a node, the network may be presented in a mass service system with two input streams. Then the average waiting time $t_{ant_{avg}}^{max}$ for the maximum priority requests will be equal to: $\lambda_1 + \lambda_2 / \mu(\mu - \lambda_1)$, and the time $t_{ant_{avg}}^{min}$ for the minimum priority demands is: $\lambda_1 + \lambda_2 / \mu(\mu - \lambda_1)(\mu - \lambda_1 - \lambda_2)$. On the other hand, for networks with two streams without priorities, the expectation will be equal: $t_{ant_{avg}} = \lambda_1 + \lambda_2 / \mu(\mu - \lambda_1 - \lambda_2)$.

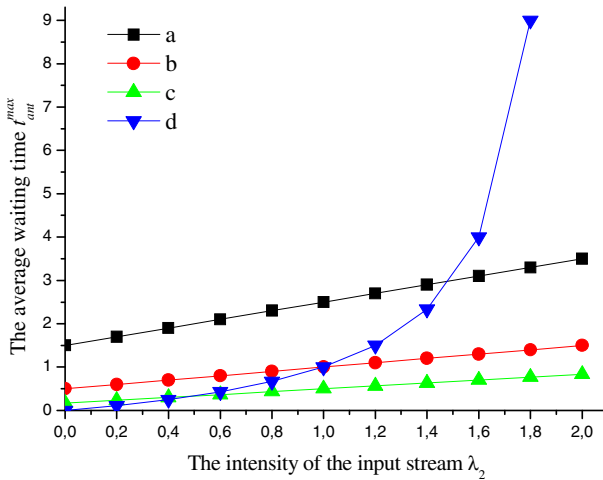


Fig. 4 The dependence of the average expectation of the intensity of the input stream

In order to illustrate the consequences of these considerations, fig 4 shows the dependence of the average waiting time $t_{ant_{avg}}^{max}$ on the intensity of the input stream λ_2 with a maximum value of priority. The various curves refer to the following values of intensity λ_1 : a. 1,5; b. 1,0; c. 0,5; d. $\lambda_1 = \lambda_2$ and $\mu = 2$. Similar studies for the stream with minimal priority are presented in Fig. 5.

The presented results show that the intensity of the stream without a major impact on the priorities of waiting for the channel with the priorities. Moreover, with increasing intensity of λ_1 to increase the flow of time is faster than waiting for its smaller values.

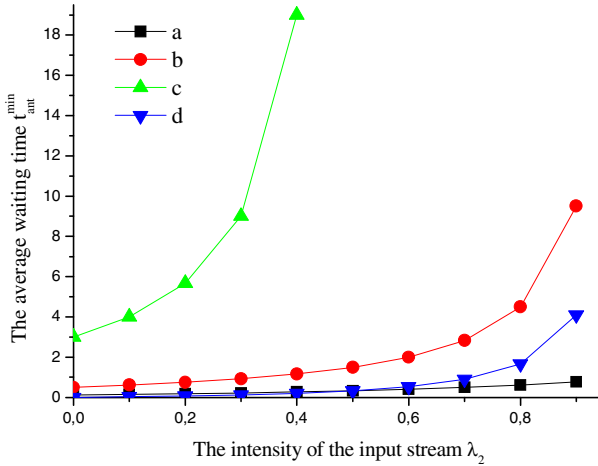


Fig. 5 The average waiting time dependence of the intensity of the input stream

2.3 Purpose and Test Methods

Aim of this study is to improve the quality of real-time interactive services, in particular the development of alternative processing algorithms and data, specific to our services.

3 Method of Dynamic Changes Priorities

3.1 The Idea of the Method

The idea of the proposed method is shown in Fig. 6.

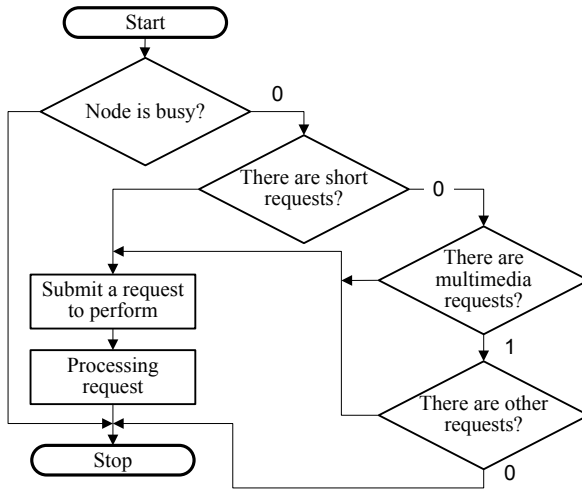


Fig. 6 Illustration of single bar operation node in the proposed method

Emerging demands are divided into classes. Demarcation criterion is the length of the data package. First, short packets are handled, which is likely to be used to manage the system. This method allows separation of many classes of short packets, varying only the length of the information. If none of the packages may not be eligible for any of the short class, it must be made sure the line input node is not a multimedia package of any length. If there are no more short or multimedia packages in queue, a check for the other requests are made. Processing shown in fig. 6 is repeated periodically, throughout the system, while the first short packets are processed, then multimedia, and finally all the others.

3.2 Delays Model

In order to evaluate the effectiveness of the proposed method, we define an analytical relationship between the delay at a given class of service demands and costs of operation and the intensity of streams. Time t_{ant} waiting for the selected service request in the system without priorities is equal to: $t_{ant} = t_0 + t_n$, where: t_0 – time to complete service requests currently being processed; t_n – time service requests contained in the input queue on arrival at issue requests.

Next, instead of instantaneous values, we use their expected values. Thus, the expected value $E(t_n)$ of time t_n XY handle requests the importance of $E(m)/\mu$, where: $E(m)$ – the expected value of the number of requests in the queue entry; μ – intensity of their support. Using the average residence time of the request in

the queue input value of $E(m)$ write as $\lambda E(t_{ant})$, where: λ – intensity of the input stream. Performing the appropriate substitutions we obtain: $E(t_{ant}) = E(t_0)/(1-\rho)$, where: ρ – ratio of traffic.

Consider a system in which service requests have been divided between dynamically allocated priorities q , with values ranging from 1 to n . In order to determine the priority value $q_k(t)$ k request at time t should be used as a function of: $q_k(t) = t_{iu} + C_k$, where: t_{iu} – time of appearance requests analyzed; C_k – cost of service.

Priorities of claims can be determined by linking any of them with the cost of their operation, while this cost increases with increasing the priority of the class claims. In this way, k priority requests at time t is defined as: $q_k(t) = (t - t_{ant})C_k$, but, for the cost of its operation there is the relationship: $C_1 \geq C_2 \geq \dots \geq C_n \geq 0$. If your system uses n different priorities, the demands placed on the input queue are by definition distributed between different priorities. The intensity of information flows for each class of priorities is equal to respectively: $\lambda_1, \lambda_2, \dots, \lambda_n$. Assume that the input stream for each class is poisson's character. For i -class ($i = 1, \dots, n$), the average service time is equal to $1/\mu_i$. Moreover, with increasing levels of i -class, its priority will be decreased.

We define the average waiting time $t_{ant,avg}$ to handle requests from any class, given the relative priorities. Consider the request of the class k ($1 \leq k \leq n$) appear at the time t_0 . Waiting time t_{ant}^k that will elapse from the time of appearance requests, until the start of its operation depends on three factors: **a.** waiting time t_0 at the end of the previous service requests; **b.** time waiting t_i^{\geq} to complete the handling of all client requests a class $i \geq k$, which were already waiting at the entrance to the emergence of the analyzed requests; **c.** time service requests $t_{i_0}^{\geq}$, each of the classes with priority higher than or equal to k , which occurred while waiting t_{ant}^k . These conditions can be written as: $t_{ant}^k = t_0 + \sum_{i=1}^n t_i^{\geq} + \sum_{i=1}^n t_{i_0}^{\geq}$.

In further considerations, instead of the actual value of the time, we will use the expected values (formerly mathematical expectation). Expected value $E(t_{ant,avg}^k)$ the average waiting time will be equal:

$$E(t_{ant,avg}^k) = E(t_0) + \sum_{i=1}^n E(t_i^{\geq}) + \sum_{i=1}^n E(t_{i_0}^{\geq}), \tag{2}$$

where: $E(t_0)$ – expected value of the average service time of the current request; $E(t_i^{\geq})$ – expected value of the average service time requests with a priority no less

k contained in the input queue; $E(t_{i_0}^{\geq})$ – expected value of the average service time requests with a priority no less k occurring during t_{ant}^k .

The scheme for the time $E(t_0)$ does not depend on the used mode of operation and is identical for all classes, assuming that the requests are processed with equal priority in order of appearance. The importance of $E(t_0)$ can be determined on the basis of the expected value of the average waiting time in the system of mass service M/G/1 type. It is equal to: $E(t_0) = \frac{1}{2} \lambda E(\tau_2)$, where: $E(\tau_2)$ – second moment of the distribution of service time, determined by the Pollaczek-Khinchine formula [Taha 2007].

Since the input stream is the sum of several streams with identical distributions, the expected importance of the average service time can be represented as:

$$E(t_0) = \frac{1}{2} \sum_{i=1}^n \lambda_i E(\tau_2^i) = \frac{1}{2} \sum_{i=1}^n \lambda_i \left(\sigma_2^i + \frac{1}{\mu_i} \right),$$

where: λ_i – intensity of the i input stream; σ_2^i – dispersion of the distribution of service time demands of the i class; μ_i – intensity of stream requests i -class (output stream).

The value of $E(t_i^{\geq})$ depends on the average number of requests $E(m_i)$ i class, waiting and served in the system before the request under. Each of them requires an average of $1/\mu_i$ units of time and therefore:

$$E(t_i^{\geq}) = \frac{f_{ik} E(m_i)}{\mu_i}, \quad (3)$$

where: f_{ik} – expected demands of i -class, supported the request under consideration; $E(m_i)$ – the average number of requests i class.

On the basis of the Little law may save:

$$E(t_i^{\geq}) = \frac{f_{ik} \lambda_i E T_{ant}^i}{\mu_i} = \rho_i f_{ik} E(t_{ant}^i), \quad (4)$$

where: ρ_i – intensity of the i stream.

The last element found in formula **Błąd! Nie można odnaleźć źródła odwołania.**, is a consequence of the emergence of an average of $E(m_i)$ requests the i class by the time of the expected value of $E(t_{ant}^k)$. Since, according to

previous findings, the intensity of the emergence of claims is equal to λ_i , and each request requires an average $1/\mu_i$ time units, we get:

$$E(t_{i_0}^{\geq}) = \frac{\lambda_i g_{ik} E(t_{ant})}{\mu_i} = \rho_i g_{ik} E(t_{ant}),$$

where: g_{ik} – expected demands of i class, declared in the time interval t_{ant}^k , and supported the request under consideration.

By definition f_{ik} and g_{ik} show that: $f_{ik} = 1$ for $i \leq k$; $g_{ik} = 0$, if $i \geq k$. Substituting above values in to expression(1), we get:

$$E(t_{ant}^k) = \frac{1}{2} \sum_{i=1}^n \lambda_i E(\tau_2^i) + \sum_{i=1}^n \rho_i f_{ik} E(t_{ant}) + \sum_{i=1}^n \rho_i g_{ik} E(t_{ant}).$$

Solving this equation terms of waiting time demands of the k class, we obtain:

$$E(t_{ant}^k) = \frac{\frac{1}{2} \sum_{i=1}^n \lambda_i E(\tau_2^i) + \sum_{i=1}^k \rho_i E(t_{ant}^i)}{1 - \sum_{i=1}^{k-1} \rho_i g_{ik}} + \frac{\sum_{i=k+1}^n \rho_i f_{ik} E(t_{ant}^i)}{1 - \sum_{i=1}^{k-1} \rho_i g_{ik}}. \tag{5}$$

To solve equation **Błąd! Nie można odnaleźć źródła odwołania.** specify the value of the coefficients of f_{ik} and g_{ik} . Since f_{ik} describes the part of the demands that arise when analyzed can be found in the input queue and are handled before it can use a simple geometric interpretation. Assume that arise at any time request the i class, stays in the queue $w(t_1)$ units of time before the appearance at the time of the request under consideration t_1 . Note that if $w(t_1) > t_1 + t_2$ is a demand of this i class has a lower priority than analyzed. At t_2 priorities of both the demands become equal. Thus, the class analyzed the cost factors determine the slope tan priority functions. In this way: $t_1 + t_2 = t_1 C_k / (C_k - C_i)$.

The value of the expected number of requests i class, served before consideration can be written as:

$$E(n_i) f_{ik} = \int_0^{\infty} \lambda_i P \left[t \leq \omega_i(t) \leq \frac{C_k}{C_k - C_i} t \right] dt,$$

where: $\lambda_i dt$ – number of claims expected value of the i class emerging in the time interval dt ;

$$P \left[t \leq \omega_i(t) \leq \frac{C_k}{C_k - C_i} t \right] - \text{the probability that demand, which appeared this time}$$

frame, the queue will spend no less than t and not more than $t C_k / (C_k - C_i)$ units of time.

Using **Błąd! Nie można odnaleźć źródła odwołania.** to transform **Błąd! Nie można odnaleźć źródła odwołania.** we obtain a new form of the expression denoting the value of interest:

$$E(n_i) f_{ip} = \lambda_i E(t_{ant}^i) - \lambda_i \left(1 - \frac{C_i}{C_k}\right) E(t_{ant}^i).$$

Given that $E(n_i) = \lambda_i E(t_{ant}^i)$, we obtain the sought value $f_{ip} = C_i/C_k$.

Now define the value of g_{ik} . Note that for i class, the following requirements must be met: $E(m_i) g_{ik} = \lambda_i t_1$, where: $E(m_i)$ – the expected value of the number of requests i class entry in the queue. The value of t_1 can be estimated using the relationship: $C_k E(t_{ant}^k) = C_i (E(t_{ant}^k) - t_1)$. Then, you can specify the wanted coefficient using the expression $g_{ik} = 1 - C_k/C_i$. After substituting this value into **Błąd! Nie można odnaleźć źródła odwołania.** we obtain:

$$E(t_{ant}^k) = \frac{\frac{E(t_0)}{(1-\rho)} - \sum_{i=k+1}^n \rho_i E(t_{ant}^i) \left(1 - \frac{C_i}{C_k}\right)}{1 - \sum_{i=1}^{k-1} \rho_i \left(1 - \frac{C_k}{C_i}\right)}, \text{ where: } \rho = \sum_{i=1}^n \rho_i.$$

An interesting characteristic is the length of the queue service system, defining the necessary buffer size for packets arriving at the node. The average length of queue \bar{L} can be written as:

$$\bar{L} = \sum_{k=1}^n \lambda_k E(t_{ant}^k).$$

4 Simulation Test

In the following simulation studies, we analyze the network protocols of the group supported TCP/IP. in which the multimedia information is transmitted. Packages are divided into classes taking into account the length of the information. Study subject information packets to part with a length of between 500 and 1500 byte with 250 byte projection.

At the beginning set the expected value of $t_{ant,avg}$ dependence of the average delay for $n = 2$ on the importance of traffic ρ . Assumed that for each $i = 1, \dots, n$, $\rho_i = \rho$ and $C_1 = 2$, $C_2 = 1$. This dependence is presented in fig. 7.

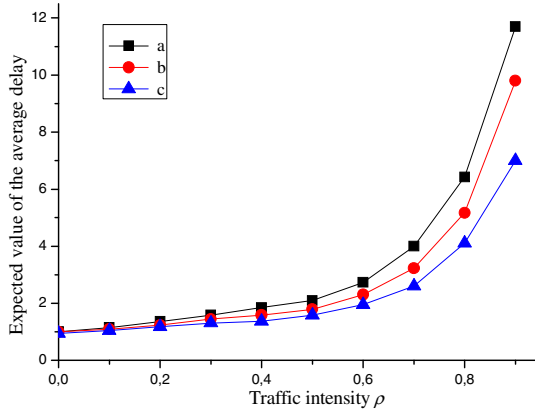


Fig. 7 Dependence of the expected value of the average delay for the amount of traffic: a. $k = 2$; b. Trails without priorities with support for FCFS; c. $k = 1$

The above figure shows that the expected value of the average delay for packets with higher priority is reduced as compared with the standard operation way.

Fig 8. presents the discussed dependence for n above 3 and $C_i/C_{i+1} = 4$.

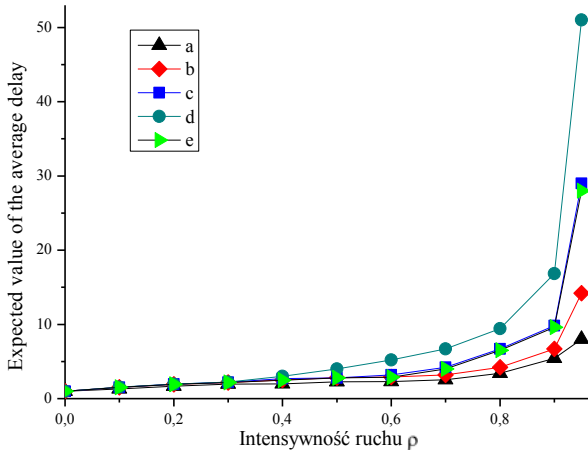


Fig. 8 Dependence of the expected value of the average delay from the intensity of traffic for: a. $k = 1$; b. $k = 2$; c. $k = 3$; d. $k = 4$; e. queue without priorities with support for FCFS

As in the previous case, for some classes of traffic ($k=1$, $k=2$), the proposed solution proves to be better, for others ($k=3$, $k=4$) worse than the standard solution. Similar results were obtained also for a greater number of traffic classes, and the various relationships operating costs.

5 Conclusions and Further Work

Packets in performed computational experiments are divided into 5 groups depending on the its length. In all cases, the proposed method improves the communication parameters. It allows communication system for significant reduction of average processing time of short packets by node. This reduction improves performance and reliability of communication. A significant increase of delay is observed when using processing modes without priorities excising 0.6 of the maximum intensity of information flow. Because the method of dynamic priorities reduces the processing time by node, the qualitative and quantitative characteristics of the transmission are improved.

The drawback of systems with traditional service modes is faster increase of delay for shorter packet than for longer one caused by the increase of stream intensity. Implementation of the proposed method can minimize this differences of processing time. Moreover, increasing of packet size causes increasing of communication system efficiency. For a low intensities (below 0.3) of information stream, the method does not introduce noticeable delay compared to traditional solutions.

Further work will focus on: a. implementation of the proposed method for a group of selected, real communication technologies b. modification of the method involving the use of its only in under loaded systems.

In addition, the methods of adapting communication network to dynamically changing of traffic pattern will be analyzed. Note that a significant part of sensitive media traffic in the transmission is not only one problem of the remote, man and machine communication. Interactive systems are characterized by high dynamics of operating parameters changes, and life-time of optimal or quasioptimal connecting networks still becomes shorter. Network load reflected as the volume of traffic and used the services is not static but dynamically changes. Therefore designing and implementation of a network which satisfies at least the medium-term requirements is difficult. This problem is another topic for further research.

References

- [Bartczak et al. 2009] Bartczak, T., Paszczyński, S., Korniak, J.: Traffic reduction in computer networks by agent technology. In: Proc. of 2nd Int. Conf. on Human System Interaction, Catania, Italy, pp. 730–734 (2009)
- [Calyam et al. 2004] Calyam, P., Sridharan, M., Mandrawa, W., Schopis, P.: Performance measurement and analysis of H. In: Traffic in Passive & Active Measurement Workshop, Antibes Juan-les-Pins (2004)

- [Clark 1998] Clark, M.P.: Networks and telecommunications: design and operation, 2nd edn. John Wiley & Sons, Chichester (1998)
- [Dutta et al. 2009] Dutta, R., Kamal, A.E., Rouskas, A.E.: Traffic grooming for optical networks. Springer, Berlin (2009)
- [Dutta et al. 2004] Dutta, K.A., Dutta, N.K., Fujiwara, M.: WDM technologies: optical networks, vol. III. Elsevier, Amsterdam (2004)
- [Fische 2004] Fische, G.: Systems and communicating networks: traffic and performance. Kogan Page Science, London (2004)
- [Hajder et al. 2002] Hajder, M., Loutskii, H., Stręciwilk, W.: Science. Virtual Journey into the World of Computer Systems and Networks (2002)
- [Hajder and Kielbus 2007] Hajder, M., Kielbus, M.: Mathematical model of delay in packet switched networks. In: The 15th Conf. of the Network and Information Systems, Łódź, Poland, pp. 47–50 (2007) (in Polish)
- [Haverkort 1999] Haverkort, B.R.: Performance of computer communication systems: A model-based approach. John Wiley & Sons Ltd., Chichester (1999)
- [Taha 2007] Taha, H.A.: Operations research: An introduction, 7th edn. Pearson Education, Inc., Upper Saddle River (2007)

An Autocatalytic Emergence Swarm Algorithm in the Decision-Making Task of Managing the Process of Creation of Intellectual Capital

A. Lewicki¹ and R. Tadeusiewicz²

¹ University of Information Technology and Management, Rzeszow, Poland
alewicki@wsiz.rzeszow.pl

² AGH University of Science and Technology, Krakow, Poland
rtad@agh.edu.pl

Abstract. This paper describes proposal for the application modified Ant Colony Optimization Algorithm in the task for recruitment and selection of employees. After analyzing the combinatorial problem involving multicriterial process of recruitment and selection model proposed non-compensating its solution using the modified ACO heuristic strategy, showing a lack of opportunities to receive appropriate the resulting matrix, related to the accurate prediction of the decision at an acceptable as satisfactory for implementation only available deterministic algorithms.

1 Introduction

In a modern society in which human work is not only a source of measures to ensure his daily maintenance, but also is an important stimulator that is used to meet the individual needs of the individual and society as a whole, the continuing desire of the industrialized countries to continuously upgrade the quality of life always involves economic development. This development in turn is determined by factors such as work, both in qualitative and quantitative, material resources in the form of natural capital, financial and human resources and assets, including inter alia, delineating technological progress. No operator shall not use, however, without adequate technological development of intellectual capital accumulated in the minds of people employed. It depends on him being an organization that is firm.

Proper selection and appointment of staff is therefore a critical task. It depends on many factors, both external and internal, and is therefore costly and time-consuming process, which may become totally ineffective and useless, if the actual decision-making activities in this sphere of business management will be conducted in a thoughtful and skilful. Although there are unquestionably talented

managers who are relying on their extensive experience, comprehensive knowledge, and also on intuition - they can make the selection of personnel in a way that even a perfect, unfortunately they are few, and in addition the result of their work is never predictable or sure, because they use subjective criteria may eventually fail - and you never know when this happens. It is therefore difficult, complex and very responsible decision problems related to the recruitment of personnel and policies promoted in a company associated with the need to assist those who must fulfill these tasks, and it tends to reach for the computer aided decision-making systems. However, due to a complex and sensitive matter, which is the recruitment of staff and general staff policy, it is here the use of such information technologies that are characterized mainly by the ability of adaptation, namely the ability to flexibly adapt actions to new determinants of the established process.

The proposed decision support systems in this regard should therefore incorporate not only the mechanisms for collecting and processing large amounts of data, but also mechanisms to ensure the use of different models and intelligent use of both data that is collected, how and expertise within the meaning of the expert. These tools, however, is currently very little, and when they do exist, due to their complex nature and the use of complex database engines - the costs of their implementation and licensing are very large [Yakubovich 2006; Jassim 2007]. Hence the need to continually explore new methods and build new tools to make computer-aided decision considered the more effective and more credible decision-makers. One of these tools in the aforementioned areas may be developed by the author of the publication [Lewicki 2010] multicriterial decision support system in the problem of staff selection. This system was based on a metaheuristic Ant Colony Optimization algorithms [Dorigo and Socha 2006], modified so that it could take into account both these arguments, as well as those factors which make up the data sets used in the typical process of recruitment and selection. The studies modeled both situations, as well as real data sets such arrangements were primarily to answer the question whether the proposed strategy will check in this type of optimization problems, and what will be its efficiency.

2 Formalizing the Problem

Regardless of the method of carrying out the process of recruitment and selection of the final step should always return the expected information in the form of desired personal data. Review and analysis made by the authors and not yet implemented the methods used reveals that the problem is multidimensional, because of the multiplicity of contexts, with which we deal here speaking of the forecasting ability of candidates to carry out their tasks. Prediction must be given not only associated with a rational and fair assessment, but also to the maximization of the quality of the whole process, taking into account while also minimizing costs. This,

in turn, implies that we get a complex task to optimize a vector (multiobjective optimization), including not only linear discrete optimization, but also data mining and classification. With both because they know the requirements and have the expected degree of suitability, skills and competences of candidates in the applicant-recruited jobs, which can result in a graph mapped on the trigeminal $G = (K, C, S, E)$ (shown in Figure 1), composed of collections: $K = \{k_1, k_2, k_3, \dots, k_m\}$, $S = \{s_1, s_2, s_3, \dots, s_n\}$, $C = \{c_1, c_2, c_3, \dots, c_x\}$ and edge E , where: m – represents the number of jobs, n – is a finite number of jobs, which a company can be both decrement and increment, provided that $S \notin \Phi$, x - consideration is the number of candidate traits, where $x > l$ represents the fact that we consider multicriterial selection of employees, while w_{Mki} – is the weight on the edge of the transition between the anthill, and k_i -th candidate, and for each existing connection is 1, because it proves that until a global solution to each of the candidates can be useful for one of the offered positions within the company, w_{kicd} – provides a degree of knowledge possessed c_{ij} ability, aptitude or competence for the k_i -th candidate, w_{cdsj} – is the weight representing the lowest acceptable value for a given position (index) expected predisposition and it is a negative value, because it means gaining the competency need to find the best set of solutions, within the meaning of global optimum, associated not only with how best choice for a single position, but also to the maximization of the number of associations with the candidate positions so that you can get to maximize profit function, we can also save functionally:

$$F(x) = \sum_i^n \sum_j^m x_{ij} * \sum_{i=1}^n \sum_{j=1}^m z_{ij} x_{ij} \rightarrow \max \tag{1}$$

$$\forall_{l \in \langle 1, 2, \dots, o \rangle} \sum_{i=1}^n \sum_{j=1}^m c_{il} \geq k_{lj} \tag{2}$$

$$\forall_{i \in \langle 1, 2, \dots, n \rangle} \sum_{j=1}^m x_{ij} \tag{3}$$

where i – represents the j-cantata for the workplace, l - represents possession predisposition, representing the desired criterion for a single or group of jobs, n – is the number of applicants for the job, m - maximum number of posts to the cast, o - maximum number of criterial features, z_{ij} - means choosing the best candidate for the position j , x_{ij} - takes the value 1, for the selection of the candidate or the value 0, if it is rejected, c_{il} - disposition possessed by the candidate i , k_{lj} - criterion for selecting candidates for the position j , max - maximize the number of positions filled with the greatest profit in the form of skills competence. This problem has been graphically depicted in Figure No. 1

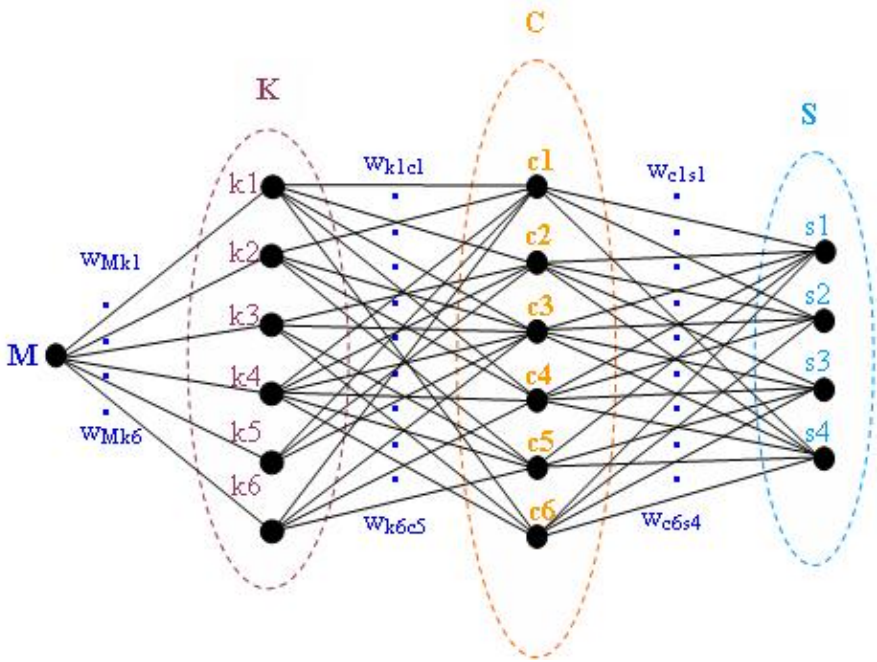


Fig. 1 A graphical representation of the quest for an optimal set of solutions in the form of vector optimization job recruitment and selection

From the analysis of combinatorial optimization problem [Pang-Ning 2006], on how to best fit a set of characteristics possessed by each of the candidates to a set of behavioral indicators of positions related to the recruitment process shows that the main barrier preventing the use of an efficient deterministic algorithm, both in the case of bringing him back to the problem of finding maximum flow in web, as well as in the case of attempts to implement the standard method of Ford-Fulkerson, or its modification due to the instability of such approaches is the multitude of potential solutions related to the occurrence of the uneven distribution of the coefficients of respondents ranked applicants. But this is not the only factor in determining the expected complexity of decision-making system, assisting the staff recruitment and selection company. Such a system is in the process of allocation of human resources should in fact provide a very time-consuming identification of the appropriate set of indicators of quality possession, so that you can for each candidate to create a profile and personal competence. As shown by studies conducted by the authors [Lewicki and Tadeusiewicz 2010], methods of selection, having a limit even to pay for the SME sector with a high rate of accuracy in the form of a competence-based forms are used infrequently, precisely because of the fact that their analysis takes a very long time. Therefore, the lack of sufficient support in the very area in the form of appropriate data mining tools, their classification and the assignment means that in this case are chosen and used methods of prognostically poor performance. However, you can build such a decision support

system, which notwithstanding the form of structuring documents is not only sufficient to identify the right set of features that characterize each of the candidates, but that the purpose of having the vector will also pay a set of possible solutions that meet well-defined criteria, taking into account all constraints strictly. Such a system may be in fact based on a novel approach, which is modified by Ant Colony Optimization algorithm [Decastro and Zuben 2004; Dowsland and Thompson 2005; Dorigo et al. 2005; Sendova-Franks 2004]. The studies proposed tool, taking into account both the test data and real data showed that the use of solutions in the form of a modified Ant System algorithm to form the ACO, screening, even when searching large, complex data sets is precisely the right approach, and gives businesses a good tool optimize profits in the form of raising the intellectual capital of the desired quality.

3 Method of Solution

Made by the authors of the analysis of complex combinatorial optimization problem, multiple criteria for the selection of m tasks to n workers you a set of positions (whatever the degree of structuring of the data obtained), and the proposed model, its solution no compensation demonstrated inability to obtain the correct matrix resulting in an acceptable time as satisfactory, with only the use of deterministic algorithms available. The study showed, however, that the right approach in this regard may be making use of the tactics of the study, represented by the ant algorithms.

In carrying out the implementation of these algorithms in knowledge acquisition module and the module has been seeking solutions to their collection of modifications to the form of improved (ACO), which took into account both the new rules update an osmotic media communications, and developed through experience new approaches to research, so that the transition to the next of the problem has always provided a compromise in the form of a choice between the exploitation associated with the accepted criteria, within the meaning already introduced restrictions on the stage of initialization and exploration based on the heuristic function of the probability calculations. In this way not only get a guarantee of quality solutions increase with increasing the solution space, but also thanks to the new rules, not only global but also local pheromone update paths, increasing the probability of finding a solution to at least close to accurately.

Ant task of decision support system based on heuristic mechanisms characterized was to determine a discrete set of feasible solutions, so that you can obtain the extreme criterion function described by formula (1). This can be achieved only when based on the available repository of data acquired knowledge about the degree of fulfillment of each of the possible limitations of the stand in the company leading the recruiting process. It is therefore considered and implemented in a non-deterministic system of positive feedback mechanisms have been associated with the algorithm based on a strategy of acquiring knowledge, and an algorithm based on the optimum search strategy for a given objective function.

A number of already performed in the problem analysis experiments showed that the algorithms used in the stochastic mechanism of the transition to a new state, associated with the likelihood function, calculated using the formula:

$$P_{ij}(t) = \begin{cases} \frac{[\tau_{ij}(t)] \cdot [\eta_{ij}(t)]^\beta}{\sum_{j \in TABU} [\tau_{ij}(t)] \cdot [\eta_{ij}(t)]^\beta} & , \text{for } j \notin TABU \\ 0 & , \text{for } j \in TABU \end{cases} \quad (4)$$

where $\tau_{ij}(t)$ - the intensity of pheromone trail, located on the edge $E(i, j)$ at time t , $\eta_{ij}(t)$ - local value of the criterion function associated with the weight of edge $E(i, j)$, β - parameter allows you to control the relative importance of the parameter being the weight of edges connecting the two states of the system, while $TABU$ – ants informing memory of her past, which allows both for the construction of feasible solutions, and the assessment have already been generated, must be subject to a random draw of parameter q , but compared with the permanent and well established criterion for parameter q_0 . It is in fact a guarantee of the quality of the solution, as too little exploration of space causes the premature convergence of solutions to a suboptimal result obtained is not always satisfactory. Too much exploration and makes the process of this convergence is in turn too slowly, which indicates a lack of proper knowledge about the use of the available set of solutions. Therefore, if the lottery number q is greater than the initial intensity of pheromone on that edge, it should make use of state space and to choose the next state according to the formula:

$$\arg \max_{j \in TABU} \{ [\tau_{ij}(t)] \cdot [\eta_{ij}(t)]^\beta \} \quad (5)$$

in any other case, the transition should be related to the probability of a specific formula (4).

The transition from the available set of edges E , a graph representing the problem in the cases should be considered to update the osmotic flow on that edge according to the principle of local updates, which can be represented as:

$$\tau_{ij}(t, t + 1) = (1 - \rho) \cdot \tau_{ij}(t) + \rho \cdot \Delta \tau_{ij}(t, t + 1) \quad (6)$$

where ρ - is a factor associated with decreasing (evaporation), the intensity of pheromone, which represents the value from the se (0; 1], $\tau_{ij}(t)$ - represents the amount of pheromone trail at the time t , $\Delta \tau_{ij}(t, t + 1)$ – is the initial intensity of pheromone on the edges of the graph, and the completion of the iterative cycle should be to reward your best solution in this phase, reinforcing the intensity of the smell only to its constituents, in accordance with the calculated value, which in all cases will be:

$$\tau_{ij}(t, t + n) = (1 - \rho) \cdot \tau_{ij}(t) + \rho \cdot \Delta \tau_{ij}(t, t + n) \quad (7)$$

$$\Delta \tau_{ij}(t, t+n) = \frac{Q}{L^+} \quad (8)$$

where n – algorithm is the number of cycles, Q - in this case is a constant value, which has shown a lot of experience and analysis of the available literature on this topic [Dorigo 2006] may be taken as a value equal to 1, since this parameter has little impact on the quality of the global solution, while L^+ - best result corresponds to the function of the total weighted edge transitions in the cycle of generating solutions.

The first of the proposed implementation of the heuristic described mechanism is to obtain an array of abilities possessed by the applicant:

$$\{c \in C : R(c)\} \quad (9)$$

where $R(c)$ – is a function (rule) the classification of features c .

A collection of these data may be gained by building and sharing on the Web the appropriate forms, based on descriptions or competency tests. However, due to the fact that every element of the set C , which is also a criterion of optimization functions, is described by a vector of indicators of well-defined competency bench, which is why this area can also propose, and then examine the use of nondeterministic mechanism characterized above, a comparison of evidence that gave the candidates have already laid down by the employer of the behavioral indicators.

Conducted by the authors of the study revealed that each of the presented evidence in the recruitment process is a document which is a set from 3200 to 4000 characters, which is approximately about 500 words in various forms, inflectional and derivational, but not always consistent with the pattern, which is a description of the job profile. Therefore, the implemented solution should focus mainly on the rejection affix and assume only a basis for comparing the formative each word. Therefore, it is the set of elements that are expressions without affix will represent a definition in the form of the characteristics of each criterion is understood as a collection of O , which is a subset of C :

$$O \subset C, \text{ for } \forall o \in O \Rightarrow o \in C \quad (10)$$

where O – denote the set of elements that describe a set of criteria C , C – criterion represents the set of constraints, depending on the job profile, for which recruitment is carried out, while o – element of the set O , which is an expression without affix.

The task of the first of the proposed ACO algorithm is to determine which elements of the set O defined determining predisposition, are aggregated with the available documents, a telecommunications company, so that in accordance with established criteria for classification can be paid the appropriate matrix of suitability. For this purpose, a colony consisting of mr ants in each new cycle is randomly deployed at points of indices which are complex expressions of the document, unless the information about the status of a phrase that has been fixed has not been previously saved in global memory. Then, each worker can perform the move to

one of the vertex set of O , described by the parameter d represents the number of characters of a phrase and the parameter ϕ signifying the importance of expression, provided that:

$$d(wr_i) \geq d(o_j) \quad (11)$$

where $d(wr_i)$ – is the number of characters to express the lot fell, $d(o_j)$ – is the number of characters belonging to a set of expressions O . The importance of this expression is the value such that:

$$\phi \in \left\langle \frac{1}{h}, 1 \right\rangle \quad (12)$$

where h - the number of characters representing a particular trait criterion, with a value of 1 means that the candidate has a particular predisposition, and its rate is 4. Such a situation occurs, for example, when a competency is certified. Goal displacement is also determined according to the rule:

$$R = \begin{cases} \arg \max_{j \in \Omega} \{ \tau_j(t) \cdot [\phi_{oj}]^\beta \}, & \text{when } q \leq q_0 \\ S, & \text{otherwise} \end{cases} \quad (13)$$

where τ_j - the intensity of pheromone trail, located on the top representing the element of the set O , compared with the census areas belonging to the term under review document, β - parameter allows you to control the relative importance weight of the next expression, q – random number from the interval $\langle 0; 1 \rangle$, q_0 is a factor determining the method of determining the next node transition, Ω - ant-cycle memory (local memory), while S – vertex drawn with probability:

$$S = \begin{cases} \frac{\tau_j(t) \cdot [\phi_{oj}]^\beta}{\sum_{j \in \Omega} [\tau_j] \cdot [\phi_{oj}]^\beta}, & \text{for } j \in \Omega \\ 0, & \text{for } j \notin \Omega \end{cases} \quad (14)$$

If the expression is drawn in accordance with the comparator element from the set O , it is updated also features an array of pointers, which include this item on the value $\phi * 100$. This process is repeated iteratively, but up to as much as the number of elements of set O . Updating the pheromone trail according to the formula:

$$\tau_j(t, t + 1) = (1 - \rho) \cdot \tau_j(t) + \rho \cdot \Delta \tau_j(t, t + 1) \quad (15)$$

where ρ - is a factor associated with decreasing (evaporation), the intensity of pheromone, which represents the value from the set $(0; 1]$, $\tau_j(t)$ - volume of pheromone trail left on the top of the set O at time t , $\Delta \tau_j(t, t + 1)$ – is the initial intensity of the pheromone left on the tops of the set O , however, relates to all the good elements that were not compatible with the test component of the application. This is because to reward other expressions which may be compatible with the other subset of the possible solutions. Also updated after each global iterative

cycle will apply only to promote new possible links. Therefore, the trace will be left to in this case is not on the edges of the transition, and the set of nodes O that cycle were not compatible with any expression. The growth of the trace will be:

$$\Delta\tau_j(t, t + m) = \frac{1}{L * W} \tag{16}$$

where L – is the number of vertices that are not correlated with any expression of the cycle, W – represents the sum of weights of these vertices.

Implementation of all modeled cycles should be accompanied with the return of the instrument cluster characteristics that determine suitability of the candidate. However, these indicators should be expressed in a five-point scale assessment in accordance with the principle of classification presented in Table 1.

Table 1 The criterion for the classification of characteristics that determine the suitability of indicators in the scale of five candidates on the basis of the percentages

The range of values [%]	Rating
<0; 19>	0
<20; 39>	1
<40; 59>	2
<60; 79>	3
<80; 100>	4

With expertise in both requirements and degree of expected to have predispositions, skills and competence of the candidates recruited on the applicant-job, find the best set of solutions so that you can get to maximize the profit function.

The proposed modification of previously presented a general idea of the ACO algorithm, requesting in this case, the expected boolean matrix solutions to staffing vacancies available, take into account the profit function concerns the competence of each of the posts in the company carrying out the recruitment and selection of employees. Therefore, the solution described in the process for finding the global optimum provides two main sequences of instructions:

- the sequence of construction of the matrix gains
- the sequence of construction of the matrix solutions in the stand.

In the case of the first of them the effort of moving each of the mr ant nest departing from M to a food source is located in the set S is also undertaken by successive ants and only if it involves prize in the form of profit iterative $Z(k, c_d s_j)$ defined as:

$$Z(k, c_d s_j) = w_{Mki} + w_{kicd} + w_{cdsj} \tag{17}$$

where w_{Mki} – mean weight on the edge of the transition between the anthill, and k -th candidate, in this case, for each existing connection is 1, because it proves that, until a global solution to each of the candidates can be useful for one of the

offered positions within the company, w_{kicd} – provides a degree of knowledge possessed c_d skills, abilities or competence of a candidate k_i , w_{cdsj} – is the weight representing the lowest acceptable value for a given position (index) expected predisposition and it is a negative value, because it means gaining the competency, the profit is understood here as a value greater than 0 and equal to at least 1 In other cases, the transition $M-k_i-c_d-s_j$ is marked as unattractive, which involves following the rejection of such a solution also by the other ants. For this purpose, the decision variable is defined, which, to be able to conduct further discussion and analysis will be called attractive candidate factor for the position, and we denote it as γ_{ij} :

$$\gamma_{ij} = \begin{cases} 1, & \text{when } Z(k_i c_d s_j) \geq 1 \\ 0, & \text{otherwise} \end{cases} \tag{18}$$

where k_i – mean i -th applicator candidate for a job at one of the available positions in the business, c_d – represents the x -selection criterion for the j -th position, s_j – is one of the positions of the recruitment process in the company.

At the beginning of each new cycle factor γ_{ij} is determined on the binary value 1 and can not be changed until any of the iteration of that cycle of ants forming a solution does not receive the value of $Z(k_i c_d s_j)$ less than the acceptable minimum, equal to the first Thanks to one of the assumed failure criterion constraints forced the ants to check other association between a set of candidates, and a set of positions, thus ensuring the implementation of the non-compensation mechanism for the selection of available space solutions. The current rate is therefore a decisive impact on the earnings record information on z_{ij} the cast of the i -th candidate for j -th position in the long-term memory of the whole process of searching. This gain is:

$$z_{ij} = \gamma_{ij} * \sum_{d=1}^x Z(k_i c_d s_j) \tag{19}$$

for $i \in \langle 1; n \rangle$ and $j \in \langle 1; m \rangle$, where n - is the number of applicants who applied to work, x - criterion represents the number of constraints for each of the positions recruited, m - stores information about the number of posts for which the selection of staff, γ_{ij} – rate the attractiveness of the candidate and the position j , which in each cycle is initialized to a value of 1 and does not change to 0, as long as the profit of each iteration $Z(k_i c_d s_j)$ is the value of $\langle 1; 5 \rangle$.

The described sequence of instructions is repeated periodically until the matrix is filled with a profit, informing about the perk value resulting from the cast of the i th candidate for the j -th position. In this moment the initialization of topics related to the setting of global state matrix solution x_{ij} :

$$x_{ij} = \begin{cases} 1, & \text{when the time allotted to the } j\text{-th candidate} \\ 0, & \text{otherwise} \end{cases} \tag{20}$$

In view of the fact that every vertex of S can be associated with at most one vertex of K we expect the greatest combination of numbers and as far as earnings summary Z_p :

$$Z_p = \sum_{i=1}^n \sum_{j=1}^m x_{ij} * \sum_{i=1}^n \sum_{j=1}^m z_{ij} x_{ij} \tag{21}$$

for x_{ij} – which is the state of the decision matrix in each cycle transition, z_{ij} – indicates the value associated with the choice i candidate for the position j .

Update local pheromone content at the edges of the transition are:

$$\tau_0 = \frac{1}{n * m * z_{ij\max}} \tag{23}$$

where n – represents the number of test applicants to your company, m – is the number of posts available in the recruitment process, $z_{ij\max}$ – the largest volume gain transitions between nodes i and j , determined in the initialization phase of the algorithm.

After each cycle, followed by the global update. It concerns not only the association, which suggests a solution not yet the best quality stock, but also of related searches the structure of K , because this can eliminate the likelihood of stuck in local minima. Therefore, it carries only the ant, which received the largest value Z_p profit increases and the trace is:

$$\Delta \tau_{ij}(t, t + m) = z_{ij\max} * Q \tag{24}$$

$$Q = \frac{1}{\sum_{i=1}^n \sum_{j=1}^m z_{ij}} \tag{25}$$

where Z_p – means the sum of association with the largest possible cardinality, Q – This is the size of the pheromone, the ant has, n – represents the number of candidates in the process of staff selection, m – is the number of posts for which recruitment is made, a z_{ij} – an indication of the quality of association i, j .

4 Case Study and Results

To conduct experiments related to the testing of selected two groups of data sets containing objects such as profiles and application documents bench of candidates. The first group of test objects of the recruitment process and selection of abstract data concerned, selected so that you can make the optimal selection of system parameters, and then check both the usefulness and the quality of heuristic solutions set. Each of the sets (FTM1, FTM2, FTM3, FTM4, FTM5) of the test data consisted of a number of items, so you can easily perform the verification of the correctness of returned results for different model situations defined combinatorial problem. The second group of data were made available to the author by the Office of Career Services University of Information Technology and Management in the actual data of 4 selected sector companies SME (Small-Medium Enterprises), which in 2008 conducted a recruitment process in Podkarpackie more than one position, and then revealed the university for statistical purposes the results.

These collections were used to compare the quality of measures taken by the company with the optimum global decision proposed by the system. In this process, however, were taken into account not only sets the requirements which are criterial constraints defined by the company but also those which, in the pursuit of objectivity in staff selection process are derived from public tools (such as tabs descriptions of competence), allowing the production of the correct profile for the vacant position. This approach was therefore to compare the results obtained, and then determine the degree of convergence with the decisions.

The main aim of the research conducted as part of a first group of experiments with the model data was to determine the values of the parameter q_0 , β and ρ to using the proposed decision-making system you can get the best results for sets of different frequencies.

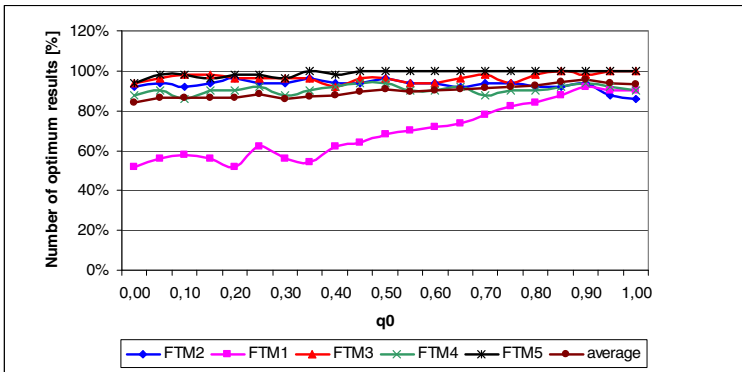


Fig. 2 Dependence as a result the percentage of optimal solutions generated by the system from the parameter being changed q_0 experiment with regard to harvesting

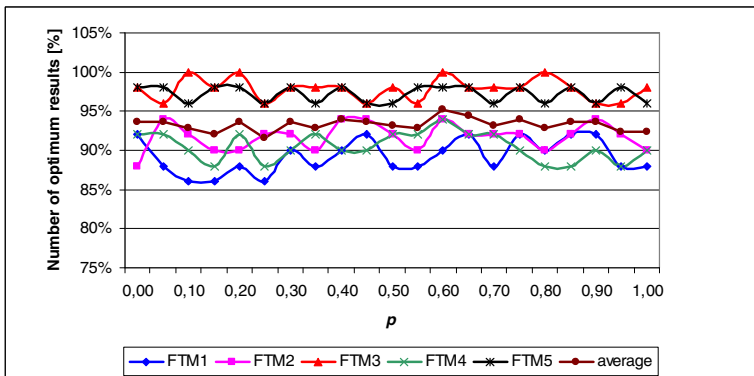


Fig. 3 Dependence as a result the percentage of optimal solutions generated by the system from the parameter being changed ρ experiment with regard to harvesting.

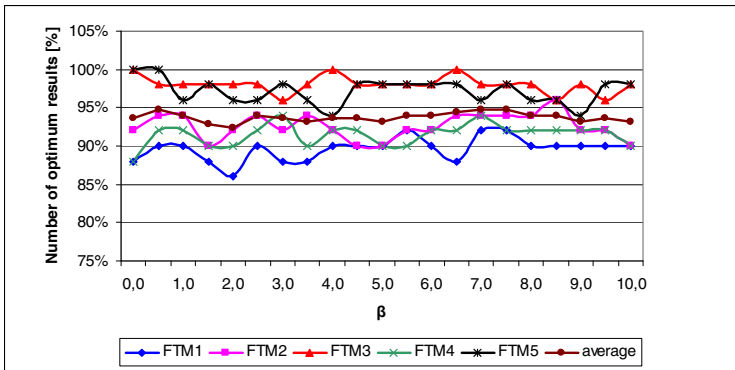


Fig. 4 Dependence as a result the percentage of optimal solutions generated by the system from the parameter being changed β experiment with regard to harvesting

Determination of the best values of parameters (as shown in the diagrams depicted in Figures 2, 3 and 4) designed and built decision support system ant staff selection process by the company, and to determine the optimal number of iterations made it possible to conduct the appropriate experiments. These experiments, which used four large real data sets have a high coefficient of combinatorial complexity, were generated to investigate the value of relative error generated solutions so as to assess the effectiveness, and thus the quality of the proposed mechanism. In addition, the final stage of this research was to attempt to assess and compare the proposed system solutions and made personal decisions already known by each of the companies represented by an appropriate set of data (FA, FB, FC i FD). Just as in the case of collections of model even now, despite the high degree of difficulty of each test set failed to appoint a satisfactory solution is optimal. However, in order to emphasize the complexity of test problems should be noted that collections of FB and FC is the combinatorial problem concerning the proper indication of an association between 90 and 92 objects, so that you can not only achieve the highest possible profit, but also cast the greatest number of positions with candidates who meet only accepted no compensation selection model (which does not satisfy one of the criteria can not be averaged by the other). Despite such a system-defined problems, however, proved to be very effective. This means, therefore, that the proposed modifications to the Ant can be successfully applied also to solve much larger problems associated with the selection of staff for the position sought.

5 Summary

Modern approach to the task of staff selection is based on the fact that the course continues to benefit from the knowledge of experienced workers and institutions associated with department personnel and payroll companies, but also should be

used as an objective computer techniques and tools available to business intelligence. The proposed evaluation process and subjected to a solution in that area is to build decision-making system, based on a modified tactical the Ant System. Analysis of problem areas, the characteristics of the conventional solutions and verification of the results obtained in the course of experiments that they make it to the following conclusions:

- Optimization problem of recruitment and selection can be mapped in the form of a formal indication of a problem for as many positions of S only candidates meeting the minimum expectations of K criterial set S , thus forming a tripartite graph $G = (K, C, S, E)$. In this notation C denote the set of constraints criterion, depending on the job profile for which recruitment is carried out, while E is the set of edges of the graph.
- Regardless of the complexity of the problem of staff selection, the best results were achieved in the case of the system which has not only be fine-tuned control parameters (experiments showed that the best value for the parameter q_0 to 0.9, parameter β is 7.0, while for the parameter ρ is 0.6), but also taking into account (often selected by the tactics of exploratory ants) as heuristic information.
- Each of the ants seeking the best possible result contributes to the rapid achievement of global optimum sought only in cases where there is also the possibility to use shared memory, consisting of a list of discovered solutions in the form of associative access (TABU_S list) and lists a set of vertices K , making these solutions (TABU_K).
- To achieve the lowest possible value of the relative error in the so-designed system, you must also determine the number of iterations. Proposed and established (in the course of experiments) for 300 iterations the value of results from research on complex systems, a maximum of 92 objects. In working with larger collections may however that this value must be reconsidered.

We conclude that the proposed solution as a heuristic strategy for the Ant Colony Optimization, in spite of the combinatorial nature of the examined problems related to the recruitment and selection, allows to obtain solutions to at least close to globally optimal. In addition, choosing appropriate strategy in forming a solution algorithm can eliminate the achievement of local minima, thus accelerating the achievement of global optimum. Therefore, the proposed approach can be successfully recommended for use.

References

- [Decastro and Zuben 2004] Decastro, L., von Zuben, F.: Recent developments in biologically inspired computing. Idea Group Publishing, Hershey (2004)
- [Dorigo et al. 2005] Dorigo, M., Handl, J., Knowles, J.: Ant-based clustering and topographic mapping. *Artificial Life* (2005)
- [Dorigo and Socha 2006] Dorigo, M., Socha, K.: An introduction to ant colony optimization. Technical Report (2006)

- [Dowsland and Thompson 2005] Dowsland, K., Thompson, J.: Ant colony optimization for the examination scheduling problem. *Journal of the Operational Research Society* (2005)
- [Jassim 2007] Jassim, R.K.: *Competitive advantage through the employees*, CCH, Australia (2007)
- [Lewicki 2010] Lewicki, A.: Non-Euclidean metric in multi-objective ant colony optimization algorithms. In: *Information Systems Architecture and Technology. System Analysis Approach to the Design*, Wroclaw (2010)
- [Lewicki and Tadeusiewicz 2010] Lewicki, A., Tadeusiewicz, R.: The ant colony optimization algorithm for multiobjective optimization non-compensation model problem staff selection. LNCS. ISICA, Wuhan (2010)
- [Pang-Ning 2006] Pang-Ning, T.: *Introduction to data mining*. Addison Wesley Publication, Reading (2006)
- [Sendova-Franks 2004] Sendova-Franks, A.: Brood sorting by ants: two phases and differential diffusion. *Animal Behaviour* (2004)
- [Yakubovich 2006] Yakubovich, V.: Stages of the recruitment process and the referrer's performance effect. *Informs*, Maryland (2006)

Implementation of the Progressive Meshes for Distributed Virtual Museum

K. Skabek¹, R. Winiarczyk^{1,2}, and A Sochan¹

¹Institute of Theoretical and Applied Informatics, Polish Academy of Sciences,
Gliwice, Poland
{kskabek, arek}@iitis.pl

²Institute of Informatics, Silesian University of Technology, Gliwice, Poland
ryszard.winiarczyk@polsl.pl

Abstract. This article presents a specific part of the idea of distributed virtual museum. The intention of the authors is to fulfill one of the scenarios for Future Internet and to give a set of requirements for future properties of the global network. A progressive mesh is a lossless approximation of the original mesh model at any, smoothly adjusted, level of detail. The main part of this work is devoted to the extension of progressive meshes for hierarchical representation associated the position of observation. This allows the selective adjustment of the level of detail, measured in the number of triangles that appear in any fragment of the model, independently from the rest. Sample solution was presented, as well as the method of construction and the main criteria for the refinement.

1 Introduction

The technology for virtual museum makes it possible to gather information about available historical objects. Its shapes, dimensions and colors are converted to digital form by the 3D scanner using laser triangulation method, then processed and stored. Thus the created file is the mostly detailed representation, yet its size is negligible when it comes to the average server storage capacity, but it becomes a serious problem, if it must be sent to the client side by the internet, especially when the number of visitors is constantly growing. The solution introduces a hierarchical representation for each exhibit, so that in the first step it can be presented to the museum visitor as coarsen mesh that has the lower level of details and therefore is much smaller. The remaining details will be downloaded and displayed, if the observed object excites the greater interest to the viewer. Such an approach allows reducing the amount of data exchanged between client and server, to relieve the link and reduce the amount of data processed on both sides, while improving the usability of the service.

For that purpose a software tool was built extending the existing editor ME3D for 3D mesh models. The application was developed in the Institute and described in [Skabek and Ząbik 2009].

2 Architecture of Distributed Virtual Museum

Architecture of distributed virtual museum (as shown in fig. 1) is based on three components:

Exhibition Server - contains information about the exhibition, including compositions of scenes. The particular scenes of the exhibition (exhibition rooms) may be situated on different exhibition servers. Description of the scene contains the general characteristics (information what it refers to, audio commentary, etc.), information about the 3D appearance (mesh, lighting, texture, etc.), scenarios of interaction, and information about the objects necessary to the instantiation (localization or identifiers or metadata - depending on the network functionality, e.g. content aware network - CAN).

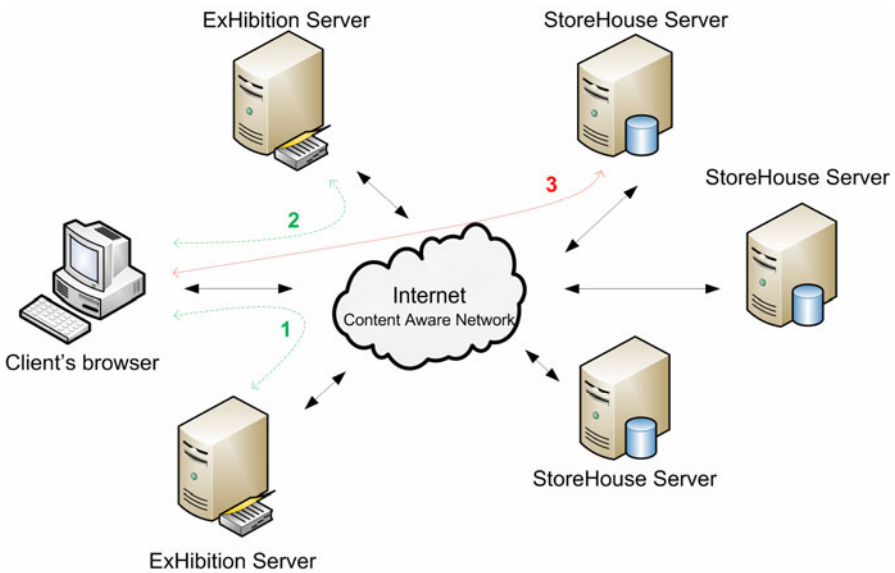


Fig. 1 Architecture of distributed virtual museum

Storage server - contains information about objects used in certain scenes. Similarly as for the scene is general characteristics, information on the 3D appearance (mesh, material, texture), however, objects can be retrieved using progressive encoding.

The client's browser – client finds an exhibition on the network (step 1), downloads information about the scene to display and make the 3D visualization (step 2), depending on the requirements (virtual position) progressively retrieves information about particular items on the stage and performs the 3D visualization (step 3). Objects can be localized in different storage servers. The peer-to-peer

network architecture is intended to be realized for the distribution of objects. The browser provides some interaction with objects according to rules defined in the scene description.

3 Progressive Representation – An Overview

Many techniques have been proposed to compress and transmit mesh data. They can be divided into nonprogressive and progressive methods. The first group comprises methods which encode the entire data as a whole. They can either use the interlocking trees (vertex spanning tree and triangle spanning tree) or utilize the breadth-first traversal method to compress meshes. On the other hand there are methods that perform mesh compressing progressively. The solution proposed by [Hoppe 1996] enables continuous transition from the coarsest to the finest resolution. In such case a hierarchy of level-of-detailed approximation is built. Also the efficient quadric algorithm for mesh decimation by proposed in [Hoppe et al. 1993].

In the article [Yang and Kim 2004] a view-dependent graphics streaming scheme was proposed. 3D models are split into several partitions and they are simplified and coded separately. The compressed data is sent on the user request. The partitioning of the model is done arbitrary and the separated partitions are simplified by merging inner vertices into a single vertex.

The article [Giola et al. 2004] shows the recent technique based on zerotree (wavelet) compression as a loopy mesh compression method. Such representation was proposed in part 16: AFX (Animation Framework Extension) of the MPEG-4 international standard for 3D models encoding.

The similar technique of the wavelet-based progressive mesh coder was presented in the article [Guan et al. 2010]. The mesh coder converts an irregular mesh into a semi-regular mesh and directly applies the zerotree-like image coders to compress the wavelet vectors. The efficient technique to the multiresolution adaptive parametrization of surfaces (MAPS) was implemented. Also a rate-distortion (R-D) optimized view-dependent mesh streaming was adapted.

Despite constantly increasing computing performance of present-day personal computers, displaying complex meshes or large quantities of meshes can lead to a situation when all available computing power is insufficient. One possible solution is to replace the models of objects located in the distance by their counterparts, containing far fewer triangles. Despite the lower level of detail the overall shape is retained, and due to its large distance it becomes impossible to perceive any difference. These models are usually created by hand and there are only a few, representing pre-defined levels of decimation.

Progressive meshes do not have such restrictions, the user specifies only the minimum acceptable level of detail, but it still has a smooth transition to any other level between the original and the most sparse mesh, through the execution of elementary operations. In order to achieve this effect it is necessary to extend the way to represent the mesh. The originator of the described approach is Hugues

Hoppe [Hoppe 1996]. Original mesh, denoted as \hat{M} , at first is decimated by the cyclical performance of operations, resulting in the reduction of detail. In the case of progressive meshes the only operation used is removal of the edge.

3.1 Operations on Progressive Meshes

Each execution of a single operation is associated with the creation of structures that store information about the changes introduced in the mesh topology. The most important feature of the edge removal operation, also called the *ecol*, is its complete reversibility. With the parameters describing it, it is possible to add to the mesh previously removed edge, exactly at the point where it was originally - this is the vertex split operation: *vsplit*. Among the necessary parameters are indices v_l, v_r, v_t, v_s , and optionally may include information about the attributes associated with the appearance of the material: its identifier, texture coordinates and a change of normal values in the surface area under consideration. At each change of the mesh there are removed or added two neighboring triangles (v_l, v_s, v_t) and (v_r, v_t, v_s) and respectively deleted or inserted one point v_t .

```

class Vsplit {
public:
    s32 flclw;
    s8 vlr_rot;
    s32 fl1_id;
    s32 f21_id;
    s32 f22_id;
    s8 fl_vt_i;
    s8 f2_vt_i;
    struct {
        s8 vs_i;
        s8 corners;
        s8 ii;
        s8 matid_predict;
    }code;
    s8 fl_matid;
    VertexAttribD vad_l, vad_s;
    E3DVectorWedgeAttribD wads;
};
    
```

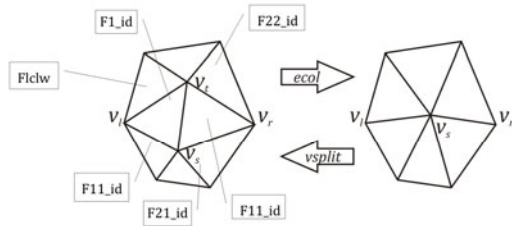


Fig. 2 The structure for operations: edge removal (*ecol*) and the reverse - adding vertex (*vsplit*)

The set of all modifications, written in the order of execution, creates a sequence of *vsplit* operations and is stored together with the resulting from the decimation of the base mesh, denoted as M^0 . The formula below shows the original mesh decimation mechanism:

$$\begin{aligned}
 (\hat{M} = M^n) &\xrightarrow{ecol_{n-1}} M^{n-1} \xrightarrow{ecol_{n-2}} \dots \\
 \dots &\xrightarrow{ecol_2} M^2 \xrightarrow{ecol_1} M^1 \xrightarrow{ecol_0} M^0
 \end{aligned}
 \tag{1}$$

In order to return from base mesh to the original, one must perform certain number of refinement operations. It should be noted that the progressive mesh representation is lossless, so the last mesh M^n is identical to the original one.

$$M^0 \xrightarrow{vsplit_0} M^1 \xrightarrow{vsplit_1} M^2 \xrightarrow{vsplit_2} \dots \xrightarrow{ecol_{n-2}} M^{n-1} \xrightarrow{ecol_{n-2}} (M^n = \hat{M}) \quad (2)$$

Corresponding opposite operations $ecol_n$ and $vsplit_n$ are stored in one and the same structure, because they have the same parameters and differ only in the nature of operation.

3.2 Strategy of the Network Transmission

The mesh data for progressive representation is stored as a special structure **Vsplit**. It stores the transition data for operations *Vertex Split* and *Edge Collapse*. The class is based on the structure proposed by Hoppe [Hoppe 1996] and was extended by certain elements. Particularly the fields of **Vsplit** structure are described in fig.~1. Faces in this structure are described by its identifiers, e.g. $f_{11_{id}}$ (see fig.~1). Variables $f_{1_{v_i}}$ and $f_{2_{v_i}}$ concern faces $f_{1_{id}}$ and $f_{2_{id}}$ accordingly. These are indexes of vertex v_i in vertex table for faces f_1 or f_2 , they can assume values 0, 1 or 2.

A single **Vsplit** instance takes 80 Bytes. The transmission of **Vsplit** packets is proportional to the number of vertices incoming to the structure and must be followed by some additional information such as number of objects, which only slightly increase the volume.

The initial transmission of the progressive mesh is arranged as a sequence of the base mesh M^0 and the **Vsplit** records up to the assumed level-of-detail. Other transmissions include only the **Vsplit** records.

4 View-Dependent Progressive Representation

The purpose for building view-dependent progressive mesh is to allow the creation of a model with different levels of detail in various parts of its surface, unlike in progressive mesh, where the level of detail was the same for the whole object. All designed solutions are extension of mechanism for progressive meshes with additional parameters, allowing dynamic adjustment of the characteristics of the model. The intended end result is such a manipulation of the displayed mesh so that from the observer's point of view it's impossible to identify the difference between observed fragment and its counterpart in the original mesh.

Hugues Hoppe, in his solution [Hoppe 1997] based on the idea of progressive meshes introduces some changes to the operation of the edge removal (see Fig. 3).

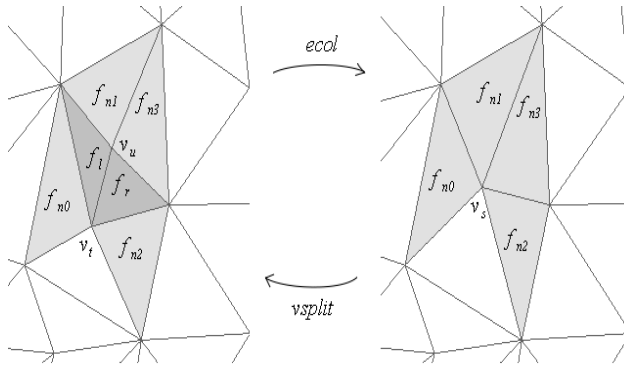


Fig. 3 *Ecol* and *vsplit* operations in view-dependent progressive meshes

In the case of progressive mesh, parameters that were sufficient to precisely locate modified area were vertices: v_b , v_r , v_p , v_s . However in the case of view-dependent extension, required vertices are only: v_w , v_b , v_s , but there are additional triangles needed: f_b , f_r , f_{n0} , f_{n1} , f_{n2} , f_{n3} . During refinement operations vertex v_s is replaced by its descendants: v_p , v_u lying between neighboring triangles f_{n0} and f_{n1} , and then face f_l is inserted. Moreover, by analogy between f_{n2} and f_{n3} , face f_r is added.

4.1 Vertex Hierarchy

Algorithm to build a hierarchy of vertices, as input data requires a view-independent progressive mesh, however with changes included in the parameters of the operations. Structure defining a single vertex is extended with additional fields, which are references to another vertices: ancestor, descendants, and two points: the next and previous on the list, coupling active nodes. This list contains only vertices which are displayed in mesh at a given time. Also, the triangles are connected in two-way list, used to determine their activity and displayed if only they satisfy relevant conditions. Vertices composing the base mesh M^0 are becoming tree roots, which contain information about relations between parent and child. At the same time these vertices are added to the list of active nodes. Hierarchy constructed this way can associate each node in the tree, with its inherent descendants and ancestors. The drawback is the need to double the size of set of points as a result of performed operations, while the number of triangles is not changed. However, the advantage of this approach is a fixed number of faces, which determine the feasibility of the operation, as opposed to another approach [Xia and Varshney 1996], where the number of subsidiaries triangles can be different for each triangle.

4.2 Refinement Conditions

In order to view the selection of the appropriate vertices Hoppe proposed a number of tests to reject those which at any given time are irrelevant to an observer. They consider: *the view frustum* and *the surface direction*.

The view frustum is the area in which objects are viewed and can be approximated by a rectangular pyramid. If object is completely outside the lump then it is completely invisible to the observer. In case of meshes, vertex contained in a space limited by frustum becomes a candidate for refinement.

The surface direction influences the content of the observed view. The triangles that form 3D models have two sides, known as front and rear. If the position of the observer clearly indicates that he sees only rear surface then refinement is not executed.

4.3 Process of Refinement

Each vertex in the activity list can be refined if it meets all the criteria of user-selected combinations. If the operation is legal, the performing precision causes vertex v_s to be replaced by its descendants: v_r and v_u , in both the mesh and the activity list. In the next step newly added descendants will be analyzed recursively. With the update on the vertices list analogous updates will be applied on corresponding triangle activity list due to addition or removal of triangles. If the pending operation cannot be performed because of the absence of any priory established triangles, then recursive search through the hierarchy starts. Only necessary operations, which ultimately lead to enable the initial operation, are executed. As a result, refined area can always be presented with maximum level-of-detail, regardless the correlations existing among various operations. If refinement condition is not fulfilled then the decimation of the mesh is considered. Requirement for decimation is that the refinement condition of the given parent vertex is not fulfilled as well.

5 Implementation and Software Environment

ME3D is an application designed for MS Windows systems, it was written mainly in C# programming language using MS Visual Studio .NET environment. In order to work properly it requires following libraries: .NET Framework and Tao Framework. The main window of ME3D system is presented in Fig. 3.

Plugin modules are created as dynamically linked libraries DLL. The environment architecture makes it possible to implement the user interface part of the module in Managed C++, and the remaining part that executes all the time consuming calculations and other operations has to be coded in traditional C++. This approach assures high overall performance and simultaneously preserves reasonable creation time.

After loading the plug-in, in the main window of ME3D a floating panel is added (see Fig. 4), which provides the functionality of progressive meshes. Now it is possible to build view-dependent progressive mesh from displayed object or read it from the file. Parameter *target number of triangles*, given by the user, specifies a stop condition for decimation algorithm. If the indicated value is reached the process is terminated, otherwise it is continued until the removal of all possible edges.

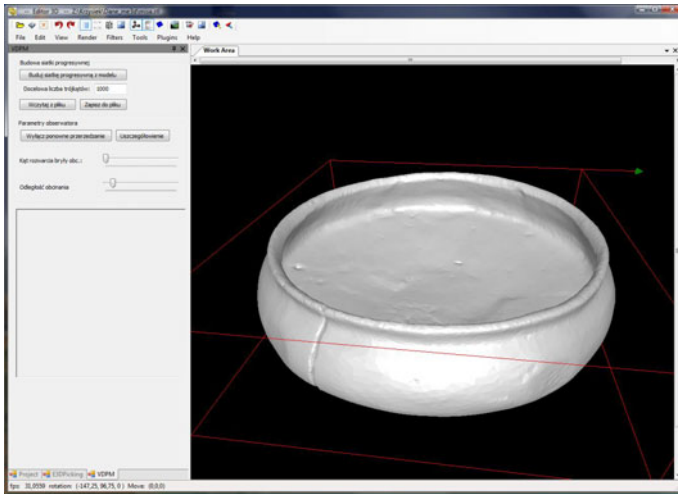


Fig. 4 ME3D editor with view-dependent progressive mesh plug-in loaded

There is also possibility to change some parameters connected with observer or the way model is processed. *Disable / Enable decimation* toggle button determines whether in the process of refinement nodes that do not meet the criteria for display, but were already on the list of active points, are not / are removed. The default setting causes decimation. Angle of view frustum allows to simply expand or narrow the field of view. Distance cutoff determines maximum distance that observer is able to see. Mathematically it is the height of rectangular pyramid that creates view frustum. Another factor is a aspect ratio, which is the proportion between two adjacent edges of pyramid's base.

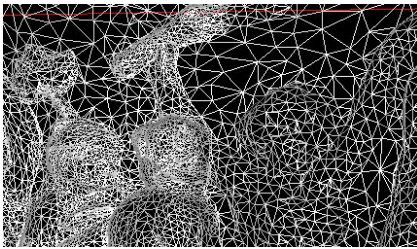
The last element on form is the small window, that is responsible for manipulating virtual observer's position and the view. There is an equivalent of the appropriate stage with drawn coordinate system and bounding box of the object. Actual model cannot be displayed once again, because of decrease in performance. This solution is quite comfortable and intuitive, allowing the observation of scene from two independent perspectives, which is extremely important at the stage of building the mesh and verification of its correctness and quality.

6 Experimental Results

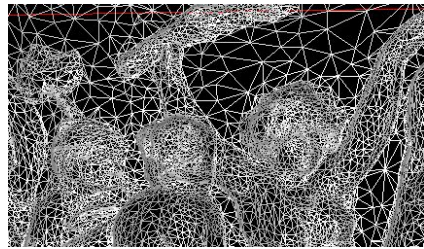
Since ME3D is an editor for general purposes, so the most important is the flexibility, not the highest graphics performance. To display object it uses software renderer. This means that any translations and rotations are performed by CPU rather than the GPU, thus it's easier to see performance impact of specific actions. Rendering performance of displayed objects highly depends on number of triangles.



a) 102915 triangles (original)



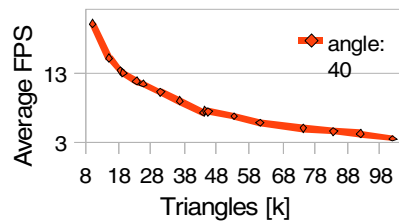
b) 37377 triangles



c) 44213 triangles



d) 29179 triangles



e) rendering performance

Fig. 5 Sample object – Bacchus Procession from the collection of Museum in Gliwice : a) original mesh, b-d) various level of detail, e) rendering performance

The more faces appear in mesh the lower average FPS drops. Therefore it is very important to discard as many triangles irrelevant to the observer as possible. Tests were performed using processor Athlon XP 2400+ 2.0 GHz, 1.5 GB RAM and graphics card Radeon HD2600 Pro 512MB RAM.

Large objects (Bacchus, Fig. 5), which are mainly viewed by parts, in case of incremental refinement, soon enforces the need for processing too much detail outside the field of view. Then again decimation should be performed, aimed at maintaining high and constant average FPS. In the decimation process only the shortest edges are removed, new point becomes the midpoint of removed edge, so the introduced distortion is relatively small. The effect of applying such criteria can be seen in Fig. 5b and Fig. 5c, where the right-hand sides are respectively decimated and refined. The differences relate only to the area consisting of small faces, while the flat region represented by much larger triangles is not altered.

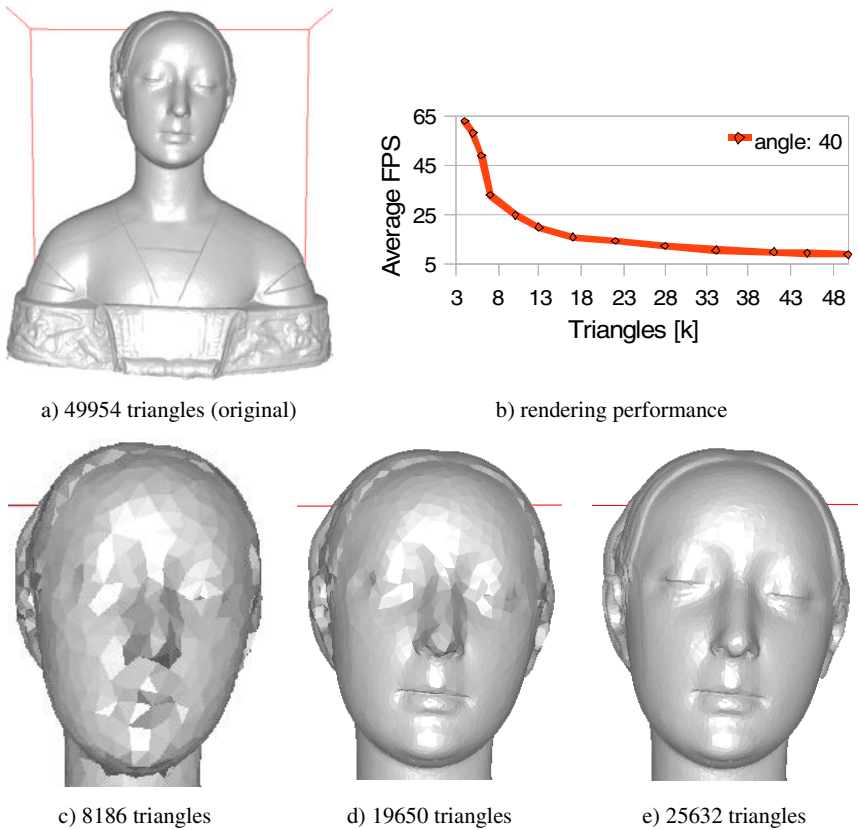


Fig. 6 Sample object – Laura: a) original mesh, b) rendering performance, c-e) various view-dependent level of detail

The case of another object (Laura, Fig. 6) clearly shows fundamental differences between the view-dependent progressive mesh and the standard progressive mesh. To present fully refined face (Fig. 6e) more than 8k triangles are needed, which is 16% of total. Progressive mesh of the same size (Fig. 6c) is distinctly distorted in a corresponding fragment, while other areas, which are negligible for observer are clearly more detailed.

Objects characterized by irregular meshes are the important problem, where a small area contains a lot of small faces, while the other triangles are much greater. As a result, a little change in the observation parameters results in a significant change in the number of triangles in the displayed mesh. This is especially disadvantageous for progressive transmission, because mostly the transmission would consume very little bandwidth, but sometimes the amount of data to be sent would cause very heavy traffic. One possible solution would be to download some parts of the mesh during the idle time.

7 Conclusions and Further Works

Progressive meshes and their view-dependent extension offer numerous capabilities. They allow to limit computing power necessary for representation of object on virtual scene and to adjust level of detail to meet actual requirements. When the mesh is downloaded from network it is possible to better balance network link load, so that the object can be displayed earlier, than when it would be downloaded completely. The view-dependent progressive meshes play important role in distributed 3D visualization systems. The idea of distributed virtual museum is being developed in the Future of Internet Engineering project as an example of such 3D visualization system.

Adjustment of the existing algorithms and data structures to take advantage of multi-core processors has become the main purpose lately. Unfortunately, still the least importance is placed on real utilization of capabilities, related with progressive transmission over the network. Especially usage of external sources, such as websites, would lead to dynamic propagation of idea of progressive meshes, which are still rarely used.

Further work includes full consideration of visual aspects of the mesh, such as processing of normals and texture coordinates, in order to improve the final effect. Additionally, to smooth transitions between two meshes by implementing geomorphs would be helpful. Also, it would be useful, so that the user could specify the mesh areas in which decimation is unnecessary or particularly important.

Acknowledgment

This work was developed i.a. in the research project Future of Internet Engineering (POIG.01.01.02-00-045/09-00).

References

- [Giola et al. 2004] Giola, P., Aubault, O., Bouville, C.: Real-Time reconstruction of wavelet-encoded meshes for view-dependent transmission and visualization. *IEEE Transaction on Circuits for Video Technology* 14(7), 1009–1020 (2004)
- [Guan et al. 2010] Guan, W., Cai, J., Zhang, J., Zheng, J.: Progressive coding and illumination and view-dependent transmission of 3-D meshes using R-D optimization
- [Hoppe 1996] Hoppe, H.: Progressive meshes. *Computer Graphics*, 99–108 (1996)
- [Hoppe 1997] Hoppe, H.: View-dependent refinement of progressive meshes. *Microsoft Research* (1997)
- [Hoppe 1998] Hoppe, H.: Efficient implementation of progressive meshes. *Computer & Graphics* 22(1), 27–36 (1998)
- [Hoppe et al. 1993] Hoppe, H., DeRose, T., Duchamp, T., McDonald, J., Stuetzle, W.: Mesh optimization. *Computer Graphics*. In: *Proc. of SIGGRAPH*, pp. 19–26 (1993)
- [Luebke et al. 2003] Luebke, D., Reddy, M., Cohen, J.D., Varshney, A., Watson, B., Huebner, R.: *Level of Details for 3D Graphics*. Morgan Kaufmann Publishers, San Francisco (2003)
- [Samet 2006] Samet, H.: *Foundations of multidimensional and metric data structure*. Morgan Kaufmann Publishers, San Francisco (2006)
- [Skabek and Ząbik 2009] Skabek, K., Ząbik, Ł.: Implementation of progressive meshes for hierarchical representation of cultural artifacts. In: Bolc, L., Kulikowski, J.L., Wojciechowski, K. (eds.) *ICCVG 2008*. LNCS, vol. 5337, pp. 123–132. Springer, Heidelberg (2009)
- [Xia and Varshney 1996] Xia, J., Varshney, A.: Dynamic view-dependent simplification for polygonal models. In: *Proc. of Visualization 1996*, pp. 327–334 (1996)
- [Yang and Kim 2004] Yang, S., Kin, C.-S., Kuo, J.: A progressive view dependent technique for interactive 3-D mesh transmission. *IEEE Trans. on Circuits for Video Technology* 14(11), 1249–1264 (2004)

The Impact of the Control Plane Architecture on the QoS in the GMPLS Network

P. Rozycki, J. Korniak, and J. Kolbusz

Department of Electronics and Telecommunications,
University of Information Technology and Management, Rzeszow, Poland
{prozycki, jkorniak, jkolbusz}@wsiz.rzeszow.pl

Abstract. The influence of the control plane architecture on the quality of services offered by the GMPLS network is considered in this paper. Several simulation experiments are prepared to verify this influence. The simulations assume separation of the functional planes of GMPLS and two typical topologies ring and mesh. The GMPLS control plane simulator, prepared by authors, simulate most of the control plane behavior by including OSPF-TE, RSVP-TE, protection and restoration procedures. The simulation results confirm quality of services improvement by providing additional interconnection in the control plane.

1 Introduction

An interaction between system more often occurs over Internet network. New expectations require serious investments in the network and new technology, especially next generation network technology. Two of the most promising technology for the Internet backbone are The Generalized Multiprotocol Label Switching (GMPLS) [Manie et al.2004] and The Automatically Switched Optical Networks (ASON) [Jajszczyk 2005]. These technologies satisfy growing demand for bandwidth and high quality of service. In order to accommodate growing Internet traffic optical crossconnects (OCXs) must be implemented to reduce IP switching. The GMPLS is considered and proposed as the control plane implementation for ASON networks and is responsible for ensuring quality of services. Conjunction of these two technology can offer previously unavailable opportunities.

The idea of the GMPLS based on inheritance of MPLS benefits and generalization of the label switching concept to all types of multiplexing such as WDM, TDM, packet switching etc. Moreover GMPLS is characterized by separation of functional planes:

- data plane is responsible for switching data traffic,
- common control plane is responsible for exchanging signaling and routing information;
- management plane is the centralized or distributed supervise system that allows, for example, to employ provider's policy.

In traditional network technologies all three functional planes share one physical network infrastructure. In the case of the ASON network this approach is not acceptable. There are two methods of signal transmission in the GMPLS:

- in-band – control plane is logically separated but signaling information uses the same media as data plane;
- out-of-band – signaling traffic is carried over dedicated network separated from the data plane.

The benefit of out-of-band signaling is the possibility to dedicate all-optical data plane infrastructure only to user data traffic. While control plane traffic is traditionally IP switched over separated network. In this way, high volume of data traffic can be optically switched by node because signaling traffic is not carried by dedicated, data plane links. However, signaling information is processed by nodes and used to control the data plane transmission thanks to the control plane network.

In the case of in-band signaling, each control plane node and link has corresponding data plane node and data plane link. This architecture is called symmetrical. The GMPLS architecture with out-of-band signaling and physically separated functional planes can be symmetrical and asymmetrical. An asymmetrical topology occurs when some functional planes have more or less connections than others. This situation is possible as a result of failure or network designer plan.

The reliability of the GMPLS network with separated functional planes is frequently considered an issue [Li et al. 2002, Perello et al. 2007]. Especially, the reliability of the control plane is discussed and the influence on the data plane transmission is considered.

The paper [Rozycki et al. 2007] prepared by authors discusses some aspects of the failure detection and notification, also for different types of architecture and applied mechanisms. Similar research described in [Komolafe et al. 2008] shows the impact of control messages lost to the data plane of the GMPLS network.

In this paper the problem of impact of the control plane topology to the quality of services offered by the GMPLS network is considered. In order to verify this influence, simulation experiment is proposed. The results presented in this paper are closely connected with results shown in [Rozycki and Korniak 2008] and they are enhanced by new results referred to the influence of signaling procedure used in the protection mechanisms. The results of this research are presented in section 3.

2 Quality of Services

2.1 Measurement of Quality of Services

The quality of services can be measured by many ways. The most important parameters, useful and easy in interpretation for measure of quality of service are [Cholda 2005]:

- *Mean Time To Failure* (MTTF) – defined as mean period of time when device is working without failure;

- *Mean Time Between Failures* (MTBF) – defined as mean period of time between failures;
- *Mean Time to First Failure* (MTFF);
- *Mean Time To Recovery* or *Mean Time To Repair* or *Mean Time To Restoration* (MTTR) – defined as mean period of time needed to repair failure or recover service;
- probability of failure p ;
- availability A (or unavailability defined as $U = 1 - A$)

An availability expresses that system preformed desired functions. This parameter depends on MTTF, MTBF and MTTR according to the following formula:

$$A = \frac{MTTF}{MTTF + MTTR} = \frac{MTTF}{MTBF} = \frac{MTBF - MTTR}{MTBF} = 1 - \frac{MTTR}{MTBF} \quad (1)$$

There are two ways to increase availability. One is to maximize MTTF and the second is to minimize MTTR. However, in this paper recovery after failure is mentioned and a method improving recovery time is proposed. Therefore MTTR parameter is selected as primary parameter used to measure QoS in the simulation experiments described in next sections.

2.2 The Protection as the Method for QoS Guarantee

Traffic engendering is powerful tool offered by GMPLS and supported protocols like RSVP-TE. One of traffic engendering crucial for optical networks is protection. This mechanism implemented in the GMPLS includes the following steps:

- failure detection,
- failure localization,
- hold off,
- notification,
- recovery operation,
- traffic recovery (switchover to backup),
- control plane state recovery.

Each step of protection procedure takes some time and influence overall protection delay. The *MTTR* parameter depends on this protection delay. There is an issue to minimize these delays. Most of these delays depend on used protocols and procedures in the control plane. The exception here is the hold-off time and switchover time (dependent mainly on techniques used in lower layers, e.g., the data link layer, and used protection mechanism).

Separation of the planes implies complication of implementation of some mechanisms, especially in the case of the asymmetrical architecture. For example, the typical method of failure detection implemented by exchanging RSVP *Hello* messages between neighbors is not sufficient. There are possible situations where the control plane nodes are separated by other host. In this case hello mechanism

cannot because *TTL* in Hello packets are set to 1. Therefore, this step of protection mechanism should be performed in other way, probably implemented as kind of in-band signaling or monitoring of forwarding process.

The separation of functional planes causes a new possibilities especially for notification and protection enhancement. The delay related to the failure notification, sent to the ingress node, depends strongly on the control plane topology and used technique of notification. The RSVP-TE protocol is preferred method for Label Switched Path (LSP) maintenance. This protocol offers two methods of notification: based on the *RSVP Notify* message and based on the *RSVP Path_Error* message. The *RSVP Path_Error* message must be sent along the LSP and processed by each RSVP-TE node. The *RSVP Notify* message may be transferred without reference to LSP. Thus, the *RSVP Notify* message may be sent over any shortest path to the destination. Moreover, additional link in the control plane can potentially decrease a notification delay. In this way the period of time needed from failure detection to switching a failed connection to backup one can be decreased. In consequence, time to restore connection and service associated with this connection (*TTR*) is also shorter. Similarly, a lack of link between selected nodes, in the control plane, may seriously increase notification delay.

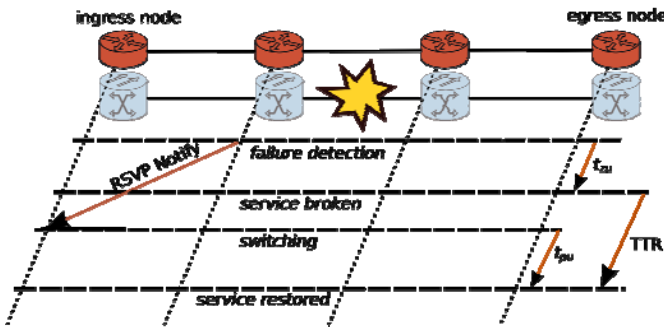


Fig. 1 Simple protection switching mechanism

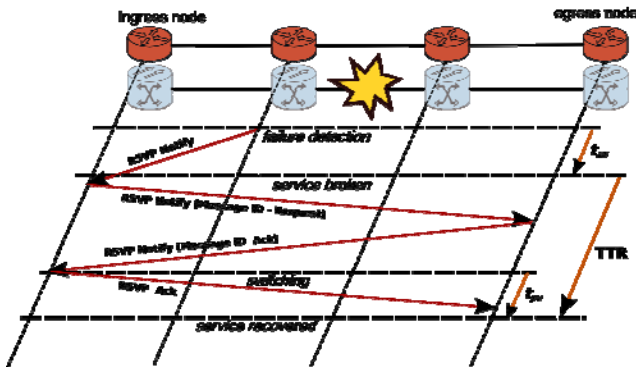


Fig. 2 Protection switching mechanism suggested for GMPLS

Besides of notification method, also other aspects of protection implementation may be important for QoS in general and for reliability in particular. Two of them are considered in this paper. The first one is the mechanism of reestablishing protection paths in the case of backup LSP failure. This feature is important especially in the case of multiple failures.

The second issue considered in this paper is a signaling procedure used during protection switching. The simplest, traditional signaling used in the protection is shown in Fig. 1. The switching to the protection resources is made by the ingress node immediately after receiving RSVP *Notify* message. t_{zu} denoted in figure is the period of time between the moment of failure and the moment of lost service detected by the end user (egress node), and t_{pu} is the period of time between the moment of switching to the backup LSP and the moment of service recovery. The relation between these delays and *TTR* parameter is shown in Fig. 1.

This type of switching, however, is not sufficient in the context of some GMPLS features such as functional planes separation and optical interfaces support. In this case, in order to agree some parameters of protection additional handshake (signaling messages exchange) between the ingress and the egress nodes is required. Such mechanism is suggested in RFC4426 and implemented in RSVP for end-to-end protection RFC4872 and segment protection RFC4873. This kind of protection signaling for end-to-end protection is shown in Fig. 2 but the same mechanism is used for the segment protection. The impact of selection of the protection signaling is bigger in the case of the end-to-end protection because protection paths are much longer than in the case of the segment protection.

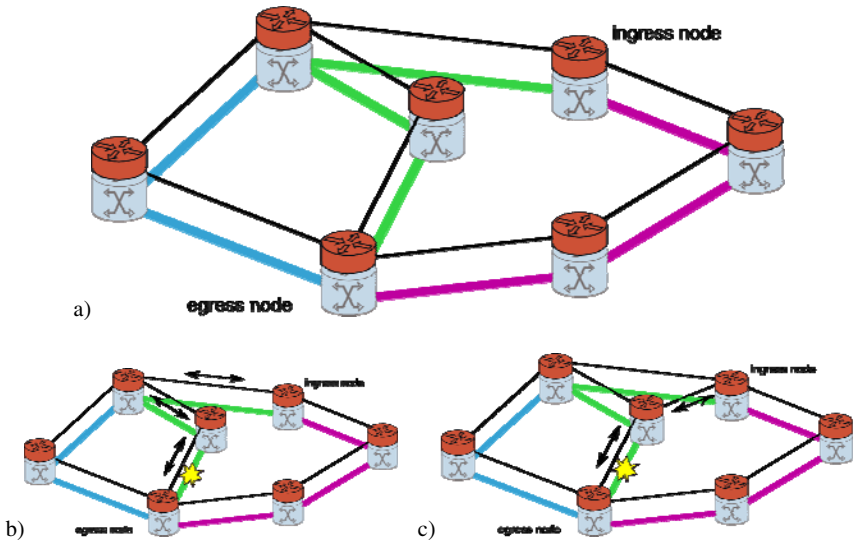


Fig. 3 Impact of control plane topology on GMPLS reliability: a) topology, b) protection signaling for symmetrical architecture, c) protection signaling for asymmetrical architecture

The last issue considered in this paper is impact of control plane topology on reliability of the GMPLS network and services supported by this network. As an example, consider topology presented on Fig. 3a. The primary LSP is assigned as green, and backup LSP is assigned as violet. In the case of failure a signaling path depends on the topology of the control plane network. If this topology is the same as topology of the data plane (Fig 3b) the protection signaling messages, notification and handshake, are transmitted usually along primary LSP. If, however, exists any additional link in the control plane such that exists shorter path to the ingress node (Fig 3c) the RSVP *Notify* messages are transmitted along this path that means shorter notification delay. The additional link However, addition link is not so important if does not provide shorter path between ingress node and egress node.

In the next section the influence of the control plane topology and described above mechanisms to the quality of services offered by the GMPLS network is proved by simulations of selected topologies and *Max TTR* and *MTTR* parameters comparison.

3 Simulation Experiments

To quantitatively analyze of influence of selected architectures and mechanisms on the QoS the several simulation experiments have been prepared. Next subsections present ns2-based tool designed and developed by authors for modeling GMPLS networks with separated functional planes, topologies used in simulations and results.

3.1 Simulation Environment

The simulations are prepared with the use of the Network Simulator (ns2) simulation environment with build-in MPLS module (MNS), additional patches with RSVP-TE and OSPF-TE [Adami et al. 2005] implementations and extensions prepared by the authors. These extensions allow to simulate the GMPLS behavior with out-of-band signaling and the asymmetrical architecture. The following functionality has been added:

- LSP setup, modification and release procedures for out-of-band signaling,
- out-of-band link-state routing,
- end-to-end and segment protection mechanisms,
- control plane failure detection by RSVP Hello,
- notifications.

All implemented procedures have been tested and verified.

Architecture of GMPLS node implemented in used tool is presented on Fig. 4. It contains two nodes that cooperate with each other. The data plane node located based on MPLS node is responsible for user data transmission based on label switching. The control plane node is the IP node with *RSVP-TE Agent* which is responsible for signaling and *rtProtoLS Agent* which is responsible for routing.

The routing table that is co-located with the *rtProtoLS Agent* in the control plane node contains only routing data for the control plane network is used to reliable transmission signaling and routing messages. The similar *rtProtoLS Agent* with the routing table located in the data plane node is responsible for routing connections established in the data plane. Due to the functional plane separation all routing and signaling messages should be transmitted in the control plane. The *rtProtoLS Agent* in the data plane cooperate, therefore, with *rtProtoLS Agent* in the control plane to transmit the data plane routing updates through the control plane network.

The *RSVP-TE Agent* is responsible for exchanging of the signaling messages and manages data plane components setting MPLS switching tables (LIB table, PFT table and ERB table), executing access control policy and resource management.

Some results of simulations prepared by this tool have been presented in [Rozycycki at al. 2007; Rozycycki and Korniak 2008].

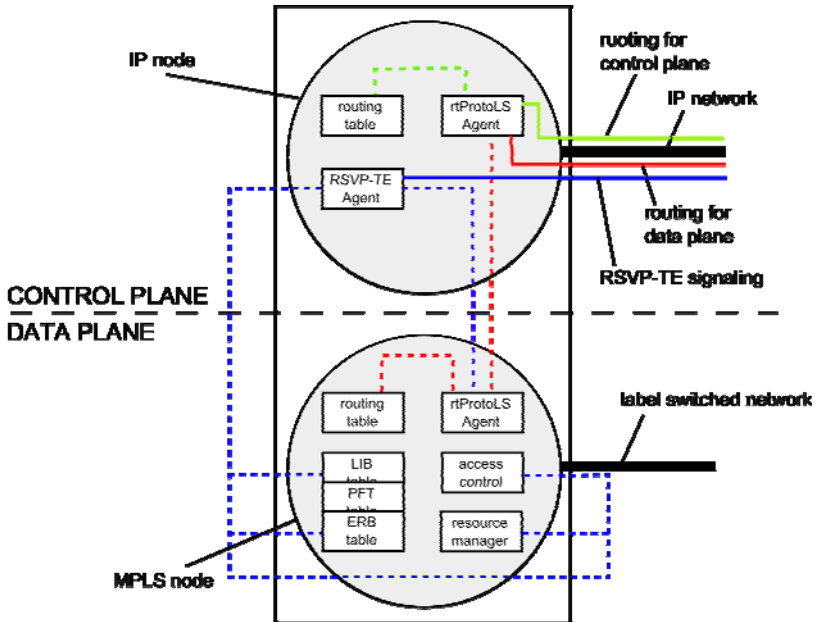


Fig. 4 Architecture of node implemented in the simulation tool

3.2 Network Topology

All simulations presented in this paper are based on GMPLS networks with separated control and data planes. The topologies of the simulated networks are presented in Fig. 5 and Fig. 6. In both cases the network contains 24 nodes: 16 core nodes and 8 access nodes. The nodes in the data plane are interconnected with bidirectional links of 155.52 Mb/s and 2 ms delay. Nodes in the control plane

are connected with 2 Mb/s bidirectional links with 4 ms delay. In particular scenarios the control plane have been configured as symmetrical or asymmetrical.

Note that all control plane topologies and their labels used in this paper are the same as in [Rozycki and Korniak 2008].

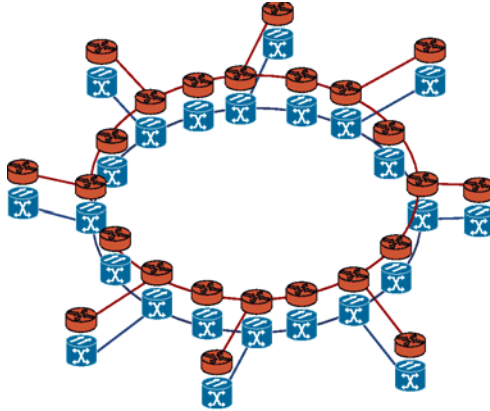


Fig. 5 Ring network topology used in the simulations

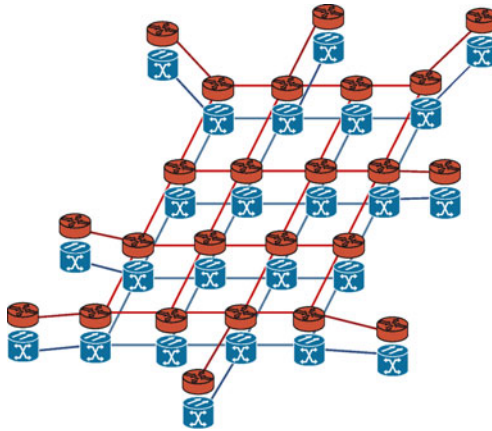


Fig. 6 Mesh network topology used in the simulations

3.3 Scenarios

To examine the protection signaling methods for both MESH and RING topologies several scenarios have been prepared with respect of the following parameters:

- control plane topologies defined in table 1;
- implemented protection signaling described in section 2.2.

For better comparison of described mechanisms, the same failure and heavy traffic pattern for all scenarios has been used. This pattern contains 10 randomly selected failures during 200 s of simulation (one failure per 20 seconds) and 800 randomly selected connections with dedicated 1:1 end-to-end backup (80 connections per failure).

The following parameters are measured:

- *MTTR* as average of *TTR* (Time To Restore) for each restored connection with 1 ms resolution;
- maximum *TTR* for each variant of scenario.

Table 1 Scenarios used in simulations

Topology	Architecture	Description	No of failures
RING	symm	symmetrical architecture - Figure 5 one link added - Figure 7a	1
	1 link ad	link added - Figure 7a	1
	2 links ad	two links added - Figure 7b	1
	2 Clinks ad	two links added - Figure 7c	1
	4 links ad	four links added - Figure 7d	1
	8 links ad	eight links added - Figure 7e	1
MESH	symm	symmetrical architecture - Figure 6	variants: 1, 2, 4
	2 links del	one link added - Figure 8a	variants: 1, 2, 4
	6 links del	two cross links added - Figure 8b	variants: 1, 2, 4

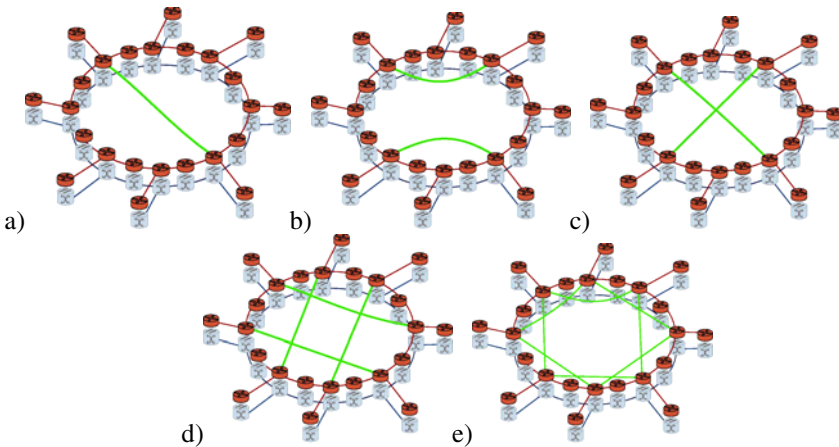


Fig. 7 Architectures used in simulation experiment for RING topology

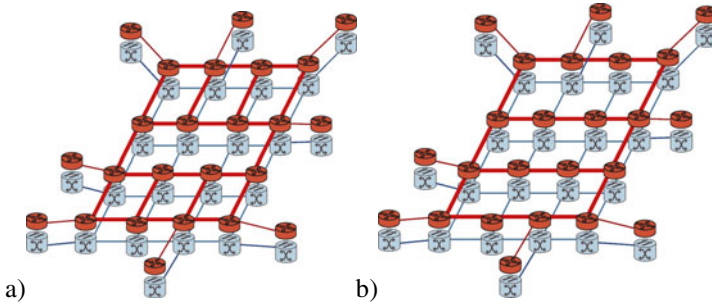


Fig. 8 Architectures used in simulation experiment for MESH topology

3.4 Results

The results of simulations presented in [Rozycki and Korniak 2008] confirm the better (shorter) the *MTTR* parameters for control plane topologies with additional links. It is important however that it should be links connecting an accurate nodes. For example, it should not be an adjacent nodes, and on the other hand is should not be a nodes located too far. In general, the additional control plane link may improve protection of LSP between given nodes but this link has no influence on protection of any other LSPs (between any other nodes). That is why nodes to be interconnected in the control plane should be carefully selected. Note that the case with 8 additional links for RING topology has insignificantly better result with probably much higher costs of that topology than topologies with 4 additional links for the same topology.

The similar results have been achieved for MESH topology. Removing of 2 links in the control plane have no influence on the *MTTR*, but even removing of 6 links have a limited influence on analyzed parameters. Each removed link interconnect nodes for which there are relatively short alternative paths (across two nodes – 3 times 4 ms – instead of directed 4 ms connection), the signaling delay is not too higher and in the consequence the *MTTR* is similar. Note that discussed results have been achieved for light traffic pattern when control plane nodes are not congested. The results for heavy traffic pattern are shown in the next part of this section.

Table 2 Restore ratio

No. of failures	Architecture	RR with H.P	RR without H.P
2	Symm	0.95	0.83
4	Symm	0.73	0.63
2	2 links del	0.92	0.86
4	2 links del	0.79	0.64
2	6 links del	0.89	0.88
4	6 links del	0.81	0.65

Table 3 Protection signaling mechanisms – RING topology (all results in milliseconds)

Architecture	Traditional		RFC4872	
	MTRR	Max TTR	MTRR	Max TTR
symm	111.2	422.0	414.3	636.0
1 link	105.9	403.0	347.1	575.0
2 links ad	100.0	402.0	344.2	581.0
2 Clinks ad	107.8	404.0	304.1	591.0
4 links ad	100.6	407.0	240.5	547.0
8 links ad	90.6	407.0	222.6	483.0

Table 4 Protection signaling mechanisms – MESH topology (all results in milliseconds)

Architecture	Traditional		RFC4872	
	MTRR	Max TTR	MTRR	Max TTR
symm	52.8	196.0	127.8	301.0
1 links del	69.0	310.0	150.7	484.0

In the cases with more than one failure it is possible that the backup connection is failed before primary connection fails. The possibilities of using mechanisms to reestablish the backup connection is important in this case. Such mechanisms, called Hold Protection, have been used in presented simulations. Table 2 shows the results of simulations (Restore Ratio) for MESH topology with multiple failures (with and without Hold Protection). As expected, in each case more failures cause the decreasing of Restore Ratio. However this parameter is higher if Hold Protection mechanism is enabled.

The comparison of *MTRR* for protection signaling mechanisms for RING topology and MESH topology are shown on Table 3 and Table 4, respectively. Due to heavy traffic pattern and more congested control plane network, achieved results are much higher than results achieved in simulations presented in [Rozycki and Korniak 2008]. This is caused by large number of supported connections and filled buffers on control plane routers, especially at the moment of failure. As expected, the additional signaling messages exchange in the protection signaling suggested for GMPLS is a reason that the *MTRR* parameter is much bigger and more dependent on architecture of the control plane. Note that for RING topology and traditional protection signaling the *MTRR* is decreased only around 20 percent in the case of 8 additional links while for RFC4872 protection signaling this parameter is decreased almost 50 percent. Moreover, the *Max TTR* is clearly decreased only for RFC4872 protection signaling and for traditional protection signaling is almost constant and not depend on architecture. Similarly, modification of the control plane for MESH topology has much more impact on reliability parameters in the case of RFC4872 protection signaling.

4 Conclusions

Presented results confirm the influence of the control plane topology to the quality of services offered by the GMPLS network. In the ring topology the additional links in the control plane can significantly improve protection mechanism and therefore the quality of services. This improvement is especially significant for signaling procedure of protection dedicated for the GMPLS network described in RFC4872 and RFC4873. In the mesh topology the less impact of topology modification on the quality of services is confirmed.

Moreover, not only the number of additional links is significant, but strongly important is also which nodes will be directly connected. Selection of the control plane nodes to be interconnected, therefore, should be a key issue in the context of the control plane network design and should be considered to develop algorithmic procedures.

References

- [Adami et al. 2005] Adami, D., et al.: Signalling protocols in diffserv-aware MPLS networks: design and implementation of RSVP-TE network simulator. In: IEEE GLOBECOM 2005, St. Louis, MO, USA (2005)
- [Cholda 2005] Cholda, P.: The reliability analysis of recovery procedures in GMPLS-based optical IP networks. PhD Thesis, Krakow (2005)
- [Komolafe et al. 2008] Komolafe, O., Sventek, J.: Impact of GMPLS control message loss. *J. of Lightwave Technology* 26(14), 2029–2036 (2008)
- [Jajszczyk 2005] Jajszczyk, A.: Automatically switched optical networks: benefits and requirements. *IEEE Communication Magazine* 43(2), S10–S15 (2005)
- [Li et al. 2002] Li, G., Yates, J., Wang, D., Kalmanek, C.: Control plane design for reliable optical networks. *IEEE Communication Magazine*, 90–96 (2002)
- [Manie et al.2004] Mannie, E., et al. (eds.): Generalized multi-protocol label switching (GMPLS) Architecture, RFC3945 (2004)
- [Perello et al. 2007] Perello, J., Spadaro, S., Comellas, J., Junyent, G.: An Analytical study of control plane failures impact on gmpls ring optical networks. *IEEE Communications Letters* 11(8), 695–697 (2007)
- [Rozycki et al. 2007] Rozycki, P., Korniak, J., Jajszczyk, A.: Failure detection and notification in GMPLS control plane. Presented at the Workshop on GMPLS Performance Evaluation: Control Plane Resilience, Glasgow (2007)
- [Rozycki and Korniak 2008] Rozycki, P., Korniak, J.: Influence of the control plane architecture on QoS in the GMPLS network. *IEEE Human System Interaction* (2008)

On Social Semantic Relations for Recommending Tags and Resources Using Folksonomies

A. Dattolo, F. Ferrara, and C. Tasso

Artificial Intelligence Lab, Department of Mathematics and Computer Science,
University of Udine, I-33100 Udine, Italy
{antonina.dattolo, felice.ferrara, carlo.tasso}@uniud.it

Abstract. Social tagging is an innovative and powerful mechanism introduced by social Web: it shifts the task of classifying resources from a reduced set of knowledge engineers to the wide set of Web users. However, due to the lack of rules for managing the tagging process and of predefined schemas or structures for inserting metadata and relationships among tags, current user generated classifications do not produce sound taxonomies. This is a strong limitation which prevents an effective and informed resource sharing; for this reason the most recent research in this area is dedicated to empower the social perspective applying semantic approaches in order to support tagging, browsing, searching, and adaptive personalization in innovative recommender systems. This paper proposes a survey on existing recommender systems, discussing how they extract social semantic relations (i.e. relations among users, resources and tags of a folksonomy), and how they utilize this knowledge for recommending tags and resources.

1 Introduction

Social Web applications provide users with a set of tools for creating, sharing, and promoting new content: users can easily leave the role of passive consumers of resources and become active producers (prosumers) of knowledge. This approach increases both the information on the Web and the number of available resources. Consequently, the growing number of resources prevents an effective access to them: a user needs to read the content of each resource for evaluating whether it is interesting for her.

An effective classification of the resources could greatly improve the access to knowledge. Although the manual process usually reaches high quality levels of classification for traditional document collections, it does not scale up to the enormous size of the Web, both in terms of cost, time, and expertise of the human personnel required [Dattolo et al. 2010].

In order to overcome this limitation, researchers proposed automatic classification tools based on ontologies, which add a semantic layer to the classification

process. But, these tools are domain dependent due the obvious difficulties to build and maintain universal ontologies covering all possible information needs.

According to Mathes¹, a possible, cheap, and domain independent solution is provided by social tagging applications, which are not constrained to a specific informative domain and distribute the task of classifying document over the set of Web 2.0 users. While approaches based on ontologies use semantic information defined by knowledge engineers, in social tagging systems semantic relations emerge from the classification process exploited by Web 2.0 users, that tagging resources generate folksonomies. This means that on one hand people can freely choose tags in order to classify resources, on the other hand meaningful relations (socially defined) between pairs of tags can be extracted by analyzing the aggregated mass of tagged content.

The tagging activity does not require significant efforts since users can associate tags to resources without following specific rules: each user applies her personal classification which then can be used by others to find resources of interest. For this reason, social tagging applications have both private and public aspects [Golder and Huberman 2006]: users may apply tags for personal aims (typically they associate labels to resources in order to find them again), or they can enjoy/exploit the classification applied by other users and browse related documents.

However, due to the freedom of social tagging systems the classification process is not rigorous. This means that the classification proposed by a user may not be useful to other users and, for this reason, tools able to adapt and personalize the access to knowledge embedded in social tagging systems are fundamental to allow users to access information in a highly effective way.

In particular, it is assumed that in order to simplify the access to information in folksonomies, the following set of recommendation tasks should be addressed:

- *Users profiling.* Given a user, create a model to describe her interests according to her tagging activities. This is the basic task for being capable to provide personalized services.
- *Finding similar people.* Given a user, find a community of people with similar interests.
- *Finding similar resources.* Given a resource, find similar items with similar features (referring the same topic or informative context).
- *Finding domain experts.* Given a resource or a set of tags, find people who classify and share relevant information in a specific topic. They can help a user to locate resources related to her interests.
- *Supporting browsing.* Suggest tags for refining the search of contents according to a given information need.
- *Tag recommendation.* Given a resource, find a set of tags, which classifies the resource in a personalized or not personalized way.
- *Content recommendation.* Given a user, filter resources according to her user profile.

¹ <http://www.adammathes.com/academic/computermediatedcommunication/folksonomies.html>

The main aim of this paper is to present the state of the art related to tag and content recommendations. In order to face these tasks, the approaches proposed in literature basically exploit two phases: (a) mining social semantic relations (i.e. similarities among users, resources, and tags) analyzing socially annotated resources; (b) computing recommendations by means of social semantic relations.

So, this paper organizes the description of the state of the art describing first current techniques to extract social semantic relations from a folksonomy, and then, presenting methods to compute tag and content recommendations by means of social semantic relations.

More specifically, the rest of this paper is organized as follows: Section 2 introduces the reader to social tagging and recommender systems, while knowledge representation and data mining techniques for extracting social semantic relation from folksonomies are described in the Section 3; Section 4 and Section 5 deepen the discussion on the use of social semantic relations for recommending respectively resources and tags. Final considerations and a look to the future conclude the paper.

2 Background

In this section we present an overview of social tagging and recommender systems and describe how users apply tags, what are the limitations connected to the tagging process, and how recommender systems can be classified.

2.1 Social Tagging Systems

By using social tagging systems users share resources within a community, upload them, and mainly introduce personal classifications, applying on them, specific tags.

A *tag* is a term freely chosen by a user as significant for a resource; it represents a metadata describing the item; so it can be useful as a keyword to identify or to find again later a document. Tags are also the main mechanism used to browse and search new resources in social tagging systems. The collection of all the tag assignments performed by a user constitutes her *personomy*, while the collection of all personomies, present in a system, is called *folksonomy*.

Folksonomies [Dattolo et al. 2010] substitute traditional hierarchical taxonomies: while taxonomies are defined by a selected set of experts which categorize resources following a strict hierarchical predefined schema, folksonomies are flat spaces of keywords freely applied by communities of users. Thanks to the systematic work of experts, taxonomies are more rigorous than folksonomies because the classification is based on a well-defined vocabulary. On the other hand, users contributing to a folksonomy are free to add tags without using terms from a specific predefined vocabulary: this allows users to possibly use more than just one term for associating a same concept to a resource, providing in such a way a potentially very rich content to folksonomies.

Taxonomies are expensive because they require a systematic work by experts, which have to follow a well-defined set of procedures and rules. On the other hand, folksonomies are cheap because the work is distributed among Web 2.0 users.

However, the freedom associated to folksonomies causes some limitations, which may hinder an effective classification of resources:

- Due to the absence of guidelines, constraints, and control, users can exploit the same tag in different ways: for example, acronyms are a potential cause of *ambiguity*, or the same tag may be written using *different lexical forms* (e.g. ‘photo’, ‘photos’, ‘web20’, ‘web_2’, ‘Web-2.0’).
- It is frequent to find *synonymy*, i.e. different words which describe, more or less, the same concept, or *polysemy*, i.e. single words associated to various different meanings.
- Users classify documents using *different levels of expertise* and *specificity*. Since relations among tags are not defined, it is difficult to understand when distinct tags are referring the same concept.

Nevertheless, tags contain rich and potentially very useful, social/semantic information, and their nature can be understood by analyzing motivations/goals that usually lead a user to perform tagging [Dattolo et al. 2010; Golder and Huberman 2006]. Common purposes are:

- *Describe the content.* Tags may be used for summarizing the content of a resource.
- *Describe the type of the document.* Some users utilize tags for identifying the kind of document. A document may be classified according to its MIME type (as, for example, ‘pdf’ or ‘doc’) or taking into account the publication form (as, for example, ‘article’, ‘blog’, ‘book’, ‘journal’).
- *Describe features and qualities.* Adjectives (such as ‘interesting’, ‘good’, and so on) may be used for expressing opinions, emotions, or qualitative judges.
- *Associate people to documents.* Tags can report the authors of a document or people involved in a particular task or event. Moreover, tags such as ‘my’, ‘my comments’, ‘mystuff’, and so on are used to define a relationship between the resources and the tagger.
- Associate events to documents. Locations, dates, conferences acronyms are widely used for associating an event to a document.
- Associate tasks to documents. Some tags, such as ‘mypaper’, ‘to read’, ‘job-search’ reveal personal matters or engagements.

These possible motivations should be considered together with the following two further factors:

1. *Heterogeneity of users.* Taggers have different levels of expertise and goals. This has several consequences: classifications exploited by some user may be not understandable (or acceptable) for other users; different users may describe the content of a resource using distinct vocabularies; different users may have

different opinions about a topic; users may not have knowledge about people, events, or tasks associated to a resource by other users.

2. *Temporal changes.* Users' knowledge, motivations, and opinions may change over time. A tag used today for describing an item can be useless in the future: emotions and opinions of people may change; reputation of people evolves; a topic may be not any more interesting to the user.

Currently, tags are mainly used in social networks, social bookmarking applications, and Web 2.0 document sharing systems. Social networks (both general purpose ones, like Facebook or domain-specific ones, such as aNobii), allow users to apply tags for expressing opinions and for defining relationships among resources and people. Social bookmarking applications, such as Delicious, extend traditional bookmarking tools allowing users to upload, label, and access bookmarks from each computer connected on the Web, simplifying the process of content sharing among peers. Finally, Web 2.0 document sharing systems allow users to upload and share file with other peers. Remarkable examples of these systems are Flickr for photo sharing, YouTube for video sharing, Last.fm for music sharing, and BibSonomy for publication sharing. However, it is known that some authors propose a taxonomy of these applications, and classify applications according to *tagging rights* (who is allowed to tag), *tagging support* (what facilities are provided to simplify the tagging process), and *support to social interaction* among users.

2.2 Recommender Systems

The increasing volume of information on the Web is the main motivation for recommender systems: they support users during their interaction with large information spaces, and direct them toward the information they need; these systems model user interests, goals, knowledge, and tastes, by monitoring and modelling the feedback provided by the user. Such user feedback can be acquired by using appropriate ratings that quantify a relation between the user and an item: the ratings may be explicit, when they require the user evaluation, or implicit when they are automatically generated by the system in terms of measures, such as, for example, the time spent by a user on a Web page. By taking into consideration the ratings provided by a user, a recommender system defines a personalized order of importance for the set of available resources.

Several classifications of recommender systems have been proposed in the literature according, for instance, to the type of data that the user profile includes (e.g. demographic recommender systems) or the data structure used to represent the user profile (e.g. graph-based recommender systems). However, recommender systems can be classified into three classes of systems, on the basis of the algorithm utilized to produce recommendations: collaborative filtering, content-based, and hybrid recommender systems.

1. *Collaborative filtering recommender systems* filter resources using the opinions of other people; in turn, they may be differentiated in two approaches:

- *Model-based approaches*, which build a probabilistic model for predicting the future rating assignments of a user, on the basis of her personal history.
 - *Memory-based approaches*, which use statistical techniques for identifying users with common behaviour (user-based approaches) or items evaluated in a similar way by the community (item-based approaches). In particular, user-based approaches look for people, called neighbours, similar to a given user, and then combine neighbours' feedbacks for generating a list of recommendations. On the other hand, item-based approaches look for resources similar to that the user liked, i.e. resources judged similarly by the community.
2. *Content-based recommender systems* analyze the past user activities looking for resources she liked; they model resources by extracting some features (for example, topics or relevant concepts) from documents. The user profile is then defined describing what features are interesting for the user. The relevance of a new resource for a user is computed by matching a representation of the resource to the user profile.
 3. *Hybrid recommender systems* combine the results produced by collaborative and content-based recommender systems.

Two main recommendation tasks can be identified: (a) recommending content (i.e. suggesting documents, references, or URL's) and (b) recommending tags. In order to fulfil these tasks, the approaches proposed in the literature analyze and extract social semantic relations from folksonomies.

Next Section 3 provides a description of the techniques used to mine social semantic relations, while Sections 4 and 5 show how these relations are used to support, respectively, content and tag recommendation.

3 Mining Social Semantic Relations

Social tagging systems merge personal and social perspectives: the personal perspectives are embedded in personomies while social ones come from the union of all personomies; for this reason, personomies and folksonomies offer two distinct levels for mining social semantic relations.

3.1 Data Mining in a Folksonomy

A folksonomy is defined on a ternary relation which maps the tagging activities of all users: for each user, the ternary relation stores information about which tags have been applied on which resources. The ternary relation, which involves users, tags, and items, is the starting point to model knowledge, relationships and similarities in a folksonomy. However, mining similarities is not trivial because the ternary relation merges relations among objects of the same type as well among objects of different types. Two approaches have been proposed to handle this scenario:

- projecting the 3-dimensional space into lower dimensional ones [Dattolo et al. 2011];
- modelling the ternary relation by a 3-order tensor.

The projection of the ternary relation into two-ways relations (throwing away information about just one dimension) allows the system to extract the following three different matrices:

1. The *User-Resource (UR)* matrix. It describes the two-way relation between users and resources. Each row of this matrix is associated to a user which is described by a binary vector: if the user u tagged the resource r then the cell $UR(u,r)$ is set to 1 (0 otherwise).
2. The *Tag-Resource (TR)* matrix. It describes the two-way relation between tags and resources. Each row of the matrix, associated to a tag, is a vector, which counts how many times a tag has been applied on each resource.
3. The *User-Tag (UT)* matrix. It describes the two-way relation between users and tags. Each row of the matrix, associated to a user, is a vector, which counts how many times a user applied each tag.

These matrices describe relations among set of heterogeneous objects. Several notions of similarity between pairs of objects of the same type can be inferred by comparing two rows or two columns of the *UR*, *TR*, and *UT* matrices. The cosine and the Pearson similarities are commonly used to assess the similarity between two vectors. By means of this approach, given a **pair of users**, we can compute:

- *UR_user_sim*. Extracted from the *UR* matrix, this measure shows how much two users are similar according to the number of shared resources.
- *UT_user_sim*. Computed from the *UT* matrix, this measure specifies that two users are similar if they show a similar tagging behaviour.

Given a **pair of resources** we can infer:

- *UR_resource_sim*. Computed from the *UR* matrix, it states that two resources are similar if they have been tagged by the same set of people;
- *TR_resource_sim*. Inferred from the *TR* matrix, it defines two resources as similar if they have been tagged in a similar way.

Finally, given a **pair of tags** we can infer:

- *TR_tag_sim*. It is calculated from the *TR* matrix and states that two tags co-occurring frequently on the same resources share a common meaning;
- *UT_tag_sim*. We report this similarity just for the sake of completeness, as it is not really significant. It is computed from the *UT* matrix and states that two tags, used by the same user, share a common meaning. However, users may have several distinct interests and for this reason they may use tags which are not in any relation.

Unfortunately, the *UT*, *UR* and *TR* matrices used to discover similarities are sparse since each user labels only a small subset of all available resources and use only few tags. This sparsity can reduce the effectiveness of the methods developed to

find social semantic relations from these matrices: for instance, *UR_re-source_sim* cannot be used to compare users who did not label the same resources.

Similarities inferred from the *UT*, *UR*, and *TR* matrices can be used to produce the *User-User (UU)* matrix, the *Resource-Resource (RR)* matrix and the *Tag-Tag (TT)* matrix in order to store respectively similarities between pairs of users, resources and tags. These matrices can be used to overcome the computational overhead needed to derive similarities in online scenarios and they represent also the starting point to develop graph-based mechanisms for extracting relevant information from a folksonomy. For instance, the *TT* matrix describes a graph where each node represents a tag and an edge connects two tags only if the similarity between them is greater than a certain threshold. For example, the PageRank algorithm and the HITS algorithm extract authoritative tags (i.e. tags semantically relevant) from this graph for a given set of input tags.

Similarly, the *RR* and the *UU* graphs can be built (using respectively similarities between resources and users) and then explored to discover new resources and new users for a given seed of resources or users.

The similarities among pairs of objects of the same type can be used to group together tags, users, and resources with similar properties. This task can be exploited, for instance, in order to create clusters of tags with a similar meaning, people with shared interests, or resources related to same topics or contexts.

Obviously, data mining techniques based on the projection of the 3-dimensional space into lower dimensional spaces lose some information. A different approach to model the ternary relation is to model the 3-dimensional space by a 3-order tensor. The HOSVD method, generalizes the SVD method to high dimensional spaces, and has been experimented to discover latent semantic association among users, tags, and resources.

3.2 Data Mining in a Personomy

A folksonomy collapses all users activities by combining all personomies, which include different personal interests and tagging strategies. On the other hand, a personomy contains information about just one user and can be analyzed to extract knowledge about the semantic relations that the user built during her tagging activities. More specifically, a personomy can be represented by a *Personal-Tag-Resource (PTR)* matrix, which stores information about how the user applied tags on resources. Starting from *PTR* matrix the *Personal-Tag-Tag (PTT)* matrix can be built and analyzed to find patterns in the user tagging activities. This matrix describes a co-occurrence graph where each node represents a tag and a weighted edge connects two tags only if the user applied these tags together. The weight associated to each edge is directly proportional to the number of times the two tags have been used together.

Graph clustering algorithms can be used to detect patterns in the user tagging strategy grouping sets of tags usually applied together to describe items: distinct group of tags can therefore reveal that the user is interested in different and disjoint topics.

4 Recommending Resources Using Tags

A system can recommend resources whenever it discovers some relevant ones (for example sending an email to the user) or whenever it receives a specific request by the *active user* (for example, a query). We call this approach *tag-aware recommendation*.

Given a query, the simplest approach is to give higher relevance to resources labelled by a large set of tags used by the active user or if the resource has been often associated to one or more tags applied by the active user.

In this way, popular resources become also the most relevant; however, although popularity is a good mean for assigning confidence to results, other parameters should also be considered, such as, for example, previous activities or habits of the user (for instance, how she usually apply tags or what resources she visited in the past).

In other approach there is suggestion that tags are a useful mean for understanding the relationship between a user and one or more resources. Following this idea, recently, several researchers proposed some attempts for providing personalized recommendations.

The following two subsections describe collaborative and content-based strategies to recommend resources using tags.

4.1 Tag-Aware Collaborative Recommender Systems

Tag-aware collaborative recommender systems extend collaborative filtering techniques using tags to model user interests and to produce personalized recommendation. In this context, tags have been used to achieve two possible goals:

1. Extending the classical collaborative filtering approach using tagging history for calculating similarities among users. Calculating user similarities by tags, the recommender system assumes that people with similar interests usually apply the same or similar tags.
2. Detecting adaptive neighbourhoods according to a specific topic or context defined by one or more tags. Each user may be interested in several topics and then she could tag resources referring distinct informative contexts. In order to obtain higher accuracy a recommender system can consider only users interested in a specific topic (i.e. users which used tags related to the specific topic) and resources associated only to the topic (i.e. resources labelled by a specific set of tags).

Both memory-based and model-based collaborative filtering approaches, aimed at comparing tagging histories to find similarities among users, have been proposed. For instance, Social Ranking is a memory-based recommender method that, given a user and a set of tags, computes a personalized ranking of resources. More specifically, Social Ranking extends the set of input tags including other similar tags by means of the *TR_tag_sim*: in this way, it discovers relevant tags for the user.

Then, it calculates a score for a resource according both to the relevance of tags associated it and to the UT_user_sim calculated between the active user and the other users who tagged the specific resource.

Alternatively, a model-based approach has been proposed in [Zhou et al. 2010], where the PTT matrix is considered to identify the distribution of user interests by clustering tags. Two distributions are then compared by means of the Kullback-Leibler divergence to assess the similarity among users.

TagiCoFi is another model-based recommender: it uses tags for facing the sparsity problem inferring some relationships among users and resources also if the users did not explicitly tagged the resource.

On the other hand, the idea of finding adaptive neighbourhood according to a topic or a context is exploited in two memory-based approaches described in [Nakamoto et al. 2007; Dattolo et al. 2009; Dattolo et al. 2011; Nakamoto et al. 2007], given a bookmark of a user, a *context* is defined by tags applied (by all users) on the specific resource. A context is used to filter documents and users: only users who applied tags in a given context and resources labelled by tags in the same context are considered to generate recommendations. In particular, the UR_user_sim is used to evaluate the relevance of other users for a given context. Then, the relevance of a resource depends on the relevance of users who bookmarked it.

In [Dattolo et al. 2009] the authors use tags to distinguish different topics of interest for the active user: this task is performed by clustering tags with similar meanings, identified by using the TR_tag_sim . A cluster of tags allows the system to split resources tagged by the active user into different collections associated to distinct topics: a *topic of interest* is defined by a set of similar tags (applied by the active user) and the set of resources labelled by these tags. Given a topic of interest, the UR_user_sim and the TR_tag_sim are used to compute the relevance of new resources. In particular, resources labelled by tags, which are evaluated as more similar to the tags included in the topic, are considered more relevant than other resources as well as resources bookmarked by users more similar to the active user are more relevant than others.

Finally, in [Dattolo et al. 2011] the authors firstly operate on the disambiguation of tags and tag sets, by taking into account their synonyms, homonyms, and basic level variations; then they use the results of the disambiguation process to enhance both search and recommendation: in fact, tags sharing the same semantics are merged into one, while ambiguous ones are split according to their different contexts.

4.2 Tag-Aware Content-Based Recommender Systems

Generally speaking, tag-aware content-based recommender systems use tags in order to go deeper into a semantics-based approach. More specifically, they exploit tags for modelling interests, classifying documents, and comparing document representation to user profiles.

Meaningful examples of this trend have been described in [Shepitsen et al. 2008; De Gemmis et al. 2008; Shepitsen et al. 2008] the authors describe a recommender system representing users on the rows of the UT matrix and resources on the

columns of the TR matrix. Tag clustering is used to group tags with similar meanings. Each cluster of tags can be seen as a bridge between users and resources; in fact, looking at the user profile, it is possible to understand what tag cluster is relevant for the user and, on the other hand, the description of resources is used to detect resources relevant for a specific cluster. The recommendation algorithm uses as input a tag, a user profile and tag clusters, and produces an ordered set of items. In order to generate a personalized order of items, it computes, for each tag cluster, a score that is associated to both the cluster and the resources labelled by tags in the cluster. More specifically, the score assigned to the cluster depends on the number of times the active user applied the tags in the cluster (this step allows to personalize results), while the score of a resource depends on the number of times users associated to it tags which are in the specific cluster. By using this information, the relevance of a resource for a specific cluster is computed as the product of the score assigned to the cluster by the score assigned to the resource. Finally, given a resource, its relevance is computed by summing the relevance of the resource over all tag clusters. In [De Gemmis et al. 2008], the authors present a different approach where both the textual description of items and tags are used to build the user profile. This approach uses the synsets of Wordnet, structures defined as sets of words with a similar meaning and used for defining a semantic indexing of documents. A disambiguation strategy associates a synset to each word in the document looking at words that precede and follow it. Similarly, tags are also disambiguated using the textual content of the resource. In this way, a document is defined as a bag-of-synsets in opposition to the classical bag-of-words. Using this descriptive model, a Bayesian classifier considers the resources bookmarked by the user in order to learn about the synsets, which are relevant to her. Matching the synset representation of documents with the synsets in the user profile, the recommender system calculates a relevance value for each resource.

5 Recommending Tags

Tag recommendation is the second task we consider in this paper. This task can improve the usage of social tagging applications in several ways:

- Tag suggestions can *increase the probability* that people will assign many tags to resources. Users can just select one or more suggested tags instead of devising from scratch to meaningful tags.
- Tag suggestions can *promote a common vocabulary* among users. Proposing a well-defined set of tags, it become possible to reduce the problems connected to the absence of both guidelines and supervised methodologies for the tagging process.

The set of tags to recommend can be selected taking in account just metadata associated to the items (such as tags applied by other users, relevant keyphrases extracted from the text) or integrating the analysis of previous user tagging activities. Following these criteria, tag recommender systems can be divided into two classes:

1. *Not personalized tag recommender systems.* These systems select for each document a set of meaningful tags, ignoring the specific user's tagging habits. In this way, different users will receive the same suggestions for the same resource.
2. *Personalized tag recommender systems.* These systems suggest the set of the most relevant tags for a resource according to the specific user and her personal way to classify resources.

5.1 Not Personalized Tag Recommendations

Not personalized tag recommender systems do not follow the traditional organization of recommender systems because they do not build and maintain a user profile. Suggested tags can be extracted both from the content of specific resources and using tags applied by the whole community.

When tags are extracted from the textual content of a resource, well known techniques from information retrieval, natural language processing, and machine learning for classifying documents are applied. These approaches split the content of a textual resource into short textual slots, named n -grams (a sequence of n words), and then assess the relevance of each n -gram according to some criteria. Many examples of this approach have been proposed in literature.

For example, the usage of the $tf*idf$ metric to assess the relevance of n -grams. The same $tf*idf$ metric is used also in KEA. It is based on a Bayesian classifier, to weight the terms, but KEA takes in account also their first occurrence. In order to filter the set of extracted keyphrases in an unsupervised and domain independent way, in [Pudota et al. 2010] the authors apply a POS (Part-Of-Speech) tagger. Then, the relevance of a keyphrase is computed according to the following set of features: frequency, first occurrence, last occurrence, and lifespan (the distance between the first and the last occurrence positions).

However, all these methods suggest only terms, which appear already in the document. For overcoming this limitation a semantic approach is needed. In [Baruzzo et al. 2009], the authors propose the use of ontologies. In this approach a set of keyphrases is extracted from the document and is used for browsing a domain ontology in order to find other, more abstract and conceptual terms. However, the performance of this approach depends on the quality of the available ontology.

User generated annotations can be also used to suggest tags. The simplest approach can suggest, for instance, the most popular tags for a resource. However, due to sparsity of social tagging systems there are resources tagged by only few people and for this reason more sophisticated methods have been proposed. Auto-Tag is a tag recommender system: it suggests tags for blog posts; this framework recommends tags following a three-step process: first, it selects resources similar to the starting document (according to the $tf*idf$ measure) by retrieving the tags associated to these resources; then, it associates a weight to each tag according to the number of times the tag has been applied to the set of similar resources; and,

finally, it suggests the top ranked tags. TagAssist² outperforms AutoTag thanks to a pre-processing phase, where the Porter's stemmer is used to compress the set of tags.

Other approaches consider that some users produce more meaningful and semantically rich classifications than others. FolkRank [Jäschke et al. 2006], for example, takes in account this feature by computing a ranking for users, resources, and tags through a PageRank-like algorithm. FolkRank models a folksonomy by a tripartite graph where tags, resources, and users are represented by three sets of nodes; edges link users to their tags and their resources, moreover, edges connect each resource to tags which have been used to classify the specific resource. The algorithm is based on the idea that a node of this graph is important if it is connected to many important nodes. So, the random surfer model of PageRank is used to spread weights over the tripartite graph in order to assign a weight for users, resources and tags.

5.3 Personalized Tag Recommendations

Personalized collaborative approaches evaluate the relevance of a tag considering the specific user tagging preferences.

Personalized collaborative strategies [Gemmell et al. 2009; Symeonidis et al. 2008] use people tagging strategies to detect the set of tags, which can be suggested to the active user.

In [Gemmell et al. 2009], the authors adapt the classical K -nearest neighbour algorithm to the task of generating a list of recommended tags: given a resource, a set of K neighbours is defined evaluating both the UR_user_sim and UT_user_sim over users which tagged the same resource. Tags assigned by similar users will be more relevant than others.

The ternary relation among tags, users, and items is modelled as a 3-order tensor in [Symeonidis et al. 2008]. Latent semantic analysis is performed on tensors to capture the latent association among users, resources, and tags. This approach builds a set of quadruplets $(u, r, t, likeliness)$ where each quadruplet describes the probability that the user u will tag the resource r with the tag t .

Personalized content-based strategies analyze the relationship between the content of a resource and the tags applied by the active user in order to predict tags for new resources. Examples of this approach are provided in [Basile et al. 2007] and in [Musto et al. 2009].

The system proposed in [Basile et al. 2007] uses a Bayesian classifier for each tag employed by the user. Each classifier is trained using the textual content of documents tagged by the specific tag. In this way the text of a new document can be used for evaluating whether a tag can be suggested for that document.

STaR (Social Tag Recommender System) [Musto et al. 2009] is based on an approach similar to AutoTag (Section 4.1). The main difference is that STaR provides personalized tag suggestions. This framework collects two sets of documents similar to a starting resource: the set containing resources tagged by the active user and

² <http://infolab.northwestern.edu/media/papers/paper10163.pdf>

the set containing documents tagged by other users. Tags applied by the active user are weighted according to the similarity of the tagged resources to the starting one. In a similar way, a weight is assigned to tags applied by the other users. Finally, the two sets of tags are merged and a ranking of the tags is computed as a linear combination of the two scores associated to each tag.

6 Final Considerations and Future Work

In this paper, we analyzed current methods for finding social semantic relations (i.e. similarities among users, tags, and resources) in folksonomies and then we showed the ways in which these relations have been used to develop tag recommender systems and content recommender systems.

In order to extend collaborative and content-based approaches, social semantic relations have been introduced: in fact, on one hand, collaborative approaches can find similarities among users by looking at their tagging habits. On the other hand, content-based approaches use tags to model resources and build a user profile and then suggest resources, which appear relevant for the specific user profile. But, both the approaches have limitations mainly due to the ambiguity of tags and, for this reason, more semantics-based approaches are needed.

Some recommender systems use clustering for fighting redundancy of folksonomies [Shepitsen et al. 2008], but these approaches also show some open issue: during the time a user may apply the same tag for expressing different concepts; a tag may be in several clusters. Again, a deeper understanding of tags and resources can facilitate the disambiguation of tags.

Other works attempt to tackle the absence of semantic relationships among tags by associating a context to them [Nakamoto et al. 2007; Dattolo et al. 2009; De Gemmis et al. 2008; Dattolo et al. 2011]. However, it is hard to extract a context from some generic tag, such as, for example, ‘to read’, ‘my paper’, ‘job’.

A chance for reducing ambiguity is to support the manual tagging process using tag recommender systems: users do not need to define a concise description of the resource but they can just select among the suggested tags. Not personalized strategies ignore the heterogeneity of users suggesting the set of tags, which appear more representative for a resource according to a global measure. On the other hand, personalized approaches tailor the selection of recommended tags taking into account the user’s past tagging activities.

However, there is not evidence in the literature that personalized tag recommendation approaches outperform not personalized strategies, or vice versa, that not personalized techniques improve the user satisfaction. The evaluation of all these recommender systems is still an open challenge: results of different systems have been evaluated using different dataset and following different evaluation methodologies and procedures.

Other interesting lines of research on tags are actually ongoing. An interesting use of tags have been proposed in [Vig et al. 2009], where the authors introduce the concept of tagsplanations which are explanations based on community tags. Explaining the motivations at the basis of a recommendation improves the user satisfaction. Tagsplanations take into account two components: tag relevance and

tag preference. Tag relevance refers to the representativeness of a tag for describing a resource, while tag preference defines the user's sentiment toward a tag. Another line of research is concerned with the extraction of basic semantic relations from folksonomies or the augmentation of social tagging with more ontology-like features [Baruzzo et al. 2009]. For example, Folk2onto [Sotomayor 2006] maps social tags (taken from delicious) to ontological categories (using a Dublin Core-based ontology) in order to classify and give a proper structure to the tagged resources. However, the task of associating semantic to tags, and extracting semantic relation among them is still far from a final solution.

References

- [Basile et al. 2007] Basile, P., Gendarmi, D., Lanubile, F., Semeraro, G.: Recommending smart tags in a social bookmarking system. In: Proc. of the Workshop on Bridging the Gap between Semantic Web and Web 2.0 the 4th European Semantic Web Conference, Innsbruck, Austria, pp. 22–29 (2007)
- [Baruzzo et al. 2009] Baruzzo, A., Dattolo, A., Pudota, N., Tasso, C.: Recommending new tags using domain-ontologies. In: Proc of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology, Milan, Italy, pp. 409–412 (2009)
- [Dattolo et al. 2009] Dattolo, A., Ferrara, F., Tasso, C.: Neighbor selection and recommendations in social bookmarking tools. In: Proc. of the 9th International Conference on Intelligent Systems Design and Applications, Pisa, Italy, pp. 267–272 (2009)
- [Dattolo et al. 2010] Dattolo, A., Tomasi, F., Vitali, F.: Towards disambiguating social tagging systems. ch.20, vol. 1, pp. 349–369. IGI-Global (2010)
- [Dattolo et al. 2011] Dattolo, A., Eynard, D., Mazzola, L.: An integrated approach to discover tag semantics. In: Proc. of the 26th Symposium on Applied Computing. Thungai University, Taiwan (2011)
- [De Gemmis et al. 2008] De Gemmis, M., Lops, P., Semeraro, G., Basile, P.: Integrating tags in a semantic content-based recommender. In: Proc. of the 2nd ACM International Conference on Recommender Systems, Lausanne, Switzerland, pp. 163–170 (2008)
- [Gemmell et al. 2009] Gemmell, J., Schimoler, T., Ramezani, M., Mobasher, B.: Adapting k-nearest neighbour for tag recommendation in folksonomies. In: Proc. of the 7th Workshop on Intelligent Techniques for Web Personalization and 21th International Joint Conference on Recommender Systems in Conjunction, Artificial Intelligence, Pasadena, California, USA, pp. 51–62 (2009)
- [Golder and Huberman 2006] Golder, S., Huberman, A.: The structure of collaborative tagging systems. *Journal of Information Science* 32(2), 198–208 (2006)
- [Jäschke et al. 2006] Jäschke, R., Hotho, A., Schmidt-Thieme, L., Stumme, G.: FolkRank: A ranking algorithm for folksonomies. In: Proc. of FGIR 2006, Hildesheim, Germany, pp. 111–114 (2006)
- [Musto et al. 2009] Musto, C., Narducci, F., De Gemmis, M., Lops, P., Semeraro, G.: STaR: a social tag recommender system. In: Proc. of the ECML/PKDD 2009 Discovery Challenge Workshop at the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, CEUR Workshop Proceedings, CEUR-WS.org, Bled, Slovenia, vol. 497 (2009)

- [Nakamoto et al. 2007] Nakamoto, R., Nakajima, S., Miyazaki, J., Uemura, S.: Tag-based contextual collaborative filtering. *IAENG International Journal of Computer Science* 34(2), 214–219 (2008)
- [Pudota et al. 2010] Pudota, N., Dattolo, A., Baruzzo, A., Ferrara, F., Tasso, C.: Automatic keyphrase extraction and ontology mining for content-based tag recommendation. *International Journal of Intelligent Systems, Special Issue on New Trends for Ontology-Based Knowledge Discovery* 25(12), 1158–1186 (2010)
- [Shepitsen et al. 2008] Shepitsen, A., Gemmell, J., Mobasher, B., Burke, R.: Personalized recommendation in collaborative tagging systems using hierarchical clustering. In: *Proc. of the 2nd ACM International Conference on Recommender Systems, Lausanne, Switzerland*, pp. 259–266 (2008)
- [Sotomayor 2006] Sotomayor, B.: *Folk2onto: Mapping social tags into ontological categories* (2006), <http://www.deli.deusto.es/Resources/Documents/folk2onto.pdf> (accessed October 29, 2010)
- [Symeonidis et al. 2008] Symeonidis, P., Nanopoulos, A., Manolopoulos, Y.: Tag recommendations based on tensor dimensionality reduction. In: *Proc. of the 2nd ACM International Conference on Recommender Systems, Lausanne, Switzerland*, pp. 43–50 (2008)
- [Vig et al. 2009] Vig, J., Sen, S., Riedl, J.: Tagsplanations: explaining recommendations using tags. In: *Proc. of the 13th ACM International Conference on Intelligent User Interfaces, New York, NY, USA*, pp. 47–56 (2009)
- [Zhou et al. 2010] Zhou, T., Ma, H., Lyu, M., King, I.: UserRec: A User Recommendation Framework in Social Tagging Systems. In: *Proc. of the 24th AAAI Conference on Artificial Intelligence, Atlanta, Georgia, USA*, pp. 1486–1491 (2010)

The Simulation of Malicious Traffic Using Self-similar Traffic Model

J. Kolbusz, P. Rozycki, and J. Korniak

Department of Electronics and Telecommunications, University of Information Technology and Management, Rzeszow, Poland
{prozycki, jkorniak}@wsiz.rzeszow.pl

Abstract. Detection of malicious activity in the network still is a challenge. The self-similarity feature of traffic can be used in an anomaly detection method. The influence of traffic generated by intruder who performs access attack is analyzed. In the other hands the simulation of threads is useful in designing and testing processes of a network. For this purpose a multi-layer 'on-off' model of traffic source is developed and a traffic generator is implemented according this model. Finally the real traffic including attacker flow is compared to the traffic generated by generator. This comparison proves that it is possible to simulate traffic similar to malicious one.

1 Introduction

Modeling of modern computer networks still is a challenge due to the integration of services, grow of bandwidth demand and more sophisticated flow control methods. Optimization, designing and research of the networks require to have appropriate network and the traffic flow models which reflect real infrastructure, used mechanisms and the pattern of traffic flow. Additionally traffic generated by malicious activity should be considered in that models.

Especially, modeling of traffic flow becomes more important for performance evaluation of computer networks. The reason of this importance is diversity growing of traffic flow as a result of services integration. Development of a new protocols and technology, optimization of them are ones of possible application of such models. Precise models are necessary in many phases of network design and further operation: in initial phase of network mechanisms design, and final tuning of network mechanisms. The specific applications, networks testing and security requirements are also motivation for improvement of network and traffic flow models. Simulation of network performance with the use of network model allows administrator evaluate expected reliability and service availability level.

In the paper [Mello et al 2007] is shown that old methods of modeling of traffic flow do not take into account some aspects. The use of that models generate results significantly different from the observed in real network.

Willinger, Taqqu, Sherman and Wilson [Willinger et al 1997] have proved that traffic flow has self-similar property. Anomaly detection is one of the application of the self-similarity effect of network traffic. The model of network traffic with self-similarity helps to simulate data flow not only for normal operation but also for malicious activity, malfunctioning devices and network overload. Analyzes of network traffic [Rohani et al 2009], [Cheng et al 2009] show significant change of self-similarity when such events occur.

A multi-layer ‘on-off’ model of traffic source is developed by authors and shown in the next sections. This model is used to implement traffic generator which reflects real traffic including attacker flow. The correctness of proposed model is verified experimentally by comparison real traffic with simulated one.

2 Background

2.1 Self-similarity

Since the beginning of the Internet, network traffic properties have been changed significantly. Many modern traffic analysis lead to the conclusion that there is a stronger correlation in the stream of events than previously observed. At different sampling time (milliseconds, seconds, hours) some correlations can be observed and described by the term of self-similarity. Self-similarity is a property commonly known from fractals. It means that an object appears the same regardless of the scale at which it is viewed. In a self-similar phenomenon, observation looks the same or process behaves the same when viewed at different degrees of magnification or different dimension scales. Self-similarity is defined as follow.

A stream of events:

$$\{t_n\}_{n=1}^{\infty} = t_1, t_2, \dots, t_n, \dots \tag{1}$$

is self similar [Willinger et al 2002], if the statistical properties of events are similar independently on the used time scale:

$$\{t_n^{(s)}\}_{n=1}^{\infty} = \{t_1^{(s)}, t_2^{(s)}, \dots, t_n^{(s)}, \dots\} \tag{2}$$

where $s = 1, 2, 3, \dots$ is the scale parameter. Streams of events are similar (not identical) if

$$\{t_n^{(s)}\}_{n=1}^{\infty} \stackrel{d}{=} \{t_n\}_{n=1}^{\infty} \tag{3}$$

$$t^{(s)} = \frac{1}{s} \sum_{i=1}^s t_i = \frac{1}{s} (t_1 + t_2 + \dots + t_s) \tag{4}$$

which is slower than $1/s$, therefore:

$$Var\{t^{(s)}\} \sim s^{-\beta} = \frac{1}{s^\beta} \tag{5}$$

where $0 < \beta < 1$ is a coefficient describing function disappearance of variance.

A stream of events (1) is exactly self-similar with parameter $0 < \beta < 1$, if the correlation function fulfills the following condition

$$\begin{cases} \text{Var}\{t^{(s)}\} = \frac{\sigma_t^2}{s^\beta} \\ \hat{\rho}_t^{(s)}(k) \equiv \hat{\rho}_t(k) \end{cases} \tag{6}$$

for every $s, k = 1, 2, \dots$, where s is the time scale for subsequent k time intervals between events.

Stream of events (1) is asymptotically self-similar with parameter $0 < \beta < 1$, if

$$\begin{cases} \text{Var}\{t^{(s)}\} = \frac{\sigma_t^2}{s^\beta} \\ \hat{\rho}_t^{(s)}(k) \xrightarrow{s \rightarrow \infty} \hat{\rho}_t(k) \end{cases} \tag{7}$$

The measure of self-similarity is the Hurst parameter introduced by H. E. Hurst [Kettani and Gubner 2002]. The Hurst parameter (H) for self-similar processes can change in the range from 0.5 to 1 . For two identical processes $H=1$. Lower values of Hurst parameter indicate larger differences in processes and for $H=0.5$ processes are not correlated (eg. white noise). The Hurst parameter can be evaluated in several ways, primarily using R/S statistics [Wallis 1969]:

- using a rescaled adjusted range plot of R/S as a function of time,
- using a variance-time plot of R/S as a function of time,
- using a periodogram,
- using Wittle’s estimator.

2.2 On-Off Model

Simple “on-off” models are commonly used to describe the random nature of the network traffic [Willinger et al 2002]. The “on-off” model successfully captures the second-order correlations of traffic, in particular their Long Range Dependence (LRD). The “on” states are interlaced by the “off” states. The source transmits the packet in the “on” state, while the “off” state exists when there is no packet transmission. Therefore, the working principle of source is switching between the following active and inactive states.

Let $X(t), t >= 0$ be the stationary process in which [Likhanov et al 1995]:

$$X_i(t) = \begin{cases} 1 & \text{for interval "on"} \\ 0 & \text{for interval "off"} \end{cases} \quad i = 1, 2, 3, \dots, M \tag{8}$$

The distribution of the time interval for the “on” states can be described by Pareto distribution and is given by:

$$\begin{cases} F(t) = 0 & t \leq d \\ 1 - F(t) = \left(\frac{d}{t}\right)^\alpha & t > d, 1 < \alpha < 2 \end{cases} \tag{9}$$

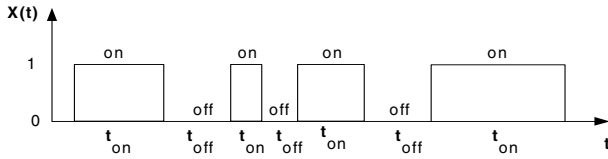


Fig. 1 An example of an “on-off” process

where d is the minimal time interval for an “on” state. This distribution is heavy-tailed if large values of statistical variables occur with high probability and the following condition is met:

$$\Pr\{X > x\} = 1 - F(x) \sim \frac{1}{x^\alpha}, \quad x \rightarrow \infty, \quad \alpha > 0 \tag{10}$$

The average time interval for an “on” state for $1 < \alpha < 2$ is:

$$\bar{t}_{on} = \frac{\alpha d}{\alpha - 1} \tag{11}$$

The time interval for an “off” state is described by normal distribution with the mean value \bar{t}_{off} . The probability that the stream is in the “on” state is given by:

$$p = \frac{\bar{t}_{on}}{\bar{t}_{on} + \bar{t}_{off}} \tag{12}$$

The average intensity of stream components is:

$$E\{X_i(t)\} = p, \quad i = 1, 2, \dots, M \tag{13}$$

while the resultant intensity is:

$$E\left\{\sum_{i=1}^M X_i(t)\right\} = Mp \tag{14}$$

In the case when the function

$$x(t) = \sum_{i=1}^M X_i(t) \tag{15}$$

is asymptotically self-similar (7):

$$\lim_{s \rightarrow \infty} \hat{\rho}_i^{(s)}(k) = \hat{\rho}_i(k) = \frac{1}{2} [(k+1)^{3-\alpha} - 2k^{3-\alpha} + (k-1)^{3-\alpha}], \quad k > 0 \tag{16}$$

and

$$\hat{\rho}_i(-k) = \hat{\rho}_i(k), \quad k < 0 \tag{17}$$

then

$$\lim_{|k| \rightarrow \infty} \hat{\rho}_t(k) = \frac{1}{|k|^{\alpha-1}} \quad (18)$$

Assuming that the correlation function is exponentially distributed

$$\hat{\rho}_t(k) \sim \frac{1}{|k|^\beta}, \quad 0 < \beta < 1 \quad (19)$$

and

$$H = 1 - \frac{\beta}{2} \quad (20)$$

then according to [Willinger et al 1997]:

$$H = \frac{3-\alpha}{2} \quad (21)$$

Typical switching between 1 and 0 states (“on” and “off” states) is random. This randomness reflects user behavior at a computer as well as the applications employing the network services. The “on-off” model of traffic is also applicable in the client-server model. In the “on” state, the client sends a request to the server and the server is expecting a request. In the “off” state, the client waits for a response and the server generates a response. Lengths of the “on” and “off” states are random, as well as the length of neighboring “on-off” intervals.

The superposition of such sources results the generation of traffic in which the LRD is observed. This behavior has been described in [Willinger et al 2002] where the “on-off” traffic has been derived as a superposition of a large number of “on-off” sources, with heavy-tailed “on” and/or “off” periods. A new model called “alpha-beta on-off model” built as a composition of two different “on-off” models has been presented [Sarvotham et al 2005].

3 Network Traffic Model

Human behavior has significant impact on network traffic by accessing to remote multimedia libraries (video and sound), web-page searching and using e-business applications. The usage of the network is mainly based on the analysis of the information resources and on their transfer. The whole process can be represented by a simple model of “on” (connecting) and “off” processes (listening).

It is obvious that the peak of LAN traffic is generated when users start their work. At the beginning of the day most of the internet traffic is related to e-mail services. Later, other kinds of traffic like web browsing, file transferring, and remote computer operation become dominant. Another characteristic feature is the tendency to repeat the day’s schedule caused by periodicity of a company internal

rules and work time set up. In consequence of wide scale marketing actions or the latest trend, in a very short period of time the number of visitors on one server grows exponentially.

The traffic burst, for example, can be caused by user preferences influenced by an advertisement. Consequently, human behavior is a very important aspect to understand the source of the self-similar phenomenon originating in network traffic. Therefore, self-similarity in network traffic cannot be explained without deep analysis how individuals use the network. Self-similarity is not only conditioned by transmission protocols and computer systems, but it is also strictly related to such disciplines like psychology and sociology.

In order to better describe network traffic, the model must consider all essential factors, which influence that traffic. To achieve a correct description of network traffic several factors such as: human behavior, the properties of the operating system, the process scheduling algorithm, and features of transmission protocols have to be analyzed. Most network applications use the TCP/IP protocol, what has a significant impact on the shape of network traffic. The information between the user's application and the physical layer is sequentially converted by several processes such as coding, fragmentation, buffering and encapsulation. Thus, communication between processes has significant influence on the traffic. The proposed model, as shown in Fig. 2, is based on the ISO/OSI reference model.

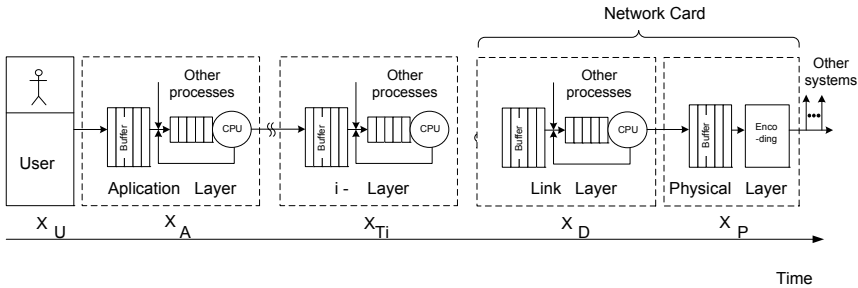


Fig. 2 The layered model of traffic source. X_U , X_A , X_i , X_L , X_P are “on-off” processes of individual layers

Individual components of the model represent network traffic generators with different periods of “on – off” processes. Superposition of subsequent sub-models is the multilayer model of traffic source.

The Hurst parameter of network traffic for different sampling rate (s) changing from 103s to 10-3s. Higher sampling frequency (>10 Hz) shows an influence of protocol algorithms, the operating system and information buffering. The influence of the particular components on the shape of corresponding function $X(t)$ is presented in Fig. 3.

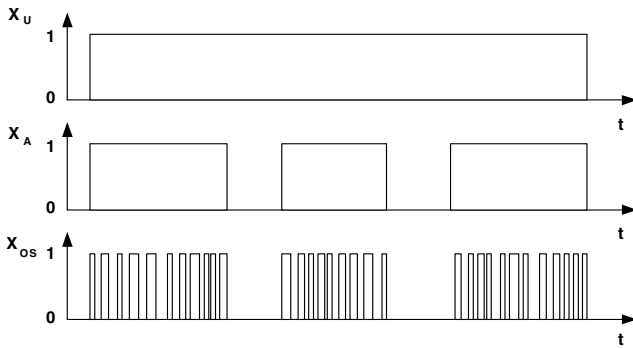


Fig. 3 Superposition of “on-off” functions modeling the source of network traffic: X_U - the user, X_A – applications and application layer and X_{OS} – operating system

In the model shown in Fig. 2 the components can be characterized by “on-off” functions in which states are switched with different frequencies (f). For example:

X_U – human behavior with $f \leq 10^{-2}$ Hz,

X_A – the network processes embedded in a network application with $10^{-2} \leq f \leq 1$ Hz,

X_{OS} – the queuing of the processes to the CPU with $10 \leq f \leq 102$ Hz,

X_i – the ith process of information conversion accordingly to network layer protocols with $10 \leq f \leq 103$ Hz,

X_L – link layer with $f \geq 103$ Hz

X_P – signal in the physical layer.

Assuming that in the physical layer observed traffic is a superposition of the model components mentioned above, the following formula describes this process:

$$X_P(t + \Delta t_{ps}) = \prod_{i=1}^n [X_i(t + \Delta t_{pi}) + X_i^b(t + \Delta t_{pi}^b)] \tag{22}$$

where

X_i – function describing behavior of ith layer.

X_i^b – function describing behavior of ith buffer.

Δt_{pi} – time of signal propagation in the computer system,

Δt_{pi}^b – time of signal propagation in the buffers.

A traffic in various communication layers of the “on-off” model is generated in the following way:

- Traffic in each layer is generated according to the “on-off” model for the given probability distribution, ex. Pareto distribution, and system behavior (processes queuing and buffering).

- In the case of the first “on” period, the starting time for sub-layers is randomly chosen in order to avoid the case that the first period starts at the beginning of the “on” period of the higher layer.
- The output traffic is a superposition of processes of all layers of the model.
- For the last layer, the size of the packet for the “on” process is generated using an adequate probability distribution such as Pareto, exponential, etc.

Assuming that the time of signal propagation at each layer is much smaller than the period of the “on” state, one may write:

$$\Delta t_{pi}, \Delta t_{pi}^b \ll t_{ON}, t_{OFF} \quad , \tag{23}$$

Thus, from (22) and ignoring the propagation times Δt_{pi} and Δt_{pi}^b :

$$X_p(t) \approx \prod_{i=1}^n [X_i(t) + X_i^b(t)] \tag{24}$$

Considering components of the model:

$$X_p(t) \approx [X_U(t)] \cdot [X_A(t) + X_A^b(t)] \cdot [X_{OS}(t) + X_{OS}^b(t)] \cdot [X_L(t) + X_L^b(t)] \tag{25}$$

and assuming that bandwidth of communication channel is equal to p_{max} , throughput $I(t)$ can be obtained from (25) as

$$I(t) \approx X_p(t) \cdot p_{max} \tag{26}$$

An average throughput $\overline{I(t)}$ can be obtained from:

$$\overline{I(t)} = \frac{1}{t_2 - t_1} \int_{t_1}^{t_2} I(t) dt \tag{27}$$

where $[t_1, t_2]$ is the range of observation time.

The effect of the delay of the operational system can be introduced by generation of delays dependent on the time required for information processing, setting queues and buffering.

The proposed model of traffic source represents single machine (computer) generating traffic to a local computer network. For a simulation of network traffic with long-term correlation, an “on-off” model can generate transmission of “on” packets and “off” silence according to the distributions: exponential, Pareto, Fisk and Frechet.

The algorithm of traffic generation for the particular layers of a single machine model is presented in Fig. 4. A traffic in a computer networks is generated from many sources. Therefore developed network traffic model should consider this fact by multiplication of algorithm for single source what is shown in Fig. 4

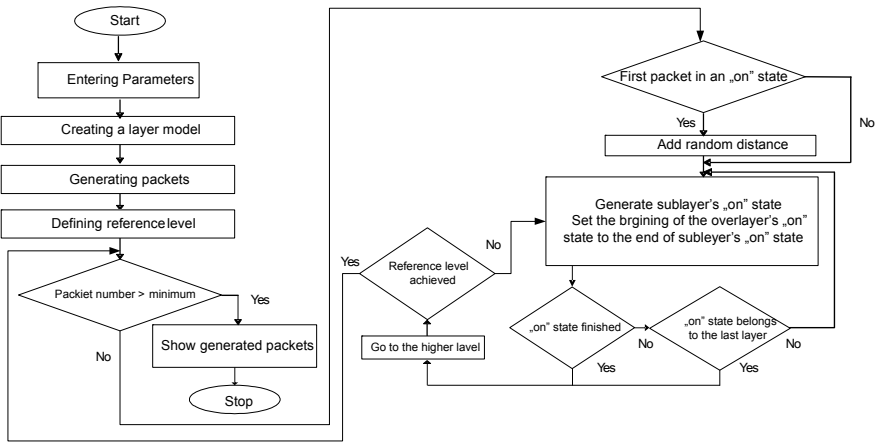


Fig. 4 “On-off” process generation algorithm for the particular layers of a single computer model

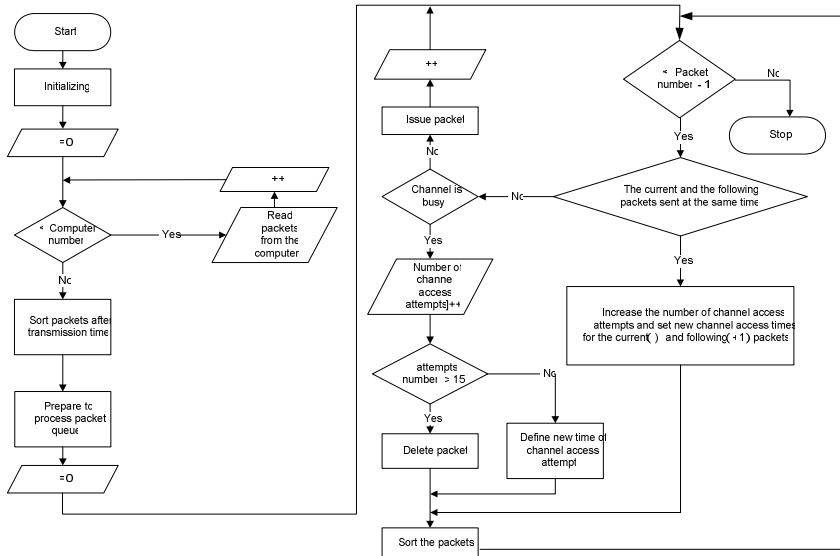


Fig. 5 Traffic generator algorithm for an n-node network

The algorithm of traffic generation for n-nods based on the proposed model is presented in Fig. 5.

The output of the program is a pair of values: transmission time of packet and packet size. The output is “on-off” traffic originated from many generators. The

pattern of this traffic depends on the parameters of random distributions used for generating “on” and “off” states in the particular sub-models.

4 Experiment and Analyses

Presented model is implemented by authors as a traffic generator software called *LanTraffic*.

In order to verify correctness of the model following experiment is performed. The experimental networks have been prepared with 5,10,30 and 50 computers connected to the Ethernet switch Cisco 2950 with port mirroring feature. Port mirroring allows capture all traffic passing the switch by frame replication to the selected port which is used to connect computer with network analyzer – LinkView Classic software. Traffic has been generated by hosts located in the experimental networks. The VLC media player installed on two hosts is the source of audio and video streams addressed to the host clients.

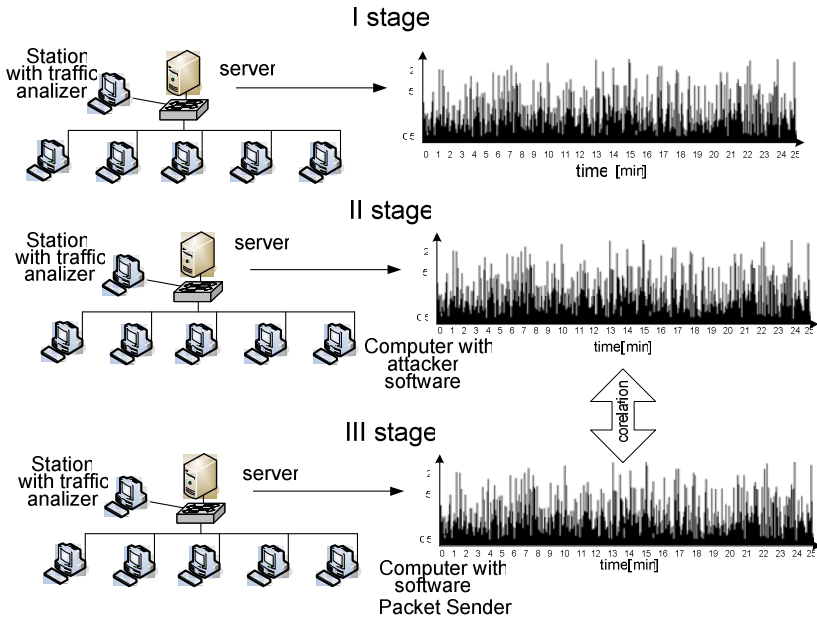


Fig. 6 Steps of performed experiment

The HTTP Traffic Generator software has been installed on other hosts. This software simulates user activity by generating different type of packets to the servers located in the experimental network. Thus, in addition to the audio and video traffic the http, ftp, smtp, pop3 ones is generated between hosts and servers. The traffic is captured with different sampling rate by mentioned, advance network interface card LinkView Classic software.

The same experiment has been repeated with the same traffic pattern but additionally with malicious activity. An access attack to the mail server has been performed with the use of Burst tool. This software supports brute-force and dictionary attacks.

The main goal of this stage of experiment is to measure the influence of the attack to the level of self-similarity. The gathered data has been analyzed and the results presented in Table 1 to 3. The Hurst parameter shows increasing of self-similarity when the malicious attack occurred.

Table 1 Self-similarity of typical LAN Traffic

Number of hosts in LAN	The method of Hurst parameter estimation	
	R/S	Variance Time Plot
5	0,792	0,788
10	0,821	0,832
30	0,797	0,81
50	0,814	0,83

Table 2 Self-similarity of LAN traffic with the presence of brute-force attack

Number of hosts in LAN	The method of Hurst parameter estimation	
	R/S	Variance Time Plot
5	0,815	0,817
10	0,845	0,862
30	0,817	0,813
50	0,827	0,831

Table 3 Self-similarity of LAN traffic with the presence of dictionary attack

Number of hosts in LAN	The method of Hurst parameter estimation	
	R/S	Variance Time Plot
5	0,823	0,819
10	0,855	0,872
30	0,826	0,818
50	0,844	0,851

When Hurst parameters from all three tables, are compared, values for all cases in Tables 2 and 3 are higher than corresponding values in table I. For example, for 5 hosts in LAN Hurst parameter estimated with the use R/S method are 0,792 (Table 1) and 0,823 (Table 3). In other cases, (more hosts) and both estimation methods of Hurst parameter the same regularity is observed.

The variance-time plots are obtained by plotting dependence of process variance on time in logarithmic scale, and then by using the least square method to find line going through the points which represent this dependence. In this analysis method small values of k are ignored.

For large values of k , the points in the plot are expected to be scattered around a straight line with a negative slope parameter equal to $2H-2$. For short-range dependence or independence among the observations, the slope parameter of the straight line is equal to -1 .

Self-similarity can be calculated from the values of the estimated slope parameters which are change asymptotically between -1 and 0 , The estimator of self-similarity is given by:

$$\hat{H} = 1 + \frac{1}{2}(\text{slope}) \quad (28)$$

The R/S and variance-time plot analysis of traffic flow prove increasing of self-similarity in the case of brute-force or dictionary attack. Hurst parameter is higher than $0,5$.

The self-similarity analysis of selected malicious traffic is also performed. A traffic of the brute-force and dictionary attack is analyzed. It shows significant level of self-similarity in this case. Figure 6 presents dependence of the variance of the process on the time in logarithmic scale for the traffic generated by host which performs brute-force and dictionary attack.

In the last step of prepared experiment the host with attacker software has been replaced to the host with the traffic generator implemented by authors. This traffic generator, called PacketSender sends the traffic flow, according the proposed model, in LAN. The generated data flow is similar to the traffic generated by real attacker software. In this case generator uses four layers and Pareto distribution for "on-off" processes. The exponential distribution is used for generation of packet length according to [Papapanagiotou et al 2007]. This step of experiment has been prepared for the LAN with 30 hosts. In this case the analyze of gathered data confirm the increase of self-similarity. The Hurst parameter increases similar like in the case of real attack is in the range of $0,6 < H < 0,8$.

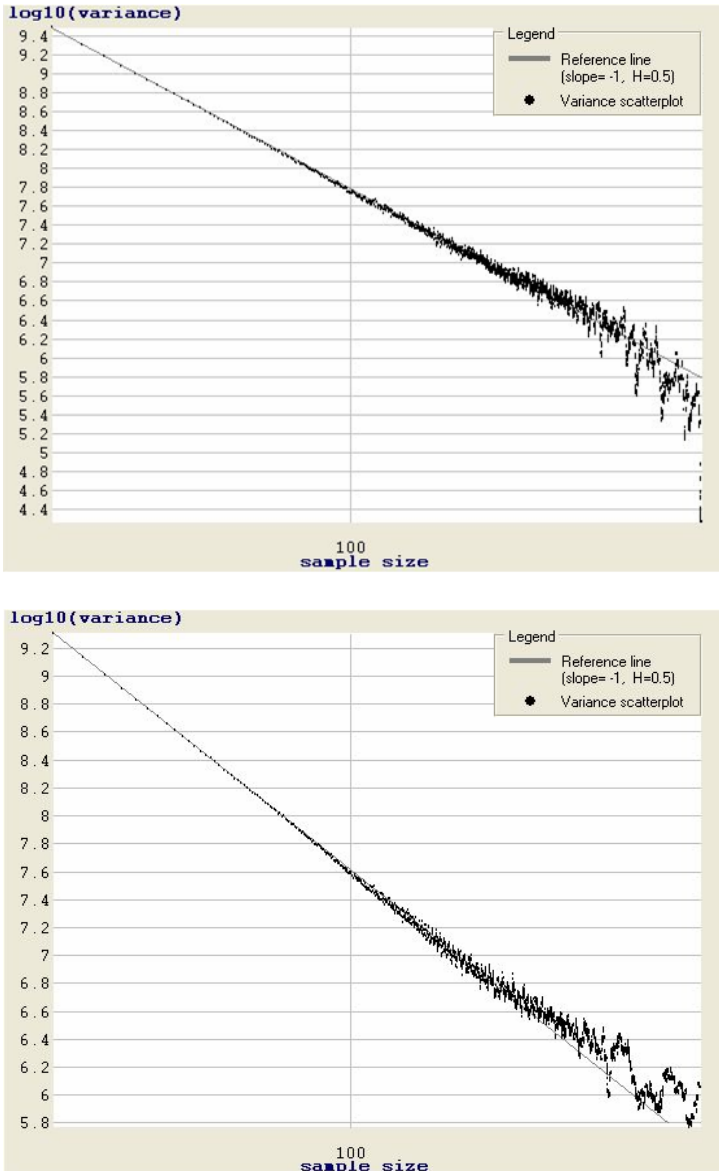


Fig. 7 Variance-time plots for the traffic with malicious activity (bottom) and without (top)

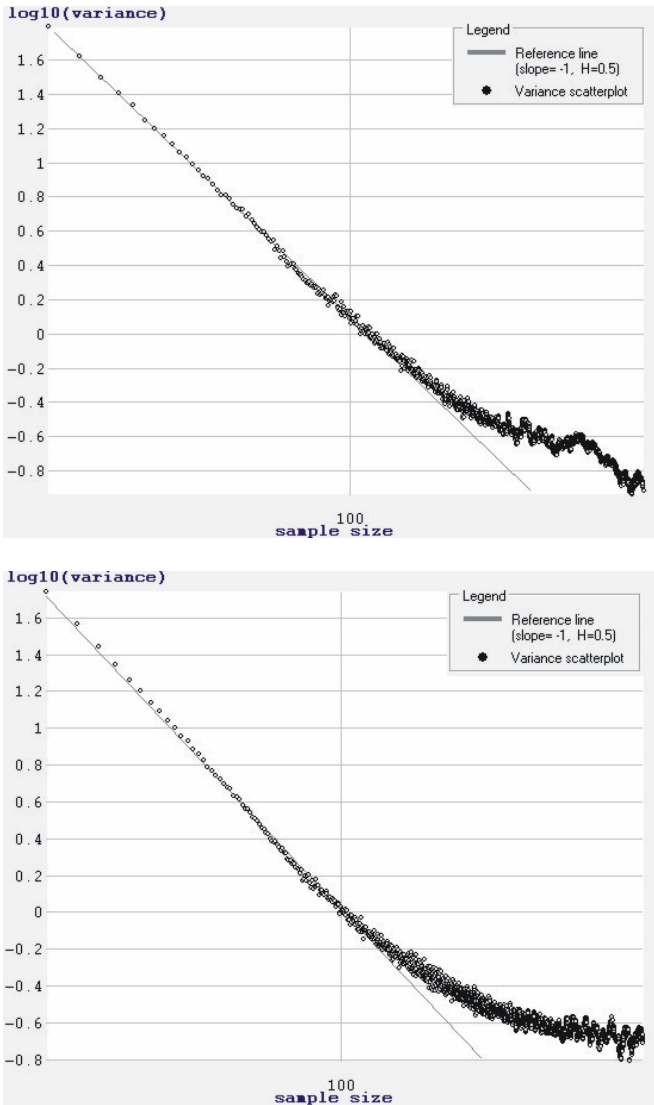


Fig. 8 Variance-time plots for the traffic generated by Brute-force (top) and dictionary (bottom) attacks

5 Conclusions

Presented model allows generate “on-off” processes for each layer with specific level of self-similarity. Traffic generator with implemented model can simulate traffic similar to different type of real data flow. Especially it can simulate malicious traffic. This model can be applied for the simulation of traffic in the case of

attack, network failure and other anomaly. Moreover, it can be used in the simulation of network operation including different anomaly and allows gather experiment data without affecting real network infrastructure.

References

- [Cheng et al 2009] Cheng, Xie K., Wang, D.: Network traffic anomaly detection based on self-similarity using hht and wavelet transform. information assurance and security. In: Proc. 5th Inter. Conf. on Information Assurance and Security, vol. (1), pp. 710–713 (2009)
- [Kettani and Gubner 2002] Kettani, H., Gubner, J.A.: Novel approach to the estimation of the hurst parameter in self-similar traffic. In: IEEE Conference on Local Computer Networks, pp. 1–6 (2002)
- [Likhanov et al 1995] Likhanov, M., Tsybakow, B., Georganas, N.D.: Analysis of an ATM buffer with self-similar (fractal) input traffic. In: Proc IEEE INFOCOM 1995, Boston, pp. 982–985 (1995)
- [Mello et al. 2007] Mello, F.L., Lima, A.B., Lipas, M., Almeida Amazonas, J.R.: Generation of self-similar Gaussian series via wavelets for use in traffic simulations (in portugese). IEEE Latin America Transactions 5(1), 9–20 (2007)
- [Papapanagiotou et al 2007] Papapanagiotou, I., Vardakas, J.S., Paschos, G.S., Logothetis, M.D., Kotsopoulos, S.A.: Performance evaluation of IEEE 802.11e based on ON-OFF traffic model. In: Proc. of the 3rd International Conference on Mobile Multimedia Communications, Nafpaktos, Greece, vol. 329, Article no. 17 (2007)
- [Rohani et al 2009] Rohani, M.F., Maarof, M.A., Selamat, A., Kettani, H.: Loss of self-similarity detection using exact and asymptotic self-similarity models. J. of Information Assurance and Security, 571–581 (2009)
- [Sarvotham et al 2005] Sarvotham, S., Riedi, R.H., Baraniuk, R.G.: Network and user driven alpha-beta On-Off source model for network traffic. Computer Networks 48(3), 335–350 (2004)
- [Wallis 1969] Wallis, J.R.: Robustness of the rescaled range R/S in the measurement of noncyclic long-run statistical dependence. Water Resources Research 5, 967–988 (1969)
- [Willinger et al 1997] Willinger, W., Taqqu, M., Sherman, R., Wilson, D.: Selfsimilarity through high-variability: Statistical analysis of Ethernet LAN traffic at the source level. IEEE/ACM Trans. Networking (Extended Version) 5(1), 71–86 (1997)
- [Willinger et al 2002] Willinger, W., Paxson, V., Riedi, R., Taqqu, M.: Long Range Dependence And Data Network Traffic. In: Long Range Dependence: Theory and Applications. Wiley, Chichester (2002)

A Cooperative Approach to Web Crawler URL Ordering

A. Chandramouli¹, S. Gauch², and J. Eno²

¹ Department of Computer Science, University of Kansas, Lawrence, KS
aravindc@ku.edu

² Department of Computer Science, University of Arkansas, Fayetteville, AR
{sgauch, jeno}@uark.edu

Abstract. Uniform Resource Locator (URL) ordering algorithms are used by Web crawlers to determine the order in which to download pages from the Web. The current approaches for URL ordering based on link structure are expensive and/or miss many good pages, particularly in social network environments. In this paper, we present a novel URL ordering system that relies on a cooperative approach between crawlers and web servers based on file system and Web log information. In particular, we develop algorithms based on file timestamps and Web log internal and external counts. By using this change and popularity information for URL ordering, we are able to retrieve high quality pages earlier in the crawl while avoiding requests for pages that are unchanged or no longer available. We perform our experiments on two data sets using the Web logs from university and CiteSeer websites. On these data sets, we achieve a statistically significant improvement in the ordering of the high quality pages (as indicated by Google's PageRank) of 57.2% and 65.7% over that of a breadth-first search crawl while increasing the number of unique pages gathered by skipping unchanged or deleted pages.

1 Introduction

Search engines use crawlers to collect Web pages from Web servers distributed across the Internet. Crawlers are programs that automatically collect Web pages by starting with a Uniform Resource Locator, URL, downloading the Web page at that location, and recursively retrieving all the pages pointed to by the hyperlinks on the page. In contrast to traditional crawlers based on link structure, recent efforts [Brandman et al. 2000] have focused on providing support for crawlers by exploiting Web server information like Web logs and file system to provide the list of URLs on the website. Along similar lines, Google has developed site maps (<http://www.sitemaps.org>) to allow web sites to provide hints to the Google crawler.

In contrast to providing the list of all URLs on a website, we proposed a cooperative architecture that uses Web log and file system timestamps to provide a

ranked list of new, modified, and deleted pages since the last visit from a crawler. In our Web services-based co-operative approach, the individual websites make use of their own Web logs and file systems to gather this information. This information can be used to increase the number of pages collected compared to traditional crawling and achieve significant bandwidth savings for the same set of pages collected over multiple crawls. Additionally, hidden web content accessible only through forms can be added to the index based on the URL encoded arguments from a GET form request.

Even with these improvements, the size of the Web combined with the necessarily limited resources available to a crawler and the limited bandwidth on the websites, crawlers do not collect all pages from the Web. As the size of the Web keeps increasing, crawlers typically try to download the “important” pages for indexing by search engines. To determine the important pages, crawlers make use of URL ordering algorithms. The connectivity-based document quality ordering [Cho et al. 1998] and breadth-first search crawling [Nojork and Wiener 2001] are two such well-known URL ordering algorithms. However, both techniques have drawbacks. A connectivity-based metric penalizes new pages and is expensive to compute. On the other hand, breadth-first ordering is relatively inexpensive to implement, but this technique misses good pages deeper in the hierarchy of the site.

Social web content presents challenges for either connectivity-based or breadth-first URL ordering strategies. Social media tends to be fast-moving and may be out of date by the time it develops the robust link structure that would bring it to prominence in a connectivity-based algorithms. Another problem is that many of the links of interest will be shared on private social network profiles, limiting the ability to use either connectivity or breadth-first crawls to identify resources.

As part of the co-operative file system approach, we have developed an URL ordering algorithm based on popularity information extracted from Web logs. This algorithm is inexpensive to compute because it distributes the URL ordering calculation overhead among the participating websites, and identifies pages that users of the website find important enough to view. Because it does not rely on identifying external links to resources to discover content pages, it can identify pages before they are widely linked in the broader Web as well as within social networks. In this paper, we present both timestamp-based and access count algorithms, discuss the advantages and drawbacks of these approaches, and empirically compare them with the computationally similar breadth-first search crawl using Google’s PageRank as the metric to indicate each page’s importance.

2 Related Works

Researchers have investigated several ways to provide improved support for search engine crawlers. Most of these focus on exploiting Web server information and processing power. [Brandman et al. 2000] suggest the creation of a file on the

Web server that provides a list of all the URLs and their meta-data. In their approach, the crawler would download this file to identify modified pages. They make use of the file system to create their meta-data file. Although the Web server is a partner in the crawling process by providing digested information, the bulk of the processing to detect new, modified, and deleted pages is still left to the crawler. In contrast to the pull strategy employed by crawlers today, [Gupta and Campbell 2001] describe an algorithm that would push updates on popular pages from Web servers to search engines. [Castillo 2004] discusses both pull and push architectures to support cooperation between a Web server and a crawler. They also implemented a cooperation scheme that created an XML file storing update information based on the file system, and demonstrated that a crawler making use of this information would experience 40% bandwidth savings when compared to traditional crawling. [Buzzi 2003] describe a similar approach to [Castillo 2004] and propose the creation of a text file that has information about Web pages such as the last update time, file size, local request frequency, and local update frequency. The paper discussed the type of information that should be provided, but they do not discuss how this information would be gathered by the Web site, nor how it would be used by the crawler. More recently, Google introduced Google sitemaps, which is essentially an extension of the approach proposed in [Brandman et al. 2000]. Webmasters can install a program on their website that creates a text or XML file containing the URLs on the website, called a sitemap. Google sitemaps make use of both file system and Web logs to create the list of URLs.

The earliest work on URL ordering algorithms was by [Cho et al. 1998]. They used connectivity-based metrics to identify the “important” pages to download, and their experiments showed that using the PageRank metric downloaded important pages earlier than the other algorithms. [Najork and Wiener 2001] extended the work of [Cho et al. 1998] and demonstrated that it was possible to discover the important pages early in the crawl by using a breadth-first ordering. However, they do not compare breadth-first crawl with other techniques. More recent work on URL ordering by [Castillo et al. 2004] compared different ordering techniques, namely Optimal, Depth, Length, Batch, and Partial, for long term and short term scheduling for crawling on the Web.

As discussed above, the breadth-first search and the PageRank are the two of the most popular URL ordering techniques reported in literature. However, using PageRank to compute the URL ordering can be very expensive. In fact, [Najork and Wiener 2001] observe that performing the PageRank computations for all the Web pages in real time is not feasible. [Cho and Schonfeld 2007] demonstrate a more efficient method of computing a partial PageRank by relying on a vector of trusted pages to compute a lower bound on the true rank of queued pages. For a crawl of 80 million pages, the technique took only three times as long as a breadth-first ordering. However, new pages and pages outside the trusted set and pages without PageRank are still penalized by this metric. Recent work has

focused on modifying PageRank to use last-modified date or more complex web graph modeling to improve performance for new pages [Cho et al. 2005], but these algorithms are still computationally expensive. On the other hand, a major drawback with breadth-first ordering is that important pages deeper in the hierarchy of the websites will not be collected.

As an alternative to the above two techniques, we propose a URL ordering of pages on individual websites calculated using popularity information extracted from web logs. A major advantage of such an algorithm is that it is relatively inexpensive to compute when compared to PageRank and, since the ordered list of URLs is produced by the individual websites, the workload for the search engines is reduced. Because the websites can process their own file systems and Web logs efficiently, and the results of this effort can be shared with multiple search engine crawlers, the burden on the individual websites is acceptable. This upfront work also decreases the amount of effort the websites must spend serving pages to crawlers.

Another factor that may affect the optimal ordering of URLs is the probability that a new page will improve the existing index. [Pandey and Olston 2008] developed a model that uses sample queries and the existing index state, combined with content clues such as URL text and anchor text words to order URLs for crawling. Pages that have a high likelihood of being highly ranked and relevant to a topic that is sparsely populated in the index are given greater priority in the crawl order. Although this algorithm helps to alleviate some of the problems of other connectivity-based ordering algorithms, such as focusing too much on popular topics to the detriment of niche topics, it still relies on existing links. In some ways, the existing link problem is exacerbated, since the existing links are used for both reference counts and topic discovery. Such an algorithm might struggle with the common practice of using a link shortening service such as tinyurl or bit.ly in social network links.

3 Approach and Implementation

The goal of any URL ordering algorithm is to produce an ordering of URLs so that the Web crawler can collect the most important pages first. Our approach improves this ordering by providing a web service that generates an XML document including only new or modified URLs and ranking them based on popularity as measured by access counts. During a crawl, a cooperative crawler will request a list of all new, updated, or deleted URLs since the last request. The server will generate an XML document with an entry for each qualifying URL, marking the URL as new, modified, or deleted. URLs that have not been modified will not need to be requested by the crawler, so they are not included in the XML response.

For the popularity-based URL ordering algorithms, we exploit the popularity information present in the Web logs on a website and look at a variety of ways to produce this URL ordering. We classify these approaches broadly as non-learning

algorithms that use a predetermined ordering function and learning algorithms that order URLs adaptively based on a training set of URLs with quality information.

3.1 URL List Generation

The list of URLs for the web service may be generated based on the file system, Web logs, or a combination of both. The file system approach has the advantage of providing a comprehensive list of all accessible documents along with accurate modification times. However, it will fail to discover or update dynamic pages, multiple pages specified by URL arguments, or content that may have been modified in a database or content management system (CMS) without modifying the base page. It also cannot flag deleted URLs, since they no longer exist on the file system.

The Web log approach can mitigate these omissions in three ways. First, it can recognize some dynamic pages based on URL-encoded arguments to the web server. Second, by examining the number of bytes returned by a request, it can recognize when the size of a page has changed, indicating a change even when the file system timestamp is unchanged. Finally, it can discover deleted pages based on 404 (Not Found) errors in the Web log. However, it can only recognize pages that have been accessed within the time covered by the log file. In practice, a hybrid approach that gathers data from the file system and Web logs while maintaining a history of file system and Web log activity provides a means to get the most accurate information from all possible sources without relying on long log histories.

Fig. 1 shows the architecture for our system. Periodically, the Web Log Harvester harvests the Web logs for processing. Currently, the Web server used archives its Web logs weekly, so the Web Log Harvester gathers data weekly using the Web log file name provided in a configuration file. Similarly, the File System Harvester uses a text file that contains the list of directory paths on the website and the corresponding base URL for that directory. Every week, the File System Harvester recursively retrieves the filenames. The harvesters pass their information along to the Data Parsing Module. From the Web logs, the module extracts: IP address, access time, URL, number of bytes, and status code. From the file system, the module extracts: path, filename, date of last modification and maps the filenames to its corresponding URLs. The information directly extracted is stored in a URL Database that has the following entries: URL, created date, modified date, deleted date, byte count, and the source. The first time the harvesting is performed the modified date and the created date are set to be the same, while the deleted date is empty for all the URLs collected. However, during the subsequent weeks, the Data Parsing Module uses the data gathered by the harvesters and the information present in the database to infer the modified date and the deleted date using the techniques discussed in section 3.

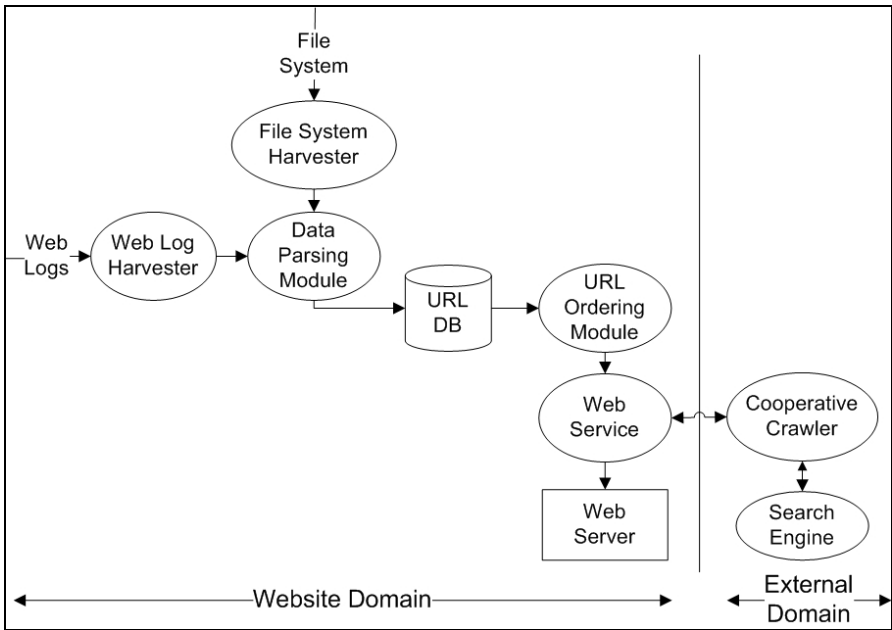


Fig. 1 System Architecture

The information stored in the database is shared with enhanced crawlers via a Web service implemented using the REST protocol. The crawler queries the Web service for information about the Web site contents. The crawler can query the Web service to find URLs that match based on the following criteria:

FileType - Text, Video, Audio, All types.

FromDate - This is the date from which it requires information.

ChangeType - Modified, Deleted, Created, All types.

Changes/Url - Give the changes to the file only or the URL of the entire file.

The ToDate is implicitly set to the current date. Currently, our system only gives the URL for the Web page, but, in future, we will explore a mechanism to provide only the changes made to a page. The crawler then parses the supplied XML file in order to identify the list of URLs that need to be collected. Finally, the Web pages are collected and passed along to the search engine for indexing.

During the experiments, the list was generated using three algorithms: file system, Web log, and file system hybrid. The file system and Web log methods used data exclusively from the file system or Web logs, respectively. However, the Web log data tended to be less reliable both because users sometimes requested pages that no longer existed and some pages were not requested at all, making them unavailable to the Web log-based system. The file system hybrid approach started with the file system list, then supplemented it with Web log data to discover hidden web content. This led to a larger list that still had high reliability.

3.2 Non-learning Algorithms for URL Ordering

The Web logs on the website register the access made to every Web page on the site. Thus, from the Web logs, the total access count for each Web page can be calculated. Then, the Web pages are sorted based on their Total Access Count (TAC). However, pages with high total access count need not necessarily indicate highly popular pages. For example, a Web page might be accessed often by its owner thus inflating the access count value. It may be possible to more accurately identify important pages by incorporating information about the number of different IP addresses from which a page is accessed.

In addition to unique IP address metrics, the hierarchical nature of IP addresses enables us to differentiate between internal and external page accesses. Hence, in order to explore a variety of URL ordering algorithms, we extract four different types of access information from the Web logs, namely, the Total External Count (TEC), the Unique External Count (UEC), the Total Internal Count (TIC), and the Unique Internal Count (UIC) where the external count refers to the requests made to a URL on the website from outside the local network and the internal count refers to the local requests made to a URL. Different URL orderings are then produced by ordering the Web pages based on different combinations of these factors. One limitation to this approach is its inability to differentiate multiple accesses originating from behind a single proxy server. However, obtaining session-level data would require more knowledge than can be derived from a typical Web access log and is beyond the scope of this paper.

The non-learning algorithms discussed so far order the URLs based on four different factors. However, they do not take into account the relative importance of each factor. Are all the parameters equally important or should they be weighed differently? To address these issues, a simple approach would be to calculate the accuracy of the different parameters to predict high quality pages and then use these accuracy values to assign different weights for the parameters. This approach, the Weighted Access Count (WAC) algorithm, has the advantage that the parameters are weighed differently based on their ability to predict high quality pages. The weighted score for each URL is calculated as shown in Equation 1 and an URL ordering is produced by sorting the URLs based on this weighted score.

$$WS = \alpha * \frac{TECacc}{Totalacc} + \beta * \frac{UECacc}{Totalacc} + \gamma * \frac{TICacc}{Totalacc} + \delta * \frac{UICacc}{Totalacc} \quad (1)$$

where

WS = Weighted Score

TECacc = TEC algorithm accuracy

UECacc = UEC algorithm accuracy

TICacc = TIC algorithm accuracy

UICacc = UIC algorithm accuracy

$Totalacc = TECacc + UECacc + TICacc + UICacc$

and α , β , γ and δ = raw external, unique external, internal, and unique internal counts for the URL.

3.3 Learning Algorithms for URL Ordering

The non-learning algorithms either make use of four different factors or a combination of these factors. In order to learn the best combination of factors, and to develop an adaptive algorithm that would work on any website, we implemented two learning algorithms, Total Access Count-Learning (TAC-L) and Split Access Count-Learning (SAC-L). The TAC-L algorithm takes the total access count for each URL as the attribute while the SAC-L algorithm takes the four different parameters discussed in Section 3.2 as the attributes for the data and they predict the PageRank categories for new URLs based on their attribute(s). Both algorithms have a training and a testing phase. In the training phase, a set of URLs with their access counts and quality information are given as input to a learning algorithm like decision trees or k-Nearest Neighbor algorithm and a model is learned. The quality information is determined using PageRank. PageRank has been used for URL ordering algorithms [Cho et al. 1998] to measure the quality of a page, that is, higher the PageRank, higher the quality of the page. In addition, by relying on the global Google PageRank value as an indicator of the 'true' importance of a page, we are able to verify our results against a much broader metric even though we only use local information to compute our ranking.

Although true PageRank values are floating-point values that provide a total ordering of URLs, we are somewhat restricted in our access to Google's PageRank values. In order to determine the true PageRank for a URL, we use the free-ware Parameter tool that determines the PageRank of a URL on a 1 to 10 scale. Since learning algorithms typically predict a category, we make use of the integer PageRank values as the categories for classification. During the testing phase, a set of URLs with their access counts are given as input and the learned model is then used to place each URL in the best matching category. The confidence factor for these assignments are used to rank order the URLs within each category, producing total ordering of the URLs.

4 Evaluation Method

In this section, we describe our experimental evaluation method used to compare popularity-based URL ordering algorithms to a breadth-first search crawl, using a PageRank ordering as a benchmark.

4.1 Data Collection

We make use of two data sets for our experiments. Data set 1 (DS1) contains the Web logs of the ITTC website over 5 weeks (<http://www.ittc.ku.edu>), to which we had access. Data set 2 (DS2) contains the Web logs of the CiteSeer website (<http://citeseer.ist.psu.edu/>) whose Web logs for a five week period was shared with us. For DS1, using the home page as a start page, we produced an URL

ordering based on the breadth-first search crawl. In contrast, since CiteSeer does not have an exposed tree hierarchy, a breadth-first crawl is essentially a random crawl. Hence, the random crawl is used as a baseline for DS2. For our popularity-based URL ordering algorithms, access count information from Web logs covering a five week period were extracted. Some of the URLs collected using a breadth-first search crawl for DS1 did not have popularity information (i.e., were not accessed during this five week period) and, similarly, breadth-first search crawl was unable to collect the hidden Web pages that were accessed but were not linked explicitly on the site. Thus, our experiments used a total of 5,480 URLs for DS1 and 102,360 URLs for DS2 that could be collected by the breadth-first search/random crawl for which we also had access information from the Web logs. It is useful to note that one of the benefits of the proposed approach is the ability to find pages that are accessible by means other than links, which is not possible with a link-graph or breadth-first approach.

4.2 Metrics

As discussed briefly in Section 3.2, PageRank has been used in literature to measure the quality of a page. PageRanks for a page are calculated recursively based on the link structure on the Web, with pages linked from many highly ranked pages receiving the highest scores. In order to enable categorization, PageRanks are assigned from a scale of 0-10, with 10 being the most important page. To find the PageRank, as outlined in Section 3.3, we make use of a freeware Parameter version 1.2.

One problem with using PageRank categories as an evaluation metric is that a URL ordering algorithm produces a total ordering on the list of URLs whereas a discretized PageRank does not. That is, although URLs with different PageRank categories can be ordered, URLs with the same PageRank is essentially an unordered set. Hence, for our evaluation, the pages with higher PageRank should be ranked higher than pages with lower PageRank but the order among the URLs with the same PageRank does not matter.

$$Accuracy = \frac{\sum_{i=1}^n Match_i}{n} * 100 \quad (2)$$

Equation 2 is the evaluation metric we use, where

n = total number of URLs,

$Match_i = 1$ if $PPR_i = APR_i$, and 0 otherwise,

PPR_i = PageRank category of URL i produced by the ordering algorithm,

APR_i = Actual PageRank category for i .

In order to illustrate this metric, consider the following URLs (denoted A-E) with their associated PageRank categories: A-6, B-6, C-5, D-5 and E-5. Since PageRank categories do not distinguish among elements in the two sets {A, B} and {C, D, E}, A and B can be in any of the first 2 slots while C, D and E can be in any of

the next 3 slots to produce an accuracy of 100%. Hence, a rank order of ABCDE will be given an accuracy of 100% while a rank order of ACDBE will be given an accuracy of 60% since B should be in one of the first 2 slots while C should be in any of the final 3 slots.

5 Results

Results are presented for both URL list generation and URL ordering systems. The URL lists are evaluated in terms of request and bandwidth savings over time, while the URL orderings are compared with PageRank rankings to determine ordering accuracy.

5.1 Evaluating the Effectiveness of URL List Generation

The list of URLs may be effective in two ways. First, it can identify additional pages that are part of the hidden web or have not been linked yet by other pages. Second, it can reduce the amount of bandwidth used to gather unmodified content. Fig. 2 shows the effect of the URL list generator in terms of additional pages collected. The results show that while the number of pages available through links was fairly constant and occasionally dropped from week to week, the file system, Web log, and file system hybrid approaches gathered significantly more pages, and increased the size of the collection from week to week. Because the pure Web log approach could only discover pages that had already been requested, it had lower performance than the approaches which included file system information. However, the combined approach was able to discover hidden web pages that were not visible to a pure file system approach.

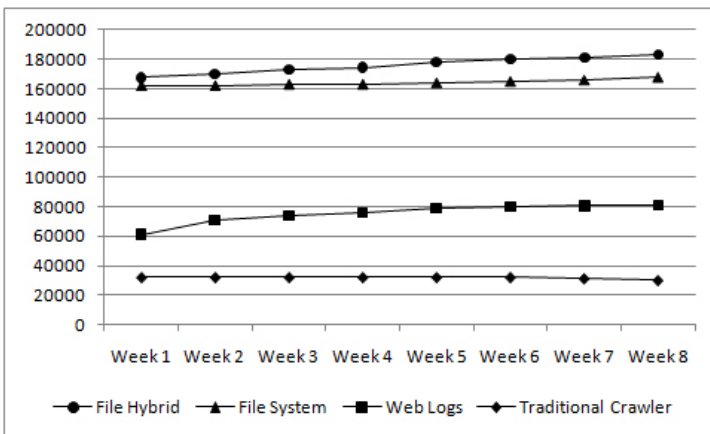


Fig. 2 Number of URLs Collected Over Time

In terms of bandwidth savings, the set of pages that were collected by the traditional crawl were largely static throughout the eight week experiment. As a result, after the first week, the cooperative approach required less than 0.2% as much bandwidth to collect the modified pages from the set of pages gathered by the traditional crawler. Because the hybrid approach discovered more pages, it gathered 159 MB compared to the 13.5 MB collected by the traditional crawler. In the process, it collected six times as many pages overall.

5.2 Evaluating the Accuracy of URL Ordering Algorithms

We evaluate the performance of our URL ordering algorithms by using a five-fold cross validation on DS1. First, the 5,480 URLs were randomly divided into 5 sets. Next, the training and testing was carried out 5 times, each time using 4 sets for training and the remaining set for testing. The numbers reported in this section are the averages obtained over the 5 trials.

We first establish the baseline with the breadth-first search crawl using the testing sets. We found that the breadth-first crawl URL ordering had a 38.8% match with the PageRank category ordering (as calculated using formula 2). Also, a random ordering of the URLs produced an accuracy of 32.9% using the same metric. Next, we used the same test sets for the popularity-based URL ordering algorithms described in Section 3. The results in Fig. 3 show that, even after 5 weeks, the total internal count (TIC) (28.7%) algorithm performs poorly when compared to the baseline. The unique internal count had similarly poor results of 28.5% accuracy. However, the total external count (TEC) (45.7%) performs better. Similarly, the total access count (TAC) algorithm also perform better than the baseline, producing an accuracy of 44.7%. The weighted access count did not improve the TAC algorithm, and neither outperformed the exclusively external count algorithm.

Fig. 3 also shows the accuracy values obtained using the total access count learning algorithms (TAC-L) and the split access count learning algorithms (SAC-L). As discussed in Section 3.3, we make use of decision trees (TAC-L_DT and SAC-L_DT) and k-Nearest Neighbors (TAC-L_kNN and SAC-L_kNN) as our learning algorithms. At the end of 5 weeks, both the TAC and SAC k-Nearest Neighbor algorithms performed slightly worse than the decision tree algorithms shown in fig. 3. The TAC-L_kNN algorithm produces an accuracy of 57.4%, compared to the TAC-L_DT algorithm produces an accuracy of 58.2%. The highest accuracy is produced by the split access count learning (SAC-L_DT) algorithm, at 64.3%. The SAC-L_kNN algorithm was also good, producing an accuracy of 63.3%. We performed a two-tailed t-test with $\alpha = 0.05$ for the SAC-L_DT algorithm and found a statistically significant improvement ($p=5.40E-12$) of 63.1% over that of a breadth-first crawl.

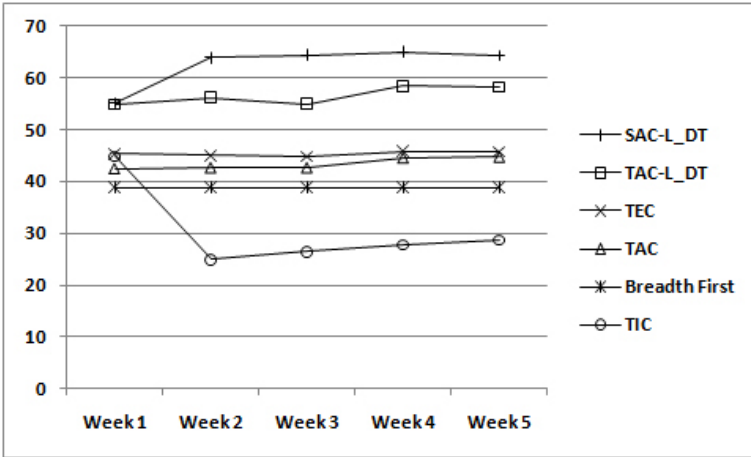


Fig. 3 URL Ordering Accuracy for DS1

5.3 Analysis of Results Obtained Using DS1

Table 1 gives the number and the average access counts of URLs/PageRank. From Table 1, we can see that the total number of URLs with PageRank 5 and 6 are much lower when compared to the number of URLs with other PageRanks. In addition, the average access counts do not increase linearly with the PageRank values. Table 1 also provides the average accuracy values/PageRank for all the popularity based URL ordering algorithms after Week 5.

Table 1 Accuracy, URL Count, and Hits per PageRank

PageRank	Avg. Accuracy	URL Count	Avg. External Hits	Avg. Unique Ext. Hits	Avg. Internal Hits	Avg. Unique Int. Hits
0	68.7	2672	20.8	14.8	3.6	2.2
1	0.6	240	8.9	8.4	1.1	1.1
2	29.5	993	22.7	19.5	2.4	1.9
3	38.7	1183	86.9	31.7	2.0	1.1
4	17.9	345	83.8	56.3	9.5	7.4
5	3.2	38	97.6	83.0	12.6	9.7
6	14.8	9	652.8	510.1	524.8	270.1

One observation from Table 1 is that all the algorithms seem to do well or badly on the same PageRank. For example, all the algorithms seem to perform better on

URLs with PageRank 0, 2 and 3 than they do on pages with higher PageRank because there are so few pages with high PageRank (only 2808 of the 5480 URLs have PageRank higher than 0 and only 392 URLs have a PageRank of 4 or higher). On the pages with moderate PageRank values, the popularity-based learning algorithms outperform the other algorithms, leading to their high overall accuracy. It is worth noting that all these PageRanks have a high number of URLs. In contrast, none of the algorithms do well for URLs with PageRanks 1, 5 and 6. For URLs with PageRank 1, we see that their average access count is much lower than the access count for URLs with PageRank 0 and 2. Although the average counts for URLs with PageRank 5 and 6 is higher than the access counts for URLs with lower PageRank, the poor performance of the popularity-based techniques may be due to the low number of URLs in this category. This is not surprising for techniques using learning algorithms since it is consistent with the axiom that the accuracy of a learning algorithm increases with more number of examples per category.

5.4 Discussion

From the results obtained in Sections 5.2 and 5.3, we conclude that, in general, the popularity-based learning algorithms order important URLs higher than breadth-first crawlers, a statistically significant result. Furthermore, our experiments show that page accesses from external domains are more important for URL ordering than page accesses from internal domains. Moreover, among the popularity-based URL ordering techniques, the methods that used learning algorithms outperformed the methods that used raw access counts. This shows that although the access count is correlated to the importance of the pages, they are not directly proportional as shown by the improved accuracy obtained by our learning algorithms (highest accuracy of 64.3%) when compared to techniques that make use of raw access counts (highest accuracy of 45.7%). In addition, this correlation may be different on different websites and learning algorithms may be able to identify this correlation and hence, find “important” pages better than non-learning algorithms.

In order to evaluate the effect of a different website, we used the best performing non-learning (TEC) and learning (SAC-L_DT) algorithms on a larger data set, the CiteSeer data set (DS2) with 102,360 URLs. Similar to DS1, we perform a five-fold cross validation and report the averages obtained. Recall that the CiteSeer data set does not have a hierarchical linking system, so we must use a random crawl to obtain a baseline. Using a random crawl baseline, we obtained an accuracy of 28.1%.

Fig. 4 provides the results obtained for TEC algorithm and the SAC-L_DT algorithm. From Fig. 4, we see that the accuracy of SAC-L_DT is 44.2% versus 28.1% for random crawl. We performed a two-tailed t-test with $\alpha = 0.05$. We achieve a statistically significant improvement ($p = 6.4E-17$) of 57.2% in our URL ordering algorithm over that of a random crawl on a large data set. This demonstrates that, once again, popularity-based URL ordering techniques outperforms a baseline (random) crawl.

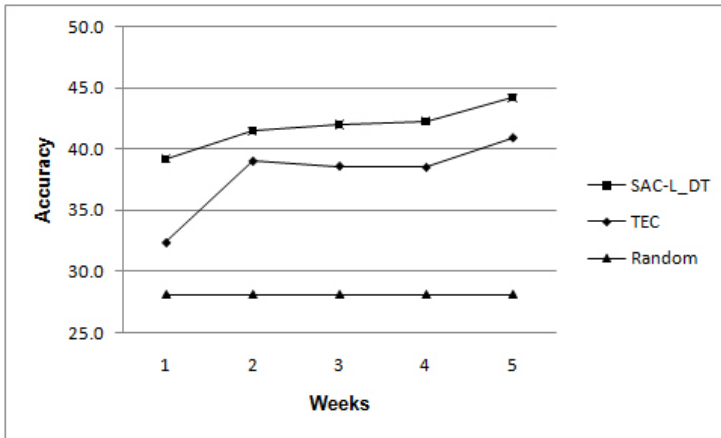


Fig. 4 URL Ordering Accuracy on DS2

6 Conclusions and Future Work

In this paper, we propose a new class of URL list creation and ordering algorithms based on file modification and popularity information from file systems and Web logs. Specifically, we present different strategies to discover new or modified pages and perform URL ordering using popularity information and compare our approaches to a breadth-first search crawls. Our evaluations show improved performance of these algorithms when compared to breadth-first search crawl.

This approach seems well-suited to social media settings where URLs may be shared in a variety of forms beyond links embedded in other web pages. In particular, the approach mitigates some of the difficulties in determining page popularity when the URL is shared in RSS feeds, text messages, or through URL shortening services. The combination of URL list generation and ordering in the system could keep up with quickly evolving social network media much more efficiently than traditional connectivity-based or breadth-first approaches.

One drawback with our approach is that new pages that have not been accessed are penalized. Adding last modified dates as a factor along with popularity information may address this issue. Another improvement to the experiment would be to rely on a larger collection where we could compute our own PageRank values, rather than relying on the discrete categories provided by Google. This would allow us to compare two full ordering algorithms to further discover how well the two methods' rankings match.

One concern with an approach that relies on web server logs for popularity ranking is the danger of manipulation to boost rankings. In the context of URL ordering, this is less of a concern, since the ordering on a site will primarily be used to modify the order of a crawl within a site rather than to request more pages from the server. In a result ranking context, one approach to mitigate the manipulation

problem would be to establish a similar ranking budget for each site, so that boosting one page only cannibalizes other resources on the site. Another option is to develop tools that can digitally sign logs and log harvesting tools to prevent external manipulation.

A final area of future work is to expand the scope of the information used for the URL list and ordering system. The URL ordering algorithms proposed in this paper are for ordering pages on a single website based only on file system and web log information. Understanding how to combine the popularity information from different logs to order pages from various websites could be another interesting problem to explore. Another possible source of information is directly from content management systems on large websites. This would allow dynamic or hidden web pages to be discovered before they appeared in Web logs.

Acknowledgment

This work was partially supported by NSF ITR 0225676 (SEEK).

References

- [Brandman et al. 2000] Brandman, O., Cho, J., Garcia-Molina, H., Shivakumar, N.: Crawler friendly Web servers. In: Proc Workshop on Performance and Architecture of Web Servers (PAWS), Santa Clara, California (2000)
- [Buzzi 2003] Buzzi, M.: Cooperative crawling. In: Proc. Latin American Conference on World Wide Web (LA-Web), Santiago, Chile, pp. 209–211 (2003)
- [Castillo 2004] Castillo, C.: Effective Web crawling PhD Thesis, University of Chile, Chile (2004)
- [Castillo et al. 2004] Castillo, C., Marin, M., Rodriguez, A., Baeza-Yates, R.: Scheduling algorithms for web crawling. In: Proc. Latin American Web Conference, Brazil, pp. 10–17 (2004)
- [Cho et al. 1998] Cho, J., Garcia-Molina, H., Page, L.: Efficient crawling through URL ordering. In: Proc. 7th World Wide Web Conference, Brisbane, Australia, pp. 161–172 (1998)
- [Cho et al. 2005] Cho, J., Roy, S., Adams, R.E.: Page quality: In search of an unbiased web ranking. In: Proc. 2005 ACM SIGMOD International Conference on Management of Data, Baltimore, Maryland, pp. 551–562 (2005)
- [Cho and Schonfeld 2007] Cho, J., Schonfeld, U.: RankMass crawler: A crawler with high PageRank coverage guarantee. In: Proc. 33rd International Conference on Very Large Data Bases, Vienna, Austria, pp. 375–396 (2007)
- [Najork and Wiener 2001] Najork, M., Wiener, J.L.: Breadth-first search crawling yields high-quality. In: Proc. 10th International World Wide Web Conference, Hong Kong, pp. 114–118 (2001)
- [Pandey and Olston 2008] Pandey, S., Olston, C.: Crawl ordering by search impact. In: Proc. of the International Conference on Web Search and Data Mining, Palo Alto, California, pp. 3–14 (2008)

Synergic Intranet: An Example of Synergic IT as the Goal of E-Engineering

K. Krzemiński¹ and I. Józwiak²

¹Institute of Physics, Wrocław University of Technology, Wrocław, Poland
Kamil.Krzeminski@pwr.wroc.pl

²Institute of Informatics, Wrocław University of Technology, Wrocław, Poland
Ireneusz.Jozwiak@pwr.wroc.pl

Abstract. Article is an attempt of formulating of synergy in IT system, synergic IT system definition and synergic intranet, relying on a design, construction, modification and maintenance of the cost effective solutions for practical problems in the field of the Information Society development within the Internet, with the use of scientific and technological knowledge. Authors introduce term of synergy in IT system, synergy types and methods of classification. They present also synergy image in Intranet. Article has to be start for formalizing new domain of informatics, called synergy in IT.

1 E-Engineering

The Information Society development resulting from a new information technology incidence, rapid development of a network communication attracts attention of a society to engineers, forerunners of the technology and Prometheus of the Information Society formation. More information about Information Society is possible to find in D. Bell's *The coming of Post-Industrial Society* [Bell 1973] and M. Castells's *The information Age: Economy, society and culture* [Castells 1997]. With full awareness we introduce the term of E-Engineering being a direction main goals of which are to provide the Information Society with a means of development.

E-Engineering relies on a design, construction, modification and maintenance of the cost effective solutions for practical problems in the field of the Information Society development within the Internet, with the use of scientific and technological knowledge. This activity requires solution of problems of different nature and scope. On the whole e-engineering deals with the development of the Internet technology and the way it can be used for satisfying the needs and requirements of the Information Society as well [Krzeminski 2010].

One of the goals of e-engineering is the development of existing solutions that support the development of information society. One such solution is an Intranet, which will devote the next subsections.

2 From Information to Synergistic Intranet

Intranet is one of the most rapidly developing Internet technologies of the last decade. During this time many Intranet definitions and classification arose.

Currently Intranet is a symbol of company modernity, certain technological vision of information management which in its whole complexity is human-dependent. The weakness of this vision is that there is no clear, unequivocal and standardized Intranet definition due to the scope of the issue and wide spectrum of researches conducted over the subject. Dynamical development of Internet technologies and changing with the same speed business expectations do not help [Guengerich 1996], [Suciu 2010], [Kelley 2010]. Hereby a proposition of definition unification is presented as well as two basic taxonomic classifications.

Intranet is a technology of development and use of IT system based on private computer network with limited quantity of strictly defined users having specified access to network services and resources.

Intranets can be classified taking into account distribution and functionality. Considering distribution two types of Intranet can be distinguished: universal and dedicated. The first one is a technology aimed on general clients and has very wide appliance, while the second one is created with the orientation on certain unique clients. Both distributions have many similar features known for all types of Intranet (Intranet Web, forums, announcement boards, video conferences, Intranet communicators etc.). The main difference lies in models, implementation and degree of business processes fulfillment.

As it was mentioned before, changing expectations on the market forced evolution of Intranet functionality. Those changes are presented on the picture below.

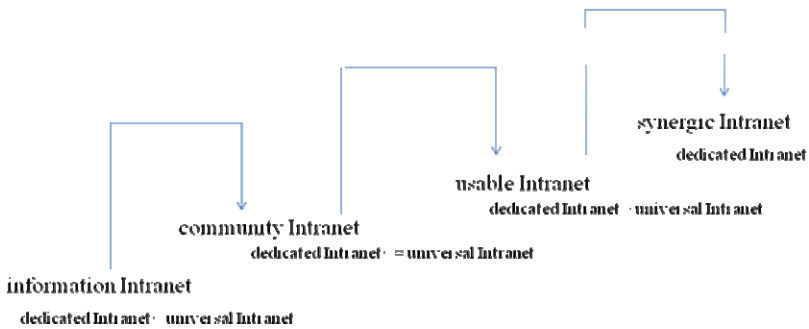


Fig. 1 Intranet evolution forced by changes of expectations on the business market

Intranet evolution forced by changes of expectations on the business market results in Intranet of fourth level. Information Intranet is a first-level Intranet. The main goal of information Intranet is information management. Among the most important activities on information the following should be listed: creation, storage, processing, download and use. Intranet of the second level is a community Intranet. Apart from first-level Intranet functionality, it was additionally enriched

by elements of Web 2.0, among which the one can mention forums, corporative knowledge portal such as wiki and Intranet community portals. In case of first- and second-level Intranets universal distribution is much more better than dedicated one, taking into account proportion of quality to price, because while both these Intranets propose the same functionality, box distribution is a cheaper solution. Usable Intranet is a third-level Intranet, enriched by usable Intranet applications such as i.e. Enterprise Resource Planning (ERP), Customer Relationship Management (CRM), Business Intelligence (BI), Human Resource Management (HRM), Material Requirements Planning etc. [Nielsen 2010], comparatively to two previous Intranets. Nowadays the majority of existing Intranets are third-level Intranets. Although business expectations force further Intranet evolution once again. This time dedicated Intranet is definitely much better solution, due to the fact Intranet usable applications are matched to company structure and process taking place in it. Universal distribution can be better solution for small companies where procedures can be easily modified or for new enterprises where there are no procedures at all and they can be prepared based on purchased IT solution. The following development level is a synergic Intranet with only dedicated distribution which is an issue of the next chapters.

3 Synergic IT System

The conception of synergic Intranet results from organic description of company structure, functioning and processes taking place in it. But before we go on with synergic Intranet, let us get closer to the idea of synergy in IT system [Sloan 2008] and synergic IT system.

To be brief, synergy in IT system is a cooperation of at least two factors which result in synergic effect being certain excess of such cooperation influence for cooperation factor with other factors, in comparison with possible benefit from individual activity of these factors. In other words, sum of effects from factors cooperation is greater than sum of separate effects of these factors when they do not cooperate with each other. Cooperating with each other factors create certain dynamic system which we will call synergic system. Its complexity depends on the amount of factors taking place in synergy. State of this system changes over time.

Synergic IT system is a system in which synergic effect takes place. Two types of factors can be distinguished in synergic IT system:

- static – do not change over chosen sufficiently long period of time, among which equipment resources can be named,
- dynamic – changing over time with big frequency, among which software, computer data, human resources, IT processes and even whole synergic systems can be named.

Besides the fact certain factors can be static, a system created by them can change over time. This influences on the complexity of an issue. Depending on correctness of cooperation of these factors, positive and negative synergy can be distinguished.

Positive synergy takes place when factors cooperation effect is subjectively useful; by analogy negative synergy takes place when subjectively unfavorable effects occur.

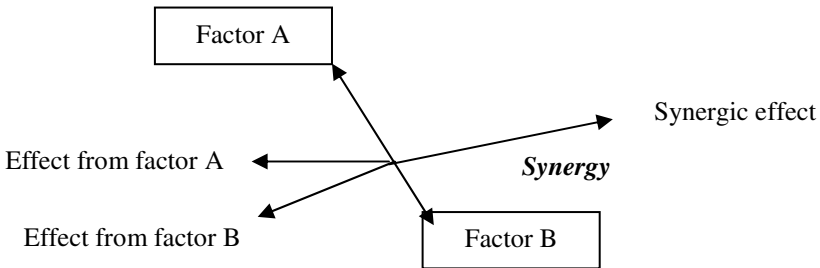


Fig. 2 Simple synergy schema in IT system

The above figure presents general schema of synergy, in which only two factors take part. Let us call two-factor system a system of second degree and by analogy a system in which n factors take part can be called a system of n degree, where n is integer. Complexity of synergic system in IT environment decides of how complicated synergy is.

Special attention should be paid to the initial state of factors because incorrect initial state of factors or small disorder of conditions for these factors in a system causes exponentially growing over time changes in behavior of the whole system. Well known name for this is a butterfly effect a meaning of which is that insignificant difference on one of the stages after some period of time can grow into huge sizes. Although a model can be deterministic (i.e. network structure, IT system architecture, business process model implemented certain application), butterfly effect can cause that in the longer time scale model seem to behave unpredictably [Ott 1997].

Synergy in IT systems can reveal in any of main elements which create these systems. Considering types of mentioned synergic factors there are several levels of synergy of IT systems. Levels of synergy are presented on a figure below.

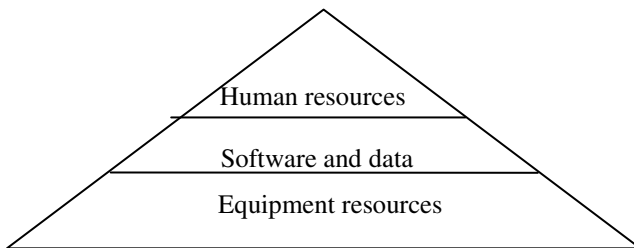


Fig. 3 Pyramid of synergy levels in IT system

The figure presents synergy levels in IT system. Well designed synergic system should be designed in accordance with ascending tendency. With the very same tendency the below synergy levels are presented:

- first-level synergy takes place when factors participating in cooperation are static which are various devices being part of computer system (computers), devices for data storage, devices for communication between hardware components of a system, devices for communication between people and computers, devices for data receiving from the outside, devices influencing systems of the external world, non-computer devices of data processing,
- second-level synergy takes place when factors taking place in cooperation are both static and dynamic which are different applications being part of IT systems such as operation systems of hardware, service programs and programs applied in certain area and information in a form of computer data, different resources,
- third-level synergy can be observed when factors taking place in cooperation are not only hardware and software resources but also human resources.

Second classification of synergy in IT systems is based on occurrence of synergic effect. Following types of synergy can be distinguished:

- intentional – synergy was designed and works properly, it can be easily controlled and managed,
- unintended – synergy wasn't planned during designed IT system but became revealed, it can be controlled but its management is not easy, sometimes even impossible,
- latent – synergy takes place in a system but system owner doesn't realize this, it cannot be controlled and managed until it is revealed. This type of synergy can become dangerous for IT system, latent cooperation of several factors at least one of which doesn't work properly will sooner or later lead to incomprehensible work of the whole system, incorrect functioning of IT system and can result in unimaginable financial losses.

Nowadays there are partially synergic applications on the market, even partially synergic IT systems can be seen although they are not called so. It is hard to count and list all the advantages of appliance of synergic IT system. Everything depends first of all on goals of use of such solutions, complexity of synergic system, applied synergy level. Advantages being synergic effect in every synergic IT systems are:

- long-term limitation of various IT costs,
- save up of time and financial costs by elimination of some processes and automation of other,
- functional complexity and essential advancing,
- increased system safety,

- system ability of evolving, rapid development and ability of extension in case an issue is well known.

Let us pay attention to disadvantages. The most important of them are given below:

- synergic system design is quite complicated, requires a lot of time and is connected with big risk which depends on system size and complexity of synergic system,
- synergic IT systems are expensive in implementation,
- it is difficult to develop synergic systems without complex knowledge,
- even small interference into one of factors taking part in cooperation causes disorder in the whole synergic system,
- incorrect work of at least one factor taking part in participation causes incorrect work of the whole synergic system,
- not-thoroughly thought removal of a factor taking part in cooperation can result in destruction of the whole synergic system or even whole IT system.

As you can see, there is no place for even tiny mistake in synergic system because in contrast to non-synergic system (where all factors are acting separately), along the given earlier rule, one small mistake in a dynamic system results in avalanche of further mistakes which may cause catastrophe of IT system. Such mistake may be difficult or even impossible to be removed from a system after some time because it is not easy to define the reason of the mistake due to the fact all the factors cooperate with each other. Time needed to find and remove a mistake is proportional to system complexity.

The above problems can scare. But synergic IT systems are our future. Following subchapter presents synergic Intranet.

4 Synergic Intranet

According to proposed unified definition, Intranet is an IT system, so all presented issues connected with synergy in IT system can be applied to Intranet. As it was mentioned earlier synergy feature can take place on several levels. First level of synergy is cooperation of static factors which are hardware resources. Symptoms of first-level synergy, where only static factors cooperate with each other, are computer stations being terminals and virtualization.

In order to optimize efficiency and reduce probability of accidents, use of terminals and virtualization require:

- use of star topology as for physical structure of LAN network. Damage of a single computer as well as adding of new computer doesn't have influence on other devices connected to a network if such solution is applied,
- synergic appliance of client-server architecture,

- atomicity which should be considered as separation of network services and resources, division of tasks on many (depending on needs resulting from IT system sizes) devices cooperating in a network playing role of domain servers (virtualization servers, services, application and data servers etc.), mass memory devices, supercomputers, terminals, network components of mass memories, network printers and so on.

Figure 4 presents the simplest structure of synergic Intranet network.

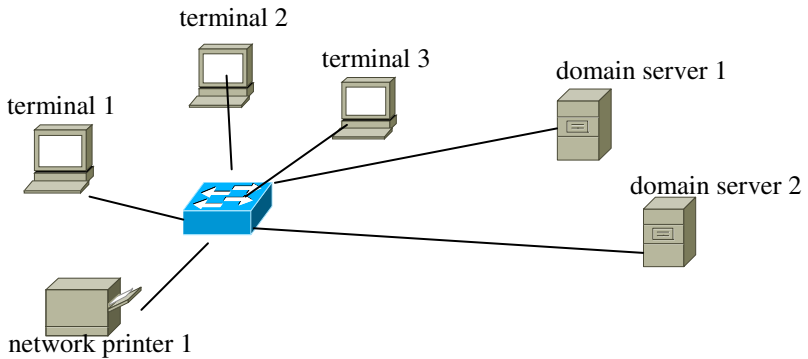


Fig. 4 Network structure of synergic Intranet

The above structure is presented in accordance with notation proposed by Cisco. More information on design of computer networks can be found in a book of Adam Józefiok [Jozefiok 2009].

Synergic cooperation can take place between following static factors:

- terminals with virtualization servers,
- virtualization servers with service servers,
- service servers with data servers.

Terminal is a device which allows to work in a synergic Intranet system. Terminal must have input device in order to input instructions and output device in order to present information to operator. Nowadays great popularity among terminal solutions have stations All-in-one. Using terminal users connects to the server where operating system is virtualized. User has impression that he works on real physical equipment. Actually references of virtualized operating system to those physical components of computer which would collide with other virtualized environments or owner's operating system, are caught by software responsible for virtualization and further emulated. Such emulation slows down work of virtualized environment that is why hardware support of virtualization is desirable.

In synergic Intranet user is not allowed to write data on his terminal, all data are stored on a data server and are accessible to other users, if there is such need, through application shared by service servers.

Cooperation of terminals, virtualization servers, service servers and data servers generate among others synergic effects such as reduction of costs, time and facilitating of different activities connected with physical structure of Intranet.

There are several examples presented below:

- long-term costs reduction (not terminals but domain servers are expanded),
- reduction of time necessary for hardware management (all workstation have the same architecture which results in among others easy administration and maintenance),
- reduction of time necessary for creation of new workplace by connecting of new terminal (save up of time connected with installation of software controlling hardware and users software),
- ease and efficiency of data management (creation, storage, processing, transmission, writing, search, use and control of information written in one place – on a data server), all pieces of information are stored on a server which influences on better data protection. Server can decide who has right to read and modify data, there are many technologies supporting operation, safety and usability of such solution.

An example of first-level synergy is very simplified situation.

There are two users in synergic Intranet, both work on stationary computers and regularly store data on their hard disks. Hard disks are of the same capacity 40 GB. Let us suppose that none of users cannot delete and archive already stored data making in this way more space on hard disk. First of them stores only 100 MB of data monthly while the second one stores 10 GB. After four months one of users won't be able to store data and suspends his work until additional mass memory is purchased. New hard disk should be purchased for him while the second user will use only 10 percent of his disk. There will be additional costs connected with purchase of mass memory.

In synergic Intranet users work on terminals and store their data on disk server where both disk of 40GB are dynamic synergic factors (their state changes over time) and cooperate with each other being connected in a disk matrix. This is the simplest synergic system of second complexity degree. Thanks to this operation the one can gain time (three additional months of work) and money earned by workers who don't suspend their work. Besides in every moment a state of mass memory can be controlled and needs can be overrun (mass memory extension), more over the one can react more quickly than in case of non-synergic systems. Once again time is gained which can be transformed to money. Using synergic system stored data can be also granted greater safety. But it is necessary to remember of dangers of i.e. incorrect operating of at least one factor taking part in cooperation. Incorrect work of one of disks would result in incorrect work of whole synergic system which in this case is created by disks. The result of this would be that both works would not be able to work further.

On the second level of synergy factors cooperating with each other are software, users application, information in form of data and different processes taking place in synergic Intranet. The above structural solutions require appliance of multilayer architecture. In order to enable synergy in Intranet system it is necessary to use mentioned type of architecture, allowed can be also appliance of distributed architecture. Hereby multilayer architecture is briefly presented and areas in which synergy can appear are indicated. In the simplest case our architecture can look as shown on figure 5.

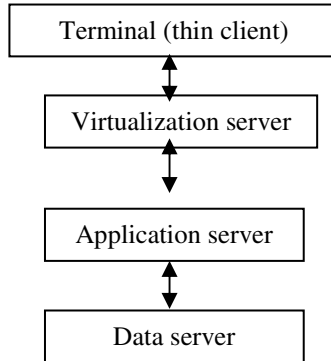


Fig. 5 Multilayer architecture in physical form

Every application used in synergic Intranet should work based on multilayer architecture in accordance with schema presented on a figure below.

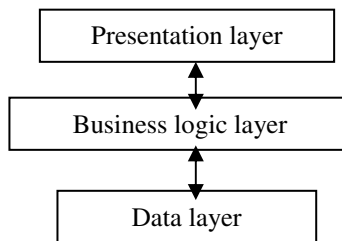


Fig. 6 Multilayer architecture of application

Intranet applications are dynamic synergic factors, their state can change over time. They can cooperate with each other and that cooperation can result in occurrence of synergic effect:

- synergy in presentation layer of two applications result in synergetic effect of information. Several separate pieces of information concerning certain event include not much of essential contents while respective information connected

with each other allow to understand the whole event through logic connection of facts,

- synergy in business logic layer, both applications proceed users requests, apply business rules and transfer data between layers cooperating with each other constantly or using module coordinating work of both these applications, which results in synergic effect of automation of business processes fulfillment,
- synergy in data layer results in synergic effect of data integration.

It is necessary to admit that not all applications should create synergic system. These systems must be created reasonably and in those areas where work can be facilitated and to make it more complicated.

Third-level synergy is the one where apart from factors participating in first- and second-level synergy, other factors take place – these are human resources. Cooperation of people (group work) causes synergic effect of human operations. To be brief, if two person make something well individually than cooperating with each other they will do it well also but apart from this as a result of synergic effect some additional thing will be gained, i.e. more time will be gained by quicker fulfillment of tasks. In order to make third-level synergy possible in an Intranet system, people must have well assigned and properly working tools on which first- and second-level factors of synergy have influence.

To conclude with, the below example presents how synergic and non-synergic Intranets fulfill business process.

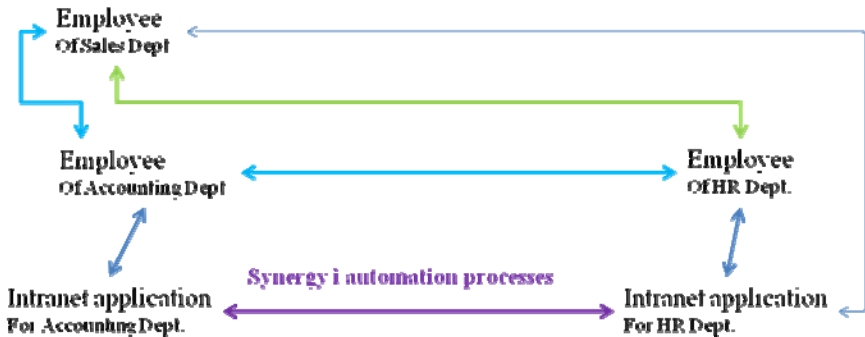


Fig. 7 Example synergy model in synergy Intranet

Reason: employee of Sales Department changes his place of residence.

Goal: change of tax office for employee of Sales Department and modification of personal data in HR Department.

In non-synergic system

Employee of Sales Department informs employee of HR Department of change of his residence place and eventually employee of Accounting Department. Employee

of HR Department modifies data about place of residence through Intranet application for HR Department. Then he informs employee of Accounting Department of this. The last one changes tax office for employee of Sales Department to proper one. This operation is done through application for Accounting Department.

In synergic system

Employee of Sales Department informs employee of HR Department of change of his residence place or employee of Accounting Department. Informed employee inputs proper data through application. Application for Accounting Department and application for HR Department cooperate with each other causing synergic effects: process automation results in reduction of amount of operations to be done which saves up time (only one employee is informed, one data update through one application is done).

In first case there are 6 steps to be done, in the second case there are only 3 steps. It is necessary to remind that this is very simple example. Let us imagine benefits generated in greater synergic systems.

5 Brief Summary

Intranet evolution tends in synergic direction. This synergy can reveal in different forms on different levels. Synergic IT systems can be powerful weapon in competitive battle giving advantage on a market which result from information synergy and synergy of operations between different divisions in company. Improvement and introduction of synergy to business processes and third-level synergy to these processes in a form of synergic components of group work and synergic systems of documents flow additionally raises the bar for competition. Synergic Intranets are developmental solutions for information society that is why development of such solutions should be overtaken by E-engineering. This technology must be accessible for public. Nowadays synergic IT systems are used by pioneers and experimenters while just tomorrow they can be widespread and become a basis of solutions in IT systems..

References

- [Bell 1973] Bell, D.: The coming of post-industrial society. New York, USA (1973)
- [Castells 1997] Castells, M.: The information age: Economy, society and culture vol. 3. Oxford, UK (1997)
- [Guengerich 1996] Guengerich, S., Graham, D., Miller, M., McDonald, S.: Building the corporate intranet. John Wiley & Sons, Chichester (1996)
- [Jozefiok 2009] Józefiok, A.: Construction of computer networks for Cisco switches and routers, Helion, Gliwice, Poland (2009)
- [Kelley 2010] Kelley, V.: INTRANETS: My company wants one-what's involved (2010), <http://www.mainstream.com/intranet.shtml>
- [Krzeminski 2010] Krzemiński, K., Józwiak, I.: Introduction to E-engineering. In: Proc. 3rd Int. Conf. on Human Systems Interaction, Rzeszów, Poland (2010)

Nielson, J.: 10 Best Intranets of (2010),

http://www.useit.com/alertbox/intranet_design.html

[Ott 1997] Ott, E.: Chaos in dynamical systems. WNT, Warszawa (1997)

[Sloan 2008] Sloan Career Cornerstone Center, Information systems. Alfred P. Sloan Foundation (2008)

[Suciu 2010] Suciu, P.: The basics: what is an intranet? (2010),

<http://technology.inc.com/networking/articles/200609/intranet.html>

Part III

Impaired Persons Aiding Systems

Electronic Systems Aiding Spatial Orientation and Mobility of the Visually Impaired

P. Strumillo

Institute of Electronics, Technical University of Lodz, Poland
pawel.strumillo@p.lodz.pl

Abstract. The problem of out-door mobility of the visually impaired and a review of key assistive technologies aiding the blind in independent travel are discussed in this paper. Space perception abilities important for mobility of the visually impaired are outlined and basic concepts such as: cognitive mapping, wayfinding and navigation are explained. Sensory substitution methods and interfaces for non-visual presentation of the obstacles and communicating navigational data are addressed. Current projects under way and available technologies aiding the blind in key mobility tasks such as: obstacle avoidance, orientation, navigation and travel in urban environments are reviewed and discussed. Special attention is paid to a class of teleassistance systems in which an operator at a remote site is capable of guiding the visually impaired. Finally, results of trials of the teleassistance system with participation of blind volunteers are reported.

1 Introduction

Common understanding of blindness and the needs of the visually impaired is poor, even in modern societies. The white cane and more rarely a guide dog are the primary mobility aids that are mainly associated with this disability. In spite of recent remarkable achievements in electronic technology and also information and communication technologies (ICT) the devices that are termed electronic travel aids (ETA) are very slowly fighting their way into the community of the visually impaired. In fact no single ETA has been widely accepted by the blind as a useful aid. Hence, it is important to focus on the current state of the problem of aiding mobility and safe travel of the visually impaired which is not marginal in terms of social and economic scale. According to recent WHO reports worldwide there are 314 million of visually impaired and 45 million of them are blind. These statistics will worsen due to aging demographics (about 82% out of all visually impaired are aged over 50). Currently, in the USA the funds expended on the costs related to blindness amounts to \$68 billion annually.

This chapter is devoted to the currently available solutions to key problems the blind indicate as the barriers complicating their every day life, i.e. independent

mobility and safe travel in urban environment. The material comprises two parts. In the first one the main definitions explaining fundamental mobility skills of the visually impaired such as space orientation, obstacle avoidance and wayfinding are explained. The second part focuses on description of various classes of electronic travel and navigation aids. Principle of operation of these, frequently complex devices, are explained and discussed. Novel navigation aids that employ RFIDs and ICT technologies will be addressed also. The chapter will conclude with description of the teleassistance system for the blind that was developed at the Technical University of Lodz, Poland.

2 Mobility of the Blind. Basic Definitions

The following concise yet comprehensive definition of mobility was given in [Hersh and Johnson 2008]: Mobility - “the ability to travel safely, comfortably, gracefully, and independently”. A frequently used term is also Orientation and Mobility (O&M) which widens the concept of mobility skills by the capabilities of being aware of one's position and heading in the surrounding space. The aim of this section is to propose a consequent definitions of many terms related to mobility that tend to be used inconsistently in literature. Short definitions of space perception, orientation, wayfinding, navigation, obstacle avoidance, etc. are given.

Space perception – is understanding the geometrical structure of the surrounding environment, awareness of self-location in that environment; knowing in terms of depth and directions the location of surrounding objects.

A number of studies have shown that the visually impaired use either body-centred or external referencing strategies for perceiving space. These studies conclude that the blind who use body-centered strategy perform better in understanding near space, whereas persons using the external referencing are more successful in understanding more complex scenes requiring information about objects located in the far space. Near space according to [Hall 1966] can be defined as a space encircling the person with approx. 4 m radius whereas the remaining part of the surrounding space is termed the far space.

There is also a significant difference in space perception by congenitally blind and those adventitiously blinded. Those who have become blind later in life, are better in understanding perspective and forming mental images of the remembered spaces or new spaces that were also examined by touch. This capability, however, tends to worsen with time or eventually can be lost.

Orientation – is an ability of being aware of one's body position (while at rest) and heading (while in motion) in relation to surrounding objects, cardinal directions and one's location in the followed path, e.g. in relation to the destination. Orientation can be subdivided into two categories [Gollage 1999]:

1. Spatial orientation – i.e. orientation in a near-space.
2. Geographic orientation – i.e. geographical orientation in a far-space.

Wayfinding – is the capability to select correct route from a network of routes that would lead from a starting point to a destination. It is predominantly a cognitive ability focusing on understanding the path layout and being constantly aware of one’s location on the path.

Cognitive mapping – is defined as a mental representation of a space important for wayfinding, referred also as “an internal wayfinding aid”.

There are three theories postulating how humans perform the wayfinding task:

1. Landmark based wayfinding – in which an ordered sequence of reference points and their spatial relationships on a path are used for wayfinding.
2. Route based wayfinding – route geometric characteristics are learned from resources available while planning a travel (e.g. from tactile maps).
3. Ordered scene views based wayfinding – a theory suggesting that wayfinding can be accomplished by learning a sequence of ordered “images” along a path.

Navigation – the process of purposeful control of ones movements, i.e. updating one's position and orientation (in a near and a far space) along a preselected route leading to a destination. Navigation can be associated with answering the questions: Where am I? Where I am going? How to get there?

Note that the terms of wayfinding and navigation are closely related and are used inconsistently in literature. The suggested distinction between the two is that wayfinding can be understood as the cognitive aspect of the travel task, whereas navigation is the operational aspect of the task.

Obstacle detection and avoidance – capability to safely detect obstacles and find a path to avoid them.

The list of the defined concepts that are related to mobility of the visually impaired can be summarized graphically. In Fig. 1 the mobility activities are subdivided into the perception based activities and into the action based activities. The former can be associated with the mental capabilities involved in mobility action whereas the latter with the operational skills of the mobility.

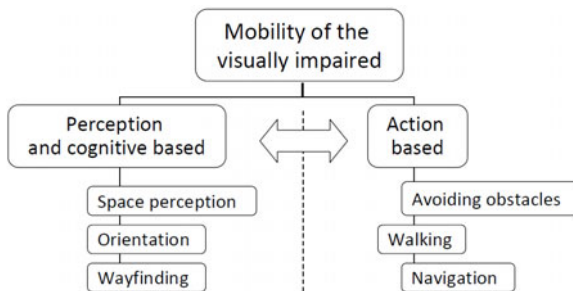


Fig. 1 The proposed model of mobility that combines perception and cognitive capabilities and other skills required for efficient mobility

3 Assistive Technologies for Aiding Mobility and Travel

Devising a useful assistive technology for aiding mobility and travel of the visually impaired has turned out to be a difficult interdisciplinary challenge. The history of electronic travel aids (ETA) reaches back over more than a century. In 1897 a Polish scientist Kazimierz Noiszewski built the electroftalm, a device that utilized the photoelectric properties of Selenium cells as an impractical, but still the first environment sensing device. The most significant advancements in ETA designs took place during the decades after the Second World War, when such technologies as ultrasounds, lasers and computer imaging became available. Currently, ETA is the general term encompassing a large class of assistive devices aiding the blind in mobility. Recently, a new class of systems are being designed and implemented that can be termed orientation and navigation systems (ONS) assisting the blind users in travelling to far and unfamiliar places. Here the following functional, rather than technological classification of these assistive devices is suggested:

1. Obstacle detectors
2. Environment imagers
3. Orientation & navigation systems

The first two classes of aids are personal (wearable) devices that scan the environment in personal and near spaces. On the other hand, the third group of aids are the systems that offer sensing of far spaces and can acquire data from larger scale distributed networks, e.g. sensor networks, digital maps or GPS.

Whatever the group of assistive systems under consideration, their main functional blocks are the ones shown in Fig. 2. Firstly, there is an environment sensing module. Its construction depends on the sensing modality (active or passive) used for acquiring information about the environment. The processing module is optional. In simpler designs the transformed modality is used directly for driving the presentation interface. Other, systems perform heavy processing of the acquired data (e.g. selection of obstacle in scene images) to select meaningful information for presentation. Finally, the presentation interface implements a predefined data conversion scheme to suit its format and physical nature for the senses (touch and/or hearing) substituting vision.

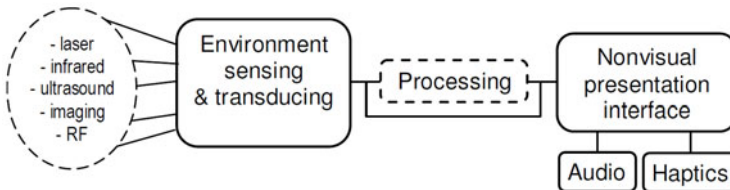


Fig. 2 A general block diagram of electronic mobility assistive devices for the blind

3.1 Electronic Obstacle Detectors

Electronic obstacle detectors are either hand held small devices or devices attached to the white cane. The principle of operation of such devices is based on measuring the time of a return path of a back-scattered light or an ultrasound beam. Examples of such devices that use the echolocation principle are: UltraCane (no longer produced), Bat 'K' Sonar Cane, MiniGuide and Palmsonar. The devices using ultrasound suffer from the effect of large divergence of centimetre acoustic waves. Consequently, location precision of obstacles is poor. The laser based devices are better in the precision of localizing obstacles, however, their operation can be interfered with ambient light. Available mobility aids of this class are: LaserCane and Teletact. The main advantage of electronic obstacle detectors is that they extend the cane protection range in distance and angle (e.g. provide head level protection). They are signaling obstacles to the user either by vibrations or special sound patterns (using different pitch or timbre of generated sounds). Computer simulations of different environment scanning scenarios used in such devices were conducted in [Bujacz 2005]. In Fig. 3 a simulation example taken from these studies is shown. An up to-date source of information on the electronic obstacle detectors can be found on the web page: www.bariery.ug.gda.pl/english/orientacja.html.

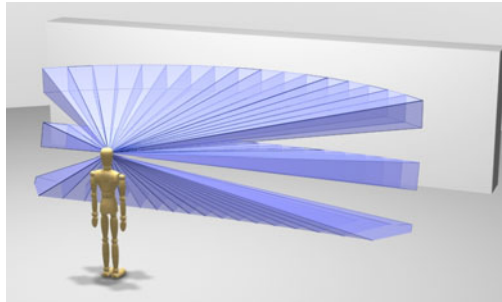


Fig. 3 Simulation of three-angle horizontal sweep scanning using a beam sensor [Bujacz 2005]

3.2 Environment Imagers

Environmental imaging systems are passive systems employing computer vision technologies to convert scene images into other modalities. These are much more complicated and more costly designs in comparison to electronics obstacle detectors. This is because a complex processing phase of image sequences need to be implemented to select the key objects (obstacles) for presentation. Recall that, information capacity of the non-visual senses is much smaller than vision.

Only one mobility aid of this class has been commercialised. This is the vOICE system. The vOICE converts grayscale images (collected with a single frame per second) into cyclic sounds repeated with each incoming image frame. However, because, no depth information is conveyed to the user, this aid is capable of representing only a flat projection of 3D scene geometry.

A large number of studies have been carried out on applying stereovision as a sensory module of the mobility aid. In such systems 3D structure of a scene can be reproduced. In [Bourbakis 2008] 3D range data is converted into a 2D vibrating array (32x32) attached to the human chest.

An innovative mobility assistive device is the BrainPort [BrainPort 2010]. It consists of a camera (mounted in sunglasses), transducer and a postage-stamp-size electrode array that is positioned on the top surface of the tongue. The recorded images are translated into gentle electrical signals activating tongue touch receptors. The generated stimulation patterns reflect key features of the recorded images. Currently, the BrainPort is an investigational device not available for sale.

Another scene presentation approaches use a technology for generating 3D sounds in stereo headphones. In the reported Spanish system the “Espacio Acustico Virtual” such an approach, i.e. stereovision to 3D sound conversion has been employed. Clipping sounds played simultaneously are used to represent elementary scene regions. However, this means of presenting spatial information is over cluttered and masks natural sounds coming from the surrounding environment.

A recent study carried out at the Technical University of Lodz implements a more advanced processing of the captured stereovision sequences. In this system a 3D scene model is developed with each recorded frame [Skulimowski and Strumillo 2007]. The model consists of planes (interpreted as scene context) and other objects interpreted as scene obstacles. The proposed scheme of presenting the environment to a blind user can be limited to a pre-selected number obstacles only. The HRTFs (Head Related Transfer Functions) are employed to obtain externalization of sounds generating an “auditory scene image”.

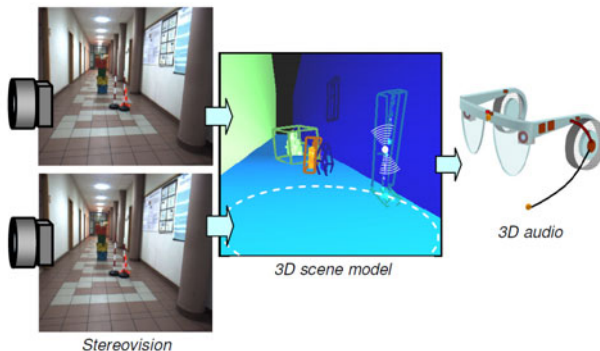


Fig. 4 A scheme for “auditory imaging” of 3D scenes

Blind users of the system report having a sensation as if the sounds are arriving from the location of the selected obstacle. Emphatic trials of the system are being run on a mobile notebook platform. The second version of the stereovision module is built into spectacles frames to improve ergonomic and aesthetic features of the rig.

3.3 Orientation and Navigation Systems (ONS)

The recent decade has brought new approaches to Orientation and Navigation Systems (ONS) for assisting the blind in mobility and travel. The technologies that have contributed to these solutions are:

- GPS (since 2000 a non-degraded signal has been made available to all users),
- GIS (Geographic Information Systems) and the technology of digital maps,
- wide access to the Internet (remote access to servers and databases),
- wireless communication networks (at near and far spaces, e.g. RFIDs, Bluetooth, WiFi, GSM and the awaited WiMax),
- miniaturized electronic inertial sensors and electronics compasses.

More importantly there are means of coupling the above technologies into integrated systems, e.g. cellular phones can be equipped with GPS receivers, inertial sensors, have digital map software installed, can be connected to other nearby devices via Bluetooth link and can communicate with the Internet.

Orientations and Navigation Systems for the visually impaired can be grouped further into a number main classes:

- Embedded infrastructures,
- GPS navigation systems,
- Urban travel aids,
- Teleassistance systems.

Embedded infrastructure is a system of electronic tags (buoys) attached to selected environment objects (e.g. signs, bank teller machines, building entrances). The tags are activated once the blind traveller enters their communication range (see Fig. 5). The tags identify themselves by sending a unique code to a reader device carried by the blind user. The received code can be converted into a sound or haptics message informing about the object type. The reader-tag communication can be implemented by using RFID (Radiofrequency Identification) technology or infrared beacons. Examples of such systems are the RFID Information Grid, and infrared bases systems: i.e. the TalkingSigns system developed at Smith-Kettlewell Eye Research Institute (www.ski.org), the Polish Blind-enT system model and the StepHear system (www.stephear.com). The cost of building the infrastructure of such an embedded orientation system is still very high, although, unit cost of the tags is relatively low.

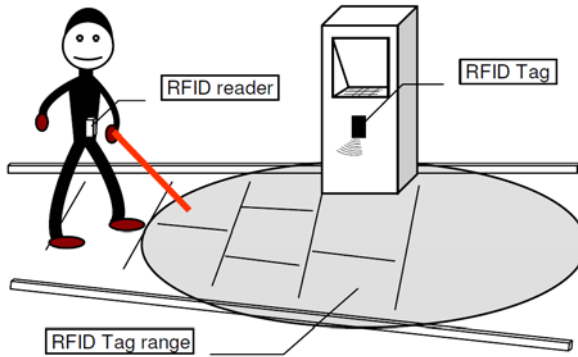


Fig. 5 Illustration of an embedded infrastructure system

Examples of commercially available systems or concluded projects on GPS navigation aids for the visually impaired are: BrailleNote, Trekker, Navigator, Easy Walk, MoBIC and Drithisi. Some of these systems are based on cellular phones and others use dedicated hardware for providing the required functionalities like: route preplanning, inserting electronic landmarks or warnings, speech synthesized information on nearby POI (Points of Interest). Note that functional requirements of GPS navigation systems for the blind should be based on different navigation software, offer better positioning accuracy and feature properly designed audio or tactile interfaces.

Recent studies on ONS focus also on aiding travel tasks of the visually impaired in the urban environments. These systems integrate a number of different ICT technologies with servers storing data on urban infrastructures (public transport timetables, important on-line information etc). The RouteOnline system combines UMTS/HSDPA, D-GPS, RFIDs (mounted in pavement tiles) and the Internet to continuously track the position of the blind traveller. Moreover, an access to a dedicated Web server enables provision of user defined on-line information on public transport and nearby points of interests (POI).

The prototype of a Finnish NOPPA navigation and guidance system offers passenger information (time tables, route planning), navigation and pedestrian guidance in urban environment. The system integrates GPS, GSM, Bluetooth and the Internet technologies. On the user side a PDA is used as a mobile terminal playing the role of the communication modem and user interface for the NOPPA.

In Poland a speech synthesized software for mobile devices is marketed as the ITINER and offers trip-planning, timetable browsing and on-line information of tram/bus locations (www.itiner.pl).

4 Teleassistance Systems

An innovative class of ONSs are based on guiding the visually impaired person by a remote human guide termed also teleassistance systems. Such a remote guidance

concept was first proposed at the Brunel University, UK [Garaj et al. 2003] and later undertaken in a simplified version by the Polish Design-Innovation-Integration Company. However, no continuation of these research studies were reported.

Since a number of years a research work has been under way at the Technical University of Lodz that has resulted in building a prototype version of a new teleassistance system. The system consists of two terminals. The mobile terminal which is a device carried by the blind traveller and the remote terminal, which is a PC operated by a sighted assistant. A schematic of this teleassistance system and the employed ICT technologies are indicated in Fig. 6. The video stream recorded by a miniature camera (mounted on the blind person's chest) is transmitted to the remote operator. The operator observes the scene in front of the blind traveller, verifies the geographic position on the digital map and instructs the guided person through spoken commands.

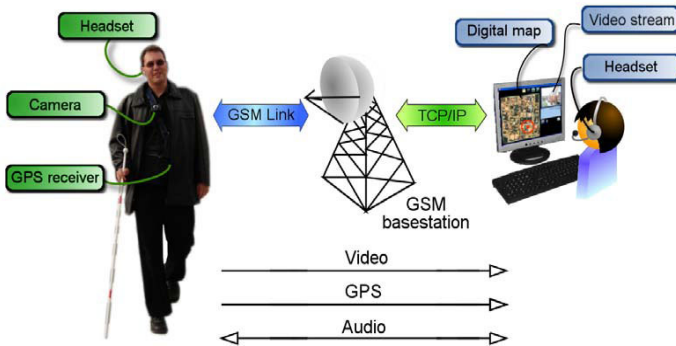


Fig. 6 The teleassistance system for aiding the blind in mobility and navigation

The first prototype of the mobile terminal of the system was based on a small laptop. Current version of the terminal is housed in a box of a PDA size (see Fig. 9 showing a blind person equipped with the mobile terminal). The device comprises an embedded processor controlling other system hardware units i.e.: a GPS receiver, a digital camera and a headset (see Fig. 6). The terminal transmits the data via a Bluetooth link to a mobile phone that plays the role of a modem of the mobile terminal. Currently a digital camera featuring the following frame rate/resolution modes is employed: 5 frames per second of 160x120 pixels image, 2.5 frames per second of 320x240 pixels image and 1 frame per second of 640x480 pixels image. The user can switch the camera modes at run time. The frames are compressed with the use of the JPEG standard.

On the remote operator's end a dedicated application is run that allows to remotely guide the blind person. A view of the graphical user interface of the remote

operator's terminal is shown in Fig. 7. Note in the upper-left panel of the interface a widow displaying the relayed video stream that is recorded by the blind person terminal. In the lower-left panel a number of control icons are viewed that allow the operator to monitor the transmission quality of the link and e.g. check the battery charge level of the remote terminal. The right had side panel displays a satellite view the terrain with an arrow indicating whereabouts of the guided person.

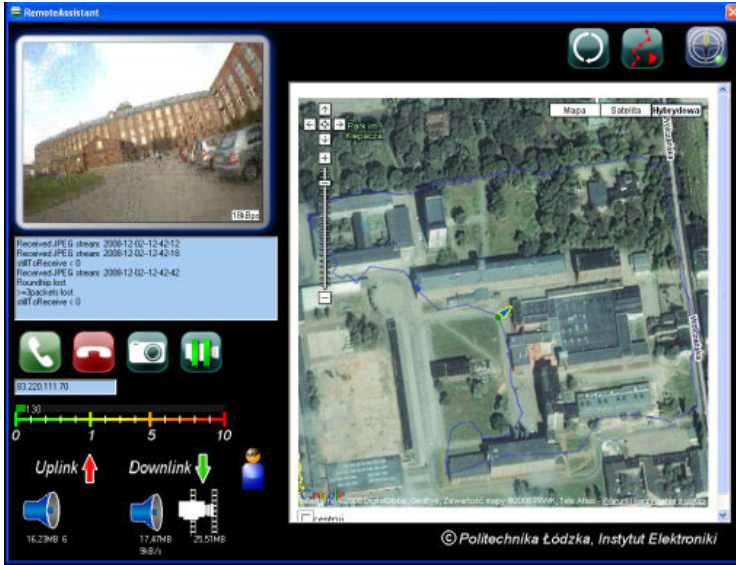


Fig. 7 Remote operator's graphical user interface (the relayed image displayed in upper-left corner features the lowest resolution mode of 160x120 pixels)

4.1 First System Trials

Initial prototype trials of the teleassistance system for the blind were conducted at the University campus with the laptop-based version of the mobile terminal [Bujacz et al. 2008]. Three blind volunteers participated in the trials. They were asked to follow two different paths of average length ca. 200 m. First the volunteers were guided along the paths with a human guide, who was commenting on the way all difficult turns, obstacles and possible landmarks. After these introductory walks, the blind volunteers considered these paths as familiar ones. Then they were asked to perform two more walks along each path (in a random order). The trial walks were supervised by a sighted observer whose role was not to guide the blind person but to take preventive action in case dangerous collision was imminent (such events were noted for the records). The first type of walks

were the independent walks in which the visually impaired used only a white cane as a primary aid and a second walk type in which the volunteers were guided by a remote operator. Results of these trials are summarised in a bar diagram shown in Fig. 8. Note a considerable reduction of unwanted events like missteps, minor and dangerous collisions and finally lost way occurrences. No lost way events were recorded for the assisted walk. The blind volunteers gave enthusiastic comments about the concept of remote assisted navigation. They followed the path more smoothly, faster and reported the feeling of being more confident and safer thanks to constant two-way voice connection with the remote operator. See also the project web page www.naviton.pl (in research outcomes link) for video recordings of the conducted trials.

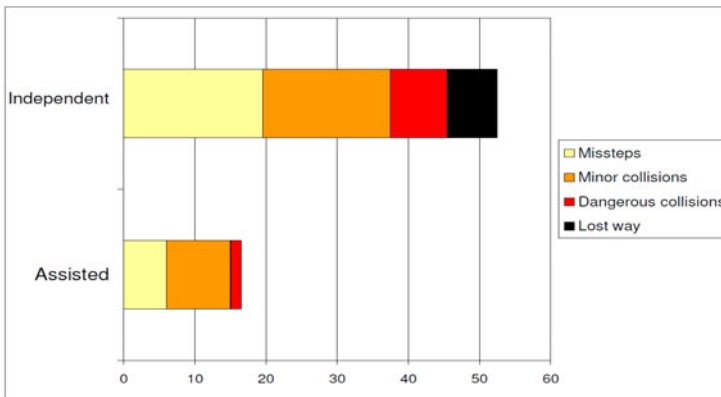


Fig. 8 Comparison of an average number of events that occurred in independent and assisted paths following (average length of the traversed paths was ca 200 m)

Recently a new version of the mobile terminal has been built [Baranski and Strumillo 2010]. The laptop from the first prototype was replaced by a dedicated hardware design of a size of a PDA. It underwent a number of seminal trials in urban terrain. See Fig. 9 showing a snapshot from the training session and trials with the blind participant.

Currently, the system undergoes a longer term test by a blind individual. The comments about the system functionality that were thus far reported from its user are the following.

Advantages:

- generally very useful aid in navigating in urban environment (particularly along unfamiliar paths), e.g. very helpful in returning to a route after straying from it,
- finding an alternative way around roadwork sites,
- finding correct platform on a station, identifying the right carriage,
- identifying entrances to buildings and doors within buildings,

- getting operators's warning message about unintentional stepping on the road lane,
- helpful in shopping (comments from the operator about the displayed items e.g. its price and colour).

Shortcomings:

- annoying antenna cable,
- to narrow view angle of the camera (approx. 50 deg),
- better camera resolution and frame rate necessary for improving the guiding precision and comfort of work of the operator.

It was also noted, that the remote operator should be familiar with the city area in which he undertakes the responsibility of guiding a blind person Hardware design of a consecutive version of the mobile terminal of the teleassistance system is aimed at further improvements of system functionality and ergonomics.



Fig. 9 A blind volunteer taking part in the trials of the teleassistance system

5 Discussion and Conclusions

Advancements of electronics technologies (e.g. miniaturization of such devices as sensors, GPS receivers, electronic compasses) and ICTs (wireless communication, 3G mobile systems such as UMTS) open new prospects offering unprecedented functionalities for mobility and orientation aids for the visually impaired.

However, the following goals need to be aimed at while designing and manufacturing the assistive aids and also training the blind persons in using new devices:

1. Identifying what information is needed for the blind person and when it is needed to enhance his mobility capabilities according to the assumed aid functionality, i.e. obstacle avoidance, orientation, navigation, urban travel etc.
2. Devising a suitable interface for nonvisual presentation of the information about the environment.
3. Proper ergonomic design of the device so that it is accepted by the blind users.
4. Providing a scheme for training the users in efficient use of the aid.

Irrespectively of the technology used for substituting vision, the wide sense of safety and ergonomics of the assistive device should also be the main design objective. This applies to the robustness of the assistive device in performing the mobility aiding task, level of attention required to handle the device, the modality of the non-visual interface used and finally the comfort of use.

Having a number of years research experience in building O&M aids I would indicate the teleassistance systems as the class of orientation and navigation systems that can shortly bring truly useful solution in supporting mobility and travel activities of the visually impaired.

As a concluding remark to this communication it is appropriate here to cite a maxim descended from the Smith-Kettlewell Eye Research Institute webpage (www.ski.org) concerning design questions one should ask before setting off to a challenging task of building assistive devices for the blind. i.e. instead of asking: “*How can this technology be adapted to the blind user?*” we should to ask: “*What information is actually needed by a visually impaired traveller and how it should be presented to him/her?*”. If we ask these questions more often, our research or any other efforts on improving quality of lives of the visually impaired will be certainly more successful.

Acknowledgment

This work has been supported by the National Centre for Research and Development of Poland grant no. NR02–0083–10 in years 2010–2013.

References

- [Baranski and Strumillo 2010] Baranski, P., Strumillo, P.: A remote guidance system for the blind. In: 12th IEEE Int. Conf. on e-Health Networking, Application & Services, Lyon, France, pp. 386–390 (2010)
- [BrainPort 2010] BrainPort Vision Technology webpage, <http://vision.wicab.com> (accessed on January 20, 2010)
- [Bujacz 2005] Bujacz, M.: Stereophonic representation of a virtual 3-D scene – simulated mobility aid for the blind. MSc Thesis, Technical University of Lodz (2005)

- [Bujacz et al. 2008] Bujacz, M., Baranski, P., Moranski, M., Strumillo, P., Materka, A.: Remote mobility and navigation aid for the visually disabled. In: Sharkey, P.M., Lopes-dos-Santos, P., Weiss, P.L., Brooks, A.L. (eds.) Proc. 7th Int. Conf. on Disability, Virtual Reality and Assoc Technologies with Art ArtAbilitation, Maia, Portugal, pp. 263–270 (2008)
- [Burbakis 2008] Burbakis, N.: Sensing surrounding 3-D space for navigation of the blind. *IEEE Eng. in Med. and Biol. Mag.*, 49–55 (2008)
- [Garaj et al. 2003] Garaj, V., Jirawimut, R., Ptasinski, P., Cecelja, F., Balachandran, W.: A system for remote sighted guidance of visually impaired pedestrians. *Br. J. of Visual Impairment* (21), 55–63 (2003)
- [Gollage 1999] Gollage, R.G. (ed) *Wayfinding Behaviour: cognitive mapping and other spatial processes*. The John Hopkins University Press, Baltimore (1999)
- [Hall 1966] Hall, E.T.: *The hidden dimension*. Anchor Books, New York (1966)
- [Hersh and Johnson 2009] Hersh, M.A., Johnson, M.A.: *Assistive technologies for visually impaired and blind people*. Springer, London (2008)
- [Skulimowski and Strumillo 2007] Skulimowski, P., Strumillo, P.: Obstacle localization in 3D scenes from stereoscopic sequences. In: Proc. of the 15th European Signal Processing Conf., Poznan, Poland, pp. 2095–2099 (2007)

Towards Vision-Based Understanding of Unknown Environments

A. Śluzek^{1,2} and M. Paradowski³

¹ Nanyang Technological University, Singapore
asssluzek@ntu.edu.sg

² Nicolaus Copernicus University, Toruń, Poland
asssluzek@fizyka.umk.pl

³ Wrocław University of Technology, Poland
mariusz.paradowski@pwr.wroc.pl

Abstract. The paper demonstrates how to transform (using a combination of techniques reported in our previous papers) a collection of random images gathered in an unknown environment into a limited-scale visual model of that environment. The model generally consists of the template images of the typical “visual objects” identified in the explored world. Both the concepts of objects and their templates are formed without any assumptions about the content of acquired images, i.e. the semantics is built using the pictorial data only (although users may subsequently identify the real-world semantics of the formed objects). From the image processing perspective, the method consists in detecting near-duplicate (i.e. photometric/geometric distortions and partial occlusions are allowed) fragments in random images. It is envisaged that such a proposal can be instrumental in assisting both autonomous agents and visually impaired humans (including both blind people and people unable to understand perceived visual data) facing unfamiliar worlds. The paper focuses on the practical aspects of the problem (exemplary results, computational efficiency, etc.) although a substantial amount of theoretical background is also included.

1 Introduction

Vision is certainly the most powerful human sense, and machine vision can become the most powerful sensor of (semi-)autonomous agents embedded in natural worlds. However, the semantic gap between the raw visual data and their meaningful (i.e. relevant to the context of performed tasks or other current needs) interpretation seems to be one of the most significant limitations of machine vision. Actually, humans can also experience such difficulties (e.g. in case of certain brain damages resulting in broken paths between retinal images and the cortex-located categorization and localization modules, [Riesenhuber and Poggio 2000]). In such cases, both people and autonomous agents can usually handle objects (once grasped or touched)

but the major difficulty is to perceive the object's presence in a visually complex environment. Moreover, the sheer understanding of such objects can be difficult, especially if they have not been encountered before.

In this paper we discuss a machine vision methodology that, in our opinion, can provide some assistance in such cases. In particular, we believe that systems equipped with such mechanisms could develop a certain "visual understanding" of unknown worlds so that they can either support visually impaired humans or autonomously explore the unknown worlds. Our objective is not to develop a universal machine vision system. Instead, we propose a limited-scale technique for detection and generalization of the most typical components in captured images.

In general, the presented methodology incorporates:

1. A machine vision module for detecting nearly identical fragments in random images. The theoretical background is presented in Section 2, and exemplary experimental results are overviewed in Section 3.
2. An adaptively modified visual database containing both recorded images and models of automatically built *visual objects*. The issues related to the database building are discussed in Section 4. Section 5 concludes the paper.

2 Background and Principles of the Method

Although *image similarity* can be differently defined in various applications of machine vision, in CBIR (content-based image retrieval) it is often based on the concept of *near-duplicate* images (e.g. [Zhao et al. 2007]). *Near-duplicates* are basically the same scenes distorted by occlusions and minor content changes, captured from a different viewpoint, under different photometric conditions and/or by a different camera, etc. If the objective is to detect near-duplicate fragments present in two or more images, the terms sub-image retrieval or image-fragment retrieval can be used instead (e.g. [Ke et al. 2004]).

Near-duplicate image fragment retrieval is the first fundamental operation of the presented method which attempts to detect fragments of database images similar to unspecified fragments of the current image (query). Thus, the method can identify that something previously seen exists in the observed scene. It should be noted that such a concept differs from typical vision-based navigation systems (e.g. assisting blind humans, [Bourbakis and Kavraki 2001]) that focus on building 3D models of the observed scenes. Our objective is to highlight the presence of "familiar" contents only.

The accumulated results of near-duplicate fragments detection are subsequently generalized. Mutually similar fragments from multiple images are grouped to form "visual objects" which are assumed to be typical components of the explored environment (explanations are in Section 4). Therefore, near-duplicate fragments in the future query images not only can be detected but also classified (if similar to one of formed "visual objects"). Thus, a certain level of visual semantics and vision-based understanding can be developed in the explored world that originally has been totally unknown.

Several images with very different contents, but containing similar fragments (which represent the same physical objects) are shown in Fig. 1. For example, Figs 1.A, 1.B and 1.D share two near-duplicates, while Fig. 1.C shares only one near-duplicate with the remaining images.



Fig. 1 Images containing near-duplicates (appearances of the same objects)

2.1 Affine-Invariant Image Fragment Matching

If two near-duplicate image fragments are appearances of a planar object, their geometric relations can be modeled by affine transformations (e.g. [Mikolajczyk and Schmid 2004]) which are considered accurate approximation of perspective projections for planar objects moving in a 3D space. Even if the surfaces of objects are not planar, they can still be locally mapped by affine transformations (using piecewise-linear approximations of the surfaces).

It should be noted that such a representation of similarities between various appearances of objects is also used in the proposed models of human vision, e.g. [Biederman 1987]. 2D appearances of 3D solids are represented by families of *geons* which contain shapes related by (approximately) affine transformations.

Therefore, the problem of near-duplicate sub-image detection can be approximated by detecting visually similar image fragments related by affine transformations. Given a query image Q and the image database S , we attempt to localize in images from S all fragments (with no assumption about presence and locations of such fragments) which are that are related by affine transformations to unspecified fragments of Q . The visual similarity between image fragments is built over similarities between sets of keypoints extracted from matched images.

2.2 Keypoint-Based Visual Similarity

Keypoints (also referred to as *interest points*, *visual saliencies*, etc.) indicate image fragments with distinctive visual properties. The distinctiveness should be prominent enough to ensure that whenever the same object appears in several images (possibly captures under diversified settings) the majority of keypoints can be always identified. Therefore, it can be assumed that by matching similar keypoints detected in a pair of images we can identify similar fragments (near-duplicates) in both images.

The original keypoint detectors (proposed almost 30 years ago) were simple corner detectors, i.e. local operators of fixed size detecting local maxima of image

intensity variations. Robustness and repeatability (under geometric and photometric distortions of images) of keypoints extraction by corner detectors were limited. However, the next generations of detectors (e.g. [Mikolajczyk and Schmid 2004; Matas et al. 2002]) can more robustly extract keypoints in images deformed by strong distortions (including scale, viewpoint and illumination changes). Such keypoints are often affine-invariants and they are typically depicted as circular or elliptical areas (representing the local scales/orientations/anisotropy of image intensities) so that the term *keyregions* is often used interchangeably.

In the presented approach, any reliable keypoint detector can be used but we generally assume (for the reasons explained in Subsection 2.4) affine-invariant detectors that extract keypoints in a form of elliptical patches. Examples of such keypoints are given in Fig. 2.



Fig. 2 Exemplary images and their keypoints (extracted by using Harris-Affine detector, [Mikolajczyk and Schmid 2004])

Keypoints are represented by various n -dimensional descriptors (characterizing local distributions of image gradients or shapes) so that similarities between keypoints can be measured as Euclidean distances between n -dimensional points. The proposed framework can incorporate any typical keypoint descriptors, but we generally apply SIFT descriptor which seems to be the most popular choice in CBIR applications.

2.3 Affine-Related Fragments as Histogram Maxima

Numerous CBIR algorithms have been recently proposed on how to derive the image similarity from the local similarities between individual keypoints. However, very few algorithms (e.g. [Zhao and Ngo 2009]) address the issue of near-duplicate fragment retrieval in unknown visual data. While [Zhao and Ngo 2009] consider only the similarity transformations, we propose a more general solution where near-duplicate fragments can be related by affine transformations. Details of our approach are presented in [Paradowski and Śluzek 2010] but the basic ideas are included in this paper for completeness.

Since any two triangles (e.g. three non-colinear pairs of similar keypoints) define an affine transformation, a large number (nearly n^3 , where n indicates the number of similar keypoint pairs) of affine transformations can be built to relate

two images. Affine transformations have six parameters so that a 6D histogram of affine parameters can be built in the corresponding parameter space from all such transformations.

Because the transformations are built using randomly selected triangles of similar keypoints, we generally expect randomly distributed (i.e. uniformly flat) histograms. However, if certain regions from both images are similar (i.e. there are two subsets of correspondingly similar keypoints related by the same affine transformation) a local maximum of the histogram would be formed because all pairs of triangles from these regions generate (approximately) the same affine transformations. Alternatively, a local maximum of the histogram defines an affine mapping relating a large number of triangle pairs. These triangles must, therefore, originate from two regions (one from each image) related by this transformation. The approximate outlines of the near-duplicate regions can be obtained by drawing the convex hulls of all triangles involved.

This observation is the central principle of the proposed technique of near-duplicate fragment detection. It should be noted that no prior knowledge about the image contents is needed to identify such near-duplicates.

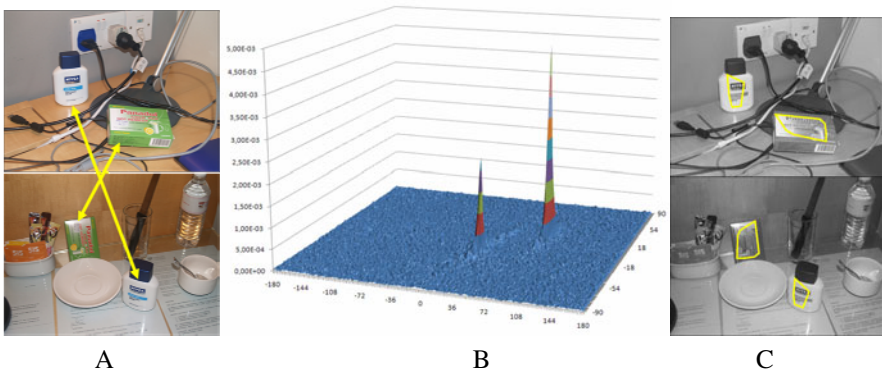


Fig. 3 A pair of images with near-duplicate fragments (A), a 2D projection of the affine transformation histogram showing two prominent spikes (B) and the corresponding outlines of the detected near-duplicates (C)

Fig. 3 shows an exemplary pair of different images containing, nevertheless, two similar objects (indicated by arrows in Fig. 3A). The histogram built over random affine transformations between triangles of similar keypoints has two prominent spikes (Fig. 3B) so that two pairs of near-duplicates regions are found (outlines shown in Fig. 3C). Since a 6D histogram cannot be visualized, only its 2D projection (onto the subspace of two affine parameters) is presented.

In practice, several factors should be taken into account before such an idea can be implemented. Most importantly, typical images of natural scenes contain thousands of keypoints (with the correspondingly large numbers of keypoint pairs

matched in two images). Thus, the complexity of the process of affine transformation building is bounded by $O(n^3)$, i.e. nearly n^3 triangles can be build using n matched keypoints, the methods is computationally intensive. However, we can preliminarily limit the number of processed triangles by:

- Ignoring very large triangles (it is usually unlikely that large parts images are affine-related and, if so, such cases can be detected as super-positions of smaller triangles) and very small/narrow triangles (which are usually inaccurately extracted).
- Building the triangles over limited neighborhoods of keypoints. If only N closest neighbors of each keypoint are used, the total number of transformations is bounded by $O(Nn^2)$. Typical values of N range from 40 to 70.

Dimensionality of the histogram space is another inconvenience. It is impossible (because of the memory capacity needed) to directly build and process 6D histograms. Instead, they are built using hash-tables. Each histogram bin is represented by a single entry in the hash table containing data about the corresponding affine transformations (including all pairs of triangles contributing to that bin). In this way, hundreds of thousands of transformations can be built and processed for a pair of images.

Finally, we have identified that the algebraic representation of affine transformations is not very descriptive in specifying deformations of the underlying image objects. Therefore, affine transformations are eventually decomposed into more meaningful forms. We actually use both SVD decomposition, and a decomposition emulating 3D motions of planar objects (more details in [Paradowski and Śluzek 2010]). Fig. 3B actually shows a histogram of two angular parameters of SVD-decomposed affine transformations built for the given pair of images. It can be clearly seen that both pairs of near-duplicate fragments are related by transformations with different values of the angles. The visual inspection of Fig. 3A confirms this mathematical conclusion.

Even with the proposed improvements, the algorithm is still computationally intensive. Using moderately optimized Java codes run on Intel Core 2 DUO CPU 2.66GHz processor, typical pairs of images are analyzed in 1-1.5sec. Such results are obviously unsatisfactory for the intended application where the algorithm should interact with a human or an autonomous agent performing in a real world.

2.4 The Ellipse-Based Fragment Matching

The alternative method of building affine transformations between two images assumes that the available keypoints are actually elliptical keyregions. Instead of triangles, only pairs of similar keypoints are used (which results in the theoretical complexity of $O(n^2)$). However, because the principles of the transformation building are unchanged, the remaining points needed to form matched triangles should be found in a different way.

This third pair of matched points is obtained from the shapes of keypoint ellipses. Although details of this solution are not discussed here (they are available in [Paradowski and Śluzek 2010]) a self-explaining illustration how to find such pairs of points is provided in Fig. 4.

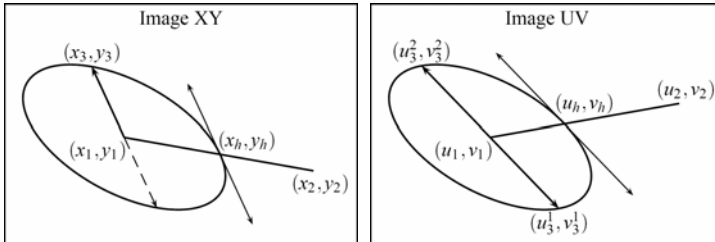


Fig. 4 By intersecting the matched ellipses with shifted tangent lines, the third pair of points needed to define the affine transformation can be obtained (two alternative options)

The actual complexity of the ellipse-based transformation building is $O(Nn)$ because we use only a limited number of keypoint neighbors to create the matched structures (see Subsection 2.3). The resulting processing time is correspondingly lower. For a given pair of images, near-duplicate fragments are typically found within 100-150msec. From the perspective of the intended applications (both in assisting humans and in autonomous agents) this can be considered a real-time performance. However, the actual timing characteristics strongly depend on the number of database images matched to the query. For larger databases, a reliable pre-retrieval mechanism is indispensable or, alternatively, the size of the database should be reduced because otherwise the system response would be unacceptably slow. The pre-retrieval module is presented in [Paradowski and Śluzek 2010a] while the effective method of reducing the number (and size) of images in the database is discussed in Section 4.

3 Performances and Results

Performances of the proposed method of detecting near-duplicate image fragments have been extensively tested on our specialized database containing 100 diversified indoor and outdoor images (available at <http://www.ii.pwr.wroc.pl/~visible>). Each image contains one or a few instances of fifteen preselected objects (which are generally the only similar objects shared by the images, although several cases of accidental local similarities are also encountered). Thus, the ground truth for image matching can be established and the system's performances can be objectively measured.

Six exemplary images (including two images already shown in Fig. 3A) are given in Fig. 5. This set of images will be used as an example in Section 4.

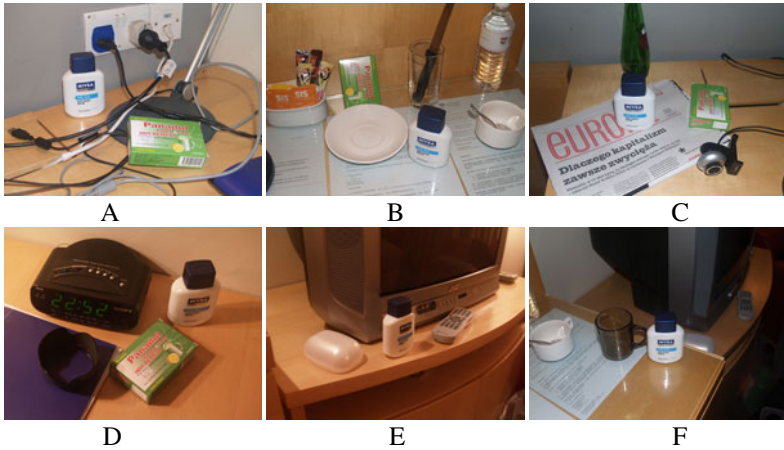


Fig. 5 Exemplary indoor images from the testbed database

Accuracy of the method is estimated based on two popular measures used in information retrieval, i.e. *precision* and *recall*. However, we consider two different aspects of these measures:

- *Object-based* accuracy indicates how many near-duplicates are correctly detected.
- *Area-based* accuracy indicates how accurately the extracted areas of near-duplicate outlines correspond to semi-automatically established ground truth outlines (more details in [Paradowski and Śluzek 2010]).

With over 100 images in the database, nearly 5,000 pairs of images have been matched using both the triangle-based and the ellipse-based approaches. A few keypoint detectors have been tested, but SIFT has been eventually the only keypoint descriptor to be used (because of its superior performances). The obtained results for two types of keypoint detectors (i.e. Harris-Affine and MSER) are summarized in Table 1.

Table 1 Average *precision* and *recall* of near-duplicate fragment detection

Detector	Triangle-based method		Ellipse-based method	
	Harris-Affine	MSER	Harris-Affine	MSER
Precision(object)	0.96	0.96	0.95	0.91
Recall(object)	0.82	0.61	0.65	0.62
Precision(area)	0.96	0.96	0.87	0.90
Recall(area)	0.65	0.47	0.49	0.45

The triangle-based approach is, in general, slightly more accurate than its ellipse-based counterpart, but the results are actually similar. Very high levels of *precision* should be noticed which means that different objects are very seldom detected as near-duplicate fragments. The values of *recall* are significantly lower (for the *area-based* accuracy in particular) i.e. similar objects are relatively often missed. However, human vision performs similarly. We seldom recognize different objects as similar, but more frequently (especially in cases of tired, distressed or absentminded individuals) do not notice that certain fragments of the observed scenes/images are actually similar.

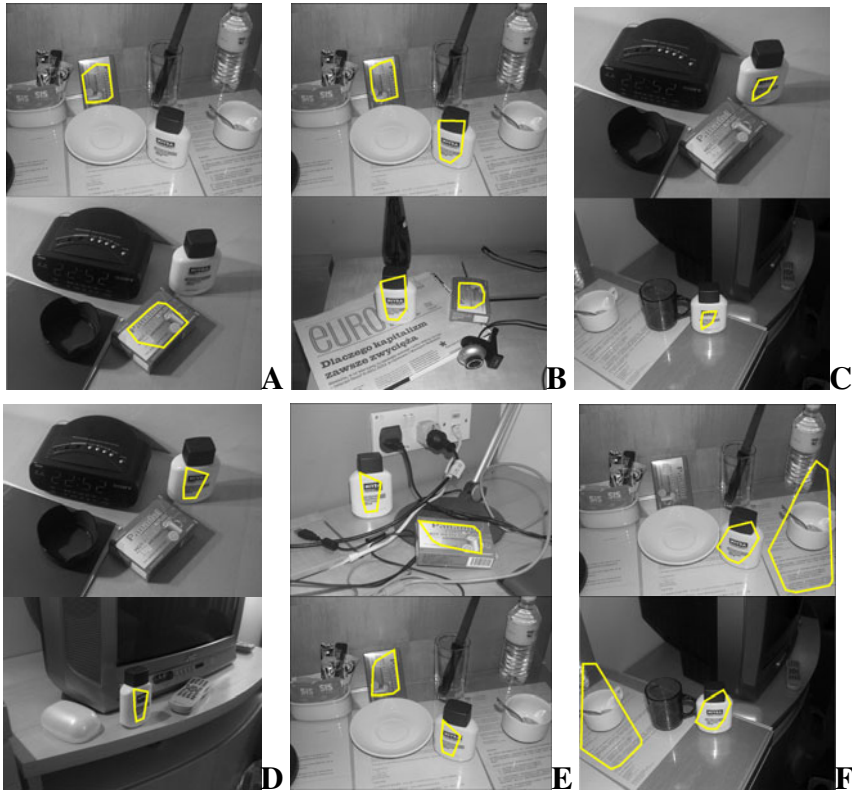


Fig. 6 Near-duplicate fragments detected in the selected pairs of images from Fig. 5

Exemplary detections of near-duplicate fragments (for the selected pairs of images from Fig. 5) are shown in Fig. 6. The examples illustrate that the method can identify near-duplicate fragments in images of diversified and unpredictable contents (although sometimes, see the example in Fig. 6A, the actual near-duplicates

are not detected). An interesting effect can be noticed in the the example in Fig. 6F where the algorithm has identified (correctly!) a complex fragment consisting of a cup, a spoon, a newspaper and a piece of bottle that are accidentally present in both images. Even though this is not actually a planar fragment, it is seen in both images from a similar viewpoints and, thus, it looks planar.

A few more examples are given in Fig. 7 to show that the technique works satisfactorily for outdoor images as well.

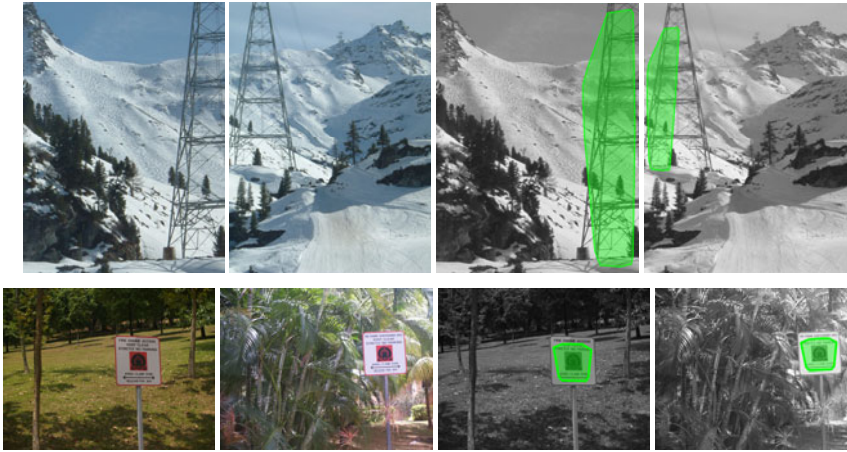


Fig. 7 Exemplary results of near-duplicate detections in outdoor scenes

Performances of the algorithm additionally depend on certain factors which are not discussed in the paper. In particular, the number of keypoints detected in matched images plays an important role. It depends on the image contents and on the capabilities of keypoint detectors. In most cases, however, the number of keypoints can be approximately controlled by specifying the cut-off threshold. As indirectly explained by the content of Table 1, our preferred keypoint detectors are Harris-Affine and MSER. SIFT is the recommended keypoint descriptor even though we experimented with the alternative options (e.g. GLOH, SURF, etc.).

Another critical factor is the value of n (the number of matched keypoint pairs) which strongly depends not only on the number of keypoints but also on the keypoint matching scheme. We generally prefer to use O2O (*one-to-one*) scheme where two keypoint are considered a match if they are mutual nearest neighbors. O2O usually provides the best precision of keypoint matching (i.e. the smallest number of outliers). However, in case of too few keypoints available (and in case of multiple copies of the same objects expected in the images) various variants of M2M (*many-to-many*) might be needed. It should be highlighted, nevertheless, that neither the choice of the keypoint detector/descriptor not the keypoint matching scheme affects the principles of the algorithm.

4 Database Building for Visual Exploration

The proposed method is able to identify near-duplicate image fragments without any prior knowledge about the image contents, numbers of objects in the images, etc. However, when more and more images are captured (and memorized) during the exploration of an unknown world, the amount of data to be matched with the query images (i.e. images depicting currently observed scenes) can grow beyond capabilities of even very powerful computational systems. At the same time, however, with such an amount of collected data, the algorithm becomes somehow “familiar” with the environment. Thus, instead of a simple retrieval of near-duplicate fragments in the database images, more advanced approaches to the visual understanding of query images can be attempted. In particular, we can exploit the fact that clusters of mutually similar near-duplicate fragments usually correspond to the same (or similarly looking) physical objects.

Therefore, in this section we introduce two mechanisms that provide a certain level of “visual understanding” of previously unknown words. First, it is proposed how to cluster mutually similar near-duplicate fragments into “visual objects” (which may but also may not represent physical objects of the environment). We focus on practical aspects of the process (and illustrate it by examples). The in-depth description of the algorithm details is available in [Paradowski and Śluzek 2010a].

Secondly, we discuss how to organize the database so that the balance between its size and the system’s performances can be maintained.

4.1 Visual Prototypes and Objects

Assume that near-duplicate fragments have been identified in a collection of images. We propose a graph representation for relations between detected near duplicates (images from Fig. 5 as an example to illustrate principles of this representation).

First, near-duplicates from the same image are grouped if they come from approximately the same location of the image. In practice, such groups of near-duplicates usually represent the same underlying physical object that is similar to objects found in several other images. In Figs 8 and 9, groups found in this manner are encircled.

Any group of such near-duplicates is referred to as a *visual prototype* (to emphasize that most probably it represents a physical object present in the image). Subsequently, *visual prototypes* form nodes of *visual similarity graph*. Two nodes are connected if they contain fragments which are mutual near-duplicates.

In images from Fig. 5, two connected sub-graphs of visual prototypes are found. They are given in Figs 8 and 9. Near-duplicates contributing to each prototype are also shown (as mentioned above) to justify why selected nodes are linked.

Visual prototypes linked by edges usually represent similar/identical objects shown in different images. Therefore, connected sub-graphs are natural candidates for defining *visual objects*. However, to provide a certain verification mechanism in handling visual data (which are often corrupted or distorted) we define *visual object* as **2-connected sub-graphs** of the visual similarity graph. In other words, a visual prototype must be similar to at least to other visual prototypes to be included into a visual object.

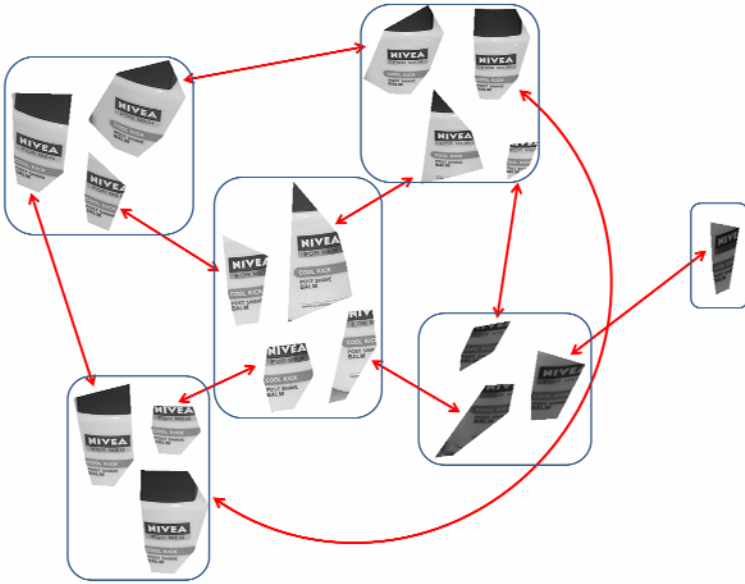


Fig. 8 A connected sub-graph of the visual similarity graph built for images from Fig. 5

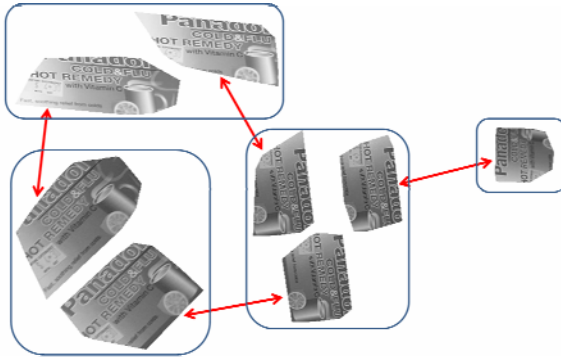


Fig. 9 Another connected sub-graph of the visual similarity graph built for images from Fig. 5

Near-duplicate fragments of the visual prototype forming a visual objects are considered template images of that object. However, because each visual prototype consists of significantly overlapping fragments of the same image, they can be merged to create a single template image from each visual prototype. Thus, the

visual object built from Fig. 8 sub-graph has five templates while the visual object based on Fig. 9 has three templates (all templates of both objects are shown in Fig. 10). It should be noted that the templates, which are unions of the corresponding convex hulls, do not have to be convex hulls.



Fig. 10 Template images for two *visual objects* automatically found in Fig. 5 images



Fig. 11 Exemplary images from a dataset of face photos of three people (the total number of images is 15-20 per person)



Fig. 12 Exemplary templates for three visual objects automatically built by the algorithm (the actual numbers of templates vary between 12 and 15 per object)

In a more advanced example, nearly 50 photos of human faces on diversified backgrounds (we actually used faces of three people from Caltech 101 dataset available at http://www.vision.caltech.edu/Image_Datasets/Caltech101/) have been processed. The system has fully automatically formed three visual objects which clearly correspond to the faces of three people. Examples of the database images are shown in Fig. 11. A few templates of three visual objects (more templates have been actually built for each object) are given in Fig. 12.

It should be highlighted that the systems has acquired certain “face identification skills” without any knowledge about the anatomy of human faces, about the number of individuals shown in the database and, in general, knowing nothing about the content of analyzed images.

4.2 Updating Databases of Images and Visual Objects

For a given database of images, visual object are built for two reasons. First, as highlighted in the previous subsection, visual objects can be considered a step towards “a vision-based understanding” in the explored world. They represent the most common components of the environments (even if these components not always are physical objects).

More importantly, however, visual objects can be a useful tool for maintaining the balance between the size of visual databases representing certain (originally unexplored and unknown) environments.

It is obvious that a vision-based exploratory system cannot collect too many images in its database (since the query image should be matched with all database images to find similar fragments) if performing in real time or under other timing constraints. However, it is difficult to predict how large the database should be if the environment is unknown and unpredictable. Therefore, we propose a dynamic database update scheme that, in our opinion, can offer a realistic solution. The scheme consists of the following tasks:

- a. **Primary database building.** Initially, a substantial number of images for the explored world should be collected. The number of images would be determined by the computational capabilities of the system, i.e. a query image should be matched with the database (possibly using pre-retrieval mechanisms) within the required time. Preliminary experiments with our pre-retrieval algorithms indicate (see [Paradowski and Śluzek 2010a]) that in typical scenarios near real-time performances are feasible for the primary database containing few hundred images.
- b. **Formation of primary visual objects.** Visual objects can be found in the primary database (as described in Subsection 4.1) and their templates are added to the database (note that templates are usually very small compared to the size of original images).
- c. **Processing query images.** A newly acquired query images is matched with database images to find near-duplicates. Then, three cases are possible:
 - i. All near-duplicates found in the query match templates of the existing visual objects – the query is ignored.

- ii. Some near-duplicates match fragments which do not belong to the templates of visual objects) – the query is added to the database.
- iii. No near-duplicates are found – the query can be randomly added to the database (with a small probability).
- d. **Periodical updates of visual objects.** Near-duplicates extracted from new database images can be added to the existing visual objects or can be used to form new visual objects (similarly to point (b)).
- e. **Periodical cleanups of the database.** If in recent queries no near-duplicates matching a fragments a database image are found (near-duplicates templates of visual objects do not count!) the image is removed from the database.

Although only a simple, limited-scale experiment on the database of human faces has been conducted so far, we believe that such a scheme can converge to a database that contains only (or mostly) templates of visual objects representing typical parts of the explored world. Random or incidental visual data would be removed. Of course, we can expect such a behavior only in environments which are not too complex and with new elements added not too rapidly (otherwise the expansion of database may continue at high speed until the system resources saturate in term of memory requirements and timing performances).

3 Conclusions

The paper proposes a framework for building automatically a certain level of vision-based semantics in unknown and unpredictable environments. Starting from a collection of (representative) images of the environment, the system identifies similar fragments in the database and forms from such fragments *visual objects*.

The subsequent images are either added to the database or contribute to formation of new visual objects (or update the existing visual objects). Images are gradually removed from the database if their contents are sufficiently well represented by the visual objects. It is envisaged then (at least of certain environments) the database can eventually contain only visual objects representing the most common elements of the explored world. Thus, a kind of “visual understanding” can be achieved without any prior knowledge about this world.

The paper discusses two aspects of the proposed solution. First, we overview previously developed tools and present their performances. Secondly, we introduce a novel scheme that, based in the solid results of the first part, is to be developed in the future (although certain preliminary experiments have been already conducted). We believe that such a proposal can encourage researches in diversified applications of autonomous machine vision systems and systems interacting with humans.

Acknowledgment

The research presented in the paper is a part of A*STAR Science & Engineering Research Council grant 072 134 0052. The financial support of SERC is gratefully acknowledged.

This work is also partially financed from the Ministry of Science and Higher Education Republic of Poland resources (under Poland–Singapore joint research project 65/N-SINGAPORE/2007/0).

References

- [Biederman 1987] Biederman, I.: Recognition-by-components: A theory of human image understanding. *Psychological Review* 94(2), 115–147 (1987)
- [Bourbakis and Kavradi 2001] Bourbakis, N.G., Kavradi, D.: An intelligent assistant for navigation of visually impaired people. In: 2nd IEEE Int. Symp. on Bioinformatics and Bioengineering, Bethesda, p. 230 (2001)
- [Ke et al. 2004] Ke, Y., Sukthankar, R., Huston, L.: Efficient near-duplicate detection and sub-image retrieval. In: ACM Multimedia Conference, New York, pp. 869–876 (2004)
- [Matas et al. 2002] Matas, J., Chum, O., Urban, M., Pajdla, T.: Robust wide baseline stereo from maximally stable extremal regions. In: British Machine Vision Conf., Cardiff, pp. 384–393 (2002)
- [Mikolajczyk and Schmid 2004] Mikolajczyk, K., Schmid, C.: Scale and affine invariant interest point detectors. *Int. J. of Computer Vision* 60, 63–86 (2004)
- [Paradowski and Śluzek 2010] Paradowski, M., Śluzek, A.: Local keypoints and global affine geometry: Triangles and ellipses for image fragment matching. In: Kwaśnicka, H., Jain, L.C. (eds.) *Innovations in Intelligent Image Analysis*. SCI, vol. 339, pp. 195–224. Springer, Heidelberg (2011)
- [Paradowski and Śluzek 2010a] Paradowski, M., Śluzek, A.: Automatic visual object formation using image fragment matching. In: 5th Int. Symp. Advances in Artificial Intelligence & Applications, Wisła, pp. 97–104 (2010)
- [Riesenhuber and Poggio 2000] Riesenhuber, M., Poggio, T.: Models of object recognition. *Nature Neuroscience* 3, 1199–1204 (2000)
- [Zhao and Ngo 2009] Zhao, W.-L., Ngo, C.-W.: Scale-rotation invariant pattern entropy for keypoint-based near-duplicate detection. *IEEE Trans. on Image Processing* 18(2), 412–423 (2009)
- [Zhao et al. 2007] Zhao, W.-L., Ngo, C.-W., Tan, H.-K., Wu, X.: Near-duplicate keyframe identification with interest point matching and pattern learning. *IEEE Trans. on Multimedia* 9(5), 1037–1048 (2007)

Recognition of Hand Posture for HCI Systems

R.S. Choraś

Department of Telecommunications & Electrical Engineering, University of Technology and Life Sciences, Bydgoszcz, Poland
choras@utp.edu.pl

Abstract. We present problem of recognizing gestures and signs executed by hands. Hand posture recognition is either the process by which gestures formed by a user interact with the computer or is the element of the special signs language to convey meaning. We propose methods for the recognition of hand gestures using Gabor wavelets (GW), Radon transform (RT) and texture features for gesture recognition. We compare these features and propose the fusion features to obtain high recognition rate.

1 Introduction

In communications between humans, gestures which are natural and intuitive form interactions and which are easy understanding by human brain, are often used. In recent Human-Computer Interaction (HCI) applications, the ability to sense, record and transmit complex hand gestures in real time and real environments is important requirement for computer/machine control. For these reasons HCI based on recognition static or dynamic, communicative or manipulative gestures is important field of research.

Literature described many systems to solution the hand detection and recognition problems. All these systems are based on two approaches. One approaches uses specialized hardware e.g. glove with sensors, specific markers on the hand, etc.. The other approach use image processing and computer vision algorithms. This approach has a common name – vision-based hand detection/recognition systems [Choras 2009]. Vision-based hand gesture recognition systems can identify different hand gestures from video input and use them as artificial commands, which computers can understand and respond to [Phung et al. 2005].

Recognition of hand gestures based on images improve communication between humans and computers/machines. In HCI human gives commands to a computer by hand shapes/blobs or hand gestures. In this case interface system uses a camera as an input acquisition device and some software to detection, analysis, feature extraction, recognition/classification hand images. Compared with traditional HCI devices, hand gestures are less intrusive and more convenient for users to interact with computers.

The main contribution of this work is the presentation of a general framework for real-time hand gesture detection and recognition which combines feature-based and image-based approaches to achieve hand recognition tasks in real time and real environment conditions.

2 Skin Color Modeling and Classification

Skin detection plays important role in hand detection. Skin color is used as information for identifying hand area and to model and classify color pixels. A skin color model is used to decide skin-pixel vs. non-skin-pixel. A human skin color model is characterized by color space and classification algorithms.

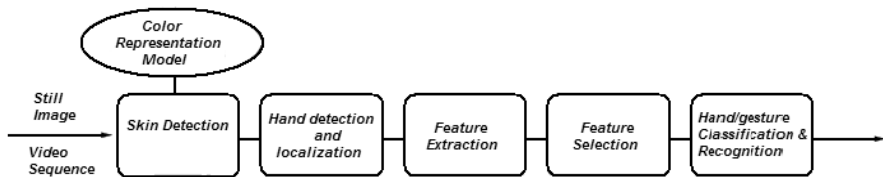


Fig. 1 Scheme of a general hand recognition system

2.1 Color Spaces

Each image is represented using three components of the color space chosen. For vision based HCI are recommended in literature several color spaces, such as RGB , XYZ , HSI , HSV , YIQ , YUV and YC_bC_r can be obtained from (Table 1). RGB is the most commonly used but in this space skin colors are sensitive to the lighting condition. RGB color space is not perceptually uniform, which implies that two colors with larger distance can be perceptually more similar than another two colors with smaller distance, or simply put, the color distance in RGB space does not represent perceptual color distance.

The first color space developed by the Commission Internationale de l'Eclairage (CIE) is the XYZ color space. The Y component is the luminance component, and the X and Z are the chromatic components. The XYZ color space is a device-independent color space, but is perceptually not uniform.

The HSI, HSV, HSL - *Hue, Saturation, Intensity (Value, Lightness)*, color spaces describe color with intuitive values. *Hue* defines the dominant color (such as red, green, purple and yellow) of an area, *Saturation* measures the colorfulness of an area in proportion to its brightness. The *Intensity, Value* and *Lightness* are related to the color luminance. *Hue* is generally related to the wavelength of a light. *Saturation* is a component that measures the "colorfulness" in HSV space.

The intuitiveness of the color space components and explicit discrimination between luminance and chrominance properties make these color spaces popular. The "Intensity", "Lightness" or "Value" is related to the color luminance. The intuitiveness of the color space components and explicit discrimination between luminance and chrominance properties make these color spaces popular in the works on skin color segmentation [Yin and Xie 2001].

The *HSV* (*Hue*, *Saturation*, and *Value*) color space is more closely related to a human color perception than the *RGB* color space, but it is still not perceptually uniform. In addition, it is device-dependent. *Hue* is the color component of the *HSV* color space. When *Saturation* is set to 0, *Hue* is undefined. The *Value* represents the gray-scale image.

The *YUV*, *YIQ*, $Y C_b C_r$ models that are basically used in color television transmission. The *U* and *V* for *YUV* and *I* and *Q* for *YIQ* are the chromatic components. The *YUV* and *YIQ* color spaces are device-dependent and not perceptually uniform. The $Y C_b C_r$ is a digital standard. These color spaces separate *RGB* into luminance and chrominance information and are useful in compression applications. Color is represented by luminance computed from nonlinear *RGB* [8], constructed as a weighted sum of the *RGB* values, and two color difference values C_r and C_b that are formed by subtracting luminance from *RGB* red and blue components.

Color spaces for skin detection are presented in Table 1 and Fig. 2.

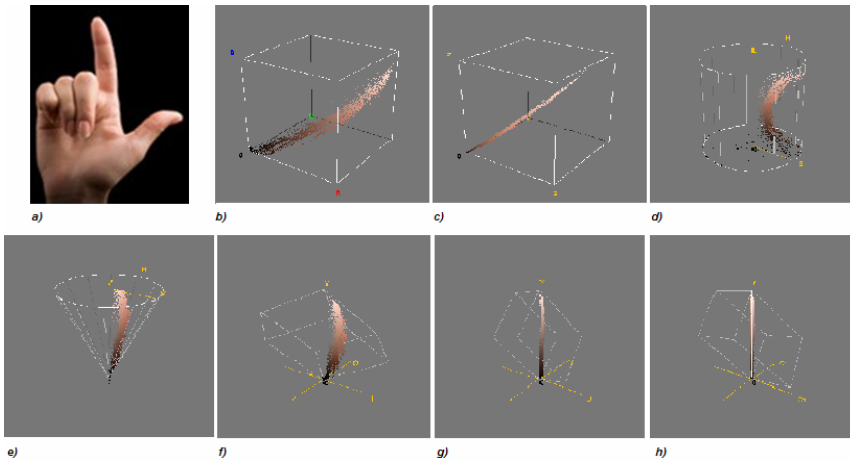


Fig. 2 Original image in *RGB* color space (a) and respectively (b-h) color space *RGB*, *XYZ*, *HSL*, *HSV*, *YIQ*, *YUV*, $Y C_b C_r$

Table 1 Characteristics of the method a skin detection

Color space	Definition
<i>XYZ</i>	$X = 0.61R + 0.176G + 0.20B$ $Y = 0.30R + 0.59G + 0.11B$ $Z = 0.00R + 0.07G + 1.12B$
<i>HSI</i>	$h = \arccos \frac{\frac{1}{2}[(R-G) + (R-B)]}{\sqrt{(R-G)^2 + (R-B)(G-B)}}$ $S = 1 - \frac{\min(R, G, B)}{I}$ $I = \frac{1}{3}(R + G + B)$ $H = \begin{cases} h & \text{if } B \leq G \\ 2\pi - h & \text{if } B > G \end{cases}$
<i>HSV</i>	$h = \arccos \frac{\frac{1}{2}[(R-G) + (R-B)]}{\sqrt{(R-G)^2 + (R-B)(G-B)}}$ $S = \frac{\max(R, G, B) - \min(R, G, B)}{\max(R, G, B)}$ $V = \max(R, G, B)$
<i>YIQ</i>	$\begin{bmatrix} Y \\ I \\ Q \end{bmatrix} = \begin{bmatrix} 0.299 & 0.578 & 0.114 \\ 0.596 & -0.275 & -0.321 \\ 0.212 & -0.523 & 0.311 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix}$
<i>YUV</i>	$\begin{bmatrix} Y \\ U \\ V \end{bmatrix} = \begin{bmatrix} 0.299 & 0.578 & 0.114 \\ -0.147 & -0.289 & -0.436 \\ -0.615 & -0.514 & -0.101 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix}$
<i>YCbCr</i>	$\begin{bmatrix} Y \\ C_b \\ C_r \end{bmatrix} = \begin{bmatrix} 0.299 & 0.578 & 0.114 \\ -0.169 & -0.331 & 0.500 \\ 0.500 & -0.419 & -0.081 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix}$

2.2 Skin Modeling

The important stage of detection of the hand is segmentation. We first use a skin color segmentation procedure based on a statistical model of human skin color that can be defined as the process of discrimination between skin and non-skin pixels. Skin color classification algorithms are presented in Table 2. We represent skin color distribution by a 2D Gaussian law and we consider a pixel to be skin if

$$p(c) = \frac{1}{(2\pi)^{\frac{1}{2}}|Cov|^{\frac{1}{2}}} e^{\left[-\frac{1}{2}(c-\mu)^T Cov^{-1}(c-\mu)\right]} \geq \tau \tag{1}$$

where τ is a threshold value for decision.

3 Features for Hand Recognition

Selecting features is main part in hand recognition process [Chang et al. 2008]. Static hand is recognized by extracting color and texture features and some geometric features.

3.1 Hand Color Features

The distribution of color forms the image’s feature vectors. The mathematical foundation of this approach is that any probability distribution is uniquely characterized by its moments. Thus, if we interpret the color distribution of an image as a probability distribution, then the color distribution can be characterized by its moments. Furthermore, because most of the information is concentrated on the low-order moments, only the first moment (mean), the second moment (variance), and the third central moment (skewness) were used. If the value of the image pixel is $f_c(x, y)$ in the c color channel and the number of pixels in the image is $M \times N$, then the moments related to this color channel are:

$$m_c = \frac{1}{M \times N} \sum_{x=1}^M \sum_{y=1}^N f_c(x, y) \tag{2a}$$

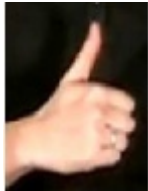

$$\sigma_c = \left(\frac{1}{M \times N} \sum_{x=1}^M \sum_{y=1}^N (f_c(x, y) - m_c)^2 \right)^{\frac{1}{2}} \tag{2b}$$

$$s_c = \left(\frac{1}{M \times N} \sum_{x=1}^M \sum_{y=1}^N (f_c(x, y) - m_c)^3 \right)^{\frac{1}{3}} \tag{2c}$$

Table 2 Characteristics of the method a skin detection

Methods	Description
Explicit defined	
1. skin region in RGB space	$R > 95$ and $G > 40$ and $B > 20$ and $\max\{R,G,B\} - \min\{R,G,B\} > 15$ and $ R - G > 15$ and $R > G$ and $R > B$.
2. skin region in YC_bC_r space	$(85 \leq C_b \leq 135 ; 135 \leq C_r \leq 180 ; Y \geq 80)$
Nonparametric skin distribution	
1. Bayes classifier	$p(c skin) = \frac{p(c skin)p(skin)}{p(c skin)p(skin) + p(c non_skin)p(non_skin)}$ <p>A pixel c is labeled as skin pixel if $p(c skin) \geq \tau$ where $skin$ and non_skin denote the classes of skin and non-skin, $p(c skin)$ and $p(c non_skin)$ are the prior probabilities of skin and non-skin</p>
Parametric skin distribution	
1. Single Gaussian model (SGM)	$p(c) = \frac{1}{2\pi \sqrt{ Cov }} e^{-\frac{1}{2} c^{-\mu} T Cov^{-1} c - \mu} \geq \tau ; \mu = \frac{1}{n} \sum_{j=1}^n c_j ; Cov = \frac{1}{n-1} \sum_{j=1}^n (c_j - \mu)(c_j - \mu)^T$ $Cov = \begin{bmatrix} \sigma_{c_r c_r} & \sigma_{c_r c_b} \\ \sigma_{c_r c_b} & \sigma_{c_b c_b} \end{bmatrix}; c_j = \begin{bmatrix} c_{rj} \\ c_{bj} \end{bmatrix}; \sigma_{c_r c_r} = \frac{1}{n} \sum_{j=1}^n x_j^2 - \mu_x^2 ; \sigma_{c_b c_b} = \frac{1}{n} \sum_{j=1}^n y_j^2 - \mu_y^2$ <p>Mean μ and covariance Cov are estimated over all the color samples c_j</p>
2. Multiple Gaussian model (GMM)	$p(c) = \sum_{i=1}^N w_i \frac{1}{2\pi \sqrt{ Cov_i }} e^{-\frac{1}{2} c - \mu_i T Cov_i^{-1} c - \mu_i} ; \mu_i, Cov_i, w_i$ are parameters of a GMM
3. Elliptical boundary model	$\Phi(c) = c - \Psi^T \Lambda^{-1} c - \Psi$ where $\Psi = \frac{1}{n} \sum_{i=1}^n c_i$ and $\Lambda = \frac{1}{N} \sum_{i=1}^n f_i (c_i - \mu)(c_i - \mu)^T$ <p>where K is the total number of samples in the training data set, f_i is the number of samples with chrominance c_i and μ is the mean of the chrominance vectors in the training data set.</p>

Table 3 Color moments

Image	Color channel	m_c	σ_c	s_c
	Y	89.398	0.526	-1.561
	C_b	117.305	-0.600	-1.204
	C_r	139.036	0.625	-1.104
	Y	50.667	0.987	-0.463
	C_b	121.227	-0.693	-1.123
	C_r	137.936	0.770	-0.902

Since only 9 (three moments for each of the three color components) numbers are used to represent the color of the hand image. This representation is a very compact compared to other color features.

3.2 Hand Feature Extraction Using Gabor Wavelets (GW)

In the spatial domain, a *GW* is a complex exponential modulated by a Gaussian function. In the most general the Gabor wavelet is defined as follows [Gabor 1946; Choras 2010]:

$$\Psi_{\omega,\theta}(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{(x\cos\theta+y\sin\theta)^2+(-x\sin\theta+y\cos\theta)^2}{2\sigma^2}} \cdot e^{i(\omega x\cos\theta+\omega y\sin\theta)} \quad (3)$$

where x, y denote the pixel position in the spatial domain, ω is the radial center frequency of the complex exponential, θ is the orientation of the *GW*, and σ is the standard deviation of the Gaussian function. By selecting different center frequencies and orientations, we can obtain a family of Gabor kernels, which can then be used to extract features from an image.

For pixel $I(x, y)$ in an image, its Gabor feature is treated as a convolution

$$Gab(x,y,\omega,\theta) = I(x,y) * \Psi_{\omega,\theta}(x,y) \quad (4)$$

For k frequencies and l orientations we have $k \cdot l$ complex coefficients for each image point.

The Gabor feature image of the hand texture image combines frequency (position) and orientation information. We calculate the sum of Gabor feature images for each orientation and the sum of Gabor feature images for each frequency. We collect the Gabor feature images respectively for each orientation and each frequency, and then we can get l Gabor feature images. Each can be described as a sum of k Gabor feature images for different frequencies (Fig. 4). We obtain $k + l$ collected Gabor feature images $Gab_sum(x, y, \omega, \theta)$ which can be used to compute energy parameter.

The features of the Gabor wavelets responses for some the hand tests images are represented by

$$E(x,y) = \frac{1}{MN} \left(\sum_{x=1}^M \sum_{y=1}^N Gab_sum^2(x,y,\omega,\theta) \right)^{\frac{1}{2}} \quad (5)$$

where M, N is image dimension.

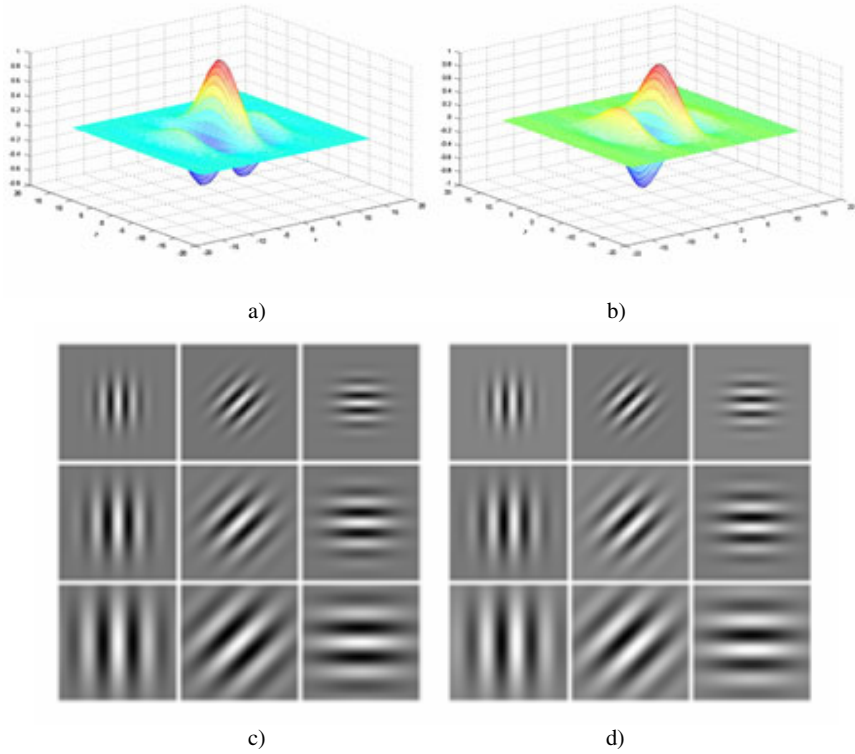


Fig. 3 The real part(a) and the imaginary part (b) of 2D Gabor wavelet and the real (c) and imaginary parts (d) of the Gabor filter for 3 scales and 3 orientations

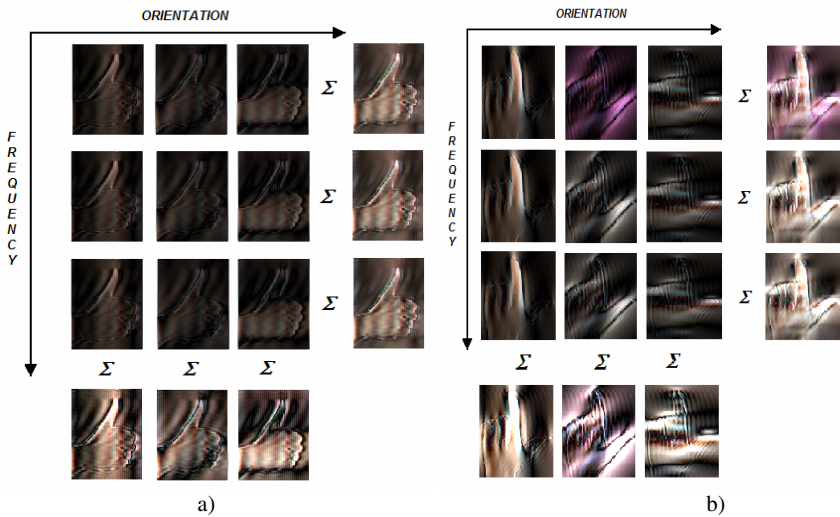


Fig. 4 Image of the Gabor features

3.3 Moment-Based Hand Features

Image can be represented by the spatial moments of its intensity function

$$m_{pq} = \sum_{x=1}^M \sum_{y=1}^N x^p y^q f(x, y) \quad (6)$$

The central moments are given by

$$m_{pq} = \sum_{x=1}^M \sum_{y=1}^N (x - \bar{x})^p (y - \bar{y})^q f(x, y) \quad (7)$$

$$\text{where } \bar{x} = \frac{m_{10}}{m_{00}} \quad ; \quad \bar{y} = \frac{m_{01}}{m_{00}}$$

Normalized central moment

$$\mu_{pq} = \frac{m_{pq}}{m_{00}^\gamma} \quad (8)$$

$$\text{where } \gamma = \frac{1}{2}(p + q) + 1 \quad \text{for } p + q = 2, 3, \dots$$

By using nonlinear combinations of the lower order moments, a set of moment invariants (usually called geometric moments), which has the desirable properties of being invariant under translation, scaling and rotation, is derived. Hu [Hu 1962] employed seven moment invariants, that are invariant under rotation as well as translation and scale change, to recognize objects independently of their position size and orientation.

$$\begin{aligned} \phi_1 &= \mu_{20} + \mu_{02} \\ \phi_2 &= [\mu_{20} - \mu_{02}]^2 + 4\mu_{11}^2 \\ \phi_3 &= [\mu_{30} - 3\mu_{02}]^2 + [3\mu_{21} - \mu_{03}]^2 \\ \phi_4 &= [\mu_{30} + \mu_{12}]^2 + [\mu_{21} + \mu_{03}]^2 \\ \phi_5 &= [\mu_{20} - 3\mu_{12}][\mu_{30} + \mu_{12}][(\mu_{30} + \mu_{12})^2 - 3(\mu_{21} + \mu_{03})^2] + \\ &+ [3\mu_{21} - \mu_{03}][\mu_{21} + \mu_{03}][3(\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2] \\ \phi_6 &= [\mu_{20} - \mu_{02}][(\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2] + 4\mu_{11}[\mu_{30} + \mu_{12}][\mu_{21} + \mu_{03}] \\ \phi_7 &= [3\mu_{21} - \mu_{03}][\mu_{30} + \mu_{12}][(\mu_{30} + \mu_{12})^2 - 3(\mu_{21} + \mu_{03})^2] \\ &- [\mu_{03} - 3\mu_{12}][\mu_{21} + \mu_{03}][3(\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2] \end{aligned} \quad (9)$$

Moment invariants for luminance component of the summary Gabor hand images with Fig. 4 are presented in Table 4.

Table 4 Hu moments

Hu moments for image with Fig. 4a							
	ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_5	ϕ_6	ϕ_7
$l=1$	19.124	24.936	4.5127	1312.85	55041.1	6555.47	266004.34
$l=2$	19.124	24.939	4.5111	1312.47	54939.6	6553.91	265915.06
$l=3$	19.123	24.933	4.5061	1312.57	54973.5	6553.65	265614.50
$k=1$	16.961	35.018	6.4993	462.944	22761.8	2580.87	105593.89
$k=2$	20.647	12.029	38.278	1254.99	-232009	4315.67	-499023.3
$k=3$	21.069	41.853	150.12	3259.50	1831407	20996.6	2499753.7

Hu moments for image with Fig. 4a							
	ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_5	ϕ_6	ϕ_7
$l=1$	13.455	16.047	0.0129	410.622	-605.82	1632.39	-24293.58
$l=2$	12.494	14.181	0.1787	353.892	2752.63	1311.97	-17425.61
$l=3$	12.493	14.179	0.1775	353.788	2744.62	1311.48	-17397.74
$k=1$	14.573	33.434	50.680	1059.57	214662	5875.13	-69601.22
$k=2$	14.123	21.407	11.03	599.995	46737.2	2413.24	-30073.54
$k=3$	14.69	12.742	4.508	259.527	-3390.4	904.213	-30921.08

3.4 Hand Feature Extraction Using Radon Transform (RT)

The Radon Transform (*RT*) of a image $f(x, y)$ is defined as [Deans 1983; Tabbone and Wendling 2002; Choras 2010]

$$RT(\rho, \theta) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) \delta(x \cos \theta + y \sin \theta - \rho) dx dy \quad (10)$$

Equation (10) can be expressed as

$$RT(\rho, \theta) = \int_{-\infty}^{\infty} f(\rho \cos \theta - u \sin \theta, \rho \sin \theta + u \cos \theta) du \quad (11)$$

where $\rho = x \cos \theta + y \sin \theta$; $u = -x \sin \theta + y \cos \theta$; $-\infty < \rho < \infty$; $0 < \theta < \pi$.

The *RT* for hand image are presented in Fig.5.

Texture is the important characteristics used in identifying objects in image. We used texture features extracted from *RT* space. A set of co-occurrence matrices for a *RT* hand image is computed and a set of texture features are extracted. The features are:

- Second Angular Moment

$$SAM = \sum_{x=1}^M \sum_{y=1}^N [P_{\delta,\theta}(x, y)]^2 \tag{12}$$

- Contrast

$$Con = \sum_{x=1}^M \sum_{y=1}^N (x - y)^2 P_{\delta,\theta}(x, y) \tag{13}$$

- Correlation

$$Corr = \frac{\sum_{x=1}^M \sum_{y=1}^N [xy P_{\delta,\theta}(x, y)] - \mu_x \mu_y}{\sigma_x \sigma_y} \tag{14}$$

- Inverse Differential moment

$$IDM = \sum_{x=1}^M \sum_{y=1}^N \frac{P_{\delta,\theta}(x, y)}{1 + (x - y)^2} \tag{15}$$

- Entropy

$$Ent = - \sum_{x=1}^M \sum_{y=1}^N P_{\delta,\theta}(x, y) \log P_{\delta,\theta}(x, y) \tag{16}$$

Table 5 Texture parameters of the *RT* hand images

Parameter	Fig.5(bottom left corner)		Fig.5(bottom right corner)	
	θ	$\delta=5$	θ	$\delta=5$
ASM	0	0.011	0	0.089
	90	0.011	0	0.079
Con	0	4271.52	0	4653.285
	90	4071.105	0	4138.023
Corr	0	1.053E-4	0	1.056E-4
	90	1.078E-4	0	1.134E-4
IDM	0	0.341	0	0.435
	90	0.264	0	0.399
Ent	0	7.323	0	5.776
	90	7.533	0	5.863

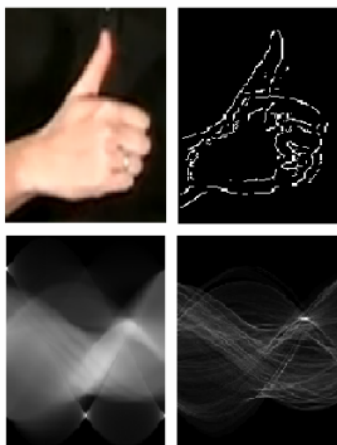


Fig. 5 Hand image, the Canny edge hand image and its Radon transform

4 Results and Conclusion

In the paper, some approaches for hand posture recognition are presented. There are two important types of errors - a false accept rate (*FAR*) and false reject rate (*FRR*). The false rejection error means that the registered data is falsely determined as a non-registered data and expressed by *FRR* (False Rejection Rate). The false acceptance error means that the system falsely determines a non-registered data as a registered data and expressed by *FAR* (False Acceptance Rate). In our case *FRR* mean that the system cannot correctly recognize the hand posture with database.

To evaluate the performance of hand posture recognition methods we use own hand sign database that consists 300 images. The results of the experiment are summarized as follows: *FRR* when *FAR* = 0 is for *GW+Hu moments +RT* ≈ 0.1 .

References

- [Chang et al. 2008] Chang, C.C., Liu, C.Y., Tai, W.K.: Feature alignment approach or hand posture recognition based on curvature scale space. *Neurocomputing* 71, 1947–1953 (2008)
- [Choras 2009] Choras, R.S.: Hand shape and hand gesture recognition. In: Proceedings of 2009 IEEE Symp. on Industrial Electronics and Applications ISIEA (2009)
- [Choras 2010] Choras, R.S.: Hand gesture recognition using gabor and radon transform with invariant moment features. In: Recent Research in Circuits, Systems, Electronics, Control & Signal Processing, pp. 93–98. WSEAS Press (2010)
- [Deans 1983] Deans, S.R.: Applications of the radon transform. Wiley Interscience Publications, New York (1983)
- [Gabor 1946] Gabor, D.: Theory of communication. *J. Inst. Elect. Eng.* 93, 429–459 (1946)

- [Hu 1962] Hu, M.K.: Visual pattern recognition by moment invariant. *IRE Trans. on Information Theory* (8), 179–187 (1962)
- [Phung et al. 2005] Phung, S.L., Bouzerdoun, A., Chai, D.: Skin segmentation using color pixel classification: analysis and comparison. *IEEE Trans. Pattern Anal. Mach. Intell.* 27(1), 148–154 (2005)
- [Tabbone and Wendling 2002] Tabbone, S., Wendling, L.: Technical symbols recognition using the two-dimensional Radon transform. In: *Proc. of the 16th ICPR*, vol. (3), pp. 200–203 (2002)
- [Yin and Xie 2001] Yin, X., Xie, M.: Hand gesture segmentation, recognition and application. In: *Proc. IEEE Int. Symp. on Computational Intelligence in Robotics and Automation*, Banff, Canada (2001)

Some Eye Tracking Solutions for Severe Motor Disabilities

M. Porta and A. Ravarelli

Dipartimento di Informatica e Sistemistica, Università di Pavia, Pavia, Italy
{marco.porta,alice.ravarelli}@unipv.it

Abstract. People affected by serious motor disabilities need proper ways to interact with the computer, which is for them an essential communication means. Thanks to recent technological advances in the field of eye tracking, it is now possible to exploit unobtrusive devices to detect the user's gaze on a screen and employ it to control graphical interfaces. In this paper we present some of the eye tracking projects we have recently developed at the University of Pavia, all aimed at providing the (disabled) user with reliable gaze-driven input modalities.

1 Introduction

Within Information and Communication Technology (ICT), the design of physical devices, software interfaces and interaction modalities for disabled people needs special attention. The theme of Accessibility is being more and more considered nowadays, especially in the Web context. According to the World Wide Web Consortium (www.w3.org, the most authoritative non-profit organization developing Web standards), “Web accessibility means that people with disabilities can use the Web. More specifically, Web accessibility means that people with disabilities can perceive, understand, navigate, and interact with the Web, and that they can contribute to the Web”. Actually, Web Accessibility includes all kinds of disabilities (visual, auditory, physical, cognitive, etc.), and also concerns limitations due to aging – increasingly relevant as more and more mature people have to use or become interested in Internet technologies.

The European Commission strongly promotes several initiatives about Accessibility. For example, *eAccessibility* — within the *e-Inclusion* activity — aims at ensuring people with disabilities and elderly people access ICTs on an equal basis with others [europa.eu 2009]. Also, the very recent *i2010* initiative on e-Inclusion (“To be part of the Information Society”) includes a strategy targeted on enhancing accessibility to the Information Society for all potentially disadvantaged groups.

Machine perception can create more natural communication modalities with the computer. By providing the machine with perceptive capabilities which are typical

of interpersonal communication, machine perception can be very helpful for Accessibility as well. In particular, an eye tracker is a device able to detect and follow the user's gaze [Duchowski 2007]. The acquired data can then be recorded for subsequent use, or (like in our interfaces) directly exploited to provide commands to the computer in an active interface. Early eye tracking techniques were very invasive; with the introduction of computer-controlled video cameras things greatly improved, but until about the end of the 1990s video-based eye tracking was still mostly characterized by intrusive systems which required special equipment to be mounted on the user's head. Fortunately, current eye trackers have evolved to the point where the user can move almost freely in front of the camera, with good accuracy.

While eye tracking is being more and more exploited for the evaluation of different kinds of interfaces, within the field of Usability, it is when it is used as a direct input source for the computer that hands-free assistive interfaces can be built, an essential requirement in all those cases where important motor impairments hinder easy hand and body motion. Eye tracking as an assistive technology is especially useful in the case of diseases which strongly and progressively limit the motor abilities of people affected by them. For instance, Amyotrophic Lateral Sclerosis (ALS), muscular dystrophy, multiple sclerosis, different kinds of head injuries, cerebral palsies, muscular spinal atrophy, Werdnig-Hoffman, Rett and locked-in syndromes, spinal cord damages. Eye tracking technology may not instead be of help for people who, besides the above-quoted diseases, have physical-visual problems, such as subjects suffering from involuntary head shifts (e.g. Athetoid cerebral palsy) or are characterized by unintentional oscillatory eye movements (nystagmus). In this respect, hurdles to be cleared at the technology level fall into two main categories. The first one refers to the need for the eye tracker to achieve a high reliability degree. In order for it to provide an actual help to disabled persons in their everyday life, it must be able to detect both the actual gaze position on the screen and real-time eye movements. The second group of problems relates to the software, which must correctly and "intelligently" interpret the input coming from the user's eyes [Donegan et al. 2005].

Although several gaze-based systems for disabled people have been developed to date, there are still many open questions connected with accessibility issues. Our activity at the University of Pavia is aimed at creating more accessible interfaces based on eye tracking. In this paper we will describe some projects we have recently carried out.

2 Some Recent Projects Developed at the University of Pavia

The eye tracker employed in our systems is the Tobii 1750 (Figure 1), one of the most widespread eye tracking devices.



Fig. 1 The Tobii 1750 eye tracker

Combining video-oculography with infrared light reflection, the system looks like a common LCD screen but is provided with five NIR-LED and an infrared CCD camera, integrated in the monitor case. Near infrared light generates corneal reflections whose locations are connected to gaze direction. Eye positions are recorded at a frequency of 50 Hz, with a precision of about 0.5° of the visual field.

In the following, we will present *Eye-S*, *WeyeB*, *ceCursor* and *e5Learning*, four gaze-driven systems for, namely, text (and generic) computer input, Web browsing, cursor control and e-learning activity monitoring.

2.1 Eye-S: A Full-Screen Input Modality for Pure Eye-Based Communication

Eye-S is a gaze-based communication system which allows computer input of both text and generic commands [Porta and Turina 2008].

The problem of writing through the eyes has been widely considered in the past, and several solutions have been proposed. The simplest approach is based on on-screen keyboards and *dwelt time*: if the user looks at a certain key for more than a predefined time interval, the key is pressed and the corresponding letter is typed. Several studies have been carried out connected to this kind of typing and related issues. Unfortunately, approaches based on virtual keyboards suffer from a major weakness: unless keys are very big, fixations may not be exactly centered on them, and this may result in a frustrating typing experience for the user. On-screen keyboards are therefore usually large. A main advantage of *Eye-S* compared to other eye writing systems is that it leaves the screen totally free for the display of applications, since it does not need a specific graphical interface.

To implement *Eye-S* we drew inspiration from the concept of "eye graffiti", first introduced by Milekic [Milekic 2003]. According to this approach, gaze gestures are used to form a vocabulary in a way similar to the text input mechanism used in personal organizers, where letters are "drawn" using a pen. Through an "eye gesture" approach, the user can create alphabet letters, punctuation marks or specific commands by means of sequences of fixations on nine (hidden) predefined areas on the screen, as shown in Figure 2.

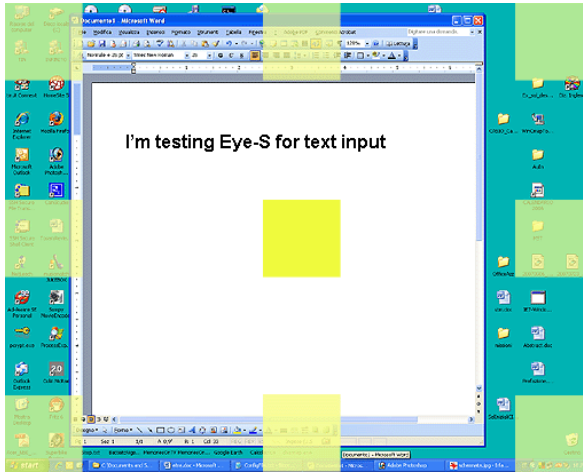


Fig. 2 *Eye-S*: explicit display of hotspots on the screen

These square areas, called *hotspots*, are placed in the four vertices of the screen, the middle of each side and the center. Even if the hotspots are not explicitly displayed (thus leaving the screen totally available for displaying any content), their position is easily predictable.

An *eye sequence* (from which the name *Eye-S* stems) is a succession of fixations on the hotspots. When the user looks at a hotspot for more than a given threshold time (e.g. 400 milliseconds), a *sequence recognition process* starts. If other hotspots are looked at after the initial one within a defined time interval, and if the succession of watched hotspots pertains to a set of predefined sequences stored in a configuration file, then a corresponding action is performed. If the system is being used for text input, the action will be the same as typing a key on a keyboard. Eye sequences can be chosen arbitrarily, but in the writing context they will resemble the form of letters (sample sequences for the ‘a’ and ‘b’ letters are shown in Figure 3).

During system use, the user can decide to get a feedback about the sequence composition process. To this purpose, when the user looks at the first hotspot for more than the defined dwell time a small green square is displayed within the hotspot itself. Such square contains a ‘1’, to indicate that this is the first hotspot of a possible sequence. If the user looks at another hotspot within a timeout, then a yellow square appears, with a ‘2’ written inside it (and the green square disappears). If the sequence which is being recognized is three segments long, the same happens for the third hotspot (orange square and ‘3’ as a sequence indicator). At last, on the final hotspot of a sequence — whether it is three or four segments long — a red square is displayed which contains the character or “action” recognized, in order for the user to immediately understand that the eye gesture has been successfully detected. The just described feedback process is exemplified in Figure 4.

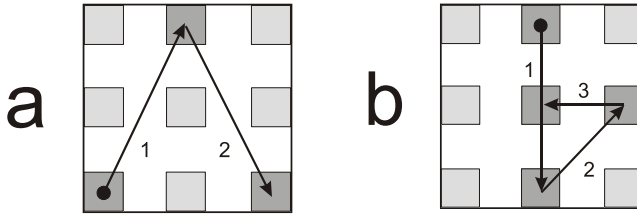


Fig. 3 Eye-S: examples of eye sequences for the ‘a’ and ‘b’ letters

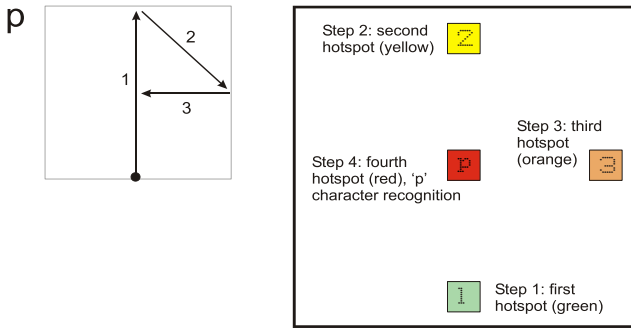


Fig. 4 Eye-S: example feedback provided for letter ‘p’

In order to better evaluate Eye-S, we have recently compared it with two other eye writing methods, namely a standard on-screen keyboard and EyeWrite [Wobbrock et al. 2007], a system in which letters are composed by looking at the corners of a square window specifically used for input, according to a predefined alphabet. Both novice and experienced testers took part in separate experiments, aimed at discovering potential qualities of Eye-S beyond its ability to leave the whole screen available for applications. After a short training period, each participant had to write nine sentences, three with each one of the three systems. Several metrics were employed for text entry research (e.g. Keystrokes per Character, Participant Conscientiousness and Total Error Rate). Among the results obtained, which are still being processed, two findings clearly stand out: (1) Eye-S, although is not as fast as the keyboard and EyeWrite, is the input system with the lowest error rate; (2) Eye-S, in spite of the need for a little longer training phase, has been the input system preferred by the majority of the testers.

2.2 WeyeB: An Eye-Controlled Web Browser for Hands-Free Navigation

WeyeB (from *Web eye Browser*) is an eye-controlled Web browser enabling the two basic activities required when surfing the Web, namely page scrolling and link selection [Porta and Ravelli 2009].

To date, very few systems have been devised for eye-controlled page scrolling and link selection. A common principle, used with some variants in different implementations, exploits “invisible threshold lines” on the screen: with reference to downward scrolling, when the user looks below a “start threshold”, in the lower part of the screen, the document begins to scroll slowly; when the user’s gaze reaches a “stop threshold” placed in the upper part, scrolling is stopped; if the user’s gaze falls below a “faster threshold” line (below the “start threshold” line), the system begins to scroll the content more rapidly.

In WeyeB, page scrolling can be activated by simply looking at one of two buttons, placed above and below the display area in the WeyeB interface. If the “watching time” (dwell time) goes beyond a predefined value, then the button is considered pressed. When this occurs, a semitransparent scroll rectangle is displayed over the page (Figure 5).

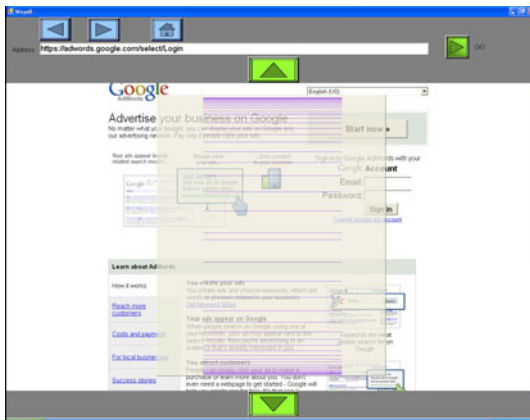


Fig. 5 WeyeB: scroll rectangle

By looking within the scroll rectangle, the content can be scrolled, with a speed that is lower (or even null) when the user’s gaze is in the central area of the rectangle, and increases progressively as it moves towards the upper or lower edges. The scroll rectangle is displayed as long as the user’s gaze is detected within it. Once the user looks anywhere outside the rectangle for more than a certain time (e.g. two seconds), it disappears. This page scroll solution gives the user full control: no scrolling can occur accidentally, since the scroll rectangle must be displayed beforehand, and the explicit display of such a graphical element represents an important visual feedback.

Also for link selection, different approaches are possible (see [Porta and Ravelli 2009] for a thorough description). For instance, links contained in the page area around the user’s fixation point may be displayed, as graphical rectangles, on the right side of the screen, and by moving the gaze there, the user can select them without interferences. For the selection of a link in WeyeB, the user must look at it for (at least) a dwell time. Subsequently, his or her gaze must be rapidly shifted

upward, and then immediately downward, again (about) on the target link. This sort of “eye gesture”, unlike other techniques, does not interfere with normal reading activities.

The detection of a selection gesture causes a “snapshot” of the area around the link to be acquired — the precision of current eye trackers is limited, and the perceived gaze may not be exactly centered on the link. The rectangular region (which has a fixed size, e.g. 200 x 100 pixels, and is centered on the fixation point) is stored in the form of a bitmap image (Figure 6).



Fig. 6 *WeyeB*: acquisition of a “snapshot” of the area around the watched link

The acquired image is then analyzed with OCR techniques to extract its textual content and identify possible (textual) links present in it (through an analysis of the HTML code). Three cases are then possible: (1) If the snapshot contains exactly one link, the corresponding page is loaded; (2) If the snapshot includes more than one link, a popup menu is shown which lists all the recognized links, easily selectable by gaze; (3) If no textual links are identified in the snapshot (e.g. because it contains an image link), the mouse cursor is simply shifted to the initial selection point and a mouse click is generated in that position.

The link selection approach adopted in *WeyeB* has several advantages. Compared to techniques that continuously display some kind of “enlarged versions” of links contained in the observed region of the page, an eye gesture eliminates potentially disturbing elements on the screen. On the contrary, an eye gesture is always fully intentional, and there is no risk that it is performed accidentally. Moreover, solutions displaying graphic elements near the link require dedicated space for their display, thus reducing the available area for actual page visualization. In *WeyeB*, the browser occupies the whole screen, therefore augmenting content accessibility.

2.3 *ceCursor*: Pointing with the Eyes in Windows Environments

People affected by severe motor impairments need effective methods for providing input to the computer, and exploiting eye gaze as a substitute for the mouse is potentially the most intuitive way to interact with a PC without using the hands: the “point-and-click” paradigm at the basis of current operative environments is universally adopted, and probably also the one most suitable for two-dimensional interfaces.

However, while pointing tasks are inherently connected with eye fixations — using the mouse, we look at a target and then move the cursor to it by means of a precise ocular-hand coordination — there are both physiological and technological obstacles which limit pure eye-based pointing. On the one hand, even during fixations the eyes are not perfectly still, but are characterized by jitters of different kinds; unless a mechanism for stabilizing the detected gaze position is employed, the eye-controlled pointer will tremble to some extent. On the other hand, even very recent eye trackers have a limited precision (typically, 0.5 degrees), and consecutive gaze samples acquired by the device cannot be exactly centered on the same point. For these reasons, the basic approach which simply displays the cursor where the user's gaze is detected on the screen is hardly practicable, since a shaking cursor is generally annoying, and precise pointing on small targets is practically impossible. We have therefore developed *ceCursor* [Porta et al. 2010], which is basically composed of a square, whose central point indicates the actual pointer position, and of four direction buttons placed around it (Figure 7).

Direction buttons are represented by triangles, and are pressed by eye gaze. The cursor is displayed on the screen with a semitransparent effect, and its size depends on the precision of the employed eye tracker, as well as on the eye pointing ability of the user (a cursor 300 pixels high and large is usually fine).

As will be explained, *ceCursor* behaves differently according to where it is at a certain moment. In any case, looking inside the central square causes a mouse click to be generated in its center after a dwell time (for instance, one second). Time lapsing is graphically represented by concentric circles progressively appearing within the square and filling it toward the center. After the first click, if another click is generated in the same position, it is interpreted as a double-click. If the user looks outside the cursor (that is, neither within the central square nor in direction buttons), after a dwell time it is shifted to a new position — the nearest icon if the cursor is on the desktop or within a folder, or the user fixation point if the cursor is within an application. A typical dwell time value is one second. The small 'M' placed in the lower-right area near *ceCursor*, when fixed for a certain time, causes the icon of a mouse to appear: looking at it, the user can change the currently active mouse button (right/left and vice versa, alternatively). The small circle located in the upper-right area near *ceCursor* is instead used to "stick" it in a certain place on the screen (it becomes more transparent and its color changes to red). This function allows the cursor not to be in the way of other user activities (e.g. reading) when not necessary.

In presence of icons, *ceCursor* is "captured" by them. This means that if the user looks at an area where there are icons, the cursor is automatically positioned on the nearest one. This behavior is in accordance with usual activities carried out within a folder or on the desktop, which necessarily involve icons. When *ceCursor* is positioned over an icon and the user looks at a direction button, the cursor "jumps" over the next icon in that direction, if there is one). This way, if the direct pointing was not successful (e.g. because the icon was small), it is very easy to shift the cursor to the right icon. Figure 8 shows an example with icons on the desktop. Within a folder, *ceCursor* can operate with any visualization mode of MS Windows (small and big icons, preview, details, etc.): the cursor is able to recognize the way icons are arranged, as well as their size, to correctly move among them.

When ceCursor is within an application window, or on the desktop but sufficiently far from icons, it can be precisely moved to the desired target. Practically, looking anywhere within an “icon free” area causes the cursor to be shifted to the fixed spot. However, since small interface elements are usually difficult to achieve at the first attempt, to exactly position the cursor the user can use direction buttons. As long as a direction button is fixed, the cursor is continuously and smoothly moved in that direction.

During continuous cursor movement, the area included in the central square is replicated within the active direction button (Figure 9). This way, the user can always be aware of what is being pointed by the cursor at a certain moment, even while constantly looking at a direction button to reach the target.

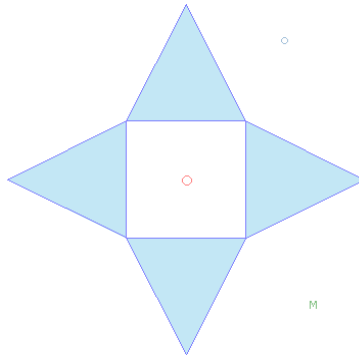


Fig. 7 ceCursor

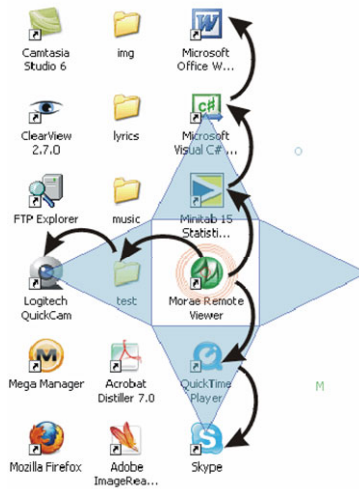


Fig. 8 ceCursor: discrete movement for icon selection on the desktop

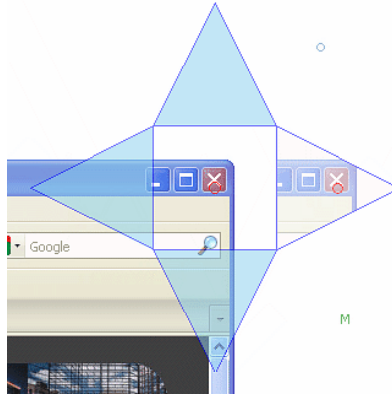


Fig. 9 *ceCursor*: replica of the currently pointed area displayed within the direction button (here the cursor is moving rightward)

2.4 e5Learning: An E-Learning Environment Based on Eye Tracking

e5Learning [Calvi et al. 2008], whose name stems from *enhanced exploitation of eyes for effective eLearning*, is an e-learning environment where eye tracking is exploited to allow the computer to get valuable data about users and their activities.

Unlike *Eye-S*, *WeyeB* and *ceCursor*, *e5Learning* is not designed specifically for disabled people; however, it can significantly improve the quality of distance learning, which is often the only possibility for a disabled person. The system is characterized by three key functionalities: (1) detection of basic user activities, such as reading text and observing multimedia content, in order to maintain a "history" of user actions; (2) generation of additional content depending on the context (e.g. textual or multimedia descriptions shown when the user is reading a specific portion of text); (3) recognition of stress, high workload and tiredness states in the user, using physiological data obtained from the eye tracker.

Thanks to the *Monitor of Accessed Screen Areas*, the author of the course can decide "how much attention" the user should pay to certain portions of content. In our prototype, a course is simply made up of Web pages. We use an ad-hoc-built Web browser which, along with page content, reads additional information defined by the author. Among other things, such information specifies the coordinates and sizes of screen rectangles (Regions of Interest, or RoIs) corresponding to relevant portions of content, and associated data. The *History Recorder* submodule relies on the *Monitor of Accessed Screen Areas* and keeps track of which portions of content (RoIs defined by the author) have already been accessed by the user, as well as "how much". If, for example, in a previous session the user did not devote sufficient time to a certain area, its content might be subsequently proposed before others, independently of its position in the logical structure of the course. Another

strategy, which is the one we have actually implemented, explicitly highlights the regions which need attention. When the user presses the ‘next’ button to load the next page in the course, if in the current page there are ROIs that have not been fully read or observed the system emphasizes them through colored rectangles (Figure 10).

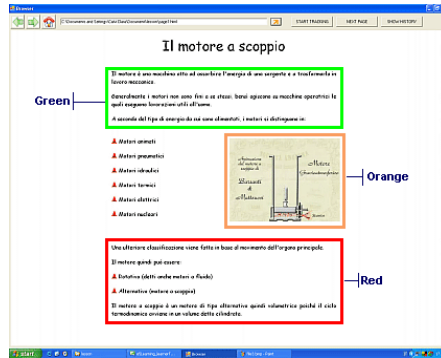


Fig. 10 *e5Learning*: colored rectangles highlighting regions of interest

Using the Contextual Content Generator, the creator of the course can associate new content to ROIs, and indicate the requirements for the additional information to be displayed (in the form of HTML pages appearing within a popup window, as shown in Figure 11). A condition for the new window to be shown is that the fixation time within a ROI is higher than a threshold.

A third module composing *e5Learning* is the *Emotion Recognizer*. Several experiments, mainly carried out in the Psychology and Physiology fields, have demonstrated that the observation of eye behaviors can reveal much information about the user emotional state. For instance, pupil size is significantly larger after highly arousing stimuli than after neutral stimuli. Other investigations (e.g. [Murata and Iwase 1998]) suggest that the mental workload can be assessed by analyzing the fluctuation rhythm of the pupil area. In particular, in our project we have considered two eye factors to (potentially) identify two main user conditions: (a) high workload or non understanding, and (b) tiredness.

For instance, if the average pupil size has progressively increased within a certain time interval, also user workload may have augmented. A decreased blink rate in the same period would further confirm such a supposition. When detected, these evidences could be used to dynamically modify the learning path, proposing a topic related to the main one but less complex (a sort of “break”). Or, if the user is potentially having problems in understanding something, extra information may be displayed.

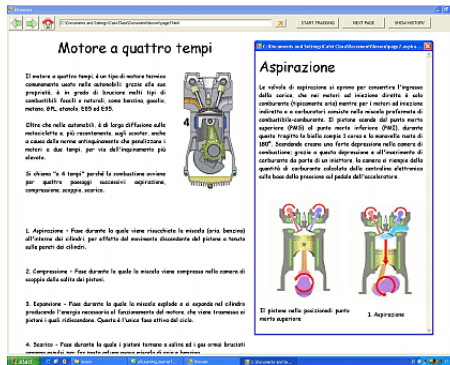


Fig. 11 *e5Learning*: additional content displayed when the user looks at a specific region of interest

3 An Application Scenario

In this section we propose a possible scenario where a person who cannot move at all (we will call him John) interacts with an eye tracker employing different kinds of applications.

For writing, John has been using an on-screen keyboard for a while, and, on average, he is now able to write about nine words per minute: a fairly good achievement. However, John feels that a common activity in computer usage like text input should be more "transparent", and should not interfere with the applications which require the input itself. Due to both technical constraints of the eye tracker used by John and his difficulty in precisely gazing at a target, the on-screen keyboard needs to be large, to allow for possible involuntary gaze shifts away from the keys. Since the functioning principle of the keyboard is the dwell time, there is no other possibility than big keys. But John, who likes very much writing, would nevertheless prefer a global overview of the page rather than having half of the screen occupied by the large keyboard.

Eye-S is potentially the right solution for John. After an initial short training period (according to our experiments, one hour is certainly enough), John can start writing by looking at the semitransparent hotspots. No matter the application requiring the input, John can display its graphic interface in full screen. Progressively, after a few days, John will be even able to remove the semitransparent hotspots, thus having the screen totally free. Text entry will probably not be as fast as with the on-screen keyboard — on average, six words per minute with *Eye-S*, as opposed to nine with the keyboard — but such a deficiency will be likely filled by a more comfortable interaction with applications.

Actually, John does not use an ordinary window-based environment. He works with a special suite of programs purposely designed to be accessible via an eye tracking device. While the suite includes all the main applications necessary for basic computer productivity (word processor, Web browser, email client, calculator, etc.), John feels himself somewhat constrained: he cannot interact with any

other program installed on his computer. Moreover, he cannot even start such programs, since the graphic interface of his Windows XP operating system is characterized by small controls (buttons, menus, etc.) and icons. It is therefore practically impossible for John to point and click a small icon or select a specific item within a drop-down menu.

Given the above explained drawbacks, it would be very useful for John to be able to exploit a mechanism allowing him to operate within a standard window-based environment. Although several gaze-driven pointers have been developed to date, they are rather difficult to control, especially if, like in the case of John, unintentional gaze shifts occur frequently.

ceCursor can be the appropriate answer to John's needs. The cursor is purposely designed to allow for errors in gazing at targets, and can be easily moved from one position to another by means of direction buttons. Furthermore, its self-adapting behavior according to the screen area where it is displayed (with or without icons) simplifies the interaction, markedly reducing the time necessary to open programs and folders. Of course, such an approach cannot be as fast as an ordinary mouse-based interaction, nor it is comparable with the ideal scenario in which the user can precisely control an eye-driven pointer using a very precise eye tracker. Nevertheless, using *ceCursor* John will be able to potentially carry out any operation within the Windows environment. This does not mean that he will not use applications from the program suite designed on purpose for gaze communication. Simply, John will be more "free": *ceCursor* will provide him with the ability to perform almost any task accomplishable with a personal computer, using standard software.

The program suite at John's disposal also includes a Web browser, which however does not satisfy him completely. Page scrolling can be performed either by looking at the upper or lower part of the screen (slow scrolling) or by means of lateral buttons (fast scrolling). Sometimes, it happens that the page is scrolled up or down inadvertently, just because John's gaze is perceived by the eye tracker as being directed toward the beginning or the end of the page. On the other hand, the need for a gaze shift outside the page — on the lateral scroll buttons — to accomplish a fast scroll forces John to continuously look back at the page, to check whether the right point has been achieved. This may be annoying for him, who would prefer a more efficient scrolling method.

The solution adopted in *WeyeB* solves these problems, as both slow and fast scrolling occurs only if the user explicitly triggers the "scrolling state" by looking at the scroll buttons. In addition, the user needs not to move his gaze away from the page during scrolling, since the semitransparent scroll rectangle allows page content to be clearly seen. Another problem with the Web browser included in the program suite used by John is that links are displayed on the right side of the screen: when he is looking at a certain portion of the page, links in that area are listed in the form of big rectangles, laterally. As a result, the region available for page display is reduced. This does not happen with *WeyeB*, because the links contained in the area where the link selection gesture has been perceived are shown in a semitransparent pop-up menu near the area itself. With *WeyeB*, John would thus be able to exploit all the screen for page display. In addition, the continuous

appearance and disappearance of link rectangles in the lateral region as John moves his eyes over the page may turn out to be irritating for him. Using WeyeB, he can instead trigger a link selection action only intentionally. The eye gesture solution is particularly suitable to prevent involuntary link activation, as it requires the user to explicitly move the eyes according to a well-defined path. With WeyeB, John can be certain that the pop-up menu with the links will be displayed only when he wants it to.

John is a student. As such, he needs to learn from didactic materials of different kinds, but of course all in electronic form. In other words, John is an e-learning user. While e-learning is now a relatively new, valuable form of education for everybody, it is especially useful for disabled persons, who cannot attend regular lectures in person. In such cases, e-learning may be the only way for a disabled student to obtain a diploma or graduate.

Current e-learning platforms are mainly Web-based systems. According to their level of complexity, they may be endowed with tracking tools and other support features aimed at helping users during their learning activity (for example, reminding them which units they have already completed, which ones are still to be started, etc., as well as recording their progress by means of tests, quizzes, exercises, etc.). However, the student is usually supposed to have the ability to freely move within the learning environment, jumping from one lesson to another, through a dynamic self-paced training process. Unfortunately, a student like John is much more limited in his choices, and his activity is slowed down by the inevitable interaction difficulties.

Even if not purposely designed for disabled people, the *e5Learning* learning environment can be a help for John. Once it will be adapted to be completely controlled by gaze, *e5Learning* will allow him to attend his virtual courses in a more comfortable way. Within each single page of the course, John will be able to easily remember what he has already read, and how much. Also, while reading specific portions of text or looking at multimedia objects, additional content will be displayed automatically, without requiring John to explicitly select links or click buttons (which would entail some effort). Moreover, stress, high workload and tiredness states, which may be critical to the health of John, can be strictly monitored by the system, thus avoiding potentially dangerous situations. In a sense, John is also advantaged in using *e5Learning* with respect to people without motor disabilities, because he is much more accustomed to interacting with an eye tracker.

4 Conclusions

The eye-based interfaces described in this paper are representative examples of the potentials of eye gaze communication for severely disabled people.

Eye-S is a system for providing eye input to the computer characterized by the fact that it does not require any graphical interface, therefore leaving all the available display area free for applications.

WeyeB is a gaze-driven system which allows Web surfing to be easily performed without using the hands, also by means of intuitive eye gestures.

ceCursor is an eye-based pointer designed with the purpose to allow potentially any interface element in MS Windows to be selected and activated, thus not constraining the user to use special program suites.

e5Learning is an e-learning environment, useful for disabled people who can only rely on e-learning for their education, in which eye data are exploited to track user activities, behaviors and “affective” states.

Thanks to recent technological advances, eye tracking devices have now become much more reliable than in the past, as well as more practical. We think that studies aimed at developing gaze-based assistive interfaces can greatly contribute to the diffusion of eye trackers among motor impaired people, considerably improving the quality of their lives.

References

- [Calvi et al. 2008] Calvi, C., Porta, M., Sacchi, D.: e5Learning, an e-learning environment based on eye tracking. In: Proc. of the 8th IEEE Int. Conf. on Advanced Learning Technologies, Santander, Cantabria, Spain (2008)
- [Donegan et al. 2005] Donegan, M., Oosthuizen, L., Bates, R., Daunys, G., Hansen, J.P., Joos, M., Majaranta, P., Signorile, I.: User requirements report with observations of difficulties users are experiencing. Communication by Gaze Interaction (COGAIN) (2005), <http://www.cogain.org/w/images/e/ef/COGAIN-D3.1.pdf> (retrieved October 6, 2010)
- [Duchowski 2007] Duchowski, A.T.: Eye tracking methodology – theory and practice, 2nd edn. Springer, London (2007)
- [europa.eu 2009] europa.eu Communication European i2010 initiative on e-Inclusion - to be part of the information society – europe’s information society (2009), http://ec.europa.eu/information_society/activities/einclusion/bepartofit/overview/ (retrieved October 6, 2010)
- [Milekic 2003] Milekic, S.: The more you look the more you get: Intention-based interface using gaze tracking. In: Proc. of the 7th Annual Museum and the Web Conference, Charlotte, North Carolina, USA (2003)
- [Murata and Iwase 1998] Murata, A., Iwase, H.: Evaluation of mental workload by fluctuation analysis of pupil area. In: Proc. of the 20th Int. Conf. of the IEEE Eng. in Medicine and Biology (1998)
- [Porta and Turina 2008] Porta, M., Turina, M.: Eye-S: a full-screen input modality for pure eye-based communication. In: Proc. of the 5th Symp. on Eye Tracking Research & Applications. ACM Press, USA (2008)
- [Porta and Ravelli 2009] Porta, M., Ravelli, A.: WeyeB, an eye-controlled web browser for hands-free navigation. In: Proc. of the 2nd IEEE Int Conf. on Human System Interaction (HSI 2009), Catania, Italy (2009)
- [Porta et al. 2010] Porta, M., Ravarelli, A., Spagnoli, G.: ceCursor, a contextual eye cursor for general pointing in windows environments. In: Proc. of the 6th Eye Tracking Research & Applications Symp. ACM Press, USA (2010)
- [Wobbrock et al. 2007] Wobbrock, J.O., Rubinstein, J., Sawyer, M., Duchowski, A.T.: Not typing but writing: Eye-based text entry using letter-like gestures. In: Proc. of COGAIN 2007, Leicester, UK (2007)

A Visual Hand Motion Detection Algorithm for Wheelchair Motion

T. Luhandjula^{1,2}, K. Djouani¹, Y. Hamam¹, B.J. van Wyk¹, and Q. Williams²

¹ French South African Technical Institute in Electronics at the Tshwane University of Technology, Pretoria, RSA
tluhandjula@gmail.com

² Meraka Institute at the Council for Scientific and Industrial Research, Pretoria, RSA
vanwyk@gmail.com, hamama@tut.ac.za

Abstract. This paper describes an algorithm for a visual human-machine interface that infers a person's intention from the motion of the hand. The context for which this solution is intended is that of wheelchair bound individuals whose intentions of interest are the direction and speed variation of the wheelchair indicated by a video sequence of the hand in rotation and in vertical motion respectively. For speed variation recognition, a symmetry based approach is used where the center of gravity of the resulting symmetry curve indicates the progressive position of the hand. For direction recognition, non-linear classification methods are used on the statistics of the symmetry curve. Results show that the symmetry property of the hand in both motions can serve as an intent indicator when a sequence of fifteen consecutive frames is used for recognition. This paper also shows less satisfactory results when fewer frames are used as an attempt to achieve faster recognition, and proposes a Brute force extrapolation algorithm to better the results.

1 Introduction

One of the challenges facing the task of realising an enabled environment where people with disabilities and the aged are independent and can therefore be active and contribute in society, is to develop systems that can assist them in performing the tasks they wish to carry out without other people's assistance [Luhandjula, Hamam et al. 2009; Luhandjula, Monacelli et al. 2009]. Good performance in a team/society environment is heavily conditioned by the awareness of people's intention within society [Kanno et al. 2003] and therefore human-machine interaction where the machine has a support role, requires that the intention of the user is well understood by the machine. This intention awareness capability is important for Human-System Interaction (HSI) and for the more specific area of the enabled environment.

Plan recognition is the term generally given to the process of inferring intentions from actions and is therefore an important component of HSI. The literature shows that the plan recognition community has spent some interest in

probabilistic network based approaches [Geib 2002]. Although plan recognition is a well-known feature of human collaboration, it has proven difficult to incorporate into practical human-computer collaboration systems due to its inherent intractability in the general case. [Lesh and Sidner 1999] describe a plan recognition algorithm which is tractable by virtue of exploiting properties of the collaborative setting, namely: the focus of attention, the use of partially elaborated hierarchical plans, and the possibility of asking for clarification. It has been shown that plan recognition can allow more efficient and natural communication between collaborators, and can do so with relatively modest computational effort. These are important results as Human-System collaboration provides a practical and useful application for plan recognition techniques.

One frequent HSI may be found in the context of a person with a physical disability whose mobility is constrained to a wheelchair. There are some solutions found in the literature such as [Christensen and Garcia 2003], where a new human-machine interface for controlling a wheelchair by head movements is presented. The position of the head is determined by the use of infrared sensors. The placements of the infrared sensors are behind the head of the user so that the field of view is not limited. Jia and Hu [Jia and Hu 2005] propose an integrated approach to real time detection, tracking and direction recognition of human faces, which is intended to be used as a human-robot interface for the intelligent wheelchair. It is implemented using Adaboost face detection and a canonical template matching to tell the nose position, therefore giving an indication of the position of the head and the direction the wheelchair must take. However these solutions are specifically dedicated to Head motion detection and are not suitable for people who would rather use their hands but not a joystick.

Many other platforms have already been devised to help people in their daily manoeuvring tasks: OMNI, Bremen autonomous wheelchair, RobChair, Senario, Drive Assistant, VAHM, Tin man, Wheelesley (stereo-vision guided), and Navchair (sonar guided) [Demeester et al. 2003]. These systems are based on “shared control” where the control of the wheelchair or any other assistive device is shared between the user and the device. Often the developed architectures consist of different algorithms that each realise specific assistance behaviour, “such as drive through door”, “follow corridor” or “avoid collision”. The presence of multiple operating modes creates the need to choose from them, and therefore makes the user responsible for selecting the appropriate mode, which might in some instances be an inconvenience.

In this paper, an alternative visual solution is proposed that infers the intention of a subject using the motion of the hand from its dorsal¹ view. The application intended for this solution is that of wheelchair bound individuals where the intentions of interest are the direction and the speed variation intended for the wheelchair. For speed variation recognition, a symmetry based approach is used where the center of gravity of the resulting symmetry curve indicates the progressive position of the hand. For direction recognition, non-linear classification methods are used on the statistics of the symmetry curve. A brute force extrapolation scheme is also proposed for faster intent recognition. This

¹ The dorsal view of the hand is referred to as the side other than the palm of the hand.

visual hand-based solution is proposed as an alternative to systems using joysticks, pneumatic switches and the motion of the head [Luhandjula, Hamam et al. 2009; Luhandjula, Monacelli et al. 2009]. The solution is non-intrusive and does not require the multiplicity of operating modes. This paper provides a contribution to the task of realising a Human System Interaction solution for the enabled environment allowing people with disabilities and the elderly to be more independent and as a result more active in society.

2 Methods

The type of data used for intention inference is visual: A sequence of images is captured by a CCD camera with a hand (from its dorsal view) in motion as the object of interest. The pre-processing step of detecting the hand in the field of view is not part of this work, and therefore it is assumed that the hand has already been detected. No visual aid or marker is provided on the hand to analyse the motion in the sequence. The hand performs two types of motion: Rotation and Vertical Motion around the wrist joint to indicate an intention in direction and speed variation of the wheelchair respectively. These intentions of interest become the commands for the motion of the wheelchair as described in Table 1 below:

Table 1 Map of hand motion to inferred intention

	Motion of Hand	Inferred Intention
Direction	Rotation to the Right	Move to the Right
	Rotation to the Left	Move to the Left
	No Rotation (Centered hand)	Move Straight
Speed Variation	Vertical motion Down	Increase speed
	Vertical motion Up	Decrease speed
	No Vertical motion (Centered hand)	Move at constant speed

For speed variation detection the method consists in extracting a symmetry curve from the input image with the hand as object of interest and the COG of the symmetry curve is calculated. A sequence of these centers of gravity is used to classify between the different types of motion namely “going at constant speed”, “going faster” and “going slower”.

For direction recognition the method consists in extracting a symmetry curve from the input image with the hand as object of interest, and a classifier (in this work the classification task has been performed using a neural network and a support vector machine: Refer to Section 2.4) is used to distinguish between the symmetry curves (using their means and standard deviations) associated to the different positions of the hand. A sequence of these positions is used to classify between the different types of motion namely “going straight”, “going left” and “going right”.

The rest of this section describes in details the above outlined approach used for intention recognition.

2.1 Symmetry-Based Approach

In previous work [Luhandjula, Hamam et al. 2009; Luhandjula, Monacelli et al. 2009] the merit of using a symmetry-based approach for intent recognition has been established. Though hands are not as symmetrical as faces, the underlying assumption is that a human hand from its dorsal view displays different symmetry properties as it moves vertically or rotates. These properties can be used to detect the motions (rotation and vertical motion) undertaken by the hand: Given a $X \times Y$ greyscale image I , the symmetry is calculated using the following expression:

$$S(y) = \sum_{\omega=1}^k \sum_{x=1}^X |I(x, y - \omega) - I(x, y + \omega)| \quad (1)$$

The symmetry-value $S(y)$ of each pixel-row in each image frame is evaluated $\forall y \in [k + 1 \ Y - k]$ by taking the sum of the differences of two pixels at a variable distance ω : $1 \leq \omega \leq k$ from it on both sides making the pixel-row the center of symmetry. This process is repeated for each column and the resulting symmetry-value is the summation of these differences. The symmetry curve is composed of these symmetry values calculated for all the pixel-row in interval $k+1 \leq y \leq Y-k$. It has been shown that the maximum distance that gives a more discriminative symmetry curve among the different positions is given by $k = 35$ for both images of size 240×200 and 145×250 of the hand in rotation and vertical motion respectively. The reason for this size difference is that the vertical motion of the hand requires more vertical space for the hand to remain in the field of view than in the case of the hand in rotation. This affects the position of the camera as well as the size of the field of view required for recognition.

2.2 Vertical Motion: Classification of Individual Positions of the Hand

The symmetry curves' center of gravity (COG) is used to classify the different positions of the hand in vertical motion. The COG is calculated as the point in the curve at which all the values of the curve can be considered centered:

$$C = \frac{y_1 S(y_1) + y_2 S(y_2) + \dots + y_n S(y_n)}{S(y_1) + S(y_2) + \dots + S(y_n)} \quad (2)$$

The symmetry curve is defined by the function $S : y \rightarrow S(y)$ with $S(y)$ given by (1), $y_i (\forall 1 \leq i \leq n$, with n the length of S) all the pixel-rows for which the symmetry curve S is calculated, and I is a 145x250 greyscale image frame as shown in Fig. 1. Fig. 2 shows the position of the COG on the symmetry curve for different positions of the hand in vertical motion as an indication of the position of the hand. Two approaches have been used to classify these different positions into three categories (center, up and down): The difference of means of the COGs, and the mean and standard deviation of the COGs in Gaussian distributions’.

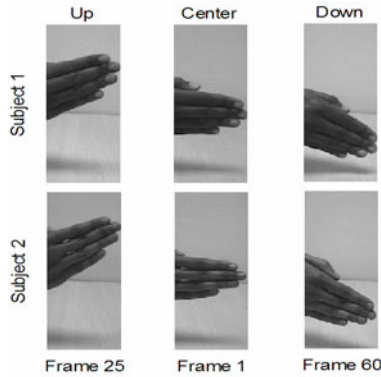


Fig. 1 Three different positions of right hand (dorsal view) in vertical motion

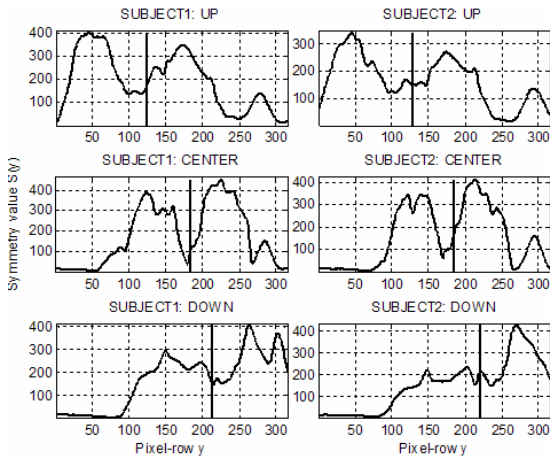


Fig. 2 Symmetry curves and the COGs for three different positions of the Hand. The COG is indicated by the vertical lines and representing the different positions of the Hand

2.3 Vertical Motion: Intention Detection for Speed Variation

The task of intent recognition involves the detection of the direction the subject intends to take and the speed variation he wishes to perform by looking at the motion of the hand. This section describes the recognition of the hand’s vertical motion indicating intent of variation in speed (increase and decrease for a down and up motion respectively). The time sequences of the symmetry curves’ COG give 15-elements vectors referred to in this work as “intention curves” and are used to recognize the different possible intentions namely: constant speed, decrease and increase in speed.

Let $E = \{I_i : I_i \text{ is the } i^{th} \text{ frame and } 1 \leq i \leq 15 \text{ frames}\}$, a sequence of fifteen consecutive image frames: $\forall I_i \in E, C_i$ is the COG of the symmetry curve (1) associated to I_i . The resulting intention curve designated by the vector $V = \{C_i : i = 1 \dots 15\}$ is shown on Fig. 3 and Fig. 4 for each scenario. In Fig. 3 both up and down vertical motions are captured from the center and in Fig. 4 the up vertical motion is captured from down to the center while the down vertical motion is captured from up to the center. These three types of motion (the hand remaining centered, the vertical motion of the hand up and the vertical motion of the hand down over time) exhibit different patterns and can therefore be easily classified.

Given the level of clarity on the difference between the intention curves associated to the three different classes of motion as shown in Fig. 3 and Fig. 4, Algorithm 3 describes the decision rules used for classification.

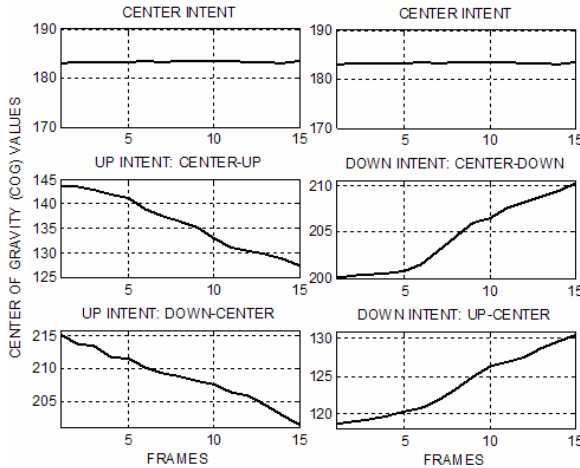


Fig. 3 Intention curves: Time sequence of the symmetry curves’ COGs for hands in vertical motion from centered position to up and down and from up and down position to the center

Algorithm 3. Decision rule for classification of intention curves. Implementation requires the use of either the difference of means or the statistics in Gaussian distribution

Let V be the intention curve to be classified:

Step1: Initialization
 $A = 0; B = 0;$ (Initialize A notifying a decrease and B notifying an increase $\forall i \in \{x : x \geq 1 \text{ and } x \leq \text{length}(V) - 1\}$,
 $D = V(i) - V(i+1)$
 If $D > 0$ $A = A + |V(i) - V(i+1)|$ (notifying a decrease in value of V, by adding the extent to which there is a decrease to the value of A)
 If $D < 0$ $B = B + |V(i) - V(i+1)|$ (notifying an increase in value of V, by adding the extent to which there is a decrease to the value of B)

Step2: Classification
 Let $\mu_{Class}, \sigma_{Class}$ be the statistics (means and standard deviations) of the difference between A and B in a training set for each class:
 $Class = \{Center, Up, Down\}.$

a) **Difference of means (DM):**
 $d_n = |(A - B) - \mu_{Class}^n|, \forall n = \{1, 2, 3\}$, $d = \min(\{d_1, d_2, d_3\})$
 If $d = d_1$ Center: Constant speed
 If $A > B$ and $d = d_2$ Up: Decreased speed
 If $A < B$ and $d = d_3$ Down: Increased speed

b) **Statistics (Mean and standard deviation) in Gaussian distribution (SGD):**

$$P_n = \frac{1}{\sqrt{2 \times \pi} \sigma_{class}} \exp\left\{-\frac{((A - B) - \mu_{class}^n)^2}{2\sigma_{class}^2}\right\}, \forall n = \{1, 2, 3\}$$
 $P = \max(\{P_1, P_2, P_3\})$
 If $P = P_1$ Center: Constant speed
 If $A > B$ and $P = P_2$ Up: Decreased speed
 If $A < B$ and $P = P_3$ Down: Increased speed

2.4 Rotation: Classification of Individual Positions of the Hand

The symmetry curves associated to images with the hand in rotation as object of interest (as shown in Fig. 5) do not display the same discriminative property as the hand in vertical motion, as well as the face in rotation and vertical motion as described in previous work [Luhandjula, Hamam et al. 2009; Luhandjula, Monacelli et al. 2009]. However, as shown in Fig. 6 they still exhibit different patterns for the different positions.

The approach proposed consists in calculating the statistics (means and standard deviation) of the symmetry curves for the different classes. Fig. 7 shows points corresponding to the statistics of the symmetry curves for the three different classes on a feature space made of mean values on the x-axis and standard deviation values for the y-axis. Given Fig. 7, two known non-linear classification methods are used: A Multi Layer Perceptron Neural network and a Support Vector Machine.

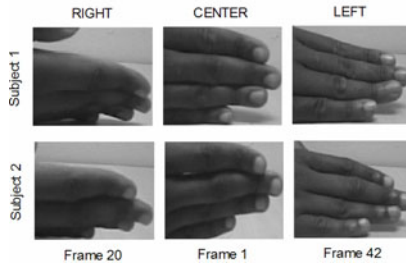


Fig. 4 Three different positions of the hand (dorsal view) in rotation of two different subjects

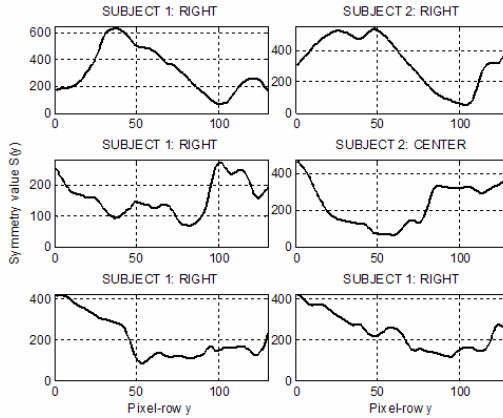


Fig. 5 Symmetry curves corresponding to the different positions of the hand in rotation in Fig. 5

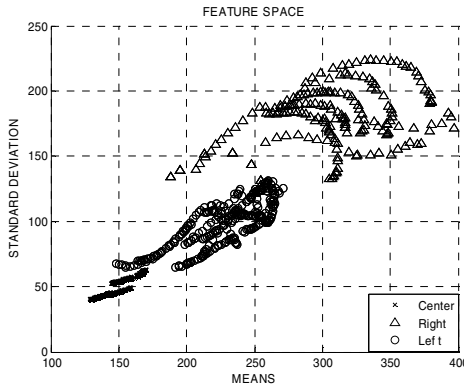


Fig. 6 Feature space with feature points representing hands in the center, right and left position. Each feature point is given by (μ_c, σ_c) , with $c = \{\text{Center, Right, Left}\}$

Neural Network: Multilayer Perception (MLP): *As a powerful data modelling tool, the neural network's ability to learn non-linear relationships [Bishop 1995] from data such as those shown in Fig. 7 is used. From empirical study conducted with the given data, the topology of the multilayer perceptron (MLP) is chosen to consist of a two neuron input layer, a 10 neuron hidden layer and the output. The training is performed using a backpropagation algorithm: Given a labelled training set consisting of a set of data points $x_c = (\mu_c, \sigma_c)$ with their accompanying labels T_c , $c = \{\text{center, right, left}\}$, the output is given by*

$$Y = f\left(\sum_{i=1}^N \omega^i x_i + b\right) \quad (3)$$

where N is the number of input neurons x_i from the previous layer ω_i is the weight associated to x_i , b is the offset from the origin of the feature space and f is the activation function chosen to be the sigmoid function:

$$f(x) = \frac{1}{1 + e^{-x}} \quad (4)$$

The weights are updated using

$$\omega_{jk}^{i+1} = \omega_{jk}^i + \Delta\omega_{jk}^i \quad (5)$$

where $\Delta\omega_{jk}^i = -\eta \frac{\partial E^i}{\partial \omega_{jk}^i}$, $E^i = \frac{1}{2} \sum_{o=1}^{N_o} (T_o - Y_o)^2$ (T_o and Y_o are the target and actual output of the network respectively).

Support Vector machine (SVM): *SVMs have become increasingly popular tools in data mining tasks such as regression, novelty detection and classification [Cristianini and Shawe-Taylor 2000] and can therefore be used for the classification problem at hand: Given a labelled training set consisting of a set of data points $x_c = (\mu_c, \sigma_c)$ with their accompanying labels T_c and $c = \{\text{center, right, left}\}$, the hyperplane expression is given by*

$$y = \langle x, \omega \rangle + b \quad (6)$$

where ω and b are the weights (giving the shape of the hyperplane) and offset from the origin respectively, and x is the data. The value of ω and b that maximizes the margin between the hyperplane and the support vectors is obtained using

$$\arg \min_{w,b} \left\{ \frac{1}{2} \|\omega\|^2 - \sum_{i=1}^N \alpha_i [y_i (\omega \cdot x_i - b) - 1] \right\} \quad (7)$$

yielding

$$\omega = \sum_{i=1}^N \alpha_i y_i x_i \quad (8)$$

$$b = \frac{1}{N_{SV}} \sum_{i=1}^{N_{SV}} (\omega \cdot x_i - y_i), \quad (9)$$

where α_i is the i^{th} Lagrange multiplier and N_{SV} are the numbers of support vectors which are found to be 63, 253 and 122 for the centered class, the right class and the left class respectively. Since this is a non-linear problem (refer to Fig. 7), the kernel trick is used to construct the hyperplane. The main idea behind the kernel trick is to map the data into a different space, and to construct a linear classifier in that space [Cristianini and Shawe-Taylor 2000]. The polynomial kernel k is used, and the solution becomes:

$$F(x) = \sum_{i=1}^N \alpha_i y_i k(x_i, x) \quad (10)$$

For this ‘three class’ problem, a one against one decomposition of the binary classifiers is used.

2.5 Rotation: Intention Detection for Direction

Let $E = \{I_i : I_i \text{ be the } i^{\text{th}} \text{ frame in a sequence of } N = 15 \text{ frames}\} : \forall I_i \in E$,

$$M_i = \frac{1}{L} \sum_{k=1}^L f_i(y) \quad (11)$$

where f_i is the symmetry curve associated to I_i . The resulting vector $V_2 = \{M_i : i = 1 \dots 15\}$ is shown on Fig. 8 and Fig. 9 for each scenario. It may be observed that rotation from the center to either side exhibits the same pattern while rotation from either side to the center also exhibits the same pattern but different from that of the previously mentioned rotation from the center to either side. It is therefore possible to distinguish between rotations from the center and those from either side. However, insufficient information is provided in V_2 to distinguish between rotation to the left and rotation to the right. To address this problem, a preliminary step is implemented that consists in getting another 15 elements vector V_1 made of the outputs of the MLP or the SVM. For a ‘center’ scenario, 15 consecutive 1s are expected, while 15 consecutive 2s and 3s are expected for right and left scenarios respectively. The Euclidean distance is calculated between the given vector V_1 and the three 15-points vectors made of ones, twos and threes respectively. The smallest distance indicates which class should be chosen (refer to Algorithm 4).

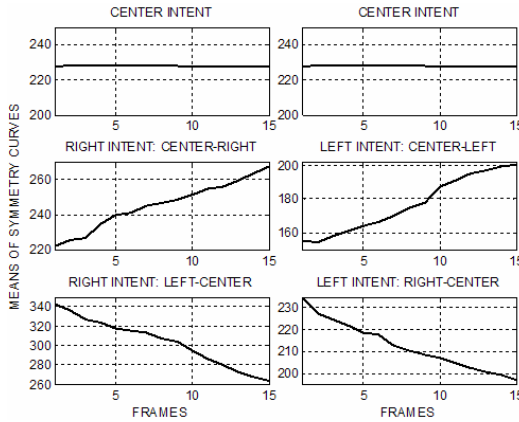


Fig. 7 Intention curves: Time sequence of the symmetry curves’ means for hands in rotation from centered position to right and left and from right and left positions to the center

Algorithm 4. Decision rule for intent recognition for the hand in rotation. The distance (dist) used is the Euclidean distance. If MLP is used in Step 1, V_2 in Step 2 is considered as an MLP-based intention curve, otherwise it is considered as an SVM-based intention curves

Step 1: Get the output of the MLP/SVM for 15 consecutive frames resulting in the symmetry curve: vector V_1 .

$d_1 = \text{dist}(V_1, \text{ones}(1,15));$
 $d_2 = \text{dist}(V_1, 2 \times \text{ones}(1,15));$
 $d_3 = \text{dist}(V_1, 3 \times \text{ones}(1,15)); d = \min(d_1, d_2, d_3);$

Step 2: Use Algorithm 3 on V_2 to classify between flat, ascending and descending V_2 .

if $d == d_1$ & V_2 flat	Centered motion
Else if $d == d_2$	Motion to the right
if V_2 ascending:	From center to the right
Else if V_2 descending:	From left to the center
Else if $d == d_3$	Motion to the Left
if V_2 ascending	From center to the left
Else if V_2 descending	From right to the center

2.6 Faster Recognition: Brute Force Extrapolation (BFE) Algorithm

This work shows (refer to section 3.2) that good results are obtained using 15 frames to extract the 15-points intention curve used for recognition. A brute force extrapolation scheme was applied for faster recognition. The aim of this approach as described in Algorithm 5 below, is to achieve recognition by extrapolating a 15-points intention curve from the n -points intention curves obtained using n frames ($n \leq 15$). The extrapolated 15-points intention curve is classified using both the DM and the SGD decision rules as described in Algorithm 1 and 2 respectively.

Algorithm 5. Brute force extrapolation algorithm

Let V be the intention curve to be classified:

Step 1: Initialization:
 $A = 0; B = 0;$
 $\forall i \in \{x: x \geq 1 \text{ and } x \leq \text{length}(V) - 1\},$
 $D = V(i) - V(i+1)$
 If $D > 0$ $A = A + 1$ (notifying a decrease in value of V)
 $D_a = V(i) - V(i+1)$
 If $D < 0$ $B = B + 1$ (notifying an increase in value of V)
 $D_b = V(i) - V(i+1)$

Step 2: D is the value to be added to get the subsequent values of the intention curve:
 if $A - B \geq n - 1$ & $V(1) > V(n)$ $d = \text{mean}(D_a)$
 if $B - A \geq n - 1$ & $V(1) < V(n)$ $d = \text{mean}(D_b)$
 else $d = 0$

Step 3: Extrapolation to a 15-points intention curve C :
 $N = 15 - \text{length}(V)$ (N is the number missing points in the intention curve to get 15), $j = \text{length}(V);$
 for $j = 1:N$
 $j = j + 1,$
 $C(j) = C(j - 1) + D$

3 Results

The experimental results have been obtained by collecting video sequences of five different subjects with three sequences each. The right hand in rotation and vertical motion viewed from its dorsal side is the object of interest. Different sets of results are given below, all aimed at demonstrating the merit of the proposed method. Section 3.1 shows the classification performance when known classes are used for individual frames indicating a specific position of the hand, section 3.2 gives the classification results for sets of 15 frames resulting in 15-points intention curves. Section 3.3 describes the results when a fewer number of frames is used thereby allowing faster intent recognition.

3.1 Results for Individual Position Classification

For classification of individual positions, the results are summarized in Tables 2 and 3. It may be observed that the up/down/center classification rate is better than the left/right/center classification. This is justified by the fact that the symmetry curves display more explicit changes for the vertical motion than for the rotation.

However, the first requires a bigger vertical region (refer to Fig. 1) of interest than the later (refer to Fig. 5) as the vertical motion scans a bigger area than the rotation. For speed variation recognition the Center class has the best classification rate and the Down classification displays the worst rate. From Table 2 it can also be observed that the SGD approach (98.3333%) performs slightly better than the

DM approach (98.1852%) again because of the added information provided by the standard deviation in the first method. From Table 3 it can be observed that for both MLP and SVM, the worst classification is that of the left class. The reason is that left and center hands are visually close to each other (Fig. 6 and Fig. 7) resulting in left hands misclassified as centered hands. This is more pronounced in the SVM giving a worse overall result than the MLP.

Table 2 Results on position classification for the hand in vertical motion (speed variation) on individual frames

Methods	Class	Training set	Testing set	Correct classification	Classification rate
DM	Center	450	900	900	100%
	Up	450	900	885	98.3333%
	Down	450	900	866	96.2222%
	Total	1350	2700	2651	98.1852%
SGD	Center	450	900	900	100%
	Up	450	900	893	99.2222%
	Down	450	900	862	95.7778%
	Total	1350	2700	2655	98.3333%

Table 3 Results on position classification for the hand in rotation (direction) on individual frames

Methods	Class	Training set	Testing set	Correct classification	Classification rate
MLP	Center	450	900	844	93.7778%
	Right	450	900	855	95%
	Left	450	900	823	91.4444%
	Total	1350	2700	2522	93.4047%
SVM	Center	450	900	900	100%
	Right	450	900	818	90.8889%
	Left	450	900	757	84.1111%
	Total	1350	2700	2475	91.6667%

3.2 Results for Intention Detection

For direction detection the decision rule described in Algorithm 4 is used where a combination of the sequence of symmetry curves' means (intention curves V_2) and the sequence of output from the MLP or SVM classifiers (intention curves V_j) constitute the input. For speed variation the intention curve is simply made of a sequence of 15 consecutive centers of gravity. The results are summarized in Tables 4 and 5.

Table 4 Results on speed variation intent recognition

Methods	Class	Training set	Testing set	Correct classification	Classification rate
DM	Center	400	600	531	88.5%
	Up	400	600	564	94%
	Up-back	400	600	489	81.5%
	Down	400	600	507	84.5%
	Down-back	400	600	565	94.1667%
	Total	2000	3000	2656	88.5333%
SGD	Center	400	600	522	87%
	Up	400	600	580	96.6667%
	Up-back	400	600	489	81.5%
	Down:	400	600	517	86.1667%
	Down-back	400	600	581	96.8333%
	Total	2000	3000	2689	89.6333%

Table 5 Results on direction intent recognition

Methods	Class	Training set	Testing set	Correct classification	Classification rate
MLP + DM	Center	400	600	573	95.5%
	Right	400	600	560	93.3333%
	Right-back	400	600	568	94.6667%
	Left	400	600	554	92.3333%
	Left-back	400	600	534	89%
	Total	2000	3000	2789	92.9667%
MLP + SGD	Center	400	600	573	95.5%
	Right	400	600	588	98%
	Right-back	400	600	564	94%
	Left	400	600	550	91.6667%
	Left-back	400	600	570	95%
	Total	2000	3000	2845	94.8333%
SVM + DM	Center	400	600	548	91.3333%
	Right	400	600	552	92%
	Right-back	400	600	522	87%
	Left	400	600	569	94.8333%
	Left-back	400	600	530	88.3333%
	Total	2000	3000	2721	90.7%
SVM+SGD	Center	400	600	548	91.3333%
	Right	400	600	586	97.6667%
	Right-back	400	600	526	87.6667%
	Left	400	600	563	93.8333%
	Left-back	400	600	564	94%
	Total	2000	3000	2787	92.9%

For direction detection the MLP classification yields better results than the SVM and the combination with the SGD approach gives the better classification rate. This is due to the added information of the standard deviation and the fact that given the data in Fig. 7.

3.3 Fast Recognition Results

This section shows the results of the proposed methods (DM, SGD as well as the BFE) using a fewer number of frames resulting in faster recognition. The processor used is an Intel(R) Pentium(R) D CPU 3.00 GHz processor. For a “25 frames per second” frame grabber the proposed solution has the ability to perform recognition in 600 ms excluding the execution time of the algorithm and the data acquisition time. For the interval ($2 \leq n \leq 15$) of the number of frames used for intention recognition, the data acquisition time required belongs to the interval $80 \text{ ms} \leq t \leq 600 \text{ ms}$. The results are shown in Tables 6, 7 and 8 (for 5 frames) for vertical motion, rotation using MLP and rotation using SVM respectively. Fig. 10, Fig. 11 and Fig. 12 show the performance for each classes, for the number of

Table 6. Recognition rate of the hand’s vertical motion for each class using the four proposed approaches (DM, SGD, BFE+DM, BFE+SGD) for intention curves obtained using 5 frames

Methods	Class	Training set	Testing set	Correct classification	Classification rate
DM	Center	300	100	100	100%
	Up	300	100	38	38%
	Up (back)	300	100	5	5%
	Down	300	100	11	11%
	Down (back)	300	100	14	14%
	Total	1500	500	154	30.8%
SGD	Center	300	100	100	100%
	Up	300	100	86	86%
	Up (back)	300	100	10	10%
	Down	300	100	33	33%
	Down (back)	300	100	3	3%
	Total	1500	500	232	46.4%
Brute force extrapolation + DM	Center	300	100	22	22%
	Up	300	100	100	100%
	Up (back)	300	100	70	70%
	Down	300	100	100	100%
	Down (back)	300	100	56	56%
	Total	1500	500	348	69.6%
Brute force extrapolation + SGD	Center	300	100	10	10%
	Up	300	100	100	100%
	Up (back)	300	100	70	70%
	Down	300	100	100	100%
	Down (back)	300	100	68	38%
	Total	1500	500	348	69.6%

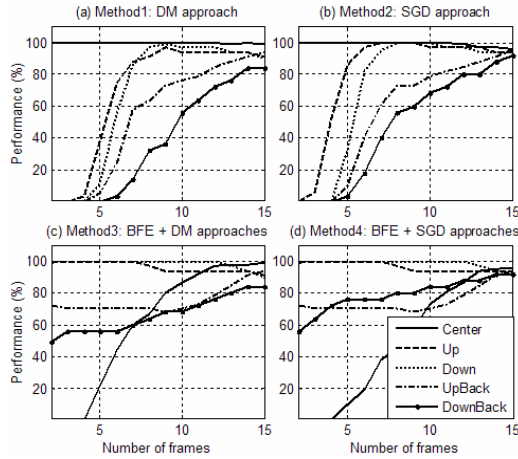


Fig. 8 Recognition rate of the hand’s vertical motion for each class using the four proposed approaches (DM, SGD, BFE+DM, BFE+SGD) for intention curves obtained using a number of frames ranging from 2 to 15

Table 7. Recognition rate of the hand’s rotation for each class using the four proposed approaches (DM, SGD, BFE+DM, BFE+SGD) for MLP-based intention curves obtained using 5 frames

Methods	Class	Training set	Testing set	Correct classification	Classification rate
DM	Center	300	100	100	100%
	Right	300	100	8	8%
	Right (back)	300	100	6	6%
	Left	300	100	0	0%
	Left (back)	300	100	7	7%
	Total	1500	500	121	24.2%
SGD	Center	300	100	100	100%
	Right	300	100	40	40%
	Right (back)	300	100	0	0%
	Left	300	100	0	0%
	Left (back)	300	100	57	57%
	Total	1500	500	197	39.4%
Brute force extrapolation + DM	Center	300	100	98	98%
	Right	300	100	92	92%
	Right (back)	300	100	100	100%
	Left	300	100	82	82%
	Left (back)	300	100	100	100%
	Total	1500	500	472	94.4%
Brute force extrapolation + SGD	Center	300	100	100	100%
	Right	300	100	96	96%
	Right (back)	300	100	100	100%
	Left	300	100	76	76%
	Left (back)	300	100	100	100%
	Total	1500	500	472	94.4%

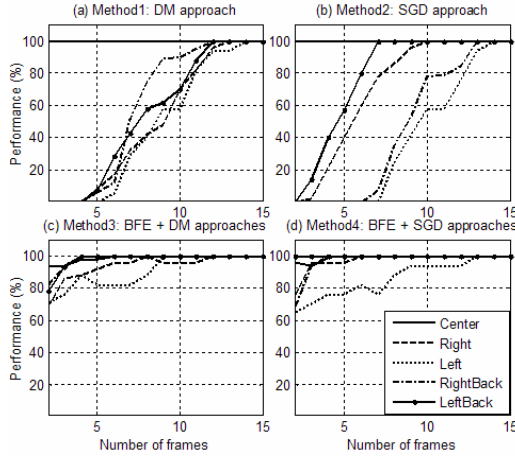


Fig. 9 Recognition rate of the hand's rotation for each class using the four proposed approaches (DM, SGD, BFE+DM, BFE+SGD) for MLP-based intention curves obtained using a number of frames ranging from 2 to 15

Table 8. Recognition rate of the hand's rotation for each class using the four proposed approaches (DM, SGD, BFE+DM, BFE+SGD) for SVM-based intention curves obtained using 5 frames

Methods	Class	Training set	Testing set	Correct classification	Classification rate
DM	Center	300	100	100	100%
	Right	300	100	4	4%
	Right (back)	300	100	0	0%
	Left	300	100	0	0%
	Left (back)	300	100	0	0%
	Total	1500	500	104	20.8%
SGD	Center	300	100	100	100%
	Right	300	100	97	97%
	Right (back)	300	100	0	0%
	Left	300	100	58	58%
	Left (back)	300	100	0	0%
	Total	1500	500	255	51%
Brute force extrapolation + DM	Center	300	100	100	100%
	Right	300	100	94	94%
	Right (back)	300	100	100	100%
	Left	300	100	82	82%
	Left (back)	300	100	48	48%
	Total	1500	500	424	84.8%
Brute force extrapolation + SGD	Center	300	100	94	94%
	Right	300	100	100	100%
	Right (back)	300	100	84	84%
	Left	300	100	88	88%
	Left (back)	300	100	64	64%
	Total	1500	500	430	86%

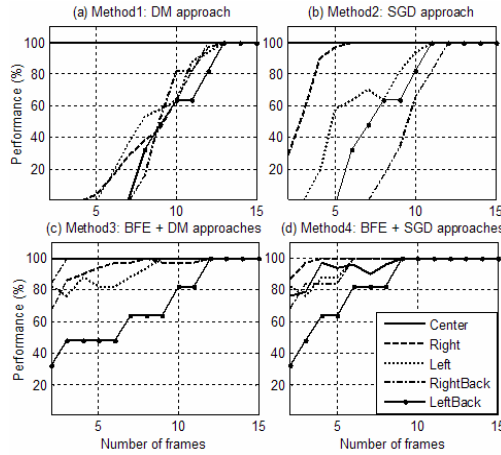


Fig. 10 Recognition rate of the hand's rotation for each class using the four proposed approaches (DM, SGD, BFE+DM, BFE+SGD) for SVM-based intention curves obtained using a number of frames ranging from 2 to 15

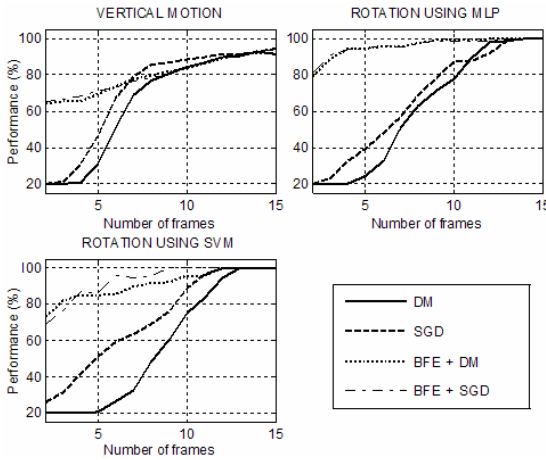


Fig. 11 Overall recognition rate of the hand's vertical motion and rotation using the four proposed approaches for intention curves obtained using a number of frames ranging from 2 to 15

frames n such that $2 \leq n \leq 15$. It can be observed in all three cases that the SGD method performs better than the DM as it displays better results for a fewer number of frames. Similarly the newly proposed brute force extrapolation (BFE) algorithm performs better than both except for the center class where the performance grows progressively as the number of frames grows, while it is constant (100%) for DM and SGD. Fig. 13 shows the overall performance for each method.

4 Conclusion

The experimental results show the validity of the proposed method. The merit of this solution is found in its simplicity where a simple symmetry property of the hand is used for rotation and vertical motion detection resulting in intention recognition. This solution targets some disabilities where the subject is only capable of moving the hand sufficiently to discern the type of motion, but not enough to manoeuvre a joystick. It can therefore be considered as an alternative among the numerous solutions found in literature for visual intention detection intended for wheelchair mobility. Ongoing work is underway to compare this method to existing works in literature in real time.

References

- [Bishop 1995] Bishop, C.: *Neural Networks for Pattern Recognition*. Oxford University Press, New York (1995)
- [Christensen and Garcia 2003] Christensen, H.V., Garcia, J.C.: Infrared Non-Contact Head Sensor, for Control of Wheelchair Movements. In: Pruski, A., Knops, H. (eds.) *Assistive Technology: From Virtuality to Reality*, pp. 336–340. IOS Press, Amsterdam (2003)
- [Cristianini and Shawe-Taylor 2000] Cristianini, N., Shawe-Taylor, J.: *An introduction to Support Vector Machines: and other kernel-based learning methods*. Cambridge University Press, New York (2000)
- [Demeester et al. 2003] Demeester, E., Nuttin, M., Vanhooydonck, D., et al.: Assessing the User's Intent Using Bayes' Rule: Application to Wheelchair Control. In: *Proc. 1st Int. Workshop on Advanced in Service Robotics*, Bardolino, Italy, pp. 117–124 (2003)
- [Geib 2002] Geib, C.W.: Problems with intent Recognition for Elder Care. In: *Proc. Association for the Advancement of Artificial Intelligence Workshop on automation as Caregiver*, Menlo Park, CA, USA, pp. 13–17 (2002)
- [Jia and Hu 2005] Jia, P., Hu, H.: Head Gesture based Control of an Intelligent Wheelchair. In: *Proc. Eleventh Annual Conference of Chinese Automation and Computing Society*, UK, Sheffield, pp. 191–203 (2005)
- [Kanno et al. 2003] Kanno, T., Nakata, K., Furuta, K.: Method for team intention inference. *Human-Computer Studies* 58, 393–413 (2003)
- [Lesh and Sidner 1999] Lesh, N., Rich, C., Sidner, C.L.: Using Plan Recognition in Human-Computer Collaboration. In: *Proc. Seventh International Conference on User Modeling*, Banff, Canada, pp. 23–32 (1999)
- [Luhandjula, Hamam et al. 2009] Luhandjula, T., Hamam, Y., van Wyk, B.J., et al.: Symmetry-based head pose estimation for intention detection. In: *Proc. 20th Annual Symposium of the Pattern Recognition Association of South Africa*, Stellenbosch, South Africa, pp. 93–98 (2009)
- [Luhandjula, Monacelli et al. 2009] Luhandjula, T., Monacelli, E., Hamam, Y., et al.: Visual Intention Detection for Wheelchair Motion. In: *Proc 5th International symposium on visual computing*, Las Vegas, USA, pp. 407–416 (2009)

Computerized Color Processing for Dichromats

J. Ruminski¹, M. Bajorek¹, J. Ruminska², J. Wtorek¹, and A. Bujnowski¹

¹ Department of Biomedical Engineering, Gdansk University of Technology, Poland
{jwr, martom, jaolel, bujnows}@biomed.eti.pg.gda.pl

² Intermedica Gdansk, Poland
jrumska@wp.pl

Abstract. Dichromacy is a serious color vision problem, where people can see and recognize only a limited number of colors. In this article we propose computerized methods and systems for processing of WWW images and pictures obtained using a camera phone. Image simulation, transformation and color vision difference visualization methods are presented as a part of two computerized systems. The paper presents also color recognition and labelling method. The results of experiments obtained by means of simulation and control group showed that both systems can be efficiently used to obtain a color name for indicated real world objects or for the image object contained on the Web.

1 Introduction

Color vision deficiency (CVD) is a functional disorder of vision resulting in troubles of color recognition and color differentiation. Persons suffering from CVD report different difficulties with colors. Abnormal color vision results from partial (very seldom complete) loss of function of cones or/and several other factors related to a human vision system (e.g. densities of macular and lens pigments). In reference to absence or dysfunction of photoreceptors the following CVD categories can be distinguished [Judd 1949]:

Monochromacy – total color blindness – all photoreceptors are absent or dysfunctional;

Dichromacy – partial color blindness – one type of photoreceptor is absent or dysfunctional,

Anomalous trichromacy – all photoreceptors are present, but functionality of one type of photoreceptor is different in reference to normal (average) observer.

Categories of abnormal color vision can be further classified using cones fundamentals of human color vision. Normal observer has three types of cones (color photoreceptors) sensitive to signals characterized by long- (L), middle- (M), and short (S) wavelengths. Combination of LMS signals is interpreted (in the human

color vision system) as a particular color. A lack or dysfunction of a cone type resulting in dichromacy [Cole 2007]: protanopia (absence of L-cones, 1% of USA population), deuteranopia (absence of M-cones, 1% of USA population), and tritanopia (absence of S-cones, rare <0,003%). In this work we will focus on dichromacy.

Color vision deficiencies are also age-related. Especially in macular degeneration the blue color perception is reduced (degeneration of S-cones, which the total number is lowest for all types of cones) [Muntean and Susan 2006]. The achromatic stage is observed as a final stage of color vision degeneration.

Color vision deficiency has been addressed in many previous works related to CVD tests, simulation of color deficiency vision, and color transformation for persons with CVD. The simulation of color deficiency vision has been investigated in [Brettel et al. 1997; Vienot et al. 1999]. Brettel, Vienot, and Mollon proposed a method using color transformation from RGB to LMS space. In LMS space a cone deficiency was simulated by collapsing one dimension by a proposed transformation. Next, modified LMS coordinates were mapped back to RGB. Some web-based image simulation software is available online at <http://www.vischeck.com> (Vischeck) and <http://www.ryobi-sol.co.jp/visolve/en/> (Visolve). Both sites offer generation of simulated images for dichromats, but details of a simulation procedure are not presented.

Knowledge of color-blindness can be used to design methods that modify the images to a more appropriate form for the target group. Recoloring objective is to change an image contrast in such a way, that individuals with abnormal color vision can distinguish objects similarly, as healthy observers. Additionally, if possible, colors should be preserved to be properly identifiable. In [Ichikawa et al. 2003; Ichikawa et al. 2004] authors proposed a method using image decomposition into interconnected regions. Spatial relations are important in pixel color modification. After calculation of optimal colors (using random bit-climber algorithm and color distance similar to CIE1976 [Seve 1991]) pixel's colors in the input image have been modified using representative and optimal color differences and distances between a pixel's color and each representative colors. Another color transformation method was presented in [Huang et al. 2007]. Authors proposed to rotate the color location in a^*b^* color space (CIE $L^*a^*b^*$) using specially designed rotation function (with a set of conditions) and the objective function which is minimized to obtain the optimal settings. The same group proposed another color transformation method in [Huang et al. 2008]. They adopted global histogram equalization using three local characteristics: a hue value at a given point, a maximum local hue difference, and local color information loss as a local color distance. Calculated data values have been used to define a hue transfer function, which is applied to recolor the input image. Procedure of recoloring an image for viewing by persons with color vision deficiency was implemented by authors of

VISICHECK system. The method is called daltonization and is available on-line (<http://www.visicheck.com>). Details are not published; however daltonization procedure is based on a histogram stretching.

The main goal of this paper is to present an integrated solution for computerized systems to simulate, transform and describe colors for individuals with dichromacy. Methods are implemented in two separate systems: a web proxy system and a camera phone system. Individuals, having problems with color discrimination can distinguish between objects in image content of web pages and obtain descriptive information about unseen/transformed colors. This is a very important feature of our proposal because the color transformation methods for persons with CVD generate false colors (which can differ between images for the same true color). Description (e.g. text labels) of unseen true colors can be valuable for web users with CVD. The camera phone system is portable and mobile so a user can be informed about a color name of the observed objects.

The remainder of the paper is structured as follows. In section 2 the methods for image simulation, transformation, and color recognition are presented. Section 3 describes implementation details of proposed systems and test scenarios. Next section demonstrates results. Final section 5 presents discussion and concludes the paper.

2 Methods

Proposed computerized systems for dichromats implements a set of image processing methods:

- simulation of images: original color transformation to a form as observed by dichromats,
- generation of image differences: creation of grayscale maps indicating color perception differences between an average, normal observer and an individual with dichromacy,
- transformation of images: color image processing to enhance the color contrast observed by dichromats,
- color identification and labeling for indicated pixel of an image.

2.1 Color Simulation

Simulation of color appearance for dichromats by color modifications can be used for better understanding of the difference in image content perception between a normal observer and an observer with abnormal vision. Normal color vision is modeled using L, M, and S cones fundamentals. The energy received by a particular type of cones can be represented as

$$[L, M, S] = \int E(\lambda) [\bar{l}, \bar{m}, \bar{s}] d\lambda, \quad (1)$$

where $E(\lambda)$ is the light power spectral density and $\bar{l}, \bar{m}, \bar{s}$ are fundamental spectral sensitivity functions for cones.

In dichromacy, one type of cone is absent or is not functioning properly so the 3D color space is narrowed or limited to a plane. In our system we used two methods to process an input image. First, an image is processed for the protanopic, deuteranopic and tritanopic simulation. Then, color difference images are calculated between the corresponding input and simulated images. Simulation procedure is based on the method introduced in [Brettel et al. 1997], [Vienot et al. 1999]. In the LMS space an observable color plane is defined as:

$$\alpha L + \beta M + \gamma S = 0 \tag{2}$$

where $\alpha, \beta,$ and γ are unknown parameters of a plane in the LMS space.

Solving equation (2) for red-green dichromacy (protanopia, deuteranopia) we used three reference points of the LMS space: origin (0, 0, 0), blue primaries (Lb, Mb, Sb), and white primaries (Lw, Mw, Sw). In case of blue-yellow dichromacy (tritanopia) we used red primaries (Lr, Mr, Sr) instead of those for blue. With known parameters new values for L, M, and S are calculated for given type of dichromacy. For example, in deuteranopia the L and S are unchanged, but new M value is calculated as a function of L and M:

$$M_D = -(\alpha L + \gamma S) / \beta. \tag{3}$$

Similar procedure is used in protanopia and tritanopia.

The following steps are applied to each image pixel represented by sRGB color coordinates:

1. Gamma correction
 $[R, G, B] = [R/255, G/255, B/255]^{2.2}$

2. Transformation of RGB to XYZ to LMS:

$$\begin{bmatrix} L \\ M \\ S \end{bmatrix} = \begin{bmatrix} 17.8824 & 43.5161 & 4.11935 \\ 3.45565 & 27.1554 & 3.86714 \\ 0.0299566 & 0.184309 & 1.46709 \end{bmatrix} \times \begin{bmatrix} R \\ G \\ B \end{bmatrix}$$

3. Transformation of 3D LMS space to dichromats 2D spaces solving the plane equation for protanopes:

$$\begin{bmatrix} L_p \\ M_p \\ S_p \end{bmatrix} = \begin{bmatrix} 0 & 2.02344 & -2.52581 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} L \\ M \\ S \end{bmatrix}$$

for deuteranopes:

$$\begin{bmatrix} L_D \\ M_D \\ S_D \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0.494207 & 0 & 1.24827 \\ 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} L \\ M \\ S \end{bmatrix}$$

for tritanopes:

$$\begin{bmatrix} L_T \\ M_T \\ S_T \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -0.012245 & 0.0720345 & 0 \end{bmatrix} \times \begin{bmatrix} L \\ M \\ S \end{bmatrix}$$

4. Inverse transform LiMiSi to XYZ to RGB, $i=\{P, D, T\}$:

$$\begin{bmatrix} R \\ G \\ B \end{bmatrix} = \begin{bmatrix} 0.080944 & -0.130504 & 0.116721 \\ -0.0102485 & 0.0540194 & -0.113615 \\ -0.000365294 & -0.00412163 & 0.693513 \end{bmatrix} \times \begin{bmatrix} L_i \\ M_i \\ S_i \end{bmatrix}$$

5. Inverse gamma correction:

$$[R, G, B] = 255 * ([R, G, B]^{1/2.2}).$$

After step No. 1 additional RGB scaling can be introduced to map colors to color gamut of the display standard (primaries, e.g. for ITU-R BT.709, scaling factor=0.992052). In our proposal we assumed that input images are in sRGB color system with gamma 2.2 (default for MS Windows and Mac Snow Leopard systems).

The color difference calculation was used to produce “error images” as pixel-to-pixel difference between the input image and each simulated image. Two color difference formulas were used CIE DE2000 [CIE 2001] and CIE 1976 [Seve 1991]. We have implemented the CIE DE2000 color difference with modification proposed in [Sharma 2004].

2.2 Color Transformation

Recoloring of images for dichromats is a very difficult problem because it is not possible to ideally map 3D color coordinates to a 2D color plane. Typical daltonization procedures use three categories of methods:

- a) color stretching (usually histogram based operations to introduce higher contrast between colors which are similarly perceived by dichromats),
- b) color remapping with minimization of an objective function (which defines the best remapping in sense of color contrast or “naturalness”),
- c) color coordinates transformation in a given color space (e.g. hue translation, rotation in $a*b*$ color space of the $L*a*b*$ color system, etc.).

Each category has disadvantages. Recoloring using the first category of methods cannot guarantee differentiation of colors for dichromats. Computational cost of methods based on optimization is very high and the same original color in two different images can be mapped to two different colors. An observer will be able to distinguish between colors however, the same colors can be presented differently

between images. The last category of methods also does not guarantee differentiation of colors because e.g. rotated color can overlap the color normally perceived by dichromats.

In our method we have used the following objectives:

preserve the normally perceived colors,
map unseen colors (color contrast) identically between images,
transform images using time effective.

We have designed a set of algorithms for red-green (protanopia, deuteranopia) and blue-yellow (tritanopia) dichromacies. The algorithms can be generalized using the following schema:

For a given image

1. Simulate a corresponding image for protanopes/deuteranopes/tritanopes,
2. Calculate and create color difference image using CIE DE 2000/CIE 1976 color difference formula (Δ_{CIE})
3. For each pixel i of an input image do
 - 3.1 normalize r , g , and b values to (0...1) range,
 - 3.2 IF ($\Delta_{CIE} [i] > K * \text{Just Noticeable Difference}$ AND $\text{color}(i)$ IS NOT gray) THEN
 - 3.2.1 IF protanopy/deuteranopy AND $\text{hue}(i)$ IS NOT blue
 - 3.2.1.1 IF $\text{distance}(b[i], \text{transformed}(r[i])) < T_d$, THEN
Shift $r[i], g[i]$ AND *transform* $r[i]$ to $b[i]$
 - ELSE
 $r[i] = g[i] = b[i] = \Delta_{CIE} [i]$,
 - 3.2.2 IF tritanopy AND $\text{hue}(i)$ IS NOT (red OR green)
 - 3.2.2.1 IF $\text{hue}(i)$ IS (yellow OR blue)
Shift $r[i], g[i]$
 - 3.2.3 limit the range of recalculated r , g , and b values (0...1).
- 3.3 ELSE DO NOTHING

The presented color processing schema is a combination of color stretching and color coordinates transformation. The “Shift” and “Transform” operation are functions of Δ_{CIE} and scaling factors (with positive or negative values chosen to separate red/green or blue/yellow colors). Details of particular algorithms and testing results (unit tests) are presented in [Rumiński et al. 2010]. In this paper we integrate the method in the computerized color processing systems for dichromats.

The “color difference image” and color transformation with shades of gray can be also very important, especially for individuals with monochromacy (e.g. in the age-related macular degeneration).

2.3 Color Recognition and Annotation

Color transformation can offer possibility to distinguish contrast between different objects in the image. However, color identification is lost when false colors (or modified ones) are introduced. In many life situations color names are used to describe entities (e.g. “a red pepper”, “a green line of a trend”). Possibility of image content labeling with color names could be very useful to prevent the exclusion of dichromats from the society.

In our research we focused on classification and labeling of image pixels. In the Web Proxy system event-based processing is used. Operator indicates the point of image which creates an event. The event object stores information about space/color coordinates of the pixel. The classification procedure uses color coordinates of the pixel or set of pixels (an average color value of pixels in a region around the indicated pixel) to find the color value and assign a name. In both systems (the web proxy system and camera phone system) we used nearest-neighbor classifier using Δ_{CIE} as distance measure (CIE DE 2000 and CIE 1997). The problem, however is to specify a training (reference) set with color names. Color labeling experiments have been performed to investigate assignment of color names by normal, average observers. The group of 37 volunteers (31 males, age 27.16 ± 4.31 , 6 females, age 29.5 ± 7.04) was asked to assign a color name (from the controlled vocabulary) to each colorful square in a chart with 64 EGA colors (the 8x8 matrix). We observed an almost full agreement (>90%) in case of fundamental colors (black, white, grays, red, green, blue, yellow, cyan, magenta) and high disagreement (about 50%) in case of color shades between subjects (i.e. the same color shade was labeled using different names, e.g. green or lime). As a conclusion of the experiment the limited set of colors was used in the final color dictionary (in Polish and English versions): black, white, grey, bright grey, red, green, blue, yellow, cyan, magenta, orange, pink, brown, purple, navy, sky/aqua. The color dictionary was implemented in both computerized systems for dichromats.

3 Systems and Experiments

Methods have been implemented in two separate systems: a web proxy system and a camera phone system. Web proxy system can be installed in-site (or can be located in the remote site) to filter image content in requested web pages. The camera phone system uses camera to capture image and recognizes a color in the center (defined region) of the image. The recognized color is indicated as a text label and as a sound message.

3.1 The Web Proxy System

Application Servers (Tomcat, Glassfish) and Java applications are used to implement the web proxy system. The web proxy is a locally or remotely installed system that filters a page content to replace every image (described in html or css files) to an applet. At the client site JavaScript code (web proxy menu) and Java Applets (image processing) are used for processing mouse events in order to modify image presentation (replacing the original image to the color difference image when a mouse cursor is over the original image) and to label image pixels (using contextual menu). As a result, web designers can simulate images for a proper site design. Individuals with dichromacy can distinguish between objects in the image content of web pages and can obtain information about unseen colors. In figure 1 screen copies of WWW browsers with active web proxy functions are presented.



Fig. 1 The main web proxy menu (left) and an image transformation menu (right) of the proxy (examples use the web site of the Polish online shop: www.alma24.pl). Colorful images for all figures are available in the electronic version of this paper

Configurable combination of keyboard/mouse events can be used to show the web proxy main menu (Fig.1, left) and an image menu (Fig.1, right). The main menu can be used to test color vision of a user, to configure the proxy (e.g. to set the type of color vision deficiency, to set a default mode for the proxy: which images should be presented by default: original, transformed, difference images, etc.), and to turn on/off the web proxy. The Firefox plug-in was prepared to integrate the web proxy functionality with the web browser. Other web browsers are currently under tests.

The most important functionality of the web proxy is accessible by indicating the chosen web image. Depending on the used combination of keys (a keyboard and/or a mouse) a user can obtain either an image transformation menu (Fig. 1, right) or a text label with a recognized color name (Fig. 2). Color recognition result can be also presented as a spoken word played from a sound file (WAV/MP3). So far Polish and English language versions have been prepared.

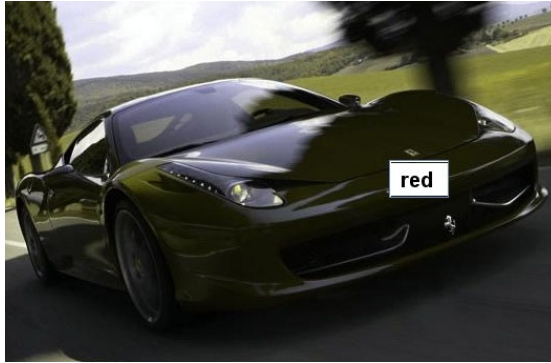


Fig. 2 The result of color recognition and labeling for the Ferrari.com web site

Each image of the original web site could be transformed to the applet. The applet code (18kB) is parameterized, so only one instance of the downloaded code (Java archive file, .jar) is required. In the evaluation version of the web proxy 4 different images are generated for each image: simulated, difference, transformed, transformed and simulated. This will lead to reduce performance of the system. Figures 3 illustrates some examples of the web proxy application.

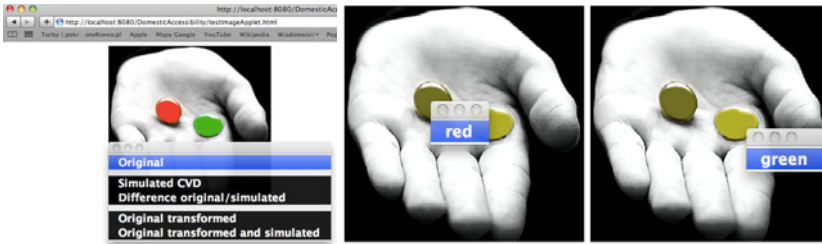


Fig. 3.“Take a red pillow” problem. Image manipulation applet in action, first simulated image is chosen, then colors of indicated pixels are recognized and labeled

The applet is responsible for loading a chosen type of an image (images are processed by the proxy system) and it performs color classification and description.

3.2 The Camera Phone System

The web proxy system is useful for web images (and using the same method for every image which is stored or can be obtained using desktop computer). However, people suffering from dichromacy need similar solution in a mobile version. For example, how to evaluate the color of tomatoes or peppers (green, yellow, orange, red) in the vegetable store? Asking a question could be not very comfortable.

The color manipulation procedures have been implemented using Java 2 Micro Edition. Implemented Java Midlet allows capturing a photo of a target and then processing the image center to get a text /spoken label of the color. It can be configured to use either a color of separated single pixel of the image or an average color of the set of pixels in the specified neighborhood of the image center. The chosen solution is limited to mobile phones which enable J2ME and JSR 135 Mobile Media API (camera phone support). In Fig. 4 examples of applications of the camera phone system are presented.

Both systems are highly configurable. For example, different color dictionaries can be used (i.e. languages, number of colors) to support internationalization of proposed computerized systems.

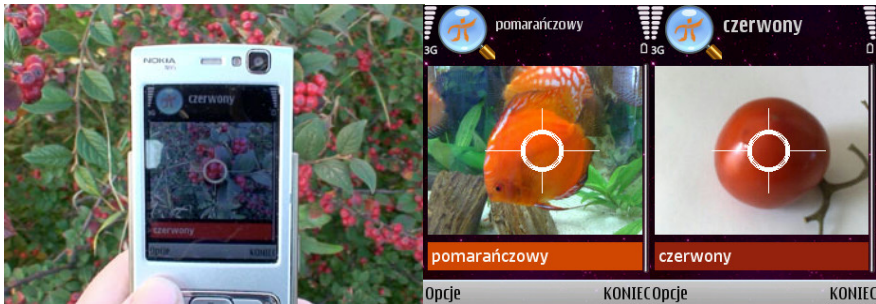


Fig. 4 Camera phone system in action; the application startup and some results. Polish language labels are assigned: "czerwony"="red", and "pomarańczowy"="orange"

3.3 Experiments

The set of experiments has been designed to validate the proposed methods and their implementations. Experiments have been performed using separate configurations for the web proxy system and the camera phone system. In first case the LAN-based system (max. 10Mbps) was used, consisting of:

- the Web Server with prepared test pages and the java archive of the applet,
- the computer (Intel Core 2 Duo CPU E6400 @ 2.4 GHz, 3.24GB RAM) equipped with a web browser (Firefox), Java Runtime Engine (JRE 1.6.0_18), and installed web-proxy for WWW content filtering.

The applet code was installed on the server site to simulate a configuration for possible remote installation of the web-proxy system (in such a case a local user downloads a page resources and additionally, the applet).

In case of the camera phone system the Nokia N95 mobile phone was used.

Both test platforms have been used to evaluate validity of color recognition and to assess the performance of the computerized systems. Validity of color recognition was measure in reference to a control group of 11 normal, average observers (mean age=25.8, std. dev=5.6; 4 females, 7 males). Text labels proposed by computerized systems were compared to those proposed by observers (using controlled vocabulary of names in the color dictionary and the same experiment conditions). The performance was measure as a time required to process a web page (with controlled content) and/or an image.

A set of images was used for experiments. First category of test images was Ishihara-based figures to evaluate simulation/transformation procedures. The second category of images was a set of artificially constructed images with colorful rectangles. Controlled color values were used. The initial test has been performed for the same colors as used in data dictionary. Another test image was constructed with 64 different colors: 16 basic colors of HTML version 3, and 48 randomly chosen color shades from X11/HTML 4 color sets (with color names).

Additional tests have been performed for common, everyday situations. The web-proxy system was used to filter hundreds of web pages (portals, on-line shops, etc.). The camera phone system was used outside to recognize colors of the surrounding objects (buildings, cars, fruits, vegetables, etc.).

Selected test images and results achieved for quantitative and qualitative experiments are presented below.

4 Results

Performed experiments have shown that each image (incl. Ishihara plates) is processed with a full agreement to the expected results described for each diagnostic test (e.g. what number a person would see if...). In Fig. 5 and Fig. 6 some results are presented for two Ishihara plates and for a picture of vegetables at the grocery shop.

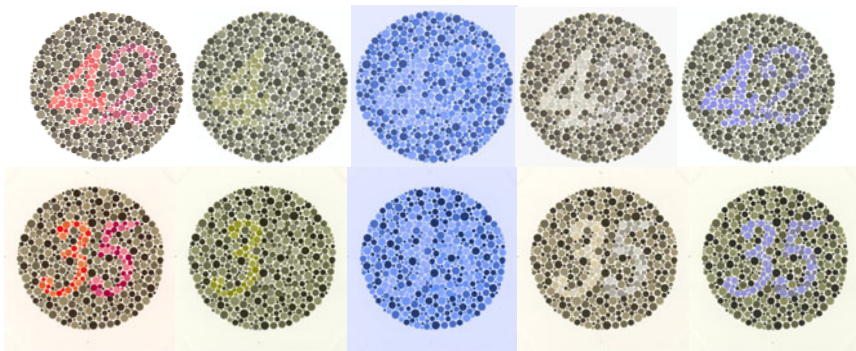


Fig. 5 Image simulation and transformation examples for Ishihara test plates 42 and 35: original, simulated, transformed and simulated by the Vischeck procedure, the Visolve procedure, and our procedure. Colorful images for all figures are available in the electronic version of this paper



Fig. 6 Original and transformed images as seen by deuteranopes: from top, left: original “vegetable” image and the results of: simulation, difference, the Vischeck procedure, the Visolve procedure, and our procedure

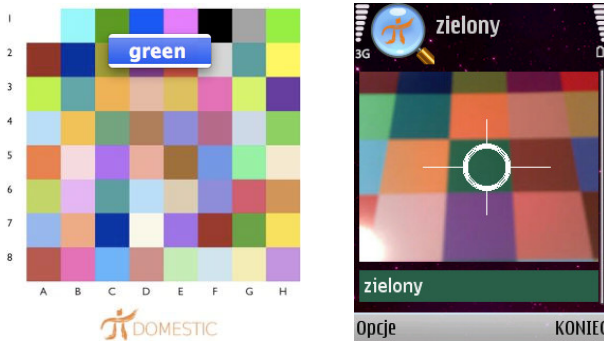


Fig. 7 The test image (“colorful rectangles”) processed by the web-proxy system and the camera phone system (“zielony”=“green”)

Validity of color recognition and labeling was performed in reference to a control group. The test image with 64 colorful rectangles (Fig. 7) was processed by the system and by normal, average observers.

Color labeling experiments with subjects produce interesting results. All subjects have given the same color name for 23 color rectangles (36%). In 12 cases exactly one color assignment was different (e.g. 10 * “green”, 1 * “yellow”), in 10 cases exactly 2 labels were different. Assigned labels for 8 rectangles were spread in two colors (about 40% one color, 60% the other). This result was observed for the shades of colors lying on the following lines: yellow-green, green-cyan, blue-navy, magenta-purple. The number of participants was not high (11) since the control group was used as a reference to the computerized systems. No subject-specific or sex-specific results have been observed.

The computerized systems (screen-based and printed pattern based tests) were tested similarly as human subjects. Both systems produced the same color labels for those 23 rectangles that have been identically labeled by all subjects. In other cases the recognized and labeled color was one of those, indicated by at least one human responder. In most cases the systems return a color name that was assigned by higher number of subjects. However some exceptions were observed. The colors characterized by high luminance and low (and similar) color coordinates (a,b) were recognized by the systems as white or bright grey, while human subjects recognized other colors with greater precision (e.g. pink, yellow). For the same high luminance (low color) cases there was a difference observed in color recognition between the web-proxy system and the camera-phone system. The mobile system captures picture of the pattern, so introduces some color modifications. This leads to different results for observed colors, which coordinates lie close to the border between colors from the used data dictionary (red-brown, blue-sky, high luminance/low hue colors).

Using the same test image (“colorful rectangles”) two color difference measures were compared (i.e. modified CIE DE 2000 and CIE 1976). Both measures produced identical color recognition and labeling results. However, about 30% different results were observed in the experiment for all possible RGB (256*256*256) colors compared to the used color dictionary. The results obtained with the modified CIE DE 2000 and CIE 1976 were different, giving similar, but not identical shade of the same base color.

In case of performance tests two groups of experiments have been done: image processing time and web page processing time. Image processing time was calculated for each implemented operation: simulation, difference, transformation, and simulation for a transformed image. Two images have been used: “vegetables” (595x662 pixels) and “colorful rectangles” (307x353 pixels). Both images have been processed to generate final test images in 3 different resolutions: 25%, 50%, and 100% of the original, proportional size. For each image ten tests have been performed for the modified CIE DE2000 and CIE 1976 color difference measures.

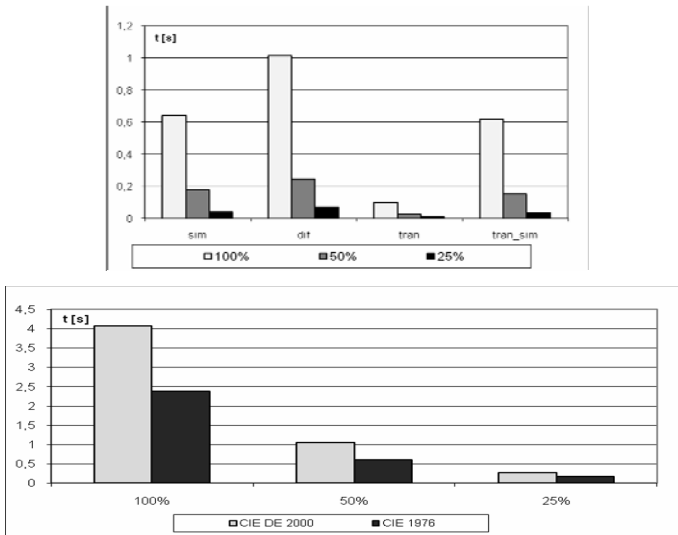


Fig. 8 Results for a single image tests: top – elementary operations (simulation, difference, transformation, simulation of transformed image) for the “colorful rectangles” test image for different resolutions; bottom – time required for all operations for the “vegetable” test image for different color difference measures

In Table 1 processing time of web pages is presented as a function of total number of images in a web page, resolution of images, and color difference measures.

Table 1. Total processing time (from the request to the complete content with all processed images) of web pages with different number of images

		Total processing time [s]			
		"vegetables"		"colorful rectangles"	
Resolution	No. of images	CIE 2000	CIE 1976	CIE 2000	CIE 1976
100%	1	4.66	3.13	1.37	0.92
	5	24.14	14.69	6.50	4.04
	10	48.46	29.19	12.40	7.95
	20	93.22	57.90	24.79	16.50
50%	1	1.27	0.90	0.42	0.35
	5	5.96	4.19	1.70	1.19
	10	11.98	7.82	3.26	2.36
	20	24.33	15.57	6.63	4.42
25%	1	0.48	0.30	0.18	0.14
	5	1.72	1.18	0.65	0.56
	10	3.33	2.25	1.22	0.94
	20	6.80	4.39	2.21	1.83

5 Discussion and Conclusions

Results for image simulation procedures were in agreement with expecting results of standard Ishihara test and also for other test images. Additionally, during experiments two subjects with red-green dichromacy indicates no difference between the original image and the simulated image. Together with image simulation procedure we calculated color difference images. Color difference was further used in color transformation methods. The performance tests have proved high computational requirements for CIE DE 2000.

Quantitative results of proposed color transformation procedures were compared to results of the Vischeck and Visolve methods. Transformation method implemented in our computerized systems produces higher gradient between colorful objects in the observed (by dichromates) objects. The Vischeck method highly modifies colors which are naturally visible for dichromats (e.g. Brussels sprouts color in Fig. 6). The proposed method uses color difference procedure, so the total computational is higher than histogram-based stretching methods but much more effective than optimization-based methods [Rumiński et al., 2010].

The most important aspect of the computerized systems for dichromates is possibility of color recognition and labeling (improving image contrast does not answer for many questions, e.g. if the fruit is ripe). Results of the validity of color recognition and labeling method are very interesting. The observed small differences between the results of individuals in the procedure of color labeling suggests that it is very difficult to map all colors (i.e. in a given sense, e.g. sRGB) to finite number of color names. The aim of this study was to assign a given color to its nearest main color (i.e. color from the color dictionary) and assigning it the relevant text label. Only 16 colors were used in the color dictionary. The computerized systems map any color to one of the color from the color dictionary (color reduction). The recognized color and the assigned color name is fully repetitive operation in comparison to different human subjects, which can vary in their decisions (due to some genetic, cultural or other reasons). Almost identical results were observed between average human results and the computerized systems results. In case of camera phone system there is high influence of image capturing procedure on the final recognition and labeling results. In an environment with poor lighting or with multiple light reflections recorded images do not accurately reflect the observed colors. Then, identified colors represent the color of the pixel image, not the color of the observed object. During experiments we also discovered two subjects with red-green dichromacy. They have suggested, that the web-proxy system offers valuable support in such cases, where textual web content refers to a color in an image, e.g. "...X feature in the image was indicated with red".

Described computerized systems are designed for dichromats who (majority of them) have never seen some colors. For example, for protanopes/deutanopes the "red" color name does not carry the same information (no "seen" reference) as for normal, average observer. Dichromats can use such names only as descriptors for

particular features (etc., fruit, painting, traffic sign, etc.). Much bigger problem, however, is related to shades of the "not perceived" color (e.g. "crimson" as a shade of red). The conclusion of the experiments is to recognize and label not only the single color name, but also category of the color, i.e. main color names (e.g. "shade of red", "shade of yellow-green", etc.).

The problem of color recognition and naming (in a given culture group) requires further research. One of the fundamental questions is how to build a universal color dictionary (including color categories – color reduction problem for dichromats). For example, National Bureau of Standards, USA (currently National Institute of Standards and Technology) specified 267 color names, which can be classified into 18 groups [Kelly and Judd 1976]. In [Berlin and Kay 1969] authors studied color terms in 100 languages and proposed that there exist 11 universal color terms: white, black, red, green, yellow, blue, brown, purple, pink, orange, and gray. Wikipedia (http://en.wikipedia.org/wiki/List_of_colors) specifies 493 color names which are divided by shade into 11 groups: white, pink, red, orange, brown, yellow, gray, green, cyan, blue, and violet (black is classified as a shade of gray). There are many other proposals (e.g. for WWW) that should be further investigated.

Color recognition and labeling procedure operates on a single pixel (or average set of pixels), so computational performance is high. The response time of the systems is almost real-time (i.e. a user does not observe important delay, usually < 1ms for both systems). In case of the camera phone system the processing performance is reduced by the "taking a photo" procedure, which depends on the used phone.

In this paper two computerized systems for dichromats were presented. The web-proxy system offers interesting features also for normal, average observers (e.g. vision simulation in dichromacy). All features were investigated in tests. The performance for the web-proxy system with full features was not satisfactory in case of rich web content (20 high resolution images on a web page) and for CIE DE 2000 distance measure. However, experiments proved that for identical recognition/labeling results have been achieved for the modified CIE DE 2000 and CIE 1976 (for tested images). In the case of a small number of colors in the color dictionary, the CIE 1976 measure is a fully sufficient. Additionally, a user with dichromacy does not need some features (e.g. simulated version of the transformed image). Instead of generating all the versions of images it will be appropriate to prepare only images with transformed colors for a given type of dichromacy (e.g. a user oriented). This can be configured in the web-proxy main menu. Other images could be available on demand using the applet menu (or other client-side programming method). Taking into account the results of the performance tests it could be possible to reduce total processing time to less than 2 seconds for rich web-page content. The proxy/applet code was not optimized, but was used only to verify the idea of web-content processing for dichromats. It can be observed that the performance is a linear function of the total number of pixels in an image. Further optimization steps in elementary procedures can improve total performance of the web-proxy system.

Future activities will concentrate on testing of color vision of an individual subject for a dedicated tuning of parameters of remapping procedures. Additionally, calibration of a graphical system and problem with different RGB color systems (primaries) of images will be considered. In this work we have assumed only sRGB color space.

Presented work is a part of the integrated system, which is under development and it is devoted to help individuals, which have web and computer accessibility problems. Another mobile system is currently under design. Color processing methods will be implemented in the electronic glasses that can introduce real-time help for individuals with dichromacy. We hope that proposed assisted living technology will be useful for persons with different types of chromatic disorders.

Acknowledgment

This work was partly supported by European Union, European Regional Development Fund concerning the project: UDA-POIG.01.03.01-22-139/09-00 -“Home assistance for elders and disabled – DOMESTIC”, Innovative Economy 2007-2013, National Cohesion Strategy.

References

- [Berlin and Kay 1969] Berlin, B., Kay, P.: Basic color terms: Their universality and evolution. University of California Press, Berkeley (1969)
- [Brettel et al. 1997] Brettel, H., Vienot, F., Mollon, J.D.: Computerized simulation of color appearance for dichromats. *J. Opt. Soc. Am.* 14(10), 2647–2655 (1997)
- [CIE 2001] CIE, Improvement to industrial colour-difference evaluation (142), CIE Publication (2001)
- [Cole 2007] Cole, B.L.: Assessment of inherited colour vision defects in clinical practice. *Clin. Exp. Optom.* 90(3), 157–175 (2007)
- [Huang et al. 2007] Huang, J., Tseng, Y.C., Wu, S.I., Wang, S.J.: Information preserving color transformation for protanopia and deuteranopia. *IEEE Signal Processing Letters* 14(10), 711–714 (2007)
- [Huang et al. 2008] Huang, J., Wu, S., Chen, C.: Enhancing color representation for the color vision impaired. In: *Int. Work on Computer Vision Applications for the Visually Impaired*, in conjunction with European Conf. on Computer Vision, Marseille (2008)
- [Ichikawa et al. 2003] Ichikawa, M., Tanaka, K., Kondo, S., Hiroshima, K., Ichikawa, K., Tanabe, S., Fukami, K.: Web-page color modification for barrier-free color vision with genetic algorithm. *LNCS*, vol. 2724, pp. 2134–2146 (2003)
- [Ichikawa et al. 2004] Ichikawa, M., Tanaka, K., Kondo, S., Hiroshima, K., Ichikawa, K., Tanabe, S., Fukami, K.: Preliminary study on color modification for still images to realize barrier-free color vision. In: *Proc. IEEE Int. Conf. Systems Man: Cybernetics*, pp. 36–41 (2004)
- [Judd 1949] Judd, D.B.: Color perceptions of deuteranopic and protanopic observers. *J. Opt. Soc. Am.* 39(3), 252 (1949)

- [Kelly and Judd 1976] Kelly, K.L., Judd, D.B.: COLOR universal language and dictionary of names. National Bureau of Standards special publication (1976)
- [Muntean and Susan 2006] Muntean, A., Susan, L.: Alterations of the color vision in elders. *Cercetări Experimentale & Medico-Chirurgicale (Anul XIII 1 Nr.1/2006)*, pp. 49–52 (2006)
- [Ruminski et al. 2010] Ruminski, J., Wtorek, J., Ruminska, J., Kaczmarek, M., Bujnowski, A., Kocejko, T., Polinski, A.: Color transformation methods for dichromats. *Human System Interactions*, 634–641 (2010)
- [Seve 1991] Seve, R.: New formula for the computation of CIE 1976 hue difference. *Color Research and Application* 16, 217–218 (1991)
- [Sharma 2004] Sharma, G., Wu, W., Dalal, E.N.: The CIEDE2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations. *Color Research & Application* (30), 21–30 (2004)
- [Vienot 1999] Vienot, F., Vienot, F., Mollon, J.: Digital video colourmaps for checking the legibility of displays by dichromats. *Color Research and Application* (24), 243–252 (1999)

Sitting Motion Assistance for a Rehabilitation Robotic Walker

D. Chugo¹, H. Ozaki², S. Yokota³, and K. Takase²

¹ School of Technology and Science, Kwansai Gakuin University, Hyogo, Japan
chugo@kwansai.ac.jp

² Graduate School of Information Technology,
The University of Electro-Communications, Tokyo, Japan
{ozaki, takase}@taka.is.uec.ac.jp

³ Faculty of Science and Engineering, Setsunan University, Osaka, Japan
Yokota@mec.setsunan.ac.jp

Abstract. In our current research, we are developing a robotic walker system with standing, walking and seating assistance function. Our developing system is based on a walker which is popular assistance device for aged person in normal daily life and realizes the standing and seating motion using the support pad which is actuated by the novel assistance manipulator mechanism with four parallel linkages. In this paper, we develop the control scheme which realizes the natural seating motion with fewer loads to the patient. For developing control scheme, we investigate the seating motion of aged people who requires to power support and typical seating motion by healthy young people. Comparing with two motions, we set the reference of seating motion with our system and we discuss the required assistance condition during seating motion. Our key ideas are two topics. One topic is analysis of condition which realizes the seating motion as young healthy people. The other topic is combination of force and position control. According to the patient's posture during seating motion, our control system select more appropriate control method from them. Using proposed control, our system reduces the patient's load and maintains his posture stably when it is necessary.

1 Introduction

In Japan, the population ratio of senior citizen who is 65 years old or more exceeds 23[%] at February 2010 and rapid aging in Japanese society will advance in the future. In aging society, many elderly people cannot perform normal daily household, work related and recreational activities because of decrease in force generating capacity of their body. Today, the 23.5[%] of elderly person who does not stay at the hospital cannot perform daily life without nursing by other people. For their independent life, they need a domestic assistance system which enable them to perform daily activities alone easily even if their strength are not enough.

Especially, standing and seating motion is the most serious and important operation in daily life for elderly person who doesn't have enough physical strength [Hughes et al. 1996]. In typical bad case, elderly person who doesn't have enough physical strength will cannot operate these motions and will falls into the wheelchair life or bedridden life. Furthermore, if once elderly person falls into such life, the decrease of physical strength will be promoted because he will not use his own physical strength [Hirvensalo et al. 2000].

Therefore, we are developing a rehabilitation walker system with standing and seating assistance device which uses a part of the remaining strength of the patient in order not to reduce their muscular strength. Our system is based on a walker which is popular assistance device for aged person in normal daily life and realizes the standing and seating motion using the support pad which is actuated by novel manipulator with three degrees of freedom.

From opinions of nursing specialists, required functions for daily assistance are (1) the force assistance for standing, (2) the posture assistance for safety and stability condition during standing, walking and seating assistance continuously, (3) the position adjustment assistance especially before seating and (4) the force assistance for seating motion to a target chair. In our previous work, we developed a force assistance scheme which realizes function (1), a patient's posture estimation scheme and support position adjustment system which realize function (2) and walking assistance scheme with indoor navigation system which realize function (3) [Chugo et al. 2009]. Therefore, in next step, for realizing function (4), we develop a seating assistance system which uses remaining physical strength of patient for his rehabilitation.

In previous works, many researchers developed assistance devices for standing motion. However, these devices are specialized in only "standing assistance" and they do not discuss on a seating motion. Some previous researchers say a seating motion is only "reverse" motion of standing [Ehara and Yamamoto 1996]. However, seating motion has high risk for falling down compared with the standing motion for elderly and other condition will be required for a seating assistance. Thus, it is difficult to realize a seating assistance using only "reverse" motion of standing.

In this paper, we develop a control scheme for our robotic walker for realizing force assistance system which realize a seating assistance using the remaining physical strength of the patients. Our key topics are discussion of required condition for seating assistance based on motion analysis and combination of force and position control. Using our control scheme, the patients can sit with fewer loads and can use their own remaining physical strength during this motion.

This paper is organized as follows: we introduce the mechanical design and controller of our system in section 2; we analyze the seating motion in section 3; we propose the new force control scheme in section 4; we show the result of experiments using our prototype in section 5; section 6 is conclusion of this paper.

2 System Configurations

Fig.1 shows overview of our proposed assistance system. Our system consists of a support pad with three degrees of freedom and an active walker system. The support pad is actuated by proposed assistance manipulator mechanism with four parallel linkages. The patient leans on this pad during assistance motion. Fig.2 shows our prototype. Our prototype can lift up the patient of 1.8[m] height and 150[kg] weight maximum.

Fig.2(b) shows our developed support pad based on the opinions of nursing specialists at a welfare event [Chugo et al. 2009]. The support pad consists of the pad with low repulsion cushion and arm holders with handles. In general, a fear of falling forward during motion reduces the ability of elderly person. Using this pad, a patient can maintain his posture easily during seating motion without a fear of falling forward. The pad has force sensors in its body for measuring applied load.

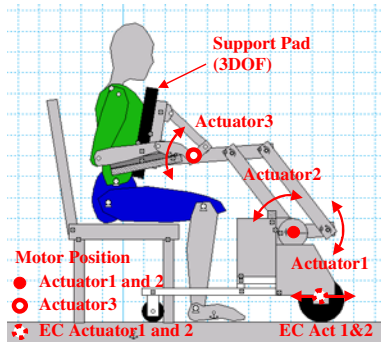


Fig. 1 Overview of our system

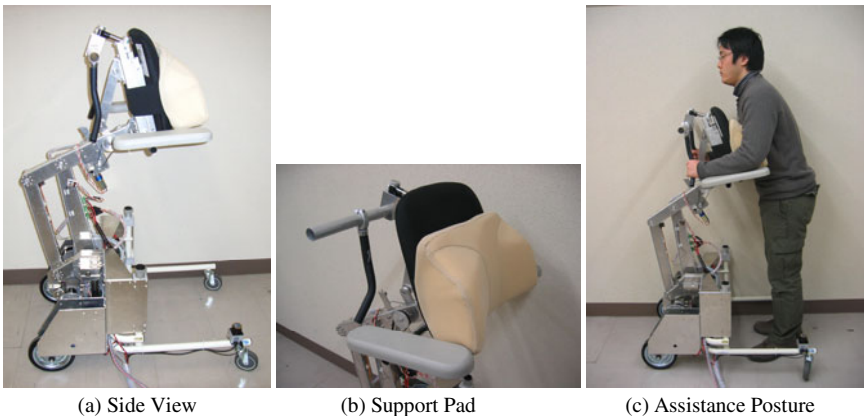


Fig. 2 Our prototype. Its weight is about 35[kg] without batteries. Our prototype requires an external power supply and control PC. (In future works, we will use batteries and built-in controller)

3 Seating Motion

3.1 Difference between Standing and Seating Motion

From previous works, a seating motion is same to “reverse” motion of standing [Ehara and Yamamoto 1996]. Therefore, we assist seating operation with this reverse motion using our prototype in a preliminary experiment. Subjects are 6 young people and 2 elderly people. As the result, all subjects feel fear of falling and a reverse motion seems to be unsuitable for seating assistance. Thus, in this paragraph, we analyze the standing motion and seating motion which nursing specialists recommends.

For analysis, we assume the standing and seating motions are symmetrical and we discuss the motion as movement of the linkages model on 2D plane [Nuzik et al. 1986]. We measure the angular values among the linkages, which reflects the relationship of body segments using motion capture system (Library Inc, GE60). The angular value is derived using the body landmark as shown in Fig.3. Furthermore, we measure the position of the center of gravity (COG) using force plate system (ANIMA Corp., MG-100). Subjects are 6 young healthy people and they operate both motions based on recommended scheme by nursing specialists.

Fig.4 shows the angular value of each joint and Fig.5 shows the position of COG during motion. Fig.4(a) is seating motion and Fig.4(b) is standing motion. (This graph shows the reverse track for easy to analysis.) The movement pattern is derived from (1).

$$\hat{s} = t/t_s \quad (1)$$

where t_s is required time for a seating operation and t is present time.

From these results, in seating motion, subject inclines his trunk and lowers it earlier than in case of standing motion. Furthermore, in seating motion, subject inclines his trunk less than in case of standing motion. These features are same to previous reports [Dubost 2005]. In general, inclining the trunk reduces the load of knee during standing [Dubost 2005], therefore, this means that in seating motion, required power is lower than standing. On the other hand, in seating motion, the position of COG moves less than in case of standing as Fig.5. It moves from foot to hip smoothly in case of seating. This means during seating motion, the body balance is stable and seating motion does not require moving the body dynamically than standing motion.

From these discussions, seating motion does not require large power because moving direction is same to the gravity. We can define seating motion is trunk-moving operation to lower position with stable body balance. Therefore, seating operation is different to standing operation essentially.

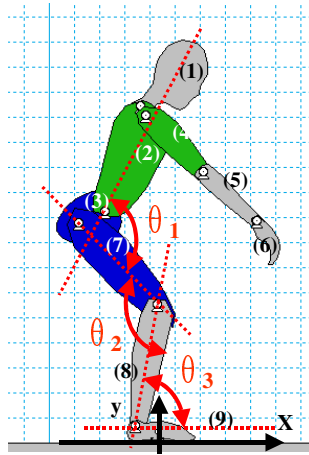


Fig. 3 Human model

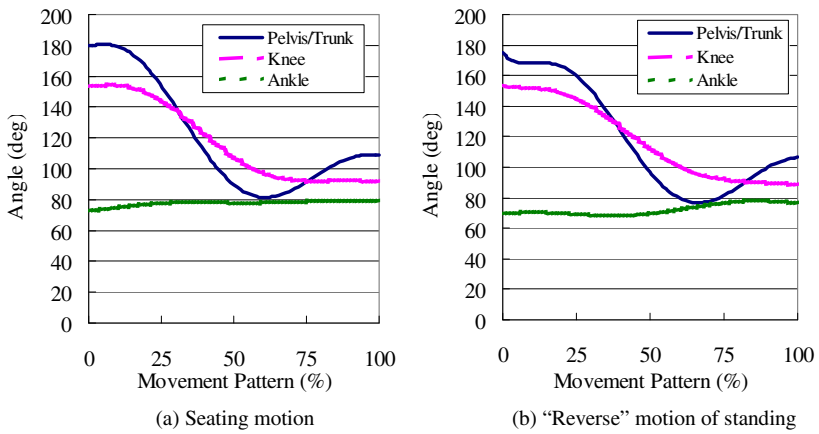


Fig. 4 Angular values of each joint during motion by a young subject

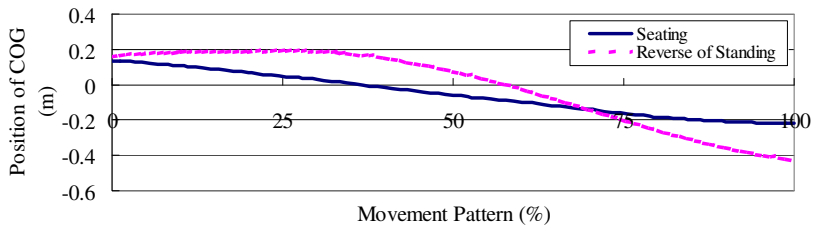


Fig. 5 Position of COG during both motions

3.2 Seating Motion of Elderly

In order to derive the condition for seating assistance, we analyze the difference between the seating motion by healthy young 6 subjects and motion by elderly 5 subjects (One man and four women from 72 to 81 years old, Japanese required care level is 1 or 2).

Fig.6 shows the typical posture of subjects, Fig.7 shows the angular value of each joint and Fig.8 shows the position of COG. From these results, we can find the following points.

- In case of young, the movement of the body is larger than in case of elderly. Especially, a trunk is inclined to forward direction during seating motion to the chair.
- In case of elderly, angular values of knee and ankle angle are same value. Furthermore, COG is fixed on the top of foot from 0 to 75[%] movement pattern. He puts his hip on the chair about 75[%] and this means elderly takes half-sitting posture during seating motion.

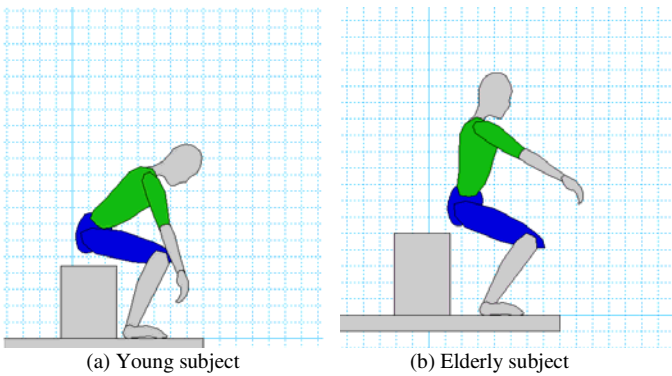


Fig. 6 Seating motion

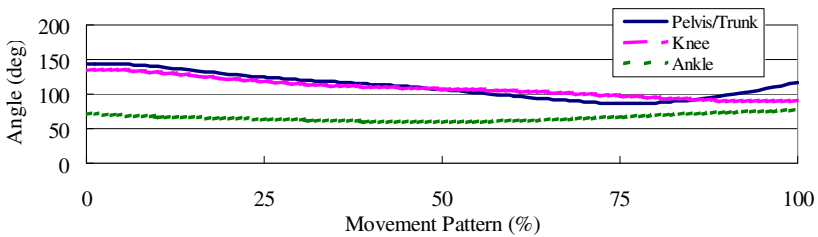


Fig. 7 Angular values of each joint during seating motion by an elderly subject. Please note angle data of a young subject during seating motion is shown in Fig.6 (a)

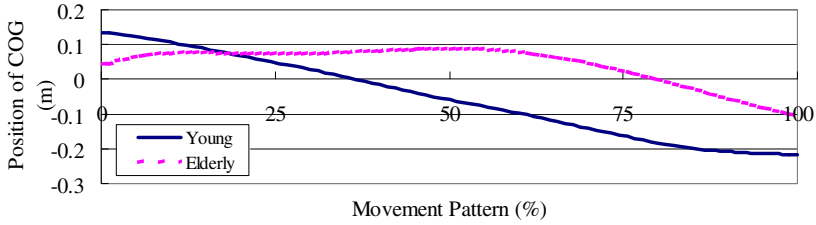


Fig. 8 Position of COG during seating motion

Furthermore, we derive the required traction output of each joint for realizing two seating motions using a computer simulation. Simulation conditions are as follows:

- The human models move each joints as Fig.4(a) (in case of young) and Fig.7 (in case of elderly).
- The parameters are chosen from a standard body data of Japanese adult male [Okada et al. 1996] as shown in Table 1.

Table 1 Human parameters

Number	Link Name	Mass [kg]	Length [m]	Width [m]
1	Head	5.9	0.28	0.21
2	Trunk	27.2	0.48	0.23
3	Hip	18.1	0.23	0.23
4	Humerus	4.5	0.39	0.12
5	Arm	2.7	0.35	0.08
6	Hand	0.5	0.2	0.07
7	Femur	9.1	0.61	0.17
8	Leg	4.5	0.56	0.16
9	Foot	0.8	0.26	0.11

*Numbers of linkage are defined in Fig.3.

From simulation results, Fig.9(a) shows the required traction of each joint in case of young and Fig.9(b) shows it in case of elderly. Table 2 shows the maximum values of traction output and required power through the motion in each case. Comparing with the results of young and elderly, in case of young, required workload for once seating motion is smaller than in case of elderly as shown in Table 2.

On the other hand, in case of young, maximum values of traction output are larger than elderly. Especially, in case of young, knee load becomes heavy from 65 to 85[%] of movement pattern. During this period, the subject puts his hip on

the chair softly and moves his weight from foot to hip. By opinions of nursing specialist, this motion is important and failure of it causes a falling down from the chair. However, knee load of this motion is heavier than 0.5[Nm/kg] and it is difficult for elderly person [19].

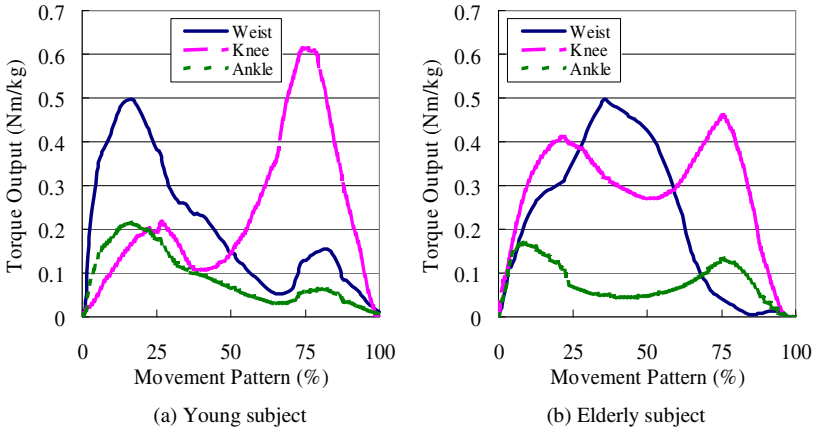


Fig. 9 Traction output of each joint during seating motion

Table 2 Maximum traction and workload of each joint

		Pelvis/Trunk	Knee	Ankle
Young	Max(Nm/kg)	0.50	0.61	0.21
	Output(Ws)	16.3	18.7	7.2
Elderly	Max(Nm/kg)	0.49	0.46	0.17
	Output(Ws)	20.4	26.2	8.1

3.3 Discussion

In previous paragraph, we obtain two findings comparing with seating motion of young and elderly.

One finding is elderly does not incline the trunk. In general, inclining the trunk reduces the load of knee during standing and sitting [Schenkman et al. 1990]. From simulation result (Table 2), we can verify that inclining trunk reduces the workload for once seating motion. However, typical elderly person doesn't incline the trunk during seating. Thus, we ask 5 elderly subjects that why they don't incline their body to forward direction during motion. Their answer is the fear of falling down. They say the balance of the body might be broken if they incline their trunk to forward direction. This means elderly cannot incline the trunk because it is required to maintain stable posture during seating.

The other finding is elderly takes half-sitting posture during seating motion. This posture is easier to keep the balance of the body [Hwang et al. 2003]. Furthermore, from simulation results (Table 2), the maximum traction output of elderly is less than one of young. This means the half-sitting posture requires less peak traction of knee joint and elderly who does not have enough strength can take this posture more easily.

From these discussions, we analyze the seating motion of elderly as follows:

- The seating motion of nursing specialists requires the less workload than in case of elderly motion. Thus, this motion helps the elderly to use own physical strength easily.
- However, it is difficult for elderly to maintain body balance during inclining the trunk.
- Furthermore, in this motion, peak traction of knee joint is larger than one of elderly. Therefore, it is difficult for elderly who has not enough physical strength.
- Thus, elderly takes half-sitting posture during seating motion reluctantly even if the workload for once motion increases. This posture prevents elderly to sit down on his own and may cause the risk of falling down accident.

4 Seating Assistance Control

4.1 Required Condition

From the discussion of previous section, the required condition for seating assistance is as follows:

- The seating motion should be based on the opinions of nursing specialists.
- The assistance system should reduce the load of knee joint when it exceeds threshold ($0.5[\text{Nm/kg}]$) from 65 to 85[%] of movement pattern.
- The assistance system should maintain stable posture of patient. Thus, assistance system is required the following function.
- Position control for maintaining the stable body shape.
- Force control for reducing the load of the patient.

4.2 Proposed Force Control

Realizing these conditions, we propose the novel control scheme as shown in Fig.10. The proposed control scheme combines damping control and position control. The damping control is suitable for the control of the objects with contact. When the required torque of each joint is small enough, the controller uses the position control. On the other hand, when required torque of knee joint is heavy, the controller uses the damping control.

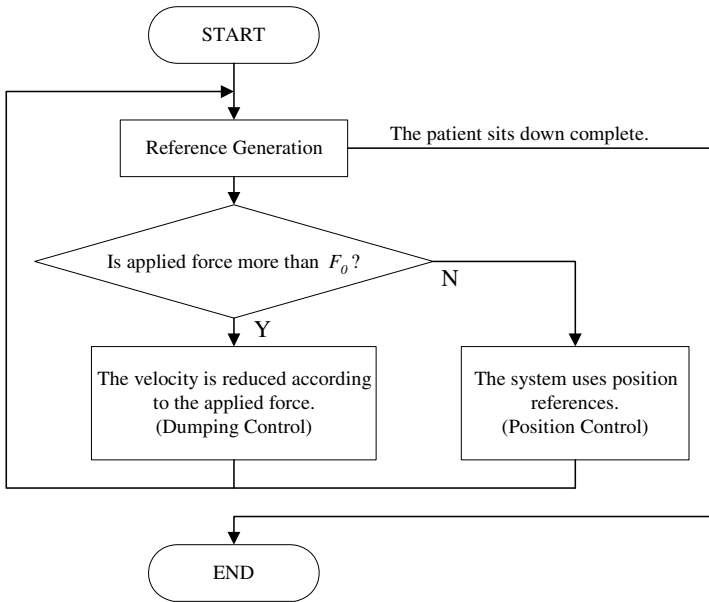


Fig. 10 The flow chart of our control scheme

We use the force sensor attached on the support pad for switching condition between the position control and the damping control. Comparing with the load of each joint and applied force in Fig.15, the applied force to the support pad shows the same tendency to the applied load of knee joint. Therefore, we can divide the situations using the measuring value of the force sensor by a threshold. Using our proposed control scheme, the controller can select more appropriate control method using the force sensor on the support pad.

Now, we explain our proposed control scheme closely. The reference generator derives the velocity control reference of each actuator from the seating motion based on opinions of nursing specialists as reference in Fig.4 (a). (We derive these references in next paragraph.)

$$\mathbf{v}_i^{ref} = [v_i^{ref}(0), \dots, v_i^{ref}(\hat{s}), \dots, v_i^{ref}(1)]^T \quad (2)$$

where $i (= 1, 2, 3)$ is identification number of actuators. v_i^{ref} is control reference (The coordination is defined as Fig.11(b)) and it is function of the movement pattern \hat{s} as (1).

The output of each actuator is derived from (3).

$$v_i = v_i^{ref} - B(F - F_0) - K(x_i - x_i^{ref}) \quad (3)$$

where $B = 0$ (if $F < F_0$), $K = 0$ (if $F > F_0$)

where F is the applied force on the support pad (Fig.11) and F_0 is the threshold which selects force or position control. v_i^{ref} is the velocity reference and x_i^{ref} is the position reference derived from track references as shown in Fig.12. v_i is the updated reference which our system uses actually during the assistance motion. B and K are coefficients.

4.3 Reference Derivation

In this paragraph, we derive the control reference of our assistance system which can realize the seating motion using a computer simulation. Fig.11 shows the simulation setup. Simulation conditions are as follows.

- The human model puts his forearm on the supporter.
- The human model leans on the pad using his arm with own enough force.
- Other conditions are same to section 3b.

From the simulation results, Fig.12(a) shows the position tracks of support pad and Fig.12(b) shows its angle tracks. In Fig.12(b), Y-axis shows the inclination angle of the support pad and X-axis shows the movement pattern \hat{s} .

In Fig.12(a), the start point is upper center and the end point is lower left. Using these tracks as the position control reference, our assistance system can realize the seating motion of nursing specialists.

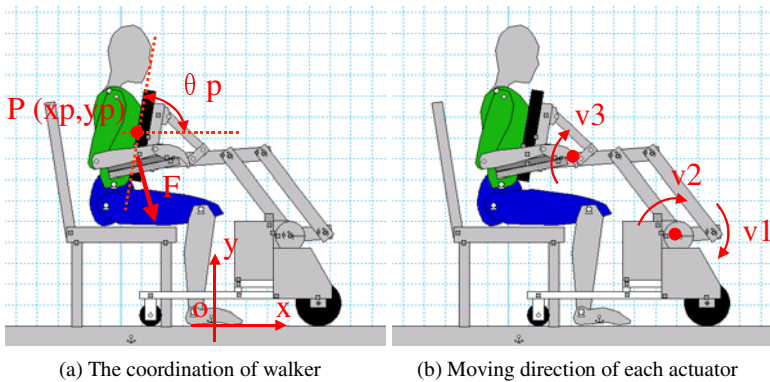


Fig. 11 Simulation setup

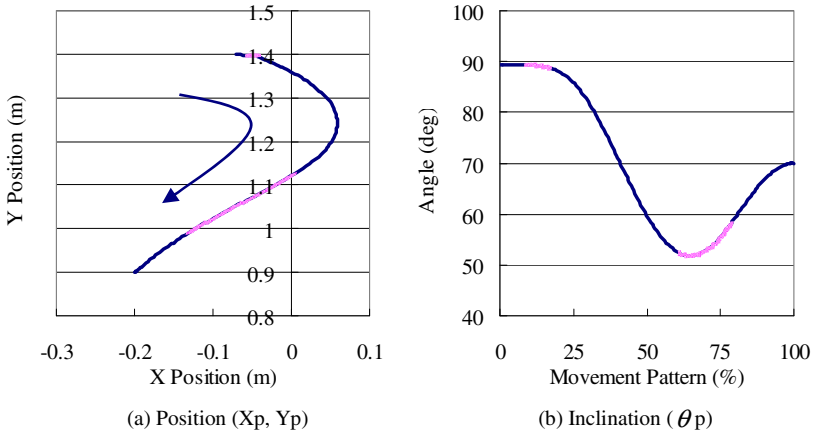


Fig. 12 Derived control references Red line shows the force control mode. (In case that the threshold is 15(kgf). We discussed in section 4b.)

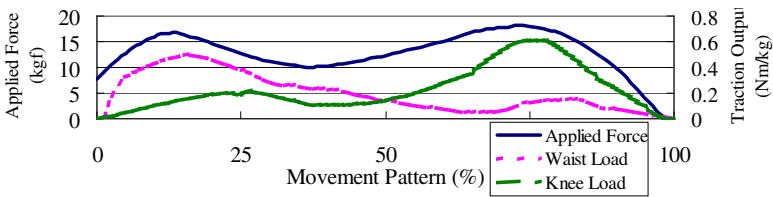


Fig. 13 Applied load and load of knee and waist joint during seating motion

5 Experiments

Here, we verify the performance of our prototype system by the experiment. In this experiment, subjects use the special wearing equipment for the experience of the elderly [Takeda et al. 2001]. This wear limits the motion of the tester body as elderly.

In this experiment, we test three cases. The first case, subjects sit down with our proposed assistance. We set $F_0 = 150[N]$ as a threshold which is derived experimentally for our proposed controller. The second case, subjects sit with only position assistance and the third case, subjects sit with only force assistance.

As the result of the experiment, our system can assist the patient as shown in Fig.14. The height of the patient is 1.7[m] and the system assists him at 30[sec]. Fig.15 shows the tracks of angular value of the patient’s waist and knee joint, and their control references. From Fig.15, both tracks are almost same line and this means our assistance system realizes the natural seating motion by nursing specialist.

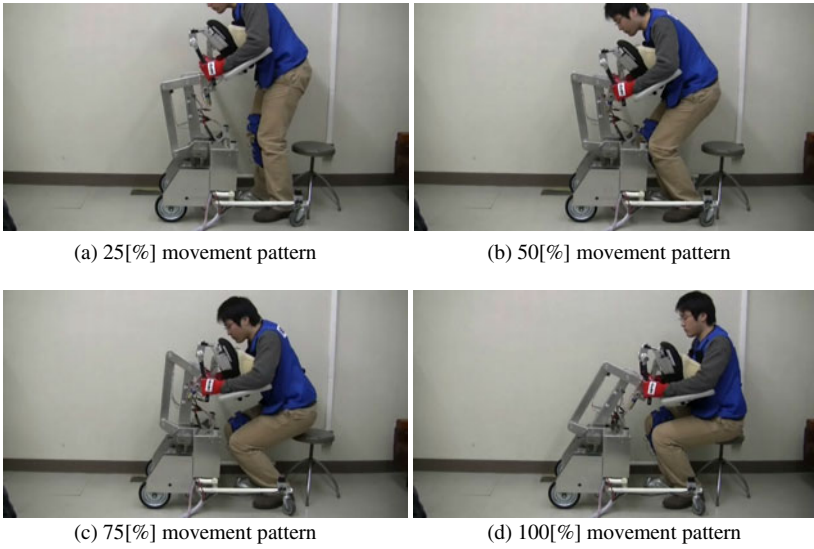


Fig. 14 Seating motion with our proposed assistance

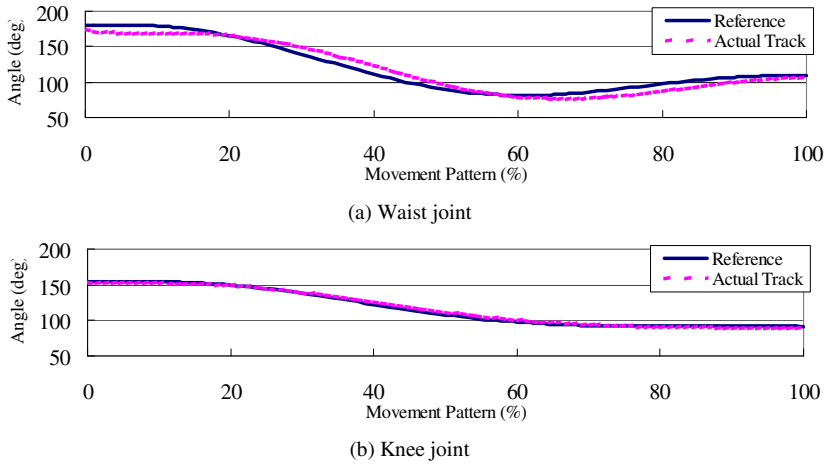


Fig. 15 Angular values of each joint during seating motion

Fig.16 shows the applied force to the support pad during seating motion. During red arrowed range, the applied force exceeds the threshold. Thus, the system exceeds force control mode and the applied force to the support pad increases.

Table 3 shows the workload of each actuator for once seating operation. From these results, required workload using our proposed control is only 112[%] comparing with the workload using position control mode. On the other hand, force control mode requires 147[%] workload. If we assume the total required workload (the patient’s workload and the assistance system’s workload) for once operation is constant, this result means that our proposed control scheme requires the patient to use own physical strength more than the case of the force control mode.

Fig.17 shows the required traction output during seating motion derived by a computer simulation based on the experimental result. From these results, we can verify that the required peak traction is reduced. On the other hand, general required traction is maintained and this means the patient has to use own strength when he sits down with our assistance.

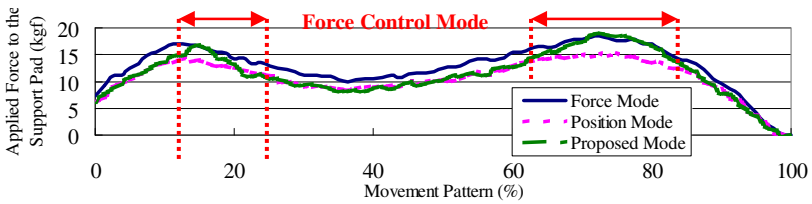
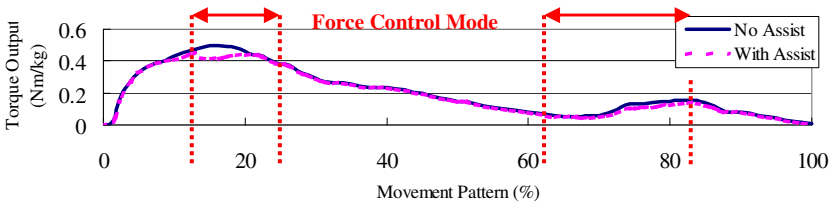
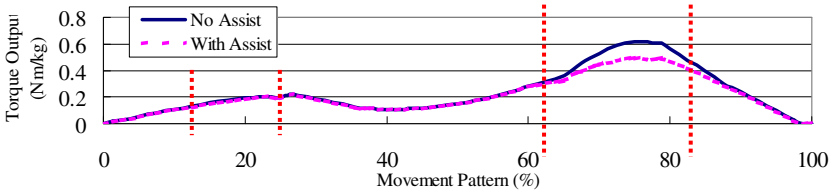


Fig. 16 Applied load to the support pad during seating motion



(a) Waist joint



(b) Knee joint

Fig. 17 Traction output of each joint during seating motion

Table 3 Workload of our system (Ws)

	Position	Proposed	Force
ACT1	68.7	77.7 (113.1%)	100.3 (146.0%)
ACT2	69.2	76.5 (110.5%)	100.8 (145.7%)
ACT3	57.5	65.3 (113.6%)	85.2 (148.2%)
Total	195.4	219.5 (112.3%)	286.3 (146.5%)

* Values in parentheses are ratio comparing with standard mode

6 Conclusion

In this paper, we develop the control scheme for our robotic walker for realizing force assistance system which realize a natural seating motion using the remaining physical strength of the patients for their rehabilitation. In order to fulfill required condition, we analyze a seating motion which is recommended by nursing specialists and typical seating motion by elderly. Based on the analysis, we propose novel control scheme, which combines force and position control.

In our future work, we will implement the individual parameter optimization scheme for our prototype.

References

- [Chugo et al. 2009] Chugo, D., Asawa, T., Kitamura, T., Songmin, J., Takase, K.: A motion control of a robotic walker for continuous assistance during standing, walking and seating motion. In: Proc of 2009 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, pp. 4487–4492 (2009)
- [Dubost et al. 2005] Dubost, V., Beauchet, O., Manckoundia, P.: Decreased trunk angular displacement during sitting down: an early feature of aging. *Physical Therapy* 85(5), 404–412 (2005)
- [Ehara and Yamamoto 1996] Ehara, Y., Yamamoto, S.: Analysis of standing up motion, pp. 65–73. Ishiyaku Pub. Inc. (1996)
- [Fisher et al. 1990] Fisher, N.M., Pendergast, D.R., Calkins, E.C.: Maximal isometric torque of knee extension as a function of muscle length in subjects of advancing age. *Arch. Phys. Med. Rehabil.* 71(10), 729–734 (1990)
- [Hirvensalo et al. 2000] Hirvensalo, M., Rantanen, T., Heikkinen, E.: Mobility difficulties and physical activity as predictors of morality and loss of independence in the community-living older population. *J. Am. Geriatric Society* 48, 493–498 (2000)
- [Hughes et al. 1996] Hughes, M.A., Schenkman, M.L.: Chair rise strategy in the functionally impaired elderly. *J. of Rehabilitation Research and Development* 33(4), 409–412 (1996)
- [Hwang et al. 2003] Hwang, Y., Inohira, E., Konno, A., Uchiyama, M.: An order n dynamic simulator for a humanoid robot with a virtual spring-damper contact model. In: Proc. of the IEEE Int. Conf. on Robotics and Automation, pp. 31–36 (2003)

- [Nuzik et al. 1986] Nuzik, S., Lamb, R., Vansant, A., Hirt, S.: Sit-to-stand movement pattern, a kinematic study. *Physical Therapy* 66(11), 1708–1713 (1986)
- [Okada et al. 1996] Okada, H., Ae, M., Fujii, N., Morioka, Y.: Body segment inertia properties of japanese elderly. *Biomechanisms* 13, 125–139 (1996)
- [Takeda et al. 2001] Takeda, K., Kanemitsu, Y., Futoyu, Y.: Understanding the problem of the elderly through a simulation experience – difference in the effect between before and after clinical practice. *Kawasaki Medical Welfare J.* 11(1), 64–73 (2001)

Pedestrian Navigation System for Indoor and Outdoor Environments

M. Popa

Department of Computer and Software Engineering, Faculty of Automation and Computers, POLITEHNICA University of Timisoara, Romania
mircea.popa@ac.upt.ro

Abstract. Mobile navigation uses smart mobile devices, such as PDAs, mobile phones or dedicated devices. In order to help the user to navigate, they receive updated data through wireless Internet, GPS, cellular phones and specialized sensors. Car navigation systems are the most known and advanced implementation of data navigation systems. They are embedded in cars or in distinct devices. Pedestrian navigation systems were less approached. This paper presents a PNS for finding a car. It was thought to be included in the car's key and to guide the user from its starting point, which is in delimited area around the car, to his/her car.

1 Introduction

Wireless communications is in a continuous and accentuated development. It covers many application areas, such as Mobile communication services, Location based services and Mobile navigation. Mobile navigation uses smart mobile devices, such as PDAs, mobile phones or dedicated devices. In order to help the user to navigate, they receive updated data through wireless Internet, GPS, cellular phones and specialized sensors.

Car navigation systems are the most known and advanced implementation of data navigation systems. They are embedded in cars or in distinct devices.

Pedestrian navigation systems (PNSs) were less approached than car navigation systems. There are some targeted achievements, such as PNSs for visiting large delimited areas (a museum or an institutional building) or PNS for monitoring the sportsmen during their training. But the most challenging task is to develop PNSs for guiding people in different areas, outdoor or/and indoor, especially in metropolitan areas. The main requirement for a PNS is, as for car navigation systems, to guide a user from a starting point to a destination point. However, PNSs have also to consider the differences between the two implementations: cars can move only on a predefined infrastructure while pedestrians have a higher degree of freedom. There are also differences in data

received: car navigation systems receive a lot of information about street networks from many countries but this information cannot be used by PNS as it is. Pedestrian navigation systems can be used for tourist reasons or/and for finding a target, for example the own car previously left in a certain place

This paper presents a PNS for finding a car. It was thought to be included in the car's key and to guide the user from its starting point, which is in delimited area around the car, to his/her car.

Section 2 presents related work, section 3 describes the proposed PNS, section 4 shows experimental results and the last section outlines the conclusions

2 Related Works

In [Beeharee and Steed 2006] an exploratory study of a guiding system that uses photographs is proposed. The photographs are extracted from existing geo-tagged photo collections from mobile phones. A user of the system sees a route description as text and a map that refers to a series of photographs. The experiment shows that presenting the right photographs helps particular types of routing instructions for users not familiar with an area.

A new pedestrian wayfinding model that addresses the problem of movement of the pedestrian under specific conditions is presented in [Gaisbauer and Frank 2008]. A graph model is created which consists of decision points and edges connecting them. The free walkable space around decision points is expanded to decision scenes where pedestrian movement is modeled in more detail, thus allowing for flexible navigation comparable to unassisted pedestrians.

In [Stark et al. 2007] the starting point is the fact that although many pedestrian navigation systems occurred, they use vocabulary and routing only appropriate to car navigation. The paper describes a field study comparing four navigational concepts for pedestrians currently available. These are: Auditory instructions plus digital, dynamic route (Audio method), Digital, dynamic route (Route method), Map with position and direction (Direction method) and Textual description by street names (Description method).

[Godha and Lachapelle 2008] presents a system based on a low-cost inertial measurement unit and high-sensitivity global positioning system receivers for personal navigation in an environment where GPS signals are degrading. The system is mounted on a pedestrian shoe and uses measurements based on the dynamics experienced by the inertial sensors on the user's foot.

[Cho and Park. 2006] describes a micro-electrical mechanical system based pedestrian navigation system. The PNS consists of a biaxial accelerometer and a biaxial magnetic compass mounted on a shoe. It detects a step using a novel

technique during the stance phase and simultaneously calculates walking information. Step length is estimated using a neural network whose inputs are the walking information. The performance is verified by experiment.

In [Miyazaki and Kamiya 2006] a pedestrian navigation system that delivers photorealistic panoramic landscape images using 3D models of a city and related information to mobile phones was described. The requirements for smooth mobile phone navigation are presented (low processing load and quick data delivery) and a server-side panoramic image generation mechanism and a divided guidance information transfer technique.

A near-complete Pocket PC implementation of a Mobile Multi-Modal Interaction (M3I) platform for pedestrian navigation is described in [Wasinger et al. 2003]. The platform easily supports indoor and outdoor navigation and uses the combination of several modalities for presenting output and user input (e.g. 2D/3D graphics, synthesized speech, gesture recognition).

3 Implementation of the Proposed PNS

The role of the PNS is to show to the user which is the way from his/her current position to his/her car previously left in a place, for example in a garage. The system is encapsulated in the car's key, without affecting the existing hardware and software and uses the existent LCD. The PNS has to be independent from the possible already existent car navigation system. The disadvantage is that the PNS has to be active when the user leaves its car in order to memorize the car's position and for that, the user must stay near the car until the PNS succeeds to obtain the current position. Dependent on the environment, this operation may last from several seconds to several minutes. The alternative would be that the system is started automatically when the car is locked and its position is saved when the user is moving away from the car. The disadvantage would be the lower accuracy of the GPS readings. The hardware of the PNS is made of a GPS receiver and a magnetic sensor. Because of the mobility requirement the minimization of the energy consume must be an important target. The software of the PNS consists of the new navigation facilities. Fig. 1 shows the sequence of operations.

3.1 Description of the Hardware

Typically, a PNS may lose rapidly the GPS connectivity due to the medium in which the user moves (tall buildings, indoor environments etc.). That is why the system has to possess a supplementary navigation system, totally different from GPS. It has to diminish the errors introduced by the GPS and to compensate the possible losses of GPS connectivity.

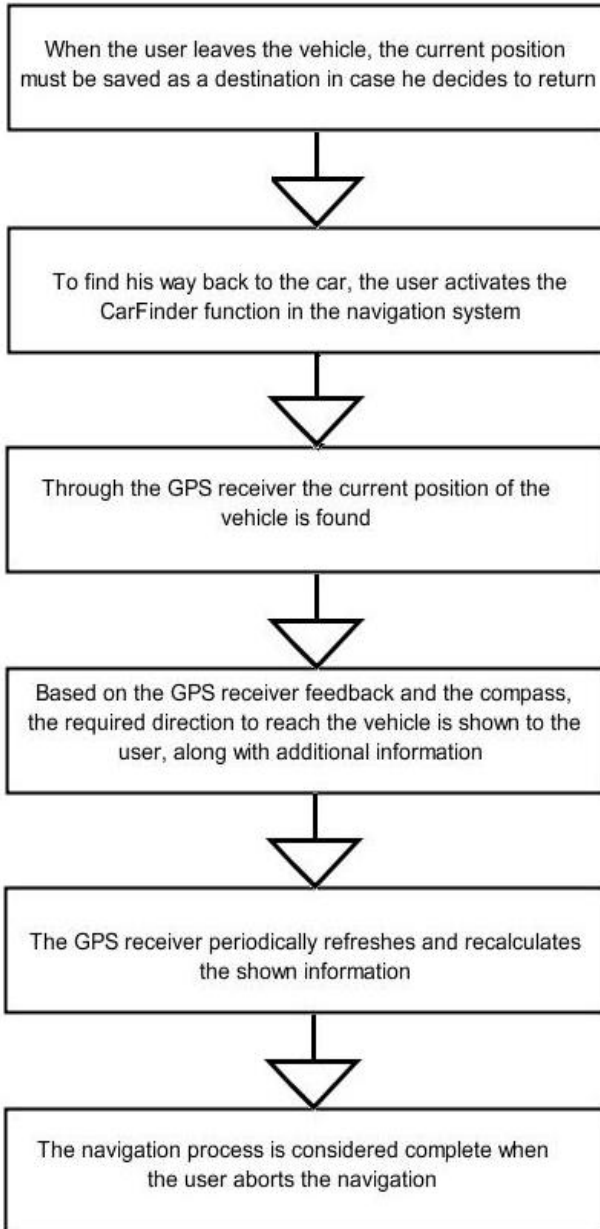


Fig. 1 Function of the PNS

A magnetic sensor was used in the proposed PNS. It indicates the current orientation of the user in the N-E-S-W axes system. For that, the magnetic sensor reads the values of the magnetic field of the Earth on the X, Y, Z axes. These readings are given to a microcontroller which transforms them in a bidimensional representation of the position of the user and guides him to the destination. Fig. 2 presents the block diagram of the hardware of the car's access key including the hardware of the PNS.

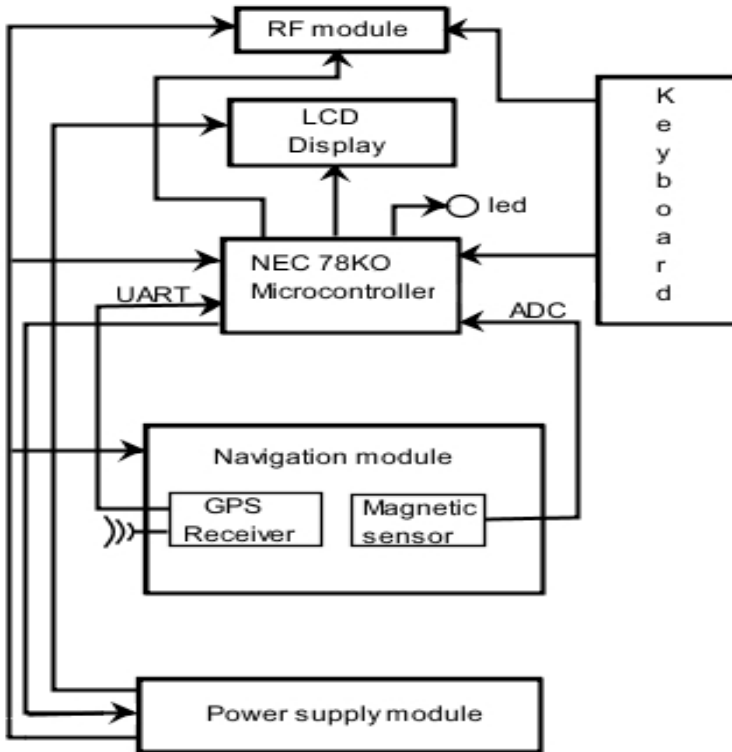


Fig. 2 Block diagram of the hardware and of the PNS

The existing hardware is made of a RF module for wirelessly commanding the car, a keyboard for opening/ closing the doors and for alarming functions, a displaying module consisting in a LCD display and a led and a battery based power supply module. All these modules are commanded by an 8 bit NEC78K0 microcontroller.

The Navigation module was added through a 9 pin connector. It consists of the uBlox NEO-4S GPS receiver and the Honeywell HMC6042 magnetic sensor. The GPS receiver is connected to the UART6 interface of another NEC78K0 microcontroller and the outputs of the magnetic sensor are connected to the ADC inputs of the microcontroller. The communication between the GPS receiver and the

microcontroller is a classical serial one, with the following parameters: 8 data bits, 1 STOP bit, without parity, without control flow, 9600 bps, NMEA protocol (which imposes a certain structure of the message).

The main features of the microcontroller are: 32 8 bit Special Function Registers, internal Flash memory, watchdog timer, 41 standard input/ output ports, 2 temporizers/ counters for counting internal clock pulses or external events, several serial interfaces: 2 UARTs, CSI and I²C, 8 channel analog/ digital converter with 10 bit resolution, power supply: 1.8 – 5.5 V.

3.2 Description of the Software

From a functional point of view the software of the PNS is made by the following modules:

- Magnetic Sensor Driver: for communicating with the magnetic sensor;
- GPS Receiver Driver: for communicating with the GPS receiver;
- Computing Module: for processing the direction and the distance until the car;
- LCD Computing Module: for displaying the direction and the distance;
- Application: for implementing the navigation-specific and user-specific interactions.

Fig. 3 presents the block diagram of the software.

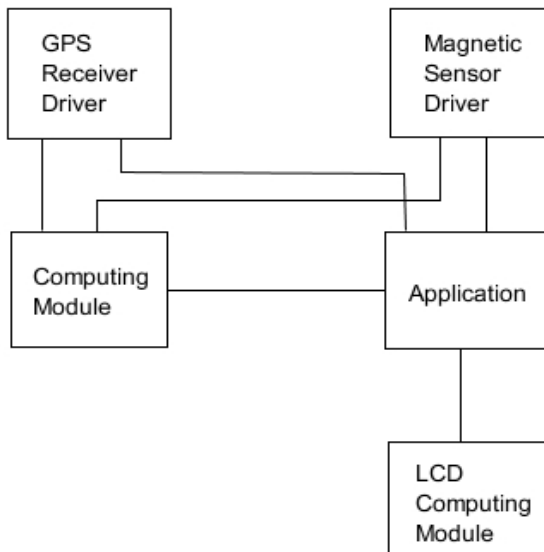


Fig. 3 Block diagram of the software

The Magnetic Sensor Driver ensures the communication with the HMC6042 magnetic sensor. The operations are:

- initialization of the communication with the GPS receiver, by configuring the microcontroller's ADC;
- connecting/disconnecting the magnetic sensor to/from the power supply;
- providing the values of the intensity of the magnetic field obtained by converting the readings of the magnetic sensor.

There are several functions implementing the mentioned tasks. For example, the function `dcompass_init` configures the microcontroller's UART6 interface. The pins 2 and 3 of the port P2 have to be configured in order to be inputs for two analogical channels, the pin 7 of the port P1 is configured as output for implementing the function Set/Reset and the pin 1 of the port P12 is configured as output. The code is:

```
unsigned int *conversionResultX, *conversionResultY;
void dcompass_init(unsigned int *x, unsigned int *y){
    conversionResultX=x;
    conversionResultY=y;
    PM2_bit.no2=1;
    PM2_bit.no3=1;
    PM1_bit.no5=0;
    PM1_bit.no7=0;
    P1_7=0;
    P1_7=1;
    PM12_bit.no1=0;
}
```

The GPS Receiver Driver is responsible with the communication with the uBlox NEO-4S GPS receiver. The tasks it has to ensure are:

- initialization of the communication with the GPS receiver by configuring the microcontroller's asynchronous serial interface;
- activation of the GPS receiver: it means the wake-up of the receiver in order to provide the positioning information;
- reception of the information from the GPS receiver, its verification and its transmission to the upper software layers;
- disconnecting the GPS receiver by putting it in its low-power mode.

The `dgps` component implements the above mentioned tasks. The `dgps_init` function initializes the microcontroller's UART6 serial interface, the `dgps_startRX` and `dgps_stopRX` functions starts, respectively stops, the reception, the `dgps_startTX` and `dgps_stopTX` functions starts, respectively

stops, the transmission, the functions `dgps_enable` and `dgps_disable` functions enables, respectively disables, the GPS receiver, the function `dgps_byteReceived` implements the mechanism for receiving the messages from the GPS receiver, the function `dgps_sendByte` transmits a string of characters to the GPS receiver and the `dgps_error` function treats an error occurred in communication. Below, a sample of the code is shown:

```
void dgps_enable(void) {
    P12_2=1
}
void dgps_disable(void) {
    P12_2=0
}
```

The Computing Module has the role to obtain the direction and the distance to the destination.

The direction is calculated from: the azimuth (it is the orientation within the magnetic poles of the Earth) and the bearing (it gives the movement sense and is the angle against the north between the line which connects the current position and the car's position). The direction is given by the so called heading value. This is obtained according with the algorithm from fig. 4.

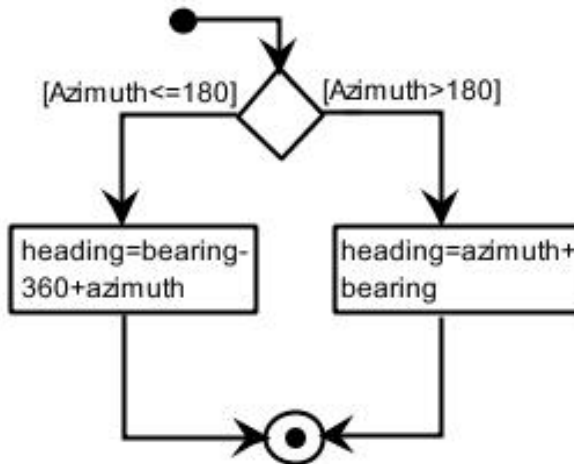


Fig. 4 Algorithm for the heading value

The function `gpsCalc_getHeading` uses the values received from the functions `gpsCalc_getAzimuth` and `gpsCalc_getBearing`. The azimuth and bearing values are calculated by using the functions `gpsCalc_convertLatitude` and `gpsCalc_convertLongitude` which convert in degrees the values received from the GPS.

The distance is calculated using the coordinates of the destination, $lat2$ and $long2$, and the coordinates of the source, $lat1$ and $long1$. There are several approaches for calculating the distance.

The distance can be calculated with the Pythagoras's theorem. Because it supposes the surface of the Earth is completely flat, it can be used only for short distances. The formula is:

$$D = R\sqrt{(\Delta lat)^2 + (\Delta long)^2} \quad (1)$$

where $\Delta lat = lat2 - lat1$ and $\Delta long = long2 - long1$ are expressed in radians and $R = 6371$ Km.

The method introduces errors for distances lower than 20 km. According to [WWW-1 2007], they are: 30 m if the latitude is lower than 700 degrees, 20 m if the latitude is lower than 500 degrees and 9 m if the latitude is lower than 300 degrees.

Another approach consists in using the Great Circle Distance method, [WWW-1 2007], which calculates the distance between two points considering the Earth as a sphere. The distance is not measured as a strait line but along circles whose centers coincide with the center of the sphere. Any two points from the sphere are positioned on a single circle having the same center as the sphere's center, except for two opposite points. The two points divides the circle in two arcs. The shortest arc represents the distance among them. The method has low accuracy for short distances. The formula is:

$$D = R * \arccos[\sin(lat2)\sin(lat1) + \cos(lat2)\cos(lat1)\cos(\Delta long)]$$

The third solution is based on the Haversine formula. It has good accuracy for long distances and for short distances too. The relationships are:

$$a = \sin^2\left(\frac{\Delta lat}{2}\right) + \cos(lat2)\cos(lat1)\sin^2\left(\frac{\Delta long}{2}\right) \quad (2)$$

$$c = 2a \tan 2(\sqrt{a}, \sqrt{1-a}) \quad (3)$$

$$D = R * c \quad (4)$$

The LCD Computing Module displays the direction and the distance. Through the LCD Computing Module, the current position provided by the GPS receiver is saved and the navigation is started or stopped. The direction is shown by an arrow on the LCD display, from 16 possible positions meaning a 22.5 degree resolution.

The application module implements the interaction between the user and the PNS. Two types of interrogations are implemented:

- navigation-specific interactions: savings of the current position and navigation information towards the car;

- user-specific interactions: interrogation of the GPS driver for obtaining the current position, interrogation of the magnetic sensor driver for obtaining the current values of the intensity of the magnetic field, interrogation of the computing module for obtaining the direction and the distance to the car and displaying the navigation information.

The navigation task is the most complete one. It uses most of the functions from this module. After activating the GPS receiver and the magnetic sensor, information from the GPS receiver is read. The latitude and longitude of the current position and accuracy are extracted. The accuracy is measured through the HDOP parameter. The coordinates of the current position are saved as coordinates of the destination. If the information from the GPS receiver has a low accuracy data from the magnetic sensor are read and used. If the information from the GPS receiver misses the distance is calculated using the previous position. The distance, direction and signal level are displayed. The signal level is given by the accuracy expressed through the HDOP parameter. Thus, a value 1 or 2 from HDOP means maximum accuracy, a value 2-4 means 50% accuracy, a value 4-6 means 25% accuracy while a value >6 mean a low accuracy.

4 Experimental Results

The tests were done for verifying the good functioning of the system and of the navigation algorithm. Several tests were done with different visibility conditions for the GPS.

The configuration of the system was: the uBlox NEO-4S GPS receiver with the ANN-MS antenna and the HMC6042 magnetic sensor connected to a NEC78K0 microcontroller. The data collected by the microcontroller are sent to a PC through an asynchronous serial interface. All the computations necessary for navigation are done on the PC.

The purpose of the tests is to verify the accuracy of the data provided by the GPS receiver and to verify the validity of the algorithm for computing the direction and the distance. The starting strategy of the GPS is Cold start.

Place of the experiments: the local town. A real route and a calculated one were obtained using the indications of the pedestrian system. The destination point has the following real coordinates: 45.740051^0 latitude and 21.238951^0 longitudes.

Support: in order to show the routes, photos from GoogleMaps were used. The photos from GoogleMaps have a certain rate of update (usual once/year) but this is not a disadvantage here because their role is only to serve as a support for showing the obtained routes.

The first set of experiments was done in an area with lower visibility for the GPS receiver. Fig. 5 presents the real route and the reconstituted one, based on the

information given by the GPS receiver. Due to the loss of visibility there are differences between the two routes. Fig. 6 presents the variation of the distance to the destination and fig. 7 shows the variation of the HDOP (Horizontal Dilution of Precision), received from the GPS receiver, parameter which gives the precision with which the GPS receiver has made the calculus. There are important variations of the precision in calculating the position.

The calculus of the distance and of the bearing value by using the data from the tests was verified. For that, if the coordinates of the destination and of the current position are known one can calculate the distance and the bearing value. If the values for distance, bearing and coordinates of the current position are known, one can calculate the coordinates of the destination point using the following formulas:

$$lat2 = \arcsin(\sin(lat1) * \cos\left(\frac{D}{R}\right) + \cos(lat1) * \sin\left(\frac{D}{R}\right) * \cos(Bearing)) \quad (5)$$

$$long2 = long1 + \arctg2(\sin(Bearing) * \sin\left(\frac{D}{R}\right) * \cos(lat1), \quad (6)$$

$$\cos\left(\frac{D}{R}\right) - \sin(lat1) * \sin(lat2))$$



Fig. 5 Real and reconstituted routes for the first set of experiments

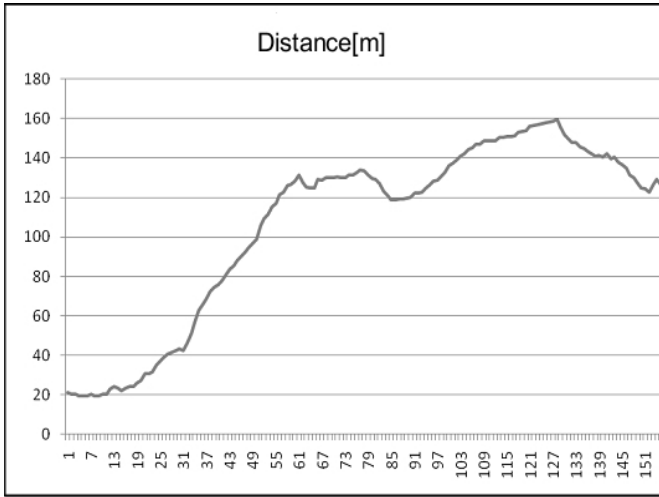


Fig. 6 Variation of the distance to the destination for the first set of experiments

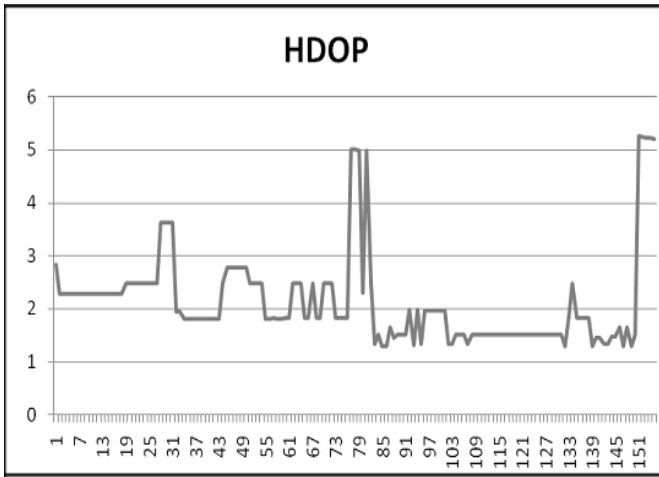


Fig. 7 Variation of HDOP for the first set of experiments

The second set of experiments was done in conditions of very low visibility for the GPS receiver with an area with better visibility. The time needed for Cold start was 150 sec. Fig. 8 presents the real route and the reconstituted one, fig. 9 presents the variation of the distance to the destination and fig. 10 presents the variation of the HDOP parameter. There are significant differences between the real route and the reconstituted one and high errors. The differences and the errors diminish considerably when the GPS reaches the zone with better visibility.



Fig. 8 Real and reconstituted routes for the second set of experiments

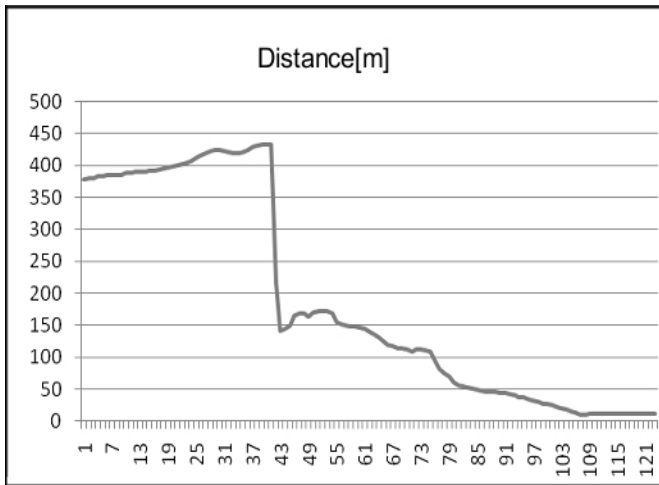


Fig. 9 Variations of the distance to the destination for the second set of experiments

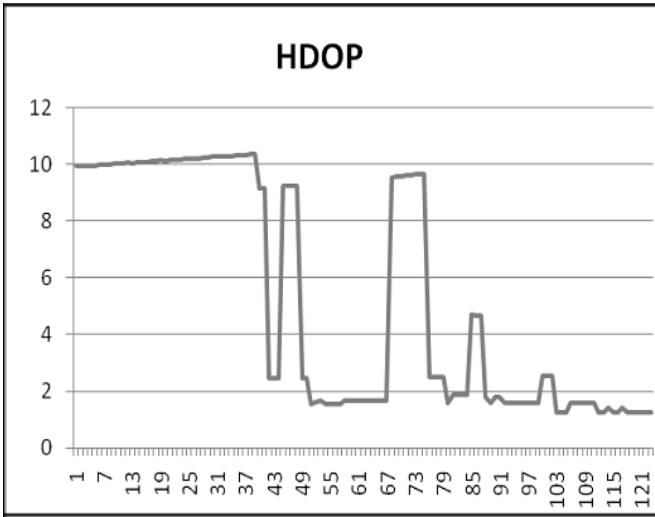


Fig. 10 Variation of HDOP for the second set of experiments

In the third set of experiments the visibility for the GPS receiver was high. The starting time was only 50 sec. Fig. 11 presents the real route and the reconstituted one, fig. 12 presents the variation of the distance to the destination and fig. 13 presents the variation of the HDOP parameter. One can observe small differences between the routes and rather small variations of the HDOP parameter.

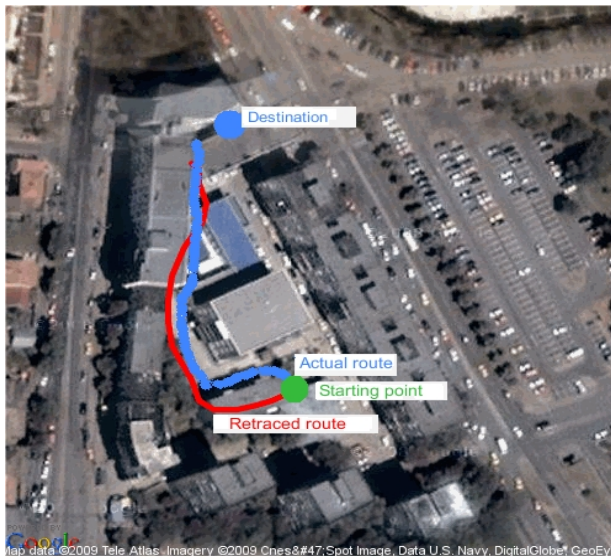


Fig. 11 Real and reconstituted routes for the third set of experiments

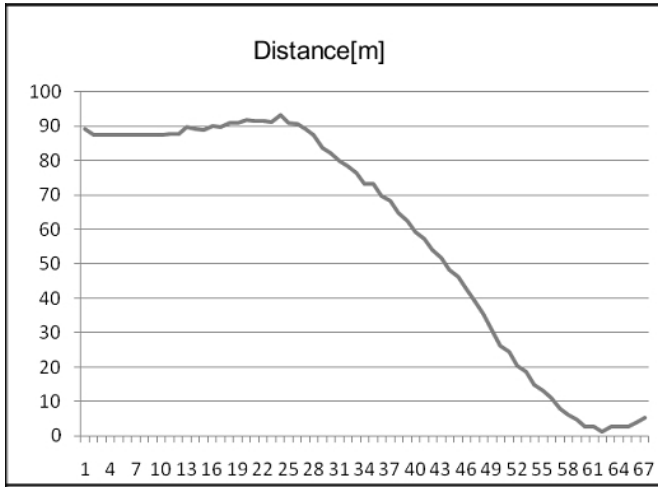


Fig. 12 Variations of the distance to the destination for the third set of experiments

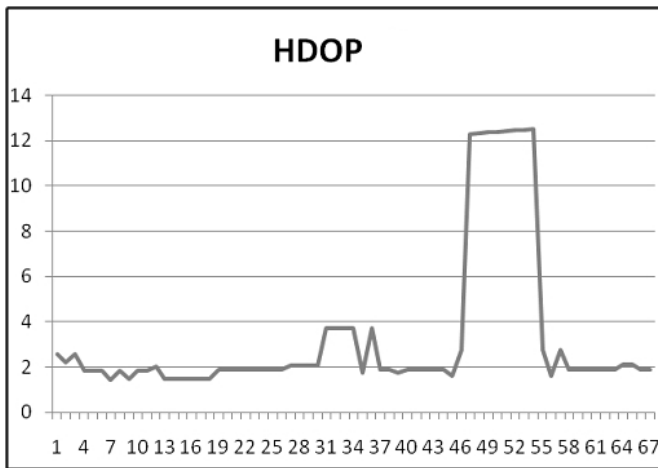


Fig. 13 Variations of HDOP for the third set of experiments

5 Discussion and Conclusions

The paper has described a pedestrian navigation system for indoor and outdoor environments. Its role is to help the user to find his/her car previously left in a certain place. The system is added to the existing hardware and software found in the car’s access key.

The research can be continued in the following directions:

- improvement of the tracking algorithms; in the current implementation the routes are constructed each time the system is activated and the improvement consists in creating a database with the routes the user has followed and to try to use the saved routes, avoiding the time and the energy needed to construct new ones;
- replacement of the magnetic sensor because of its low sensitivity which leads to often calibration; by replacing it with a more sensitive one the calibration operations will be minimized.

References

- [Beeharee and Steed 2006] Beeharee, A.K., Steed, A.: A natural wayfinding exploiting photos in pedestrian navigation systems. In: ACM International Conference Proceedings Series, Proc of the 8th Conf on Human-Computer Interaction with Mobile Devices and Services, vol. 159, pp. 81–88. Helsinki, Finland (2006)
- [Cho and Park. 2006] Cho, S.Y., Park, C.G.: MEMS based pedestrian navigation system. *J. of Navigation* 59 (2006)
- [Gaisbauer and Frank 2008] Gaisbauer, C., Frank, A.U.: Wayfinding model for pedestrian navigation. In: Proc. of 11th AGILE Int. Conf. on Geographic Information Science, University of Girona, Spain (2008)
- [Godha and Lachapelle 2008] Godha, S., Lachapelle, G.: Foot mounted inertial system for pedestrian navigation. *Measurement Science and Technology Journal* (7) (2008)
- [Miyazaki and Kamiya 2006] Miyazaki, Y., Kamiya, T.: Pedestrian navigation system for mobile phones using panoramic landscape images. In: Proc. of the 2006 Int. Symp. on Applications and the Internet, Phoenix, Arizona (2006)
- [Stark et al. 2007] Stark, A., Riebeck, M., Kawalek, J.: How to design an advanced pedestrian navigation system: Field trial results. In: Proc. of IEEE Int. Work on Intelligent Data Acquisition and Advanced Computing Systems: technology and Applications, Dortmund, Germany, pp. 690–694 (2007)
- [Wasinger et al. 2003] Wasinger, R., Stahl, C., Kruger, A.: *Mobile HCI 2003*. LNCS, vol. 2795, pp. 481–485. Springer, Heidelberg (2003)
- [WWW-1 2007] Calculate distance, bearing and more between two Latitude/Longitude points, <http://www.movable-type.co.uk/scripts/latlong.html/> (accessed May 2009)

Model Based Processing of Swabbing Movements on Touch Screens to Improve Accuracy and Efficacy for Information Input of Individuals Suffering from Kinetic Tremor

A. Mertens¹, C. Wacharamanotham², J. Hurtmanns^{1,2}, M. Kronenbuerger³, P.H. Kraus⁴, A. Hoffmann⁴, C. Schlick¹ and J. Borchers²

¹Institute of Industrial Engineering and Ergonomics,

RWTH Aachen University, Germany

{a.mertens, j.hurtmanns, c.schlick}@iaw.rwth-aachen.de

²Media Computing Group, RWTH Aachen University, Germany

{chat, borchers}@cs.rwth-aachen.de

³Department of Neurology, RWTH Aachen University,

University Hospital Aachen, Germany

mkronenbuerger@ukaachen.de

⁴Department of Neurology, Ruhr-University Bochum, St. Josef-Hospital, Germany

{peter.h.kraus, arndt.hoffmann}@ruhr-uni-bochum.de

Abstract. As a result of demographic change the average age of many western populations increases, accompanied with age-related disease patterns. Especially tremor symptoms rise accordingly, aggravating a barrier free interaction with information systems. In order to maintain a self determined lifestyle at home, new technologies and methods need to be introduced, especially for application in health care and telemedical scenarios. Hence, a new direct input technique based on wiping movements on touch screens has been developed. The combination of a new input concept and applying regular commercially available technologies helps to avoid high costs for acquisition and therefore makes it marketable. While making an input on the touch screen the precise characteristics of every wiping movement can be tracked and is used for computation of the desired entry. The efficacy of this approach was evaluated within a clinical study with n=15 subjects. The results show that the error ratio for inputs by tremor patients can be significantly reduced in comparison to a virtual keyboard, depending on tremor strength and form. The learning curve for first time users is very steep and tends to result in inputs that are only slightly steady than purposeful movements to standard buttons and keys.

1 Introduction

1.1 Motivation

Strong tremor symptoms increase the error rate of human-computer interaction and may even lead to a total refusal or inability to utilize an IT-based system for communicating individual needs. The currently available systems do not service people suffering from kinetic tremor. Those tremor patients often suffer from a level of high inaccuracy and show a worsening of precision when moving towards a virtual or real button, due to the nature of intention tremor. This inability abates efficiency, effectiveness and satisfaction of the user and even causes social isolation due to the hindered maintaining and establishing of contacts [Martínez-Martín 1998]. For this determined target group a new method of interaction has been developed and tested. Creating a sufficient interaction for people with a tremor agitation using the existing interface hardware is the key to several identified barriers occurring with the current demographic change and augments the potentials for developers. Optimizing current technology will help to maintain a personal health care system and retain the trust in it. Creating an interaction which allows user to effectively use their tremor encumbrance as part of their interaction process is highly worthwhile.

The correlation of efficiency and satisfaction for the elderly during computer based interaction has already been proven. As user acceptance, among elderly, towards new technologies is considered mediocre at best, it will help to quickly establish a user-interface relationship if the user recognizes his natural motion pattern as a required input.

1.2 Tremor in Older Adults

The word tremor is derived from Latin *tremere*, meaning „to tremble“. Tremor can be defined as involuntary oscillations of any part of the body around any plane, such oscillations being either regular or irregular in rate and amplitude and resulting from alternate or synchronous actions of groups of muscles and their antagonists. Arms are the part of the body that is most commonly affected. Tremors can happen irrespective of age but show a tendency to be prevalent in older people. About 1-4% of the total population is concerned while in the population group older than 50 years up to 14.5 are afflicted [Wenning et al. 2005]. Tremors are usually classified according to their phenomenology, most commonly “present at rest” or “present with action” (i.e. posture or movement). Occurrence and intensity can differ highly for each person affected based on physical and mental state of the day and medication. The frequency varies depending on etiologies with a range between 3 and 30 Hz. The most common action tremor is the so called “Essential Tremor”.

1.3 Classification of Tremor Strength

The quantification of motor performance of people afflicted with tremor is usually assisted with help of clinical rating scales and instrumental approaches that allow estimating the efficacy of applied therapy. While manual rating scales often are not sufficient objective the usage and appraisal of instrumented methods is mostly restricted to hospitals and medical specialists. Therefore a novel approach based on automated rating tremor strength with help of spiral drawings is applied within this study (see chapter 4.3 for details). This instrument allows a time and location independent application by non-physicians and guarantees high reliability because of the standardized interpretation by a computer-algorithm. Focus is just on quantification of tremor amplitude, regardless of clinical diagnosis and independent of tremor genesis.

1.4 Requirements for Information Input by People Suffering from Tremor

To allow a satisfying and effective interaction of people suffering from tremor with IT-systems, especially the information input has to be focused as prevalent user interfaces do not respect the specific needs and limitations. To get an overview of the different tremor pathologies and the manifold effects they have on activities of daily living, field observations and expert interviews were accomplished. Result was the deeper insight that the impact very much depends on living conditions, adequate assistance and symptomatology which can be seen in the exemplarily handwritings presented in Figure 1.

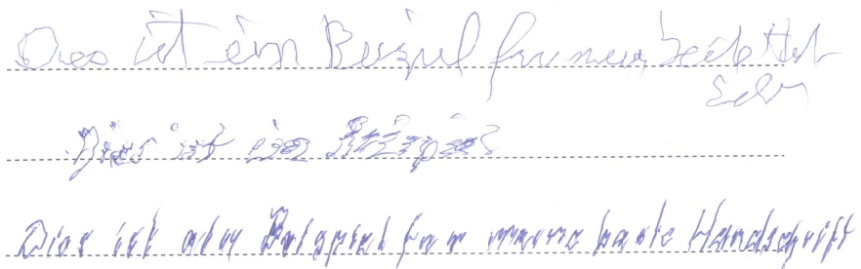


Fig. 1 Exemplary handwritings of individuals suffering from different tremor types

Workshops with the people affected by tremor, medical specialists and medical technicians resulted in several requirements that have to be satisfied by potential manual input concepts to be adequate utilities:

- Input must be identified as one gesture even when the contact while interacting with the device is lost for a short time because of jerks in z-axis
- Fault insertions because of unintentional contacts and laying down the palm have to be filtered out
- The input concept must be adaptive to the varying tremor strength depending on the form of the day as well as medication and not be depending on a calibration
- Different phenomenologies of tremor must be considered including tremor amplification while resting and with action
- Recognition of selections for non linear input patterns (e.g. staggered or sinusous actions).

1.5 Related Research

There has been a rapid development in the domain of human-computer interfaces over the past two decades. Among them are technical innovations and enhancements as motion and eye tracking devices, voice recognition and brain-computer interfaces allowing a new dimension of information input. A possible application of the devices includes everyday support for physically impaired persons. However, there are several limitations to the use of these technologies with regard to mobility, costs, privacy protection, noise immunity, training time and required calibration. These criteria constrain a broad distribution and allow feasibility only for very specific scenarios. A wide use in particular for the growing number of elderly people with limited motor skills is not realistic in the medium term.

Therefore research on the budding application of touch screens as input medium for tremor patients was accomplished. This technology has evolved from a niche product to the quasi standard input of smartphones and can already be found in e.g. ready-made computer systems, ATMs, route guidance systems and ticket machines. It turned out to be a relative untouched research field from an ergonomic point of view and with regard to the identified requirements. Research has mostly been done for medical questions concerning the rating and categorizing of tremor symptoms for different treatments and therapies.

To have a valid foundation for the development of an advanced input technique and to not reinvent the wheel, an in detail literature review was accomplished. The following section provides an overview of some “classic” approaches and concepts towards input techniques with touch screens that were appraised.

Buxton’s three state model

Buxton (1990) distinguishes three states of input: (0) out of range, (1) tracking and (2) dragging. Illustrated with mouse usage, state zero is when the mouse is not in contact with any surface (i.e. moving the mouse has no input effect); state one refers to the mouse being moved on a surface (i.e. the cursor moves according to mouse movements); finally, when selecting a target (i.e. pressing a button), the

target can be moved around the screen. However, applying this model to touch screens results in complete omission of stage 1; instead, the system jumps directly from state zero to state two. The reduction to only two states brings about advantages in terms of input time, although not necessarily reducing error rates.

In addition to Buxton's three stages, Potter et al. (1988) divide target acquisitions on touch screens into three different strategies: Land on, First contact and Take-off. Input methods utilizing the Land on strategy make their selection based on the first initial contact with the screen; all further contacts are ignored. First contact strategies are similar to Land on strategies, "but take advantage of the continuous stream of touch data. This means that the first target which is hit will also be selected.

Take-off input strategies utilize a cursor, which is moved by dragging the finger on the screen (as in First contact) but the selection is only made when the finger is removed from the screen.

Offset Cursor (First contact)

In order to avoid target occlusion by the finger, Potter et al. designed an alternative input method called Offset Cursor. When the finger is placed on the display, the selection is not made at the point of contact; instead a cursor appears above the finger and allows for precise selection of targets by dragging towards the wished location.

As much as it improves the problem of target occlusion, the main disadvantage of Offset Cursor is that the selection is always performed above the finger, making targets at the bottom area of the screen impossible to select.

Direct Touch (Land on)

Direct Touch is the most basic and straightforward input for touch screens. The principle of this technique is very intuitive: the user selects a target on the screen by directly tapping on the location of the target. In terms of selection time, Direct Touch has proved to be superior as opposed to traditional mouse. On the other hand, when the display, and accordingly the target, is small, Direct Touch results in higher selection errors due to the occlusion of the target by the finger.

Another disadvantage of Direct Touch is the limited accessibility of targets that are located close to each other: the finger occludes the targets and the width of the finger does not allow for accurate input. Further, targets that are located at the border of the screen are difficult to select as well.

TapTap (Land on)

TapTap by Roudaut et al. (2008)] is a technique that is very effective in selecting targets on small tactile displays, e.g. handhelds. The basic idea is that a first tap in the area of the target brings about an enlarged popup of this very area in the middle of the screen, where subsequently the selection is made. TapTap is an especially advanced and effective technique; unlike the Direct Touch input method, TapTap does not that much suffer from thumb occlusion because the user lifts his

finger off the screen before making the final selection. In addition to that all areas of the display are equally well accessible.

However, due to the fact that a selection is based on two contacts with the surface, this technique is not as fast as others regarding input time. (Not much information is provided on how to cancel an unwanted selection or whether selecting a target at the edge of the screen interferes with scrolling.)

MagStick (Take-off)

Another technique proposed by Roudaut et al. is MagStick. MagStick works in the following way: the user touches the screen in the area of the target he wants to select and then drags his finger in the opposite direction of the target (mirror principle).

The targets work as “magnets” so that the curser automatically jumps to the predefined targets as soon as the finger is moved in the opposite direction. A thin line indicates which target is selected at any given moment and the finger does not occlude the target at any point. Also the selection of targets that are located at the borders of the screen is possible without any restrictions.

Shift (Land on)

To address the problem of occlusion in bare finger operated touch screen inputs, the Shift method has been proposed as one solution. When targets are small and therefore occluded by the fingers, the occluded area is duplicated and projected to a free region of the screen. When targets are big enough that occlusion is not a problem, this technique is not applied and the display remains unaltered.

However Shift cannot be used when targets are located at the edge of the screen. Studies show that Shift results in faster input times and lower error rates than similar techniques, e.g. the Offset Cursor.

Escape (Take-off)

Yatani et al. [Yatani et al. 2008] present a target selection technique for mobile displays. The main advantage of Escape is that the user only needs to select a point near the target and then, by moving the finger into the desired direction, can select the target.

In essence, Escape works in the same manner as MagStick. This has proven itself to be extremely useful in selection of small targets. Yatano et al. compare the Escape technique to its alternative Shift and finds that for small targets (size between 6 and 12 pixels) selection is on average 30% faster while there is no difference in error rates.

Synopsis

The review of different input techniques regarding touch screens revealed that, although there are several approaches that try to compensate for the disadvantage of

others (e.g. thumb occlusion, difficult or impossible to reach areas), none of the discussed methods is promising for elderly people suffering from kinetic tremor or physical impairments of the neuro-musculoskeletal system. The lack of this adaptation justifies the need for the development of an enhanced technique that addresses the specific requirements of this target group.

1.6 Solution Concept

Design Pattern: SWABBING

Problem: Theoretically the problem occurring with a kinetic tremor, namely the inaccurate input, may be handled by simply increasing the size made available on the input area (paper, touch screen, etc...). This would compensate the expected deviation caused by the tremor. However, this method has clear limitations when it comes to stronger tremor deviations and limitation of input space. This leads to either the limitation of options concurrently displayed or the reduction of button size on the screen, in order to maintain the amount of choices offered. This will almost certainly increase the error ratio when handling input made with distinct tremor symptoms and resulting deviation.

Solution: In order to enable correct and independent input for the previously described target group (with smartphones, PC, telemedical systems, ticketing machines – each with a touch screen) the user user interface will be virtually enlarged by using the same area for information input/output and not restricting the input movement to the screen surface. The principle behind the enlargement is Fitt's Law. The width of the target, measured along the axis of motion on the screen, is not restricted through the screen dimensions, characterized in that the user can perform a continuous input movement beyond the borders.

This basically means that all variations or deviations are included and give the user a “free hand” to perform his input. The electronic tracking of the input appears only on the touch screen, but vital data for reproducing the movement, as direction, orientation, velocity and starting point, are collected. This very close approximation of the user movement helps to allocate the desired input from the user much more reliably than the ordinary point input method. Here only the last phase of the input movement – the contact with the touch screen – is considered.

Furthermore, precision is also increased through an increased friction on the screen surface and through this generated dumping effect, physically reducing tremor deviation. The chosen touch screen may either serve as an input only device or also as output source, meaning the input application will only be displayed on the screen when interaction is required.

2 Algorithmic Implementation

The process of one swabbing input has been identified as a three step movement. Initiated by the first “touchdown event”, every other touchdown, within certain limits, will be included in the collection of the input, due to the nature of tremor deviation. The third step is the disengagement of the user’s finger from the touch screen.

Aggregating the collected data of the whole input pattern (direction, velocity, starting point and drift) an approximation for the volitional input of the user can be accomplished. The target item will be identified regardless of whether the input finger actually directly hits the desired input location or misses it.

2.1 Regression Analysis

A swabbing motion is noted as a selecting action when a specific Euclidian distance is reached. We choose a value of 250 pixels (5.25 mm) for our user study because it is the half of the average of distance from the center of the screen (starting point) to the target (shortest distance: screen edge, longest distance: screen corner). Therefore if a touch point is more than 250 pixels away from the first touch point of the swapping motion a regression line can be calculated which runs through the desired input. This means that all collected touch points are used to calculate a line using linear regression. To increase the numerical stability of this method we decided to use every point several times for the calculation. After that the two intersection points of the line and the area lines are identified. By transforming the line into a vector the selected symbol is found.

To prevent false touches – like other unintentional touches on the screen during a motion that are caused by jerks or extreme tremor activities – a distance per time function is used that checks if a touch point is only a specific distance away from the last coordinates regarding this time slot. During early pretests was determined that tremulous users lifted their finger for a short amount of time during their swabbing motions because of their tremor. As a consequence the user had to start the whole motion again. To ensure that short and unintentional lifts of the finger are not recognized as a touch up event, a watchdog is used. If a user lifts his/her finger from the screen the watchdog is started. If a touchdown event occurs the watchdog is stopped and reset (Fig. 2).

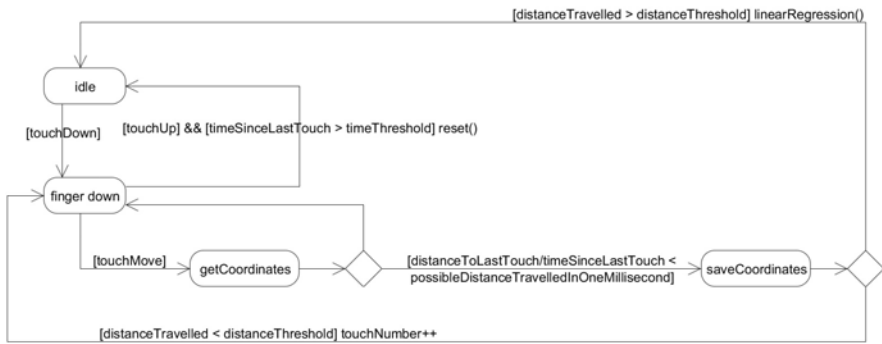


Fig. 2 Activity diagram of the algorithmic processing logic

2.2 Visual Representation and Feedback

The visualization of the areas is done by arranging them circular on the screen. There are two rationales for the layout: (1) it solves the problem of overshooting a target and (2) it maximize opening angle for a given number of targets and screen space. To ensure an equal distribution this is done by rotating two points throughout the screen. By connecting this points lines are formed which are extended to the screen border. These lines are used to assist the user by giving them a visualization of an areas corresponding “corridor”. To prevent users from having the feeling of crossing barriers while they make a swabbing motion, the “corridor lines” are dashed and a circle in the middle of the screen is left completely free. The points where these “corridor lines” hit the border of the task area define the boundary points of the “area line” of each area, which is needed for the calculation. To indicate which area should be activated next and whether the selection of an area was right or wrong we use polygons in the form of arrows. These arrows are filled with colors as follows:

- Blue: The area that is corresponding to the arrow should be selected.
- Green Blinking: Selection of the right area.
- Red Blinking: Selection of the wrong area.

3 Methods

3.1 Study Design

Testing the generated swabbing input method will face the test person with the following setup: A standard multi touch notebook is placed in front of the user, using a holding frame which enables an angular positioning (20° from desk surface)

to suit the test persons needs relative to the desk height [Müller-Tomfelde et al. 2008]. The probands chair will be adjusted so that the table height is approximately similar to the elbow height while the arm points towards the ground. On the screen the test person will encounter highlighted items he or she will have to select in order to perform the task (Fig. 3, left). The test person is asked to perform an “input” and will receive a visual feedback if the input was correctly performed. The interaction area is a square of 800 x 800 pixels (164 mm each side), and the participants have to rest the test finger on a crosshair at the same side of the hand used after each input.

The items are arranged circular (tapping & swabbing) or in grid layout (tapping) on the screen (Fig. 4). The trials are accomplished with rising resolution starting with 9 items, 16 items and finally 25 items. Each condition is repeated for 10 trials resulting in total 90 trials. The user starts by parking the finger on the crosshair before either tapping or moving the finger to the center of the screen to swab.

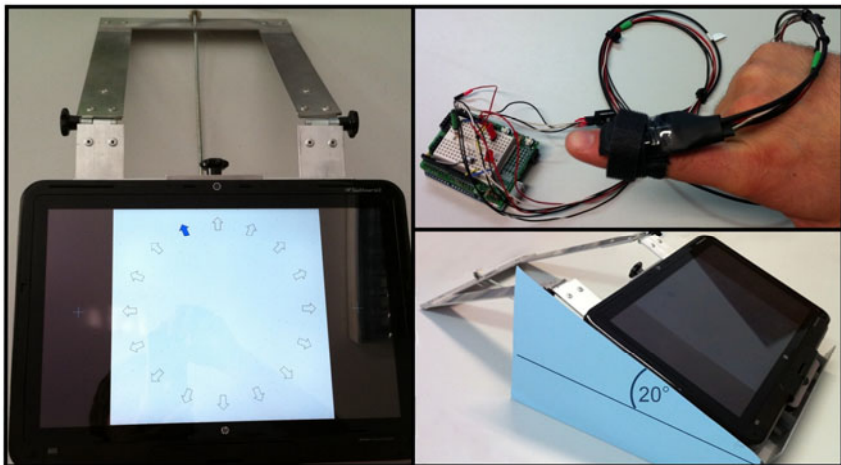


Fig. 3 Employed hardware and setup for evaluation

The screen manipulation will be done solely by the finger, no other input tool, e.g. stylus, was used. For introduction and to minimize the influence of learning effects the probands got a short hands-on demonstration for each of the three different layouts with a maximum of ten inputs per layout. Subsequently an accelerometer and gyroscope were attached to the input finger (Fig. 3, top-right) to measure oscillation for different interaction techniques as previously described by Graham [Graham 2000].

As a reference to the swabbing movements, the probands had to perform typical tapping movements (holding, resting and press-release finger moves). We decided not to provide any visual feedback to get the basic finger movement data.

Demographic factors, satisfaction about usability, cognitive and physical load as well as computer literacy were determined posterior with the help of an interview-administered questionnaire.

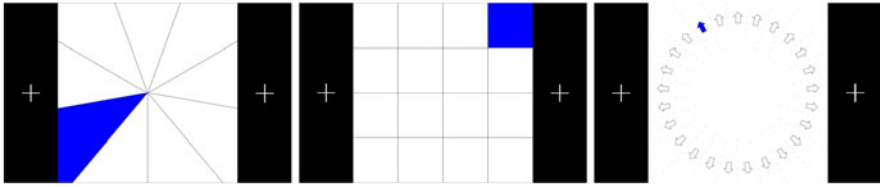


Fig. 4 Different layouts for the selection tasks with rising resolution (from left to right) radial-tapping, 9 items; grid-tapping, 16 items; swabbing, 25 items

To guarantee comparability of the results a high consistency during each individual test was conducted. A standardized test protocol helped to achieve this (Fig. 5). To prevent learning effects, we counterbalanced the order of test patterns with even-size Latin Square.

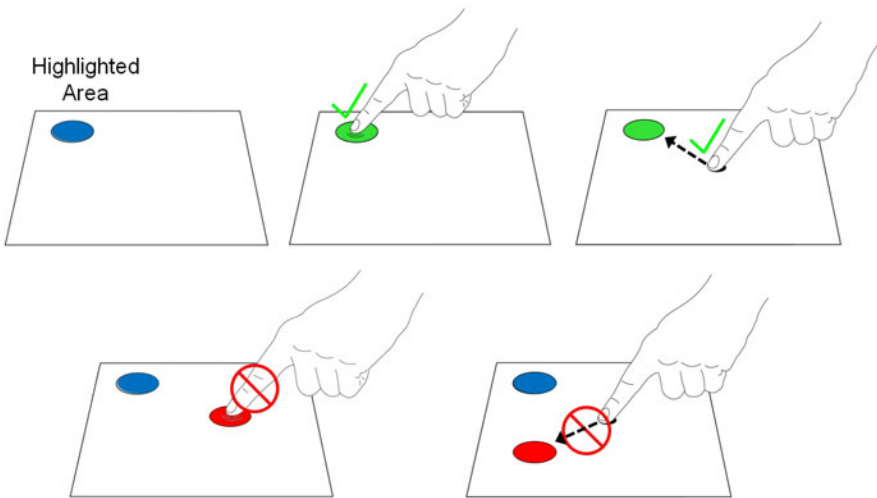


Fig. 5 Schematic illustration for introduction to the different input metaphors

3.2 Apparatus

The hardware platform is a HP TouchSmart tm2-1090eg with a 12.1 inch capacitive multi-touch screen (1280 x 800 px). The notebook was converted into tablet

mode and fixed into a customized stand. No keyboard was apparent for the participant. To measure the tremor during the interaction, we attached a tri-axis accelerometer (GForce3D-3) and a gyroscope (InvenSense IGT-3200) on the backside of the extreme joint (distal phalanges) of the test finger by a Velcro ring (Fig. 1). When the finger rests on the screen, the accelerometer's Z axis is orthogonal to the screen's plane, and X and Y axes are parallel to the respective screen axes. The equipment leaves the entire tip of the finger uncovered.

The sensors are connected with an Arduino Duemilanove which feeds the acceleration data directly to the Universal Serial Bus controller of the notebook. The data is associated with touch signals from the screen. We have made sure that the cable does not prevent the user from freely moving the hand, the arm, or any fingers.

3.3 Spiralometry: Graphimetric Classification of Tremor Strength [Kraus and Hoffmann 2010]

In order to adequately test tremor behavior of the target group, a new computerized assessment method has been chosen: spiralometry. As also recommended by the Movement Disorder Society (MDS) [Deuschl et al. 1998], drawing spirals supports the quantitative subjective evaluation of tremor amplitude. The computerized assessment of these spirals represents a blind and standardized metric measurement which is independent from subjective judgment from the investigator as well as examination's time and location. The clear guidelines of this method make it very objective and the easy setup of the experiment enables a swift process of the test person. The investigator only needs a paper and a pencil during the test, which makes the implementation of the test very feasible. The test person is asked to draw two spirals, one with each hand (Fig. 6). It is essential that the test person does not rest his writing hand and the corresponding forearm on the table. Each drawing will then be scanned by a standard scanning hardware, measuring the amplitude of each spiral in millimeters. One of important disadvantage of the paper-pencil-version is certainly the loss of time as basic information for determination of frequency. Although the pristine use of spiral drawing analysis was the control of therapy effects, it now is a useful design for capturing the current tremor amplitude of the participants when conducting the user study.

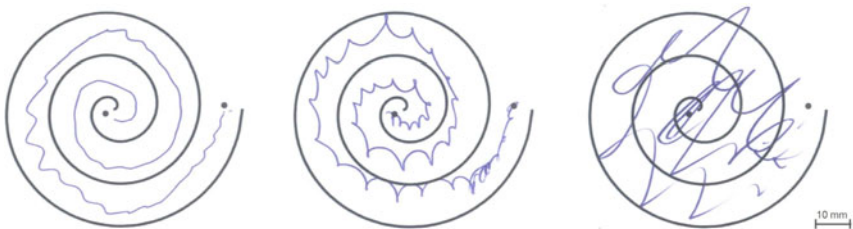


Fig. 6 Exemplarily spirals from persons with different tremor strengths

3.4 Participants

15 clinically diagnosed tremor patients were recruited from the Department of Neurosurgery at the University Hospital of the Aachen University (age: min=56, q1=67, med=75, mean=73,56, q3=78, max=83).

The participants had not used touch screens before this study. All participants used the index finger of their dominant hand in the experiment (Table 1).

Table 1 Exemplarily tremor profiles of some participants and interaction stages that tremor is most intense (see: section 5.3), axes of tremor, and effect of swabbing that lessens/worsens the tremor compared to press-release. (The axes are ordered by tremor strength from high to low. Missing values means insignificant effect.)

#	Tremor strength	Dominant hand	Gender	Most intense stage	Axes	Lessen	Worsen
1	severe > 2 cm	R	M	Rest	Y, Z, X		Z, Y, X
2	slight < 0.5 cm	R	M	Over	Y	X, Z	
3	marked 1–2 cm	R	M	Over	Y, Z	X, Z	
4	moderate 0.5–1 cm	L	F	Over	X, Z	X, Z	
5	severe > 2 cm	R	M	Over	Y, Z	X, Z	Y

4 Results

Immediate, several observations without further scientific evaluation lead to the following findings: Persistent contact with the touch screen already significantly reduced finger oscillation relative to the reference test of tapping movements for most tremor types. As also previously found by Schneider et al [Schneider et al. 2008], a touch screen serves as a very decent tool of interaction with elderly and those who have only few experiences with computer technology. This is an indication of touch screen technology being an ergonomic tool for interaction for the target group.

4.1 Attitudes towards Computers

Based on the works of Gina et al. [Gina and Sherry 1992] the attitudes of elderly people suffering from tremor towards computers were evaluated a priori. The 15 strictly positive formulated statements were rated according to a 4-point Likert Scale (Fig. 7). The correlation between the user's attitude towards a system and his /her effectivity and satisfaction is unquestioned in the field of working environments. Our results with the focus on personal assistance and aids for disabled people affirm the coherence in terms of relation between acceptance of assistive technology and the specific training time and success rate.

These parameters are of special importance when applying the swabbing-technique in (tele)medical scenarios, as first contact is usually created within the conditions of a medical necessity, rather than voluntarily.

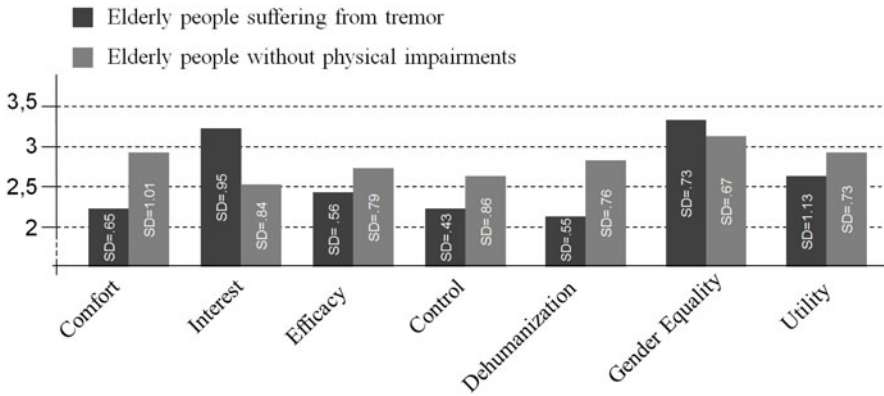


Fig. 7 Attitudes of elderly people towards computers; comparison between people with and without tremor symptoms

4.2 Computer Literacy

In order to measure computer literacy and evaluate the correlation with the test results a questionnaire from Sengpiel et al. [Sengpiel et al. 2008] with nineteen items was used. The survey is based on the symbol and term knowledge gained during human computer interaction and is generally seen as an indicator towards a person's competence in handling a computer system.

The test showed that persons with an computer handling experience of more than 7 years in average completed the trials 37% faster and made 19% less input errors. Here it was insignificant, if the usage was in a private environment or job related. 71% of the interviewees said they are not using a computer more than three times a week. As main reasons were termed the deficient usability (Software & Hardware) and unavailability.

4.3 Learning Curve

The results of a previous evaluation with 20 elderly first time users without tremor symptoms showed a steep learning curve for the self trained adaption of the swabbing input technique (Fig. 8). A time-stable input was reached after 20 to 22 input

cycles with an orientation phase from cycles 1 to 6. The average time for each input after cycles 20-22 lay at 1.52 seconds. This is a 100% input time reduction if compared to the initial input, supporting the steep learning curve argument and Schneider et al. The rise in productivity and quality (lower error ratio) are prime factors in the increase of success.

The input time for tapping input from the test persons showed a consistency after cycle 6 already. This proves the existence of a mature mental model on the standard approach of button use which for swabbing inputs yet has to evolve (Fig. 8).

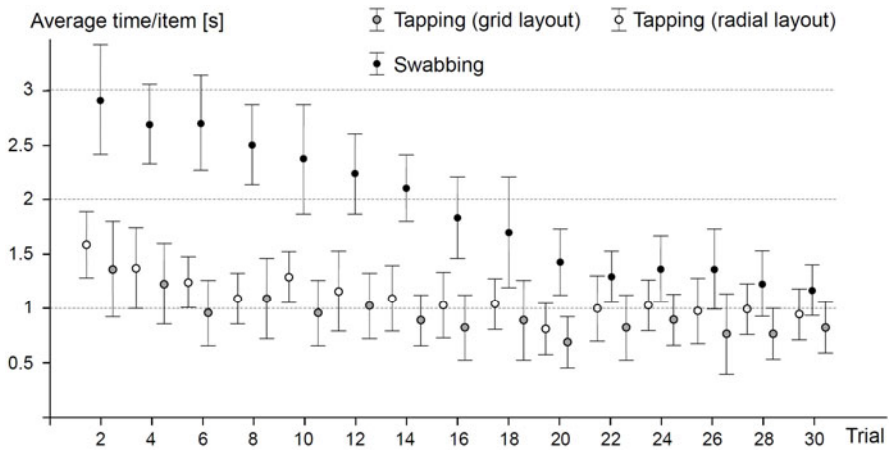


Fig. 8 Learning curve for elderly first time users for tapping/swabbing input on a touch screen

4.4 Parameters Influencing Error Ratio

In the following results, we used two-way, repeated measures ANOVA models with significance level of $\alpha=0.05$; data is normally distributed. We found no interaction between layout and resolution ($F(2,35)=0.780$, n.s.), and no significant difference between different layouts in tapping ($F(1,35)=3.128$, n.s.). Then, comparing between methods in radial layout shows significant effect of methods ($F(1,37) = 5.707$, $p<.05$). As seen in Figure 9, the error rates of swabbing in 16- and 25-buttons resolutions are lower than tapping. Post-hoc analysis with pairwise t-test with Bonferroni correction supports the effect in both resolutions (16: $p=0.0065$, 25: $p=0.042$).

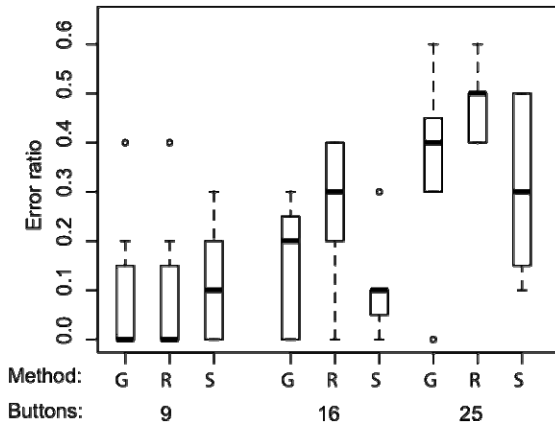


Fig. 9 Box plots of error ratio by resolutions and methods (G: tapping, grid, R: tapping, radial, S: swabbing, radial)

4.5 Tremor Characteristics

The acceleration data is Fast-Fourier transformed into frequency domain. This enables an analysis of tremor frequencies during rest and interaction with the touch screen. The frequency domain data can be visualized in spectrum plot. The frequency domain of data shown in Fig. 10 is presented in Fig. 11. Each plot represents frequency in an axis of a stage of interaction.

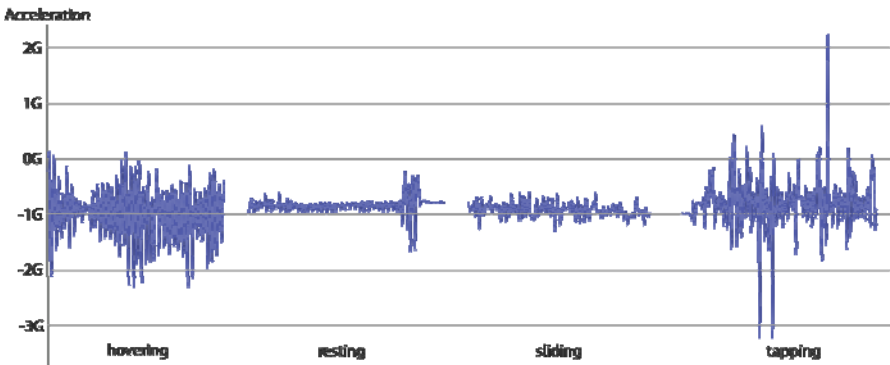


Fig. 10 An excerpt of acceleration data from each stage of interaction

We used the highest peak that is in the tremor frequency range (3–30 Hz) as a tremor frequency in each axis. In an axis, if the magnitude of tremor frequency exceeds 1 SD of the rest of frequencies in tremor range, we consider that axis a tremor axis. In Table 1, we listed the interaction stage with tremor and the respec-

tive tremor axes. Based on this data, we compared whether the tremor is worsen or lessen in swabbing compared with tapping. This result is also shown in Table 1. The investigated patterns show that measured effects are strongly dependent on the tremor type (resting, contraction, posture and intention tremor), especially regarding the axes of movement on which the tremor agitation appears (see tremor profile in Table 1 for details).

The wiping movement shows the best effect for those patients suffering from an intention and resting tremor, as here deviations among the main axes of movement are reduced, while for persons with contraction tremor sliding worsens the symptoms.

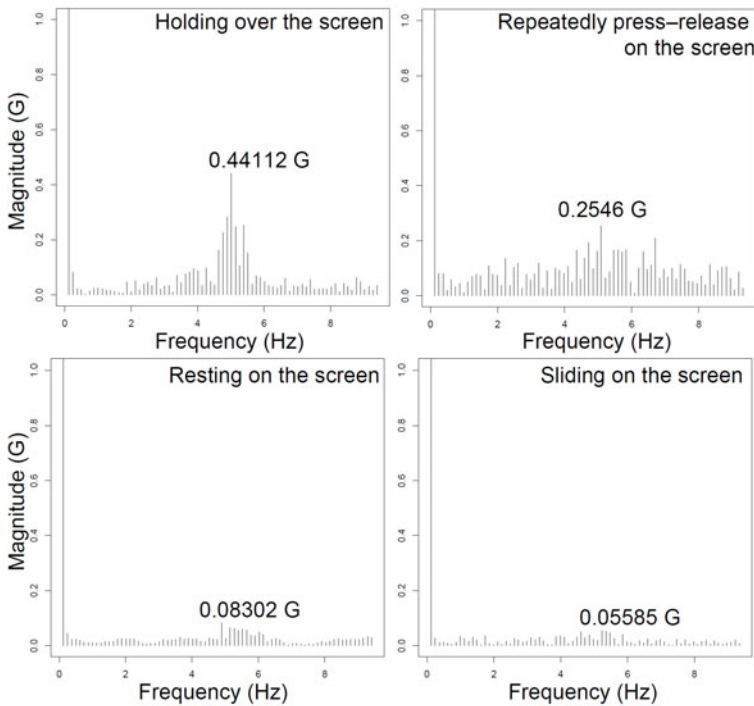


Fig. 11 Fourier plots of finger oscillation frequencies and magnitude in axis orthogonal to the ground from participant #5. The number indicates the value of the highest peak

The results further show, that people encumbered with medium and strong tremors, who use wiping movements as interaction, show a significantly reduced error rate of input compared to those touching standard button environments. It is also shown that patients suffering from a minor tremor agitation experience no noteworthy improvement during their interaction.

4.6 User Satisfaction

The satisfaction from the user while interacting with the system as well as usability was investigated by components of the “Post Study System Usability Questionnaire” (PSSUQ).

The Friedman test gave no noteworthy differences, for the given alternatives, with system satisfaction ($\chi^2 = 1$; n.s.), quality ($\chi^2 = 2$; n.s.), or usability ($\chi^2 = 0,667$; n.s.). All systems were generally evaluated positively and usage was rated as comfortable and intuitive.

4.7 Offset Analysis for Tapping Buttons

The evaluation of user input from participants suffering from tremor showed a significant offset for the core area of inputs depending on handedness. For analysis all button coordinates from grid layout were normalized in terms of size and positioning and the tracked contact points plotted separately for left handed and right handed touch screen users (Fig. 12).

The visualization shows that more than 90% of all fault insertions (not hitting the highlighted target area) are on the particular half side of the button of the person’s dominant hand.

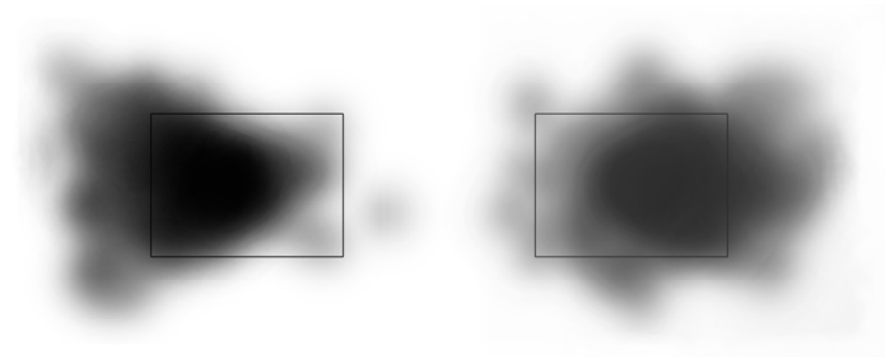


Fig. 12 Heat maps for the logged touch screen coordinates during grid-tapping, separated for left and right handed probands suffering from tremor. Higher frequency of selection is represented by darker gray tone

5 Discussion

The accomplished user study suggests that swabbing input gestures reduces error ratio for touch screen selection in older tremor patients. The result shows significant advantages of swabbing for 16 targets (button width = 41 mm). For 9 targets, buttons did slightly, but not significantly better. We speculated that the

size of the buttons were big enough to accommodate accidental movement from tremor. For 25 targets, buttons performed consistently worse, but the high variance in the swabbing results kept this from becoming statistically significant. However, the significant result in 16 targets makes swabbing worth for further study with more participants and higher resolution. Although swabbing takes more time to input than tapping, the results do not show differences in user satisfaction. This means that the trade-off between interaction time and accuracy is acceptable for tremor patients.

We believe that swabbing will make touch screen interaction more accessible to tremor patients, especially elderly persons, in the future.

Acknowledgment

- Participants of the studies for generously helping us in our research
- Parkinson Patients Self-Help Group Aachen
- Federal Ministry of Education and Research and body responsible for project DLR
- German B-IT Foundation.

References

- [Bain and Findley 1993] Bain, P.G., Findley, L.J.: Assessing tremor severity: A clinical handbook. Smith-Gordon, London (1993)
- [Deuschl et al. 1998] Deuschl, G., Bain, P., Brin, M.: Ad Hoc Scientific Committee Consensus statement of the movement disorder society on tremor. *J. of Movement Disorders* 13, 2–23 (1998)
- [Gina and Sherry 1992] Gina, M.J., Sherry, L.W.: Influence of direct computer experience on older adults' attitudes toward computers. *J. of Gerontology - Psychological Science* 47(4), 250–257 (1992)
- [Graham 2000] Graham, B.: Using an accelerometer sensor to measure human hand motion. Massachusetts Institute of Technology (2000)
- [Kraus and Hoffmann 2010] Kraus, P.H., Hoffmann, A.: Spiralometry: Computerized assessment of tremor amplitude on the basis of spiral drawing. *J. of Movement Disorders* 25(13), 2164–2170 (2010)
- [Martínez-Martín 1998] Martínez-Martín, P.: An introduction to the concept of quality of life in Parkinson's disease. *Journal of Neurology*, 2–6 (1998)
- [Müller-Tomfelde et al. 2008] Müller-Tomfelde, C., Wessels, A., Schremmer, C.: Tilted tabletops: In between horizontal and vertical workspaces. In: Proc. TABLETOP, pp. 49–56 (2008)
- [Schneider et al. 2008] Schneider, N., Wilkes, J., Grandt, M., Schlick, C.: Investigation of input devices for the age-differentiated design of human-computer interaction. In: Proc. of the Human Factors and Ergonomics Society, pp. 144–148. Mira Digital Publishing, New York (2008)

- [Sengpiel et al. 2008] Sengpiel, M., Struve, D., Dittberner, D., Wandke, H.: Entwicklung von trainingsprogrammen für ältere benutzer von IT-systemen unter berücksichtigung des computerwissens [Development of trainign programms for elderly users of IT-systems considering the computer literacy]. *Wirtschaftspsychologie, Alter und Arbeit* (3), 94–105 (2008)
- [Wenning et al. 2005] Wenning, G.K., Kiechl, S., Seppi, K., Mueller, J., Hogl, B., Saletu, M., Rungger, G., Gasperi, A., Willeit, J., Poewe, W.: Prevalence of movement disorders in men and women aged 50-89 years (Bruneck Study cohort): a population-based study. *The Lancet Neurology* 4(12), 815–820 (2005)

Compound Personal and Residential Infrastructure for Ubiquitous Health Supervision

P. Augustyniak

Institute of Automatics, AGH-University of Science and Technology, Krakow, Poland
august@agh.edu.pl

Abstract. This paper presents the concept and prototype of a compound infrastructure for ubiquitous human health monitoring. Growing range of mobile health care applications raises the question of their possible interference and cooperation. Particular benefit is expected from integration of personal and residential solutions, because of their complementary features. Consideration of various cooperation scenarios lead us to a specification of three cooperation levels depending on required integration of software. The paper presents details and experimental results of cooperation of two prototype surveillance systems lying in best result selection and conditional use of communication resources of the residential system as a carrier of messages from the personal system over the wired channel. This approach provides a cheap broadband data transfer minimizing the monitoring delays without limiting the subject mobility.

1 Introduction

Currently, the remote patient can be supported with several telemedical solutions for home care and seamless diagnosis based on selected vital parameters. Therefore the area of particular interest is the cooperation between personal and residential (i.e. building-embedded) healthcare systems. This aspect is crucial in ageing populations for maintaining the personal mobility required for today's professional activity. The two categories of applications for remote patient monitoring were developed independently and provide specific features:

- personal, body area network (BAN)-based solutions designed for active patients, providing data acquisition from multiple physiological sensors, integration and interpretation performed by a wearable server and long-distance wireless reporting to the medical surveillance center [Otto et al. 2006].
- building-embedded intelligent solutions developed for bed-ridden patients, elderly or other care-dependent people living on their own, providing continuous and discreet monitoring of disease-dependent subset of diagnostic parameters

and automatic distant interpretation of the data transmitted over a wired digital link [Liao and Yang 2008].

Wearable solutions, although providing the subject with unconstrained mobility, are limited by technological constraints (e.g. compromise of the weight to operation time, bandwidth of wireless connection, etc.) which are not justified when the subject remains within the premise. Embedded solutions, although almost unconstrained from the technological viewpoint, have precisely defined operation area, therefore lose the monitoring continuity if the subject is beyond. Accordingly to recent reports, an average, professionally active human spends 80% of his or her living time within buildings (45% in living premises, 35% in the office). This justifies the goal of our research on possible cooperation between two health monitoring systems applied for the same subject. Besides of providing wider diagnostic range and better quality, we expect to benefit from using jointly the competences of both systems and compensating for their drawbacks.

Diagnostic systems designed for a home care usage are based on a star topology network integrating multiple patient-side units considered as independent clients and managed by the central monitoring server (providing also archive and expert system services) [Wang et al 2006]. This setup assumes an exclusive connection between the clients and server without interaction with other systems or integration in the framework of a grid topology-based complex patient-oriented service [Atoui et al 2008]. The need for such services are justified by the benefit expected from a multimodal approach to gathering and integrating the health information from the subject and his or her environment with a wide range of wearable and embedded sensors. However, applying of two independent systems to the same subject also leads to a potential diagnostic ambiguity if they use various measurement methodology and signal processing techniques and yield different values of a given diagnostic result coexisting in the health record.

Due to a closed design and frequent negligence of interoperability guidelines, cooperation of two patient-side devices (e.g. personal cardiac and residential motion tracking) requires the modification of both: their embedded software, and also the respective servers' software. Considering the possible usage scenarios and the adaptation expenses, three cooperation levels were proposed in [Augustyniak 2010a]:

- sharing of the communication resources,
- overlapping measurement and interpretation competences,
- collaboration in estimation of diagnostic outcome.

This paper is focused on the details of a prototype cooperation based on context-aware best data selection and leasing of the communication resources for a personal BSN-based wearable cardiac monitor by a residential behavior tracking system embedded as a part of home automation infrastructure.

2 Materials and Methods

The idea of cooperation of two monitoring systems was applied to a design of a prototype compound infrastructure for the activity surveillance targeted to elderly but professionally active people [Augustyniak 2010b]. For the sake of usability, being here of primary importance, following the proposal of [Otto et al 2006] the set of wearable sensors was limited to an electrocardiogram monitor, three axes accelerometer and an optional SpO2 sensor. Remaining parts of the system (cameras, microphones and bed sensors) were embedded into a smart home infrastructure [Liao and Yang 2008]. Accordingly to a personalized medicine paradigm, both components are dynamically customizable by software settings.

2.1 Cooperation Scenario

In general, personal and residential monitoring systems are intended for different usage, what implies the application-specific design of communication interfaces:

- personal (wearable) systems commonly use wireless interfaces, allowing for a virtually unlimited operation range at the cost of high energy required for the omni-directional radio wave propagation and long-range transmission (GPRS, satellite, etc.),
- residential (home care) systems use wired, relatively cheap wideband connections, however this limits their operation range to the in-house patients.

When two or more different systems operate in the same area (e.g. a wearable system is used in house), there is no economical reason for continuing the usage of separate transmission channels. Moreover, the wireless transmission carrier is usually weaker in buildings, and even absent in unlucky propagation conditions.

In that case, accordingly to the *sharing the communication resources* scenario, the residential system leases its connection for sending also the data gathered by the personal system. Maintaining the subject's mobility also in house, requires an implementation of short-range wireless communication between the personal and residential systems and conditional embedding of external messages originating from the personal system within the wired communication protocol (fig. 1).

The implementation, although simple at first glance, requires the consideration of:

- selection criteria for the optimal transmission channel,
- rules and constraints of automatic initiation and termination of the leasing service,
- definition of the embedding (at the source) and dispatching (at the recipient) of external messages.

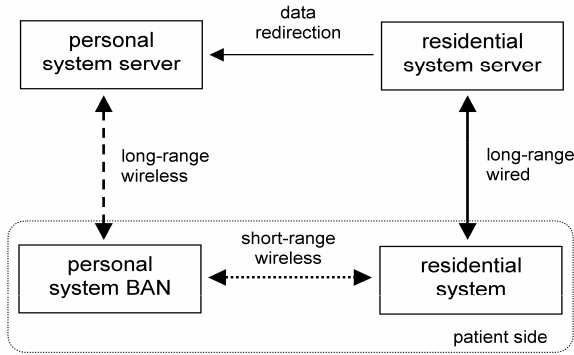


Fig. 1 Block diagram of cooperation of two monitoring systems lying in joint usage of the long-range wired communication channel

Since both considered systems provide independent estimates of subject's motion, besides the leasing of communication resources, typical for autonomous systems, next cooperation level may conditionally be applied. *Overlapping measurement and interpretation competences* enable a competition-like cooperation scenario between the concurrent systems resulting from the differences of measurement of physiological phenomena - based on different methods, or calculations of the diagnostic parameters - based on different algorithms. This process allows to select the best quality result that overrides the others, when multiple values are available at outputs from different monitoring systems (fig. 2).

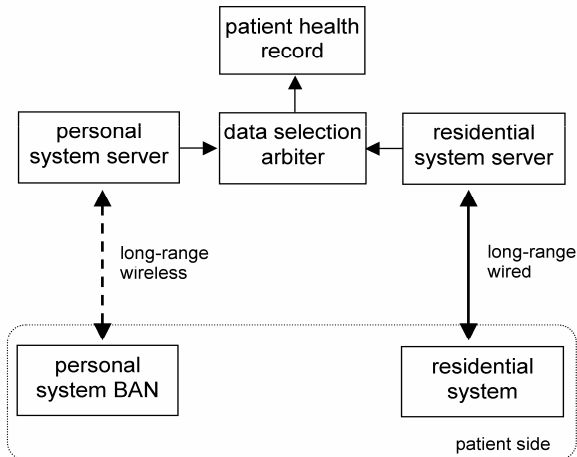


Fig. 2 Block diagram of cooperation of two monitoring systems when their measurement and interpretation competences are partly overlapping

In general, the possibility of an independent validation of the diagnostic outcome quality (e.g. a reference web service or human expert), allows for the optimization of the system outcome. Unfortunately, optimization criteria based on a disease-dependent validation of diagnostic results' quality or conditional accuracy lists for ranking diagnostic parameters are difficult to establish. Moreover, for only a few most used parameters, the international regulations require independent quality tests with use of reference data sets and the manufacturers rarely specify the results in user-accessible documents. In consequence, for *overlapping competences* of currently applied monitoring systems working with a common database (e.g. patient health record), the arbitrary selection of best result is the only practical method. In case of the presented prototype, this decision was supported by experimental results for the accuracy of motion estimation.

2.2 Personal Cardiac Monitoring System

The prototype of the personal cardiac monitoring system was designed as a cardiology-oriented monitor including MW705D GPS receiver (Mainnav), Aspekt500 12-leads ECG recorder (Aspel), TeleMyo 2400 G2 Telemetry System (Noraxon) and PXA-270 portable evaluation kit (Collibri) powered from the 4800 mAh 7.2V Li-Ion rechargeable battery pack [Augustyniak and Tadeusiewicz 2009]. Four components of wearable body sensor network (BSN) were interconnected with use of Bluetooth class II interfaces. In case of using the long-range wireless connection, the GPRS throughput limits the datastream to 2 ECG channels (upstream link of max. 16 kbps). Accelerometers placed on subject's upper and lower limbs provide a quantitative motion estimate but also precise kinematics parameters [Najafi et al 2003]. These data are particularly useful for semantic description of human motion activity and for discrimination of motion patterns.

The implementation of a short-range wireless communication between the personal and residential systems requires including of an additional communication gateway to the BSN (fig. 3). This gateway enables the alternative transmission of digital data with use of building-embedded infrastructure and wired access to the Internet. The role of data transfer service (DTS) may be fulfilled by a residential monitoring system (i.e. being a component of an intelligent house) personalized for the health-monitored subject, or by a regular wireless local area network (WLAN) often already present in offices. The bandwidth of a short-range DTS allows to send all 8 ECG channels, and with a maximum speed of 500 kbps the BSN in the *follow-up state* synchronizes the monitoring within less than 160 seconds.

During the experiments, the personal monitoring system provided raw ECG traces and the following diagnostic data: heart rate, beat type, ST-T segment parameters, time-domain Heart Rate Variability (HRV) parameters, and electrocardiogram-derived respiratory signal. Additional data streams were provided by wrist-mounted three axes accelerometers and Global Positioning coordinates updated every 5 seconds.

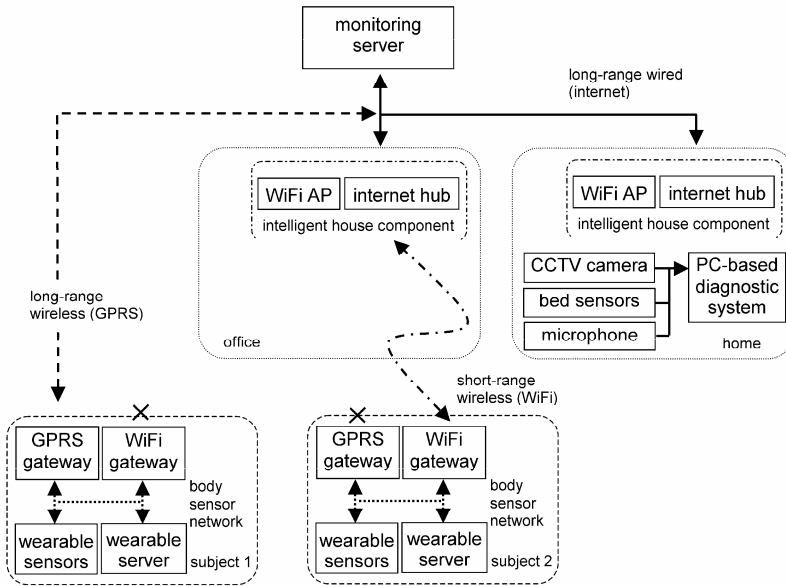


Fig. 3 Data transfer services in the compound personal and residential health monitoring infrastructure

2.3 Residential Behaviour Tracking System

A multimodal behavior tracking system is based on video and sound recordings and provides a video data stream annotated with semantic description of basic subject's states. Its principal purpose is the recognition of abnormalities in video-based behavioral patterns and classification of events into one of four categories {sleeping, resting, working and walking} implying further actions. Additionally fall detection and sleep evaluation is supported from sound recording and analysis. The recognition of alternate behavior is performed on the semantic description of the subject's status in the context of subject's personal habits record [Ślusarczyk and Augustyniak 2010].

Video-based monitoring of daily activities relies in division of the supervised living area to smaller regions in which the most common type, intensity and time duration of the human activity are specified during the setup or learning phase. Any significant deviation from the usual behavior with regard to specified spatial or temporal assumptions implies alerting. The subject's state is identified in real time as a result of motion quantification (quantity and variability) of whole human body or its selected segments, especially upper and lower limbs. The body posture is recognized with use of selected features of vertical and horizontal projections calculated on histograms of the segmented subject's silhouette.

The residential video-based presence detection and motion tracking system was designed as a component of the intelligent house infrastructure. It uses the wired broadband internet connection supporting the real-time motion picture transmission (standard datastream of 1800 kbps). This system accepts the external (non-video) data, embeds it into the packets, separates it at the recipient and redirects to the independent (i.e. cardiac) system server. For the motion monitoring, we use monochrome CCD (charge-coupled device) PAL system camera of resolution 720 by 576 pixels and additional set of nine infrared diodes placed around it. In this arrangement, the illuminators help to achieve an uniform exposure in the whole area of frame. Due to lower noise and wider analyzed volume, cameras are usually located in parallel to the longer dimension of the room. Quantitative evaluation of the human body motion is based on the absolute value of difference between each 25th video frame (i.e. in a 1 sec. time interval). For each differential frame, the value of brightness was averaged for all pixels. The motion index is then defined as percentage contribution from outlying pixels, and reveals both the value and frequency of the subject movements (fig. 4). Continuous 24-hours recording of body motion was also found useful in quantitative evaluation of behavior [Smolen et al. 2010].

Measurement of the acoustic signal for fall and snoring detection is performed with the sample rate of 44100 Hz using a microphone. The snoring phenomenon can be completely described in a frequency range of 12 kHz. Similarly to speech, snoring is produced in the vocal tract, therefore existing techniques for speech analysis may be successfully applied to identify and evaluate snoring sounds. With use of the Short-Time Fourier Transform the sound is transformed to the frequency domain in order to determine the frequency and energy distribution in local sections.

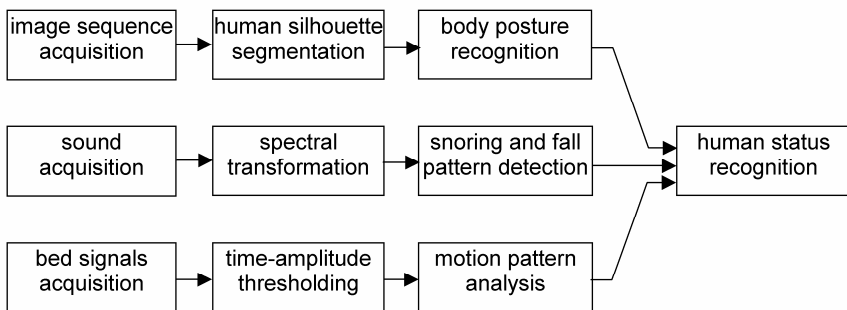


Fig. 4 Block diagram of the residential behavior tracking system

The monitoring server was a typical PC-based workstation with the static IP address. The experimental DTS infrastructure was designed with two stationary WiFi access points wired to a broadband Internet connection (Ethernet 100 Mbps).

2.4 Compound Infrastructure and Data Integration

The indispensable component of a compound infrastructure including personal and residential monitoring sensors is a device identification and negotiation protocol. It should support the temporary suspension of connection leasing, the non-availability of the final data recipient and the break of transmission in the leased transmission channel. Accordingly to the protocol, the cooperation between two systems is initiated and terminated automatically depending on the measured conditions and accordingly to the specified rules. The set of considered conditions has to cover all possible situations and the rules have to define all possible behavior of both systems. Even for this relatively simple prototype consisting of two cooperating systems and two patient states (in- and out house), we have to consider:

- detection of patient status by the presence on the video system (not concerned if the DTS is a regular WLAN access point),
- detection of patient's status by the quality of the short-range communication between systems,
- authentication and authorization for the external data support,
- detection of quality of the long-range connection,
- quality of GPS positioning (usually affected in house).

Cooperation rules for *sharing of the communication resources* between two systems are displayed in table 1.

Table 1 Prototype cooperation rules for leasing of the residential system's communication resources to the data acquired by cooperating personal monitor

rules	conditions	actions
1. patient identified as "in house"	1. patient is present on video system, 2. short-range link is established successfully, 3. authentication and authorization for external data support by home care system is successful.	1. switch off the GPS module and set the data flag to "not present", 2. use the alternative long-range communication server, 3. switch off the long-range interface. 4. use the video-based motion estimation
2. patient identified as "out house"	1. short-range link is weak or broken, 2. patient is not present on video system.	1. terminate the leased communication session and mark the last data sent, 2. try to establish long-range link, buffer the data until successful, switch to the long-range interface if possible. 3. switch on the GPS module and set the data flag to "present", 4. use the accelerometer-based motion estimation

The presentation of applied cooperation rules does not include the server software modification, in this prototype we assume that both servers continuously allow for sharing of the communication resources and the redirection of the cardiac data by the server of video surveillance system is defined.

For the reason of energy saving, when the subject is outdoor and the BSN uses the long-range gateway, the personal short-range communication module is switched to the *low-power receiver* mode. The residential DTS system is continuously propagating the premise-specific data identifying the house-embedded infrastructure and availability of connection leasing. Once these data are received and interpreted by the BSN, the personal short-range communication module is switched to the *transmitter* mode and sends the request for the DTS and the bearer identification. Since the wearable monitoring system is designed as strictly personal, detection of the BSN is equivalent to detection of the subject's presence. The purpose of presence detection is threefold:

- informs about the subject's position,
- authorizes the subject as a client of the residential monitoring system, and his personal system as a client of DTS provided by the residential system,
- identifies a subject-specific monitoring setup for downloading to the residential system.

The negotiation procedure is initiated by the residential system immediately after the detection of its personal counterpart. It is performed in three steps:

-
- verification of the personal system identifier and checking for acknowledgment from the wired network management server, preparation of server-side data dispatching routine,
- routing and verification of operation of the server being a final data recipient in the personal system-based remote diagnostic service,
- verification of quality of the short-range wireless link and the transmission bandwidth granted by the residential system.

It is noteworthy, that the bandwidth necessary for the connection state of the personal system is a rough estimate of the minimum throughput of the connection leasing service. The bandwidth used in the follow-up state may be significantly higher for short periods of time.

The authentication of cooperating devices is based on their unique identification numbers. Additionally, for each authorized data exchange session, the unique identifier is imposed by the server. These two labels, completed by the packet number and time marker are used for correct assignment of data contents, origin, destination and its temporal dependencies.

In wireless applications data continuity is particularly endangered by a variable throughput of transmission channel. For regular solutions, TCP/IP-embedded data control mechanisms are reliable enough for maintaining the continuity of diagnostic data transmission. In the proposed prototype, additional data buffering was

designed to support the acquisition when the transfer is suspended for the reason of switching from the short- to long-range data transmission and vice-versa and possible temporal absence of data carrier. Consequently, the data integrity is preserved in case of radio carrier discontinuity or unpredictable time and result of negotiations between DTSs. Besides the long- and short-range wireless links, used out- and within the reach of residential wireless connection respectively, a circular memory buffer is considered as third data recipient and works continuously in parallel as the data backup.

The supplementary mechanism for data flow management requires the active confirmation of data reception from the monitoring server. In the personal recorder the data deletion is suspended until the packet reception is granted by the positive recipient confirmation. If the delay exceeds a given time with respect to the moment of expected packet's arrival, the server issues the negative recipient confirmation with the missing packet ID which returns the read pointer to a specified location in the circular buffer. Data packets contain time markers, therefore when collected by the recipient their synchronicity may be determined as the delay between the time marker and the local clock. The borderline value allowing for interactive diagnosis and therapy is usually set to 5 s.

The capacity of the circular buffer was designed for 10 MB what corresponds to over 25 minutes of storage for 8-channel ECG data (500 sps, 12 bits per sample) with diagnostic results (ca. 48kbps). The buffer capacity determines the maximum continuous recording time with no availability of wireless DTSs (e.g. in remote areas, not WiFi-enabled buildings or air transport). The buffer contents is exclusively attributed by three separate pointers as written, read or deleted (Fig. 5).

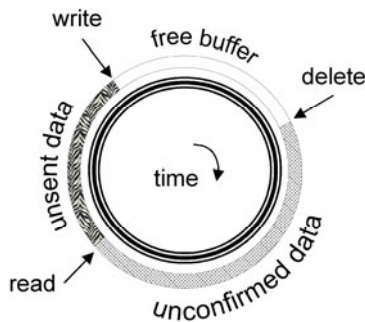


Fig. 5 Data buffering sequence in the personal health monitoring system

The management of pointers is subject to the following rules:

- the write pointer is moving forward (clockwise), each time the data are ready to sent,

- the read pointer is advancing each time the data packet is sent to the communication module, but it may be set backward by the *negative recipient confirmation*,
- the delete pointer is moving forward, by the *positive recipient confirmation*,

Writing, reading and deletion of data are asynchronous, depending on wireless carrier availability. The distance between the pointers determines one of the following states of the personal device:

- connection state - the read pointer is following the write pointer at the same average speed, the subject's health information at the recipient is synchronous,
- follow-up state - the read pointer advances faster than the write pointer reducing data delay, the subject's information at the recipient is asynchronous,
- data delay state - the read pointer advances slower than the write pointer increasing data delay, the subject's information at the recipient is asynchronous,
- low memory state - the write pointer is approaching the delete pointer causing a free memory alert, the continuity of subject's information is endangered.

When the short-range connection is broken, depending on result of video-based subject's presence detector applied in one of our prototypes, the residential system determines whether the subject is leaving the house. Until the confirmation its communication module remains active and seeks for the opportunity of restoring the connection. In the same time the personal device falls in *data delay (recording) state* (i.e. starts to buffer the data) and is seeking for the availability of long-range transmission channel. After completing a TCP-based connection via long-range channel, the device first enters in the *follow-up state* to resynchronize delayed data. Depending on the link quality, once the data are synchronized, the recorder works in *connection-* or *data delay state*. In the latter case, when the delay cumulates, the recorder may enter in *low memory state*.

Two independent measurements of the subject motion are implemented in the system based of different methodological approaches and having different characteristics. The accelerometer-based measurement:

- slightly affects the subject's comfort,
- allows for dynamic motion estimation,
- uses the subject as (relative) coordinates reference,
- is performed continuously.

The accelerometer data are integrated with the raw ECG and GPS outputs and may be sent independently from the monitoring server of the residential system.

The video-based measurement:

- is performed without the subject's action and perception,
- allows for dynamic motion estimation and static positioning,
- uses the premise as (absolute) coordinates reference,
- is possible only with subject's presence.

The video motion data are independent from the personal system operation and may be interpreted within a particular residential system.

Since both motion estimation techniques provide different information, in the prototype we applied a conditional data selection mechanism. When the subject is identified as "in house", the absolute video-based motion estimate is preferred and used for calibration of accelerometers. When the subject is outdoor, the motion patterns recorded by accelerometers in the personal device are compared to the references gathered with residential video-based system.

3 Results

Proposed prototype of a compound monitoring infrastructure was subject to tests focused on the cooperation between its personal and residential components. Two healthy volunteers (male aged 23 and female aged 26) wore the personal recorders for the total of 14 days each, moving between the "home" and "office" rooms in two different buildings at least 10 times daily. They also performed a physical exercise or resting accordingly to a predefined schedule. Within the "home" we implemented a PC-based prototype of behavior tracking system, while the "office" was a regular access point providing the short-range connection without the monitoring functionality. Besides the cardiac parameters (ST-T, HRV and EDR), subjects' data were investigated towards the identification of four basic states: {sleeping, resting, working and walking}. For each subject, transitions between states occurred 47 (+/-12) times and data carrier switching was performed 211 times.

3.1 Switching between Long- and Short-Range Carriers

The technical correctness of data carrier switching between the long-range (default, GPRS-based) and the short-range (alternative, WiFi-based) channel, was examined with particular attention to the data buffering in the personal system. The results of carrier switching delay are displayed in table 2.

Table 2 Delay time between the packets caused by switching of data carriers

switching direction	average delay time [s]	standard deviation delay time [s]	first attempt success rate [%]
long- to short-range	6.35	1.05	93
short- to long-range	17.3	8.10	71

Despite the use of relatively simple cooperation rules (see table 1), the first-attempt success rate, representing the percentage of successful switching, is far from 100% in real conditions. The common reason for inefficient switching were errors in detection of transfer conditions and poor quality of both (GPRS and WiFi) wireless links. Unsuccessful switching implies continuing of data buffering while subsequent attempts are made until a target connection is established.

3.2 Economical Savings on Telecommunication Service

Savings on the payment for the telecommunication service and on the energy consumption from the personal system's battery, determining its autonomous operation time, are main economical aspects examined during the test. The results of these tests are displayed in table 3.

Table 3. Economical benefits of conditional use of the short- instead of long-range data carriers

communication mode ratio	autonomy time [hours]	autonomy time gain [%]	communication payment [PLN] (@ 12kbps)	communication payment savings [%]
in house 100% of time	24.7	51,5	0	100
in house 80% of time	23	41.1	19	80
in house 60% of time	21.3	30,7	38	60
in house 40% of time	19.6	20.2	57	40
out house 100% of time	16.3	0	95	0

Economical savings on the telecommunication costs were estimated in Poland, but may be country-dependent. To the author's knowledge, small pocket hubs (e.g. D25HW) are in Japan a commercially available alternative providing transmission from WiFi-enabled mobile devices into long-range wireless channels. With such products, mobile systems need only to identify their WiFi-environment (SSID) to select the connection best reducing communication costs.

3.3 Correct Identification of Subject's Status

Since the volunteers strictly observed the physical exercise schedule, this can be a reference for evaluation of the subject's status as recognized by the monitoring system. Based on this reference the estimation of sensitivity was made separately for individual cardiac- and motion-based methods for subject's status recognition (table 4).

Table 4 Estimation of sensitivity [%] of individual methods for subject's status recognition

subject status detection method	volunteer 1		volunteer 2		joint methods average subject
	cardiac	motion	cardiac	motion	
sleeping	80	70	82	67	94,0
resting	84	71	87	67	95,5
working in house	59	86	62	83	93,9
walking in house	67	69	66	71	90,0
working out door	54	77	55	75	89,1
walking out door	61	93	59	91	96,8

It is noteworthy that {working} and {walking} has different performance, depending on the method used for motion estimation. In general, {walking} is more reliably recognized in outdoor subjects with use of accelerometers, whereas {working} recognition performs better indoor, when a video-based motion estimation is used.

4 Discussion

The implementation of a prototype cooperation within the compound personal and residential surveillance infrastructure revealed several practical conclusions (for 80% of the in house communication time as the most probable scenario):

- Significant (76PLN daily or 80%) cost economy due to the suspension of the GPRS connection when not necessary (see tab. 3),
- Moderate (41.1%) energy economy due to the use of the short-range wireless connection (WiFi) instead of the long-range connection based on the GPRS.
- High performance of subject's status recognition in the range defined for behavioral patterns including {sleeping, resting walking and working} with the joint use of cardiac- and motion-based parameters. Considering selection of best motion estimation method, the sensitivity of joint recognition equals 94%, 95,5%, 93,9% and 96,8% respectively.

From the technical point of view, the most disappointing outcome is a relatively long response time (6.35 or 17.3 seconds, see tab. 2) resulted from the buffering of messages in the personal system until the reception of every data packet is confirmed by the server. On the other hand, data buffering designed for the primary purpose of preventing data loss during the carrier switching is also functional in case of longer (up to 25 min) carrier absence. In conditions of our experiment the subject lost the long-range carrier at the entrance to the building, but needed another 55-75 seconds until he or she reached a WiFi-enabled premise.

This paper focuses on the design of the infrastructure combining the personal and residential parts, conditionally cooperating in surveillance of the subject. Our design considers the system behavior in any connectivity condition, even if data transmission is broken for a long period of time. Data continuity was granted thanks to the use of large circular buffer, temporarily turning the telemedical monitor into an independent recorder in dependence on the link quality. The resulting system has three automatically selected data recipients:

- wired, house-embedded telecommunication infrastructure, minimizing the operation costs, available when the subject is in house via short-range wireless link not limiting his or her mobility within the premises,
- wireless, long-range transmission service, allowing for a maximum mobility of the subject with personal monitor at the price of increased operation costs and viable connection quality,

- local storage, assuring for data continuity when no direct transmission is available or during the switching of data carriers.

The project is based on two lowest proposed cooperation levels: *sharing of the communication resources* is used for conditional leasing of the wired connection of the residential system for sending the data gathered by the personal system, and since both systems have *overlapping measurement and interpretation competences*, an arbiter procedure was applied to select best quality result.

Collaboration in estimation of diagnostic outcome needs even more open systems including the rules of use of the external diagnostic data and the rules of querying for such information. It rises many questions concerning mutual reliability, time synchronization and mutual authentication.

5 Conclusions

The presented prototype is part of a project aimed to provide a reliable solution for health parameter-based surveillance of elderly and care-dependent people. A significant novelty consists in an universal approach based on:

- adaptation of the system to the subject, and not vice-versa,
- assumption of unlimited mobility and consideration of professional activity,
- common design and flexible conversion of residential home care, personal telemonitoring and personal recording infrastructures.

The paper also points out several problems the open and cooperation-ready systems designers should solve.

Acknowledgment

The scientific work supported by Polish State Committee of Scientific Research in the years 2009-2011 under the grant no N N518 426736.

References

- [Atoui et al 2008] Atoui, H., Telisson, D., Fayn, J., et al.: Ambient intelligence and pervasive architecture designed within the EPI-MEDICS personal ECG monitor. *International Journal of Healthcare Information Systems and Informatics* 3(4) (2008)
- [Augustyniak 2010a] Augustyniak, P.: Complementary application of house-embedded and wearable infrastructures for health monitoring. In: Pardela, T., Wilamowski, B. (eds.) 3rd International Conference on Human System Interaction HSI 2010, Rzeszów, pp. 642–647 (2010)
- [Augustyniak 2010b] Augustyniak, P., Smoleń, M., Broniec, A., et al.: Data integration in multimodal home care surveillance and communication system. In: Piętka, E., Kawa, J. (eds.) *Information Technologies in Biomedicine*, vol. 2, pp. 391–402 (2010)

- [Augustyniak and Tadeusiewicz 2009] Augustyniak, P., Tadeusiewicz, R.: Ubiquitous cardiology: Emerging wireless telemedical application. IGI-Global - Hershey, London (2009)
- [Liao and Yang 2008] Liao, W.H., Yang, C.M.: Video-based Activity and movement pattern analysis in overnight sleep studies. In: Pattern Recognition, ICPR, pp. 1–4 (2008)
- [Najafi et al 2003] Najafi, B., Aminian, K., Paraschiv-Ionescu, A., et al.: Ambulatory System for human motion analysis using a kinematic sensor: Monitoring of daily physical activity in the elderly. *IEEE Trans. on Biomedical Engineering* 50(6), 711–723 (2003)
- [Otto et al. 2006] Otto, C., Milenković, A., Sanders, C., et al.: System architecture of a wireless body area sensor network for ubiquitous health monitoring. *Journal of Mobile Multimedia* 1(4), 307–326 (2006)
- [Ślusarczyk and Augustyniak 2010] Ślusarczyk, G., Augustyniak, P.: A Graph representation of the subject's time-state space. In: Piętka, E., Kawa, J. (eds.) *Information Technologies in Biomedicine*, vol. 2, pp. 379–390 (2010)
- [Smolen et al. 2010] Smoleń, M., Czopek, K., Augustyniak, P.: Sleep evaluation device for home-care. In: Piętka, E., Kawa, J. (eds.) *Information Technologies in Biomedicine*, vol. 2, pp. 367–378 (2010)
- [Wang et al. 2006] Wang, Q., Shin, W., Liu, X., et al.: I-Living: an open system architecture for assisted living. In: *Proc. on IEEE International Conference on Systems, Man and Cybernetics*, pp. 4268–4275 (2006)

Using Computer Graphics, Vision and Gesture Recognition Tools for Building Interactive Systems Supporting Therapy of Children

J. Marnik, S. Samolej, T. Kapuściński, M. Oszust, and M. Wysocki

Rzeszow University of Technology, Rzeszow, Poland

{jmarnik,ssamolej,tomekkap,moszust,mwysocki}@prz-rzeszow.pl

Abstract. The paper presents a prototype of a system which can be used as a therapeutic and educational tool for children with developmental problems. Natural body movements and gestures are used in the system to interact with virtual objects displayed on the screen. Nowadays such systems can be built with the use of widely available free software tools for both graphical and vision applications. Such tools are also shortly presented in the paper.

1 Introduction

Last years showed growing interest in using modern tools to make interaction with computer more natural. Different approaches are used to do this. For example, Herbelin, Ciger and Brooks [Herbelin et al. 2008] built various interfaces for disabled people to give them possibility to play the game Planet Penguin Racer. They used three motion sensing devices (a camera, a SoundBeam ultrasonic distance sensor, and three-axis accelerometers) to control the game. A commercial Silverfit system [WWW-1 2010], in which a motion-sensing camera is used to acquire action of human who plays one of several built-in games, is available since January 2009. The system is targeted at old people after stroke and their therapists, and it serves as a rehabilitation tool. The AuRoRa project studies whether and how robots can become a toy that might play an educational or therapeutic role for children with autism [WWW-2 2010]. Recently, gesture-based interfaces are starting to appear in computer games [Gonçalves et al. 2008]. Review of the use of modern technologies to support disabled people can be found in online proceedings of International Conference Series on Disability, Virtual Reality and Associated Technologies (ICDVRAT).

The paper presents a prototype of multimedia system, which can be used as a therapeutic tool for children with developmental problems such as Attention Deficit Hyperactivity Disorder (ADHD), Autism Spectrum Disorders (ASD), mental impairment as well as hearing impairment. The system consists of two main subsystems: vision and graphic. The graphic subsystem presents certain scenario, in

which a child accomplishes given tasks, which are recognized by the vision subsystem. The system has a 3D or 2D graphical interface. The 2D cartoon-like interface may be dedicated to younger children, whereas 3D interface may arise interest of youngsters. Tasks, which are to be executed by the child, are analyzed by the vision subsystem. These tasks can concern doing simple physical exercises or following some instructions, for example: to stand on specified place and rise hands. We assumed that graphic and vision subsystems should be able to work on different computers, so they should communicate by predefined signals. These signals are generated by the vision subsystem, which works as an input device (replaces mouse and keyboard). An external network engine connects both subsystems and supervises inter-subsystem message exchange.

The mentioned subsystems are implemented with the aid of many open source tools, available from the Internet. Such tools are shortly described in two next sections. Then an outline of the system is presented. At the end conclusions are given.

2 Computer Vision Tools

The most known software tools, which use computer vision techniques are Open Source Computer Vision Library (OpenCV) [Bradski and Kaehler 2008] and LTI-Lib [Krais 2006]. Both tools provide us with a reach set of functions, which implement advanced algorithms from the field of image processing and image recognition. Both OpenCV and LTI-Lib are freely available from the Internet and they can be used in commercial products without any cost. Short characteristics of these libraries are presented below.

2.1 OpenCV

OpenCV is a programming functions library written in two variants, C and C++, optimized and intended for real-time applications. It is independent of operating system and hardware. The library provides interface to Intel's Integrated Performance Primitives (IPP) with processor specific optimization (Intel processors). It is released under a BSD license, and free for both academic and commercial use. OpenCV can be used to resolve such tasks as human-computer interaction, object identification, segmentation and recognition, face and gesture recognition, camera and motion tracking, motion understanding, stereo and multi-camera calibration and depth computation, as well as mobile robotics. The library provides the user with various data structures, including dynamic ones (lists, queues, sets, trees, graphs), and routines for:

- image data manipulation,
- image and video I/O,
- matrix and vector manipulation and linear algebra problems solving,

- image processing (filtering, edge detection, corner detection, sampling and interpolation, color conversion, morphological operations, histograms, image pyramids),
- structural analysis,
- camera calibration,
- motion analysis (optical flow, motion segmentation, tracking),
- object recognition (eigen methods, Hidden Markov Models),
- basic GUI (display image/video, keyboard and mouse handling, scroll-bars),
- image labeling.

All procedures are divided into thematic modules:

- `cv` – contains main OpenCV functions,
- `cvaux` – comprises of auxiliary (experimental) OpenCV functions,
- `cxcore` – defines data structures and gives linear algebra support,
- `highgui` – includes GUI functions.

Many sample programs are also provided with the OpenCV package.

2.2 LTI-Lib

The LTI-Lib is an open source software library that contains a large collection of algorithms from the field of computer vision. It is targeted both for Windows and Linux platforms. It is under GNU Lesser General Public License, which allows its use in commercial products. The library is implemented in C++ with the use of object-oriented approach.

The LTI-Lib was created at the Chair of Technical Computer Science at the RWTH Aachen University. It is easy to use due to the specification of a well-defined programming interface for all classes. Its consistency is preserved by dint of the PERL-script based generator of LTI-Lib classes, named `ltiGenerator`. All algorithms in the LTI-Lib are encapsulated in so-called functor classes. They always enclose a class called `parameters` and a method `apply()`. This class contains parameters of the algorithm and methods for setting and getting their values. It can be explicitly declared or just inherited from the parent class. The method `apply()` gives access to the functionality of the class.

Besides functors, the LTI-Lib provides with classes which can be used to visualize data, read and write them from/to disc. Separate class is intended to classify objects.

More than 300 classes deal mainly with one of the following fields:

- linear algebra,
- classification and clustering,
- image processing and image analysis,
- visualization and drawing tools.

On the basis of LTI-Lib, a Graphical User Interface for Rapid Prototyping of Image Processing Systems, called IMPRESARIO, has been created [Krais 2006]. It

provides flexible and easy to use interface to test user's algorithm, which uses computer vision techniques. Image processing and computer vision algorithms are represented by blocks, equipped with icons, which correspond to parameters, input and output data of the algorithm. Separate blocks are provided for input data (e.g. image, image sequence, video stream). The output of a block related to given algorithm can be fed to the input of another algorithm by putting it through to the input of a block representing that algorithm. The IMPRESARIO main window is shown in Fig. 1. A flow chart of a sample system is visible in the document view window. Parameters' table and a figure showing an intermediate image, obtained during processing, are also visible in this window.

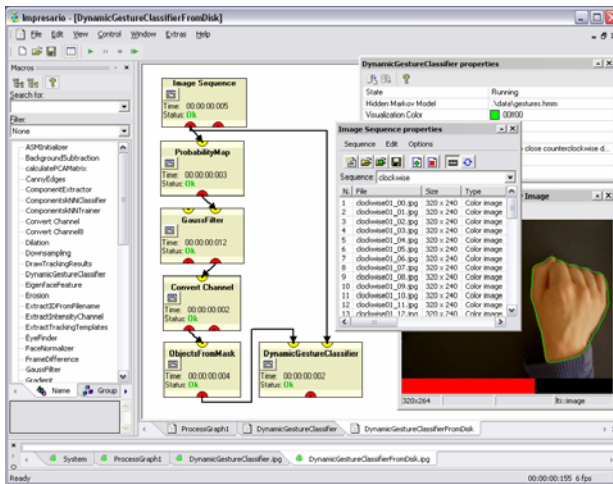


Fig. 1 The IMPRESARIO system

3 Computer Graphics Tools

The system presented in the paper has a graphic interface. The users are involved in computer game-like scenarios. To accomplish a stage of therapeutic session they have to interact with or control virtual characters. The possibility of interaction with the characters and moving into virtual space are natural for children - the target users of the system.

The important phase of the system development is selection of adequate software tools for the effective interface creation. During the selection the following aspects have been taken into consideration:

- characters and the environment development may be prepared by an artist and easily digitalized by a computer graphics designer,
- a digital library of characters as well as environment items should be easy to create and reuse,

- the user of the system would have the possibility to interact with the elements of graphic scene projected on the screen,
- the two-way communication between the vision subsystem and the graphic interface must be possible,
- the software tools should have reasonable price or, if it is possible, be available for free, even for future commercial use.

The analysis of above mentioned prerequisites brings the preliminary software tools requirements as follows. Firstly, a separate software tool should be used for the virtual environment and character creation. Quite natural candidates may be Autodesk® 3ds Max®, Autodesk® Maya® 3D, LightWave®, Autodesk® Soft-image® or Blender. All of the software toolkits can be naturally used by artists or computer graphics designers to create and animate characters or any models of 3D or 2D objects. It is also possible to export the developed items to well-known file formats for characters and objects encoding (e.g. Quake MD2, MD3) or environment models (e.g. Quake BSP).

Secondly, a separate program should be created for both: effective graphic object animation and communication with the computer vision subsystem. During the program development, two approaches may be considered. The first solution may be the attempt to produce the software from the scratch using standard graphics (e.g. OpenGL®, Direct3D®, Adobe Flash®) and communication (e.g. Win-Sock®) APIs. The second one might involve the adaptation of a graphics or game engine (e.g. OGRE, Crystal Space, XNA, Quake 3, Panda3D) to the system requirements. Although the “creating from the scratch” solution is more flexible, the “game or graphic engine adaptation” approach brings much more straightforward and rapid solution, especially for the application that is subject of our research. The game or graphic engine adaptation turns the computer graphic scene development into downloading animations or objects, arranging them in the space and managing the position of the camera without taking care about techniques of generating the objects or animations. It is possible to import externally defined (e.g. with the aid of previously mentioned graphic toolkits for computer graphics object and characters generation) graphic items into the scene. What is more, most of the graphic or game engines include or may be extended by network communication modules to naturally port them to the external applications.

After the overview of the mentioned above software tools, Blender as a graphic editor and Object-oriented Graphics Rendering Engine (OGRE) as a graphic engine have been finally chosen for the system interface development. Both of the software tools meet the assumed requirements, there exist file format exporters from Blender to OGRE, and finally the tools can be used for commercial software development for free (on GNU GPL license for Blender and on LGPL license for OGRE). The following subsections will briefly introduce the main features of the chosen software tools.

3.1 Blender

During our system interface development Blender is mainly exploited as virtual character and object generator. The characters and objects may be created using a set of graphical editors that enable:

- 3D object wireframe creation,
- 3D object texture mapping,
- skeleton definition and association its elements with selected wireframe regions,
- skeletal animation.

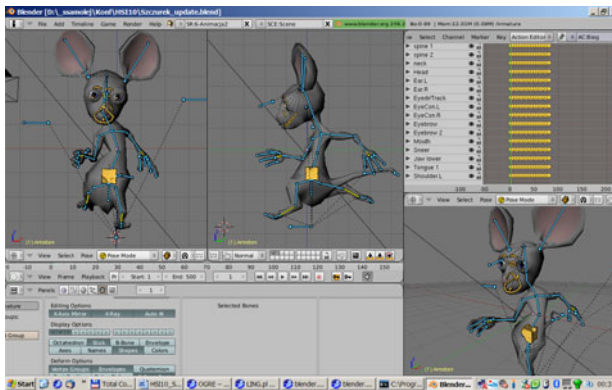


Fig. 2 Blender application for modeling exemplary character

Fig. 2 shows Blender toolkit during generation and capturing of exemplary character and its gestures.

Apart from that Blender can be effectively applied for rendering (digital picture or film generation), modeling and simulation of selected physical phenomena, such as fluid and hair motion or collisions of rigid objects.

In some scenarios the believable, realistic behavior of the animated avatar is required. This can be achieved by the technique called motion capture. It consists in capturing the movement of the person and translating that movement onto a digital model. In typical situation the performer wears markers near each joint to identify the motion by the positions or angles between the markers. The markers are tracked by the cameras. Specific hardware and special programs are required to obtain and process the data. The cost of the professional software, equipment and personnel required is prohibitive for our project, therefore authors propose to use one of the freely available motion capture databases, e.g. Carnegie Mellon University Graphics Lab Motion Capture Database. This database contains data related to interactions among people, interaction with environment, locomotion, physical activities, sports and other situations and scenarios. In particular, it contains motion

capture data describing the movements needed in our project. The files available in the database can be imported to Blender. The process consists of creating the model in Blender, then rigging it with bones, importing the BVH animation file from the data base and finally adjusting some bones if required. The motion capture data is free for use in both research projects and commercially-sold products.

3.2 OGRE

In the system presented in the paper OGRE is applied as the integration layer. On one side it executes the program graphic interface, on the other side it collects or sends data from or to the computer vision subsystem.

The typical areas of OGRE applications are 3D interactive graphic programs or games. The core of the OGRE is a set of C++ classes that constitute the platform independent layer for producing 3D interactive graphic applications. OGRE makes it possible to include most of the recent techniques dedicated to produce real-time 3D computer graphic animations, such as shaders, multitexturing, or multiple material techniques. It can effectively download and display 3D objects expressed in some selected graphic file formats. Simultaneously, it offers its own XML based, internal graphic file format for storing and rendering objects and animations. In Fig. 3 an example 3D system interface is presented. The character and its environment were created in Blender and imported into internal OGRE XML file format.

As OGRE offers only graphic interface, the additional networking engine (e.g. RakNet) must be proposed to merge with the system to perform inter – subsystems communication.



Fig. 3 Example of 3D system output interface

4 System Overview

Fig. 4 depicts the laboratory system setup. The child watches the exercise performed by the virtual character. Next the same exercise should be repeated by the child. The performance is validated by the vision subsystem. Alternatively, the child performs an exercise announced by the narrator and the virtual character repeats it. Then the next task is given. The system can use voice commands/comments generated by speech synthesizer IVONA, which is relatively cheap for private applications.

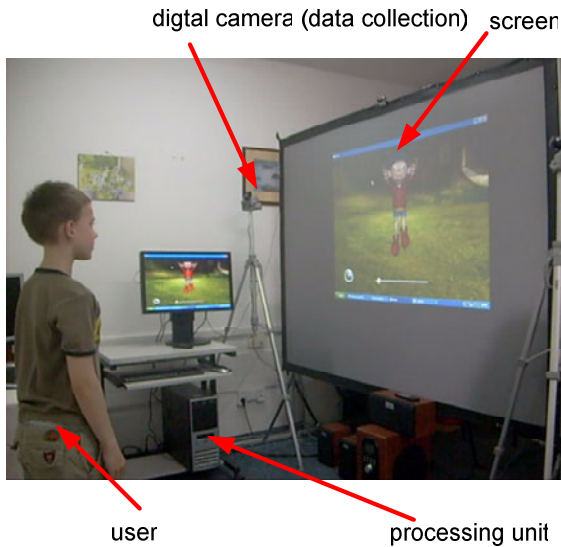


Fig. 4 Exemplary therapeutic classroom. The child is watching an exercise to perform

Apart from software tools selection, the most important phases of system development were:

- determination of therapeutic scenarios,
- choice and adaptation of computer vision algorithms for gesture recognition,
- software development and integration.

The following subsections will give some more details.

4.1 Therapeutic Scenarios

The system is designed for children with developmental problems as a therapeutic tool. Exemplary scenarios for therapeutic sessions were proposed by the therapist, who works with such children.

In these scenarios, the child has his representation in the form of a graphic character (avatar) in a virtual world. This avatar is used to perform actions suitable to the current state of the scenario development. Its behavior is dependent on the child's activity. The narrator informs the child which exercises/gestures he/she should do to achieve results which are expected at the given stage of the scenario. If the vision system recognizes one of these activities then the action in the scenario related to this activity is performed. Fig. 5 shows a typical scenario stage outline.

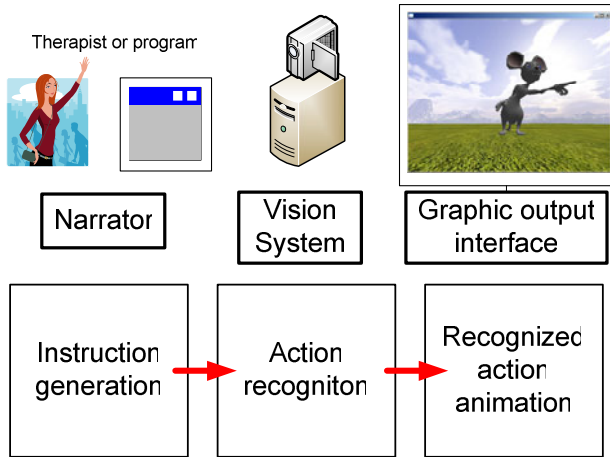


Fig. 5 Therapeutic scenario stage scheme

Some of the characters in the presented system resemble well known creatures from popular animated cartoons as Noddy or The Magic Roundabout. They were created from the scratch by members of our research team. Our experience showed that children would rather identify with the avatars that resemble the cartoon character than completely new one.

Actions provided for the therapy of children with ADHD in our system are mainly simple physical exercises, which should be performed in a fixed place. It aims at overcoming problems with increased activity of children with ADHD. Such children have also serious problem with focusing their attention. Thus, in our system directives for them are often given after telling a story related to the scenario. Children must hear out the entire story if they want to keep on playing with the scenario.

The system can be adapted to different therapeutic aims. For example, an educational aspect of children with mental impairment or deafness can be taken under consideration. Our team has rich experience with sign language and sign spelling recognition. Sign spelling can be used instead of gestures or mouse and keyboard in educational systems meant for children with hearing impairment.

There are three scenarios in the current version of the system. In the first scenario a virtual character is showing simple exercises (squat, jump, waving hands, straddle-compact jumps). Narrator asks the child to make an exercise shown by a virtual character. Accomplishing the task is rewarded with appropriate narrator voice announcement and then next exercise begins.

In the second scenario the child and the virtual character change their roles. Now the virtual character imitates exercises shown by the child. Before the therapy session is started, there is a possibility to select the virtual character in both scenarios. It can be a boy, a gnome or a girl.

In the third scenario the child meets particular situations presented on a screen. Each situation requires performing some action or solving a task. In this scenario all actions or tasks are linked with ladybug theme. The ladybug meets different creatures that live in a meadow; it flies over different objects, visits a zoo, and a town. Children should name the objects on the screen, interact with the creatures, and simultaneously do exercises suggested by the narrator. Tasks require to analyze everything what is shown on the screen or to listen to the narrator carefully. The main aim of the therapy with this scenario is training concentration and visual-motor coordination.

In order to ensure correct interpretation of child's action, calibration of the system is conducted at the beginning of the session. Then the system gets information about the range of possible changes of the size of the ellipse approximating the face. The child is asked to make one step forward and backward, and then one to the left and one to the right from the initial position. The procedure is supervised by the narrator who informs the child what it has to do.

4.2 Action Recognition Techniques

To recognize the action computer vision algorithms are used. These algorithms are implemented by our research team. We use two approaches here, namely marker-based one and markerless approach.

Using former approach we can now recognize following actions: moving right / left / forward / backward, jumping, rising the hands and straddle-compact jumping. Here, models of actions were constructed on the basis of trajectories of moving markers, acquired for specified action. The shape of particular trajectories and their relative position to each other trajectory are taken into account during action classification. In this approach, markers are detected using color model and they are tracked using camshift [Bradski and Kaehler 2008] algorithm.

With the markerless approach we have defined such actions as: waving left/right hand, waving both hands simultaneously, bowing, waving head to say "yes" (up-down) and "no" (left-right), presenting selected expressions in Polish sign language and showing some hand postures.

Here, to define and recognize all actions, except sign expressions, we used the motion templates proposed by Bobick and Davis [Bobick and Davis 2001]. Action's models are generated using Hu moments [Bradski and Kaehler 2008]. Examined action is classified by the nearest neighbor classifier. Because no markers

are required here, it is assumed that a background behind the person is uniform, and a contrast between the person and the background is relatively high.

Selected Polish sign language expressions are recognized using color images. For detection of the signer's hands and face a method based on a chrominance model of human skin is used. At the beginning of the session the signer presents open hand to the camera. A rectangular hand segment is used to build a skin-color model in the form of a 2D Gaussian distribution in the normalized RGB space. To detect skin-toned regions in a color image, the image is transformed into a gray-tone using the skin color model, where the individual pixel intensity in a new image represents a probability that the pixel belongs to a skin-toned region. After thresholding the gray-tone image is converted to a binary image. The areas of the objects toned in skin color, their centers of gravity and ranges of motion are analyzed to recognize the right hand, the left hand and the face. In order to ensure correct segmentation there are some restrictions for the background and the clothing of the signer [Kapusinski and Wysocki 2005].

Hand postures are recognized using a method based on curvature analysis of the hand boundary and the nearest neighbor classifier [Marnik 2009]. An attempt is being made to use Hierarchical Temporal Memory – a new computational paradigm based on cortical theory – for shape recognition under large variations of hand rotation [Kapusinski 2010]. To recognize signed expressions we use Hidden Markov Models [Kapusinski and Wysocki 2005; Kraiss 2006]. Actually our system recognizes 100 words and 35 sentences which can be used at the doctor's and at the post office as well as 10 hand postures corresponding to selected hand shapes occurring in Polish finger alphabet.

For both marker based and markerless approaches the number of recognized actions can be increased, by adding a new action template to the template database.

4.3 Therapeutic Classroom

To undergo therapeutic scenarios appropriate classroom preparation with necessary multimedia equipment is needed (see Fig. 4). There are four types of equipment used in the proposed system: processing units, cameras, presentation tools, lighting, and elements of the production. Cameras are connected to the processing units responsible for visual data analysis. Processing units could also handle display management tasks and speaker service in scenarios requiring the use of sounds. During the session with the child virtual reality could be displayed in various ways (e.g. LCD screens, virtual reality helmets) but to assure low cost back projection screens and multimedia projectors seem to be promising solution. Back projection eliminates glaring the camera by the projector light as well as screening the projector light by the child.

We assume that the classroom is adequately lighted to overcome problems with vision-based object recognition caused by changing lightning conditions (e.g. daylight or primary uneven lighting installed in the classroom). For this purpose halogen lamps on stands are used. As the elements of the production floor decoration, children clothes with markers, necessary toys, etc. helping the child to enter to the world presented on a screen, all related to the current scenario, are considered.

4.4 Software Components Details

The prototype system consists of two subsystems: the vision one and the graphic one. The vision subsystem is aimed to generate signals on the basis of a human action, which is captured by a camera. The graphic subsystem generates real-time animations logically related to the signals obtained from the vision subsystem. The structure of the system is distributed. It is possible to place the vision and the graphic subsystems on separate computers.

Vision subsystem is built from blocks, each of which is designed to perform determined tasks. Blocks are implemented as classes in C++. Such structure allows easy modification of the system. Thus the system can be fitted to child's age and abilities, as well as to therapist's requirements. The most important blocks of the vision subsystem are:

- image acquisition block,
- gesture recognition block,
- gesture database handling (adding a new gesture),
- marker definition block.

The image acquisition block contains tools which enable acquiring images from the camera. Methods designed for determining which gesture is performed by a person observed by the camera are defined in the gesture recognition block. This block is closely related to the gesture database. Only these gestures which are defined in this database can be recognized by methods from the gesture recognition block. Methods available from the gesture database block can be used to define a new gesture. If markers are to be used in the system then they can be defined with the use of methods contained in the marker definition block. The gesture recognition block and the gesture database block contain methods which can also handle marker-based input.

Most of tools designed to define and recognize human actions we implemented using the OpenCV library. Actions concerning sign language and hand postures were created using library of C and C++ functions/classes prepared by our research team.

The graphic subsystem consists of two main units:

- character and other graphic objects development software,
- system interface executor.

The graphic characters and their environment may be created by separate group of developers: artists or computer graphics designers. The preferred software toolset for this part of work is Blender, but it is possible to create the models using e.g. Autodesk® Softimage® or Autodesk® Maya®. The character generation involves graphic object wireframe creation, texture mapping, rigging and preparing a set of predefined character gestures (skeleton animations). The result of graphic character and environment development is a set of files including encoded structures of objects and animations. They constitute a library of graphical items that can be

used in creating different therapy scenarios. During the therapy session the user can also choose the avatar that he/she best identifies with. Separate programs can convert the Blender internal file format into the format acceptable by the graphic engine.

The system interface executor is a C++ program founded on OGRE framework that includes the following main blocks:

- graphics manager,
- network manager,
- input manager,
- animation manager.

Graphics manager sets up a scene: creates an object that includes the scene, arranges the camera and lighting, loads graphic objects that will belong to the scene, starts the animation loop. The network manager connects the vision subsystem using the external RakNet network engine and supervises inter-subsystem message exchange. Input manager reacts to some local interface signals, such as mouse or keyboard clicks and makes it possible to adjust the interface during the system run time. Animation manager executes the predefined animations according to messages acquired from the vision or local input modules. General management rule for the interface assumes, that there exist a set of separate objects that listen for the events and direct them to the main rendering module to call the proper object configuration on the scene.

Both subsystems are linked by a configuration module, which allows choosing a graphical scenario and selecting type of commands used to control this scenario. Current possibilities are:

- sign language expressions,
- signs contained in finger alphabet,
- body movements,
- keyboard commands.

If body movements interface is selected, then one of two variants of vision recognizer can be indicated, namely color marker based approach or markerless based approach. After this, default actions associated with commands used to control graphic interface appear in combo boxes related to particular commands.

These actions can be adapted to user's requirements by selecting appropriate position from relevant drop-down list. User defined configuration can be saved in user's configuration file, which later can be used to configure the scenario. At any time the settings can also be restored to the default values.

In the case, when marker based approach is checked, the additional, marker definition window appears on the screen. It allows loading the data defined earlier for specified markers, or defining new markers. Moreover, each marker can be associated to specified body part here. To provide good system behavior, the markers

should have vivid colors, which do not occur in scene background. The number of markers is not limited in the system, but too many markers can lead to incorrect system behavior.

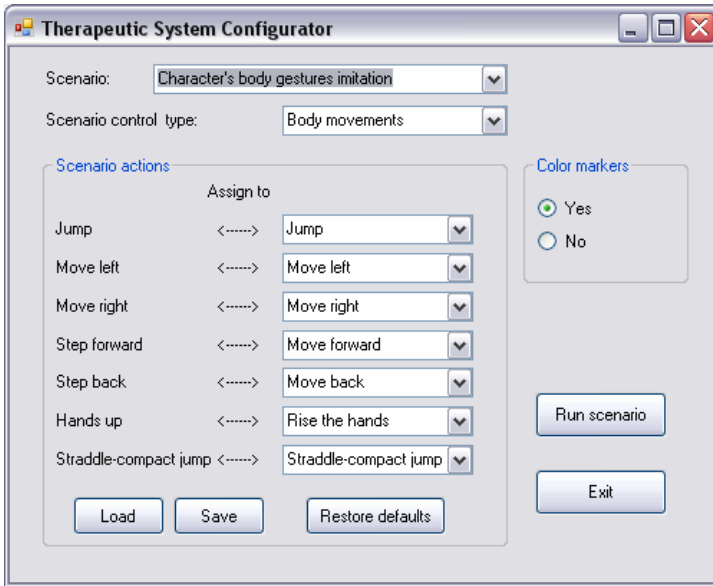


Fig. 6 Configuration module interface

The configuration module contains the button “Run scenario” designed to run selected scenario. At the beginning of each scenario the calibration procedure, related to selected options and scenario is invoked. The main window of configuration module interface is presented in Fig. 6.

5 Conclusions

The computer vision and graphics based system for interaction with mentally and physically disabled children developed by our research team is still under evolution. The main area of research concerns the effective gesture recognition algorithms using popular Internet cameras as vision data source. Separate research path includes the attempts of extending the input of the system by stereovision camera, thermovision camera and mobile accelerometers for more effective gesture and movement recognition. Moreover, various 3D and 2D graphic system interfaces will be proposed. Empirical tests will help in selection of adequate interfaces for different target users. The system is open. So, new methods, tools, modalities, scenarios, etc. can be included. The users (therapists, parents, etc.) will have the possibility to configure the system and fit it to individual possibilities and needs. Further

development will include a decision support based on the analysis of the recorded course of the therapy.

Acknowledgment

Speech synthesizer IVONA used in the research has been purchased within the Project POPW.01.03.00-18-012/0 cosponsored by UE within the Operational Program Development of East Poland 2007-20013, Priority I, Modern Economy, Action 1.3 Innovation Support.

References

- [Bobick and Davis 2001] Bobick, A., Davis, J.: The representation and recognition of action using temporal templates. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 23(3), 257–267 (2001)
- [Bradski and Kaehler 2008] Bradski, G., Kaehler, A.: *Learning OpenCV: computer vision with the OpenCV library*. O'Reilly Media, Inc., Sebastopol (2008)
- [Gonçalves et al. 2008] Gonçalves, D., Jesus, R., Grangeiro, F., et al.: Tag around: a 3D gesture game for image annotation. In: *Proc In ACE 2008*, Yokohama, Japan, pp. 259–262 (2008)
- [Herbelin et al. 2008] Herbelin, B., Ciger, J., Brooks, A.L.: Customization of gaming technology and prototyping of rehabilitation applications. In: *Proc. 7th ICDVRAT with Art Abilitation*, Maia, Portugal, pp. 211–218 (2008)
- [Kapuscinski and Wysocki 2005] Kapuscinski, T., Wysocki, M.: Automatic recognition of signed polish expressions. *Archives of Control Sciences* 15(3), 251–259 (2005)
- [Kapuscinski 2010] Kapuscinski, T.: Using hierarchical temporal memory for vision-based hand shape recognition under large variations of hand's rotation. In: Rutkowski, L., Scherer, R., Tadeusiewicz, R., Zadeh, L.A., Zurada, J.M. (eds.) *ICAISC 2010*. LNCS, vol. 6113, pp. 272–279. Springer, Heidelberg (2010)
- [Krais 2006] Krais, K.F.: *Advanced man-machine interaction. Fundamentals and Implementation*. Springer, Heidelberg (2006)
- [Marnik 2009] Marnik, J.: Hand shape recognition for human-computer interaction. In: Cyran, K.A., et al. (eds.) *Man-Machine Interactions*. AISC, vol. 59, pp. 95–102. Springer, Heidelberg (2009)
- [WWW-1 2010] SilverFit System, <http://silverfit.nl/en/index.htm> (accessed November 29, 2010)
- [WWW-2 2010] The AuRoRa Project, <http://www.aurora-project.com/> (accessed November 29, 2010)

EEG Biofeedback: Viability and Future Directions

J.P. Rodrigues and A. Rosa

Evolutionary Systems and Biomedical Engineering Lab, ISR-IST,
Technical University of Lisbon, Portugal
{jrodrigues,acrosa}@laseeb.org

Abstract. This chapter describes the structure of an EEG biofeedback platform focused on an efficient way for its user to learn how to self regulate cortical activity. A longitudinal study of how voluntary training of specific electro cortical activity produces any stable changes in the electroencephalogram is also presented. Correlations of these changes with short term memory are also hypothesized. The results from this study showed that it is possible to learn to change some rhythmic activity in the EEG, in this case the alpha activity, after a few feedback sessions. A positive relation between this frequency band and cognitive processes was also observed. A new technique based on the Hilbert Huang Transform is proposed for the analysis of EEG signals in biofeedback protocols. Initial observations of the results of this technique are presented.

1 Introduction

EEG biofeedback consists in the self regulation of cortical activity that in this case is represented by the electrical activity captured in the scalp, the electroencephalogram signal (EEG). To facilitate self regulation, the user is presented with some input regarding a certain aspect of his cortical activity. Although this input changes from study to study, it seems to depend mostly on the desired type of self regulation and population where it is applied [Gruzelier et al. 2006]. Concerning the last, usually there are two distinct populations: healthy subjects and patients with neurological disorders. Studies with mentally ill subjects often aim for the stabilization of unusual cortical activity and reduction of pathological behaviors whereas with healthy subjects the aim is testing the hypothesis of EEG biofeedback leading to cognitive improvement. Therefore, inputs given to patients depend on the unusual cortical activity location, frequency and time characteristics while inputs to healthy subjects are often based on previous studies about the relation of a certain cognitive function and cortical activations or inhibitions.

Nowadays the fact that EEG biofeedback is a viable mean to achieve self regulation and produce lasting changes in the cortical activity is practically irrefutable.

The support for this statement is present in a brief summary of EEG biofeedback protocols and their results in a later section. Nonetheless, the rationale basis that supports these techniques is still lacking in scientific detail in some areas. Some models have been suggested that support the effectiveness of certain protocols in the treatment of epilepsy [Sterman and Egner 2006] and put Long Term Potentiation (LTP) as a key factor for their success. Notwithstanding the validity of these models and except for the QEEG analysis, the interpretation of the EEG signal as a reflex of cortical activity continues to be tied to the typical view of static and wide frequency bands. The result is a higher uncertainty when relating the activity expressed in the EEG with some cortical activation or cognitive aspect and it has already been proven by Klimesch that with narrower and dynamic frequency bands, that adapt to each subject, it is possible to observe new dependencies between cognitive aspects and the EEG [Klimesch et al. 2008]. Therefore, it is of great importance for EEG biofeedback, and subsequent EEG analysis, to have a new method that is able to evaluate EEG more significantly and still, guarantee compatibility with the concept of frequency bands. This method should be able to identify every single oscillations amplitude and frequency without being affected by the non-stationary nature of EEG signals and their non-sinusoidal shape. Besides presenting a recent biofeedback study, this chapter proposes a different interpretation of the Hilbert Huang Transform (HHT) suitable for EEG biofeedback and probably other applications.

2 Biofeedback and EEG

Since the discovery of the self regulatory effects of EEG biofeedback in epilepsy that researchers and medical practitioners have been applying it to a wide span of clinical disorders that correlate with specific types of abnormal brain activity. Other studies also applied neurofeedback protocols to healthy subjects with the objective of studying how it improves certain cognitive capabilities that have been previously proved to produce changes in the subjects EEG. In most of these studies, biofeedback resulted from the EEG activity expressed by spectral power present in several frequency bins (QEEG) or in certain frequency bands (brainwaves), common to characteristic oscillatory brain activity, in a single location or between different locations. Other studies use the coherence between signals in different locations as a measure for biofeedback. This coherence measures are used to determine how functionally linked together two areas in the brain are by statistically measuring likelihood that two random signals arise from a common generator process for a certain frequency band. Very low coherence between two areas means that they are functionally disconnected while high values imply functional connection [Walker et al. 2007].

Therefore it is possible to separate these biofeedback approaches in separate groups: brainwave guided, QEEG guided and coherence guided.

2.1 Brainwave Guided Biofeedback

Usually, studies with this approach are focused on cognitive performance and memory enhancement on healthy subjects, cognitive performance improvement on patients with neurological disorders like attention deficit hyperactivity disorder (ADHD) or seizure reduction in epilepsy patients. The amplitude of the desired brainwaves can be obtained by filtering the EEG signal in the corresponding frequencies or by spectrum estimation using the Fourier Transform.

The discussion of the efficiency of these approaches is beyond the scope of this article however, detailed information concerning to positive outcomes in seizure reduction and cognitive improvement can be found in extensive reviews by Gruzelier [Gruzelier et al. 2006] and Sterman [Sterman et al. 2006]. It can be found that there exist different combinations of protocols ranging from those that reward the increase of a single frequency band in a single location to protocols that reward increases in different frequency bands in different locations and reward the decrease of other frequencies in the same or other locations at the same time.

Despite most of these protocols frequently use frequency band boundaries standardized by averages of the normative population, some studies choose to use individually adjusted boundaries in order to obtain results that are more representative. Klimesh proposed this last approach, driven by his previous findings that suggest positive correlation between EEG events and memory performance can only be observed if these individual boundaries are taken into account [Klimesh et al. 2008].

Focusing on the correlates of EEG and cognitive events it seems beneficial to support future studies with a more individualized, and therefore relevant, measure of EEG activity instead of relying on standardized averages that overlook significant individual differences.

2.2 QEEG Guided Biofeedback

The aim of QEEG guided biofeedback is to normalize EEG activity across cortical areas. Using activation maps for different frequencies it is possible to make a statistical comparison between individual cortical activity and average values from normative databases. This way, EEG abnormalities can be detected in a frequency and topographical view providing a guideline for the necessary biofeedback intervention for a successful normalization. The normalization is usually achieved after several sessions and in most cases is followed by a reduction of the disorder symptoms [Sterman et al. 2006; Walker et al. 2007]. This method is sensitive to the distribution of frequencies and their amplitudes across the cortical sites where the EEG is measured but the functional connectivity between these regions remains unclear.

2.3 Coherence Guided Biofeedback

As stated by Walker, several clinical disorders are characterized by unusual connectivity between cortical sites and QEEG guided biofeedback may not be enough to regularize cortical activity. Connectivity values can be calculated by frequency coherence measures given by the following:

$$c(f) = \frac{|S_{xy}(f)|}{[S_{xx}(f) \cdot S_{yy}(f)]^{1/2}} \quad (1)$$

where $S_{xy}(f)$ is the cross-spectral density between x and y and $S_{xx}(f)$ and $S_{yy}(f)$ are the auto-spectra.

These values are used in the same way as the QEEG activation maps but lack a normative coherence database as complete as the QEEG databases available and knowledge about regular functional connections needs to be taken into account [Walker et al. 2007].

Most of these approaches rely on the Fourier Transform for the analysis of the EEG time series. The Fourier Transform provides a general method for estimating the global power-frequency distribution of a given random process, assuming that the process is stationary but the EEG is known to be a non-stationary signal. Short-time Fourier transforms could be used with the expectation that shorter samples have a better chance to be stationary. Even so, this would impose poorer frequency resolution and, because the EEG is not composed of perfect sinusoids, estimated frequency bands would contain not only the frequency of the oscillations but also harmonic components and spectral leakage.

3 A Brainwave Guided Biofeedback Study

The main objective of this project was to develop an EEG biofeedback platform that allows advanced and flexible training protocols. Ensuring the platforms ease of use in a small number of sessions was essential for the feasibility of a longitudinal study to check its efficiency in producing tonic changes in the EEG by training the enhancement of activity in the individual alpha frequency (IAF) band and infer how these changes can relate to cognitive improvement.

According to M. Sterman and T. Egner, any biofeedback protocol must follow certain rules in order to be effective [Sterman et al. 2006]:

- Each training session should provide discrete trials separated by brief pauses.
- When the produced changes in the EEG meet the required condition (for example, increase band amplitude until a certain threshold) a reward stimulus must be presented. There must be minimum delay, less than a quarter of a second, between the reward situation and the reward stimulus for optimal learning to occur.
- The reward stimulus must have the highest reinforcement effect.

Also, the electrode placement should be determined by the “10-20 International System of Electrode Placement” since it is based on the location of cortical regions and uses relative metric, given that head sizes vary.

As can be seen in [Rodrigues et al. 2010], each training program can be composed of several sessions, based on the previous rules. Each trials length and number can be fixed (Figure 1) or controlled by its user (Figure 2). Letting the user decide the number of trials and how long each one lasts can be useful in the initial learning period where, at the end of each trial, the user can write down what cognitive strategies were used to change his brain activity for later use.

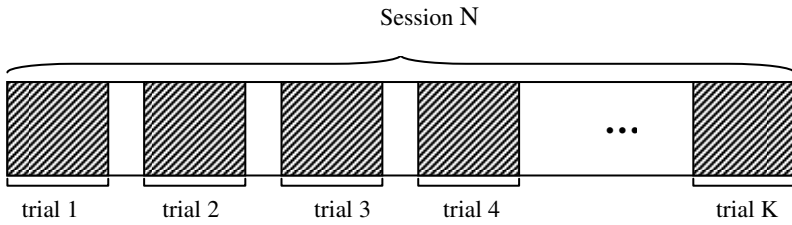


Fig. 1 Number of trials is defined in the protocol as well as their duration and intervals in between

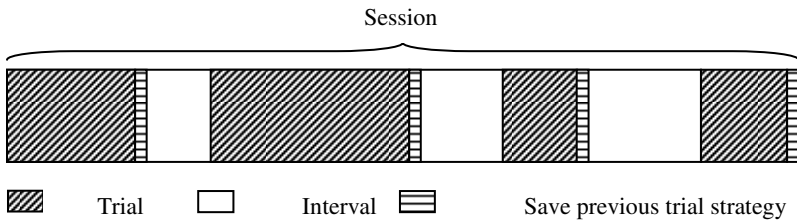


Fig. 2 Adaptation session structure. User decides when to start and end trials.

Each trial is guided by two goals. The first, Goal 1, consists in the comparison of the actual value of the feedback parameter with a predefined threshold. Here, two choices can be made in the protocol: the goal is only achieved when the feedback parameter value is above a certain threshold or it is only achieved when the feedback parameter value is below a certain threshold in case the objective is respectively enhancing or suppressing that feedback value. The second goal, Goal 2, is related to the period of time the first objective keeps being achieved continuously. If Goal 1 is being achieved continuously for more than a predefined period

of time, Goal 2 is accomplished. Each trials score is based on these two goals. It is also possible to introduce trials where the user must not try to accomplish these two goals but the feedback parameter is still being fed back to him. By comparing the results of these trials with the results from those guided by goals it is possible to see if the user is really voluntarily changing his EEG towards the objectives.

3.1 Individual Alpha Band and Short Term Memory

It has been observed that phasic synchronization in the alpha frequency band is related to cognitive idleness or cortical inhibition [Klimesch et al. 2008]. On the other hand, phasic theta synchronization can be caused by STM tasks while upper alpha desynchronization to long term memory (LTM) task. It is also possible that if STM demand is maximal, there is synchronization in the upper alpha band, suggesting active inhibition of LTM access in order to “allocate” more resources for STM. STM is measured by the digit span which is widely used in the IQ tests - Wechsler Adult Intelligence Scale.

Enhancement of alpha activity in the fixed frequency band (8.5-12.5 Hz) has already been experimented in a small group of 13 participants to test its effects on STM by using a digit span task [Vernon 2005]. The experience consisted in four sessions of one hour each. Although a significant increase in the percentage of alpha activity was achieved by the end of the four sessions, no improvements were registered in the STM tests but Vernon states this failure can result from the use of a fixed frequency band instead of individual bands based on the PAF [Vernon 2005]. Therefore in this platform it is possible to determine the new individual boundaries for each frequency band by determining the individual alpha frequency (IAF) and the peak alpha frequency (PAF). The PAF reflects the dominant or most frequent oscillation in the alpha band and it's a necessary value to adjust this frequency band between individuals. Activity in the alpha and theta band respond in different and opposite ways, when one synchronizes usually the other desynchronizes [Klimesch et al. 2008]. With increasing task demands, theta synchronizes while alpha desynchronizes the same way alpha synchronizes and theta desynchronizes when closing the eyes. This way, by plotting the EEG spectrum of a recording during a demanding task against a recording during a resting period it is possible to identify the boundaries of the individual alpha band as well as the PAF which is the frequency with the highest amplitude inside these boundaries. Another simple way to get both previous results is by plotting the spectrum of a recording where the subject has his eyes closed (alpha synchronizes and theta desynchronizes) against a recording with the subject having his eyes opened (alpha desynchronizes and theta synchronizes) like it is depicted in Figure 3.

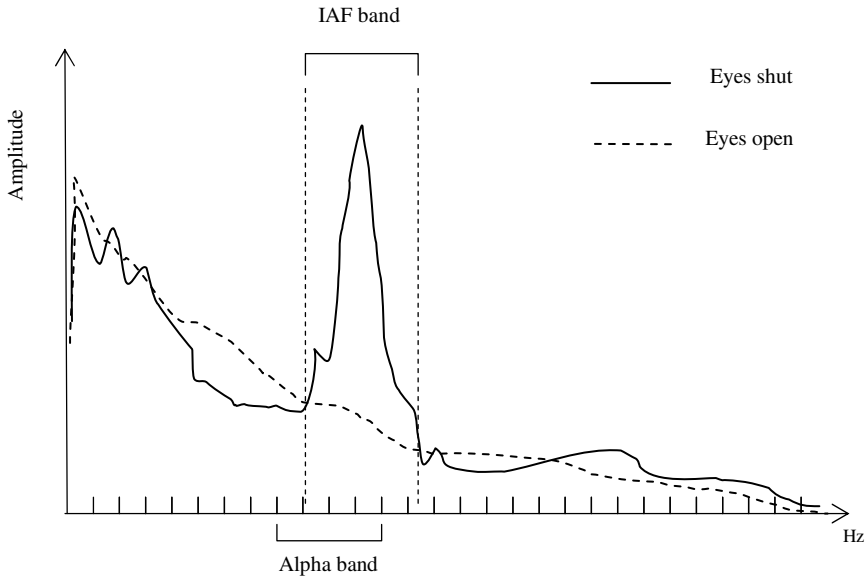


Fig. 3 Example of a case where IAF band does not coincide with the alpha band (8 – 12 Hz)

The calculation of the boundaries from the other frequency bands was based on W. Klimesch method [Klimesch et al. 2008] where fixed length bands are applied before and after the IAF band or by the same method of labeling events that induce increases or decreases in specific frequency band amplitude. For example, the individual sensorimotor rhythm (SMR) could be calculated by plotting the EEG spectrum during an event where the subject is asked to maintain motionless (SMR increases) against the EEG spectrum during an event where the subject is allowed to move his limbs if he wishes (SMR decreases) [Sterman et al. 2006]. The individualization of the frequency bands should not only be done between individuals but also between different recording sites as EEG frequencies vary between them.

Although this study [Rodrigues et al. 2010] revealed interesting results about the interaction and subsequent results of one test subject that underwent twenty biofeedback sessions, further validation with more test and control subjects is expected. For this subject, the evolution along sessions of the relative amplitude for the IAF band and three other neighbor frequency bands is plotted in Figure 4. It is evident that the amplitude in the IAF band has increased significantly more than in other frequency bands along sessions.

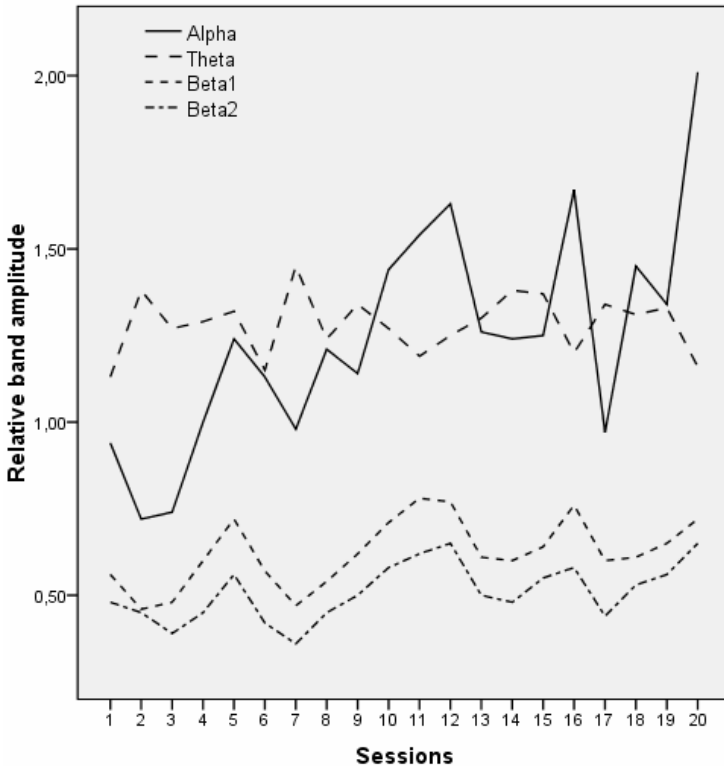


Fig. 4 Progression of the amplitude of four different frequency bands from the volunteer who participated in twenty sessions

The increasing tendency of the amplitude is confirmed applying a linear regression to the data (Figure 5). The statistically significant regression ($p < 0.05$) has an R squared value of 0.494 and is expressed as a line with a positive slope of 0.038. By the fourth session this subject started producing significantly more alpha activity and it was by that time that the subject understood how to control the feedback parameter value. After the twenty sessions baseline values of the IAF band amplitude were higher than those before the training and Goal 2 was achieved in every trial with an average delay of 8.23 seconds. Another interesting result was the increase of the digit span score found in three memory tests done along the training period and separated by approximately one month each from seven digits to ten digits. The PAF value did not change from its initial value of 10.4Hz with the training.

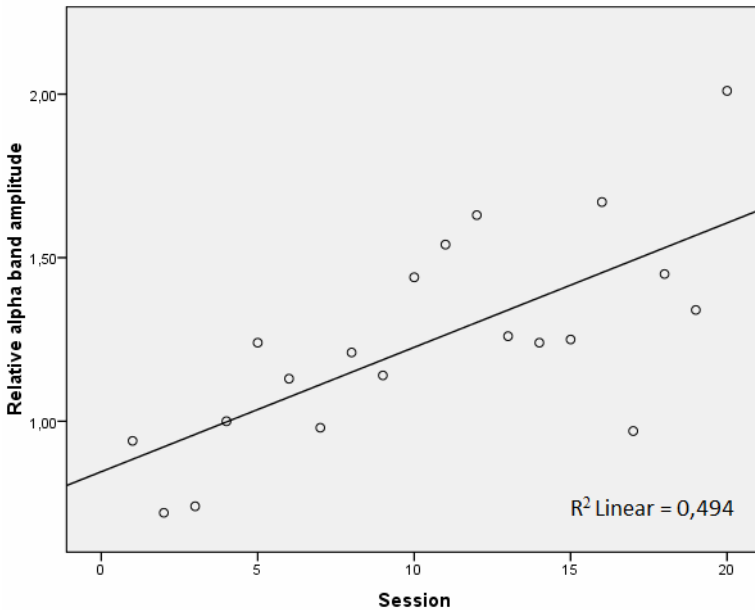


Fig. 5 Linear regression applied to the average amplitude of the IAF band data from the twenty sessions

4 Enhancing Time Frequency Resolution for EEG Biofeedback

Similarly to the Hilbert Huang Transform (HHT), our to be proposed method uses the Empirical Mode Decomposition (EMD) algorithm to decompose the time series into several modes and calculates their instantaneous phase signal. However, HHT uses this phase signal to calculate the instantaneous frequency that may not represent the frequency of the oscillation in each mode but their intrawave frequency modulation. In our method, frequency components in each mode are determined by the phase shifts in the instantaneous phase signal and this calculation does not produce intrawave frequency modulation.

4.1 Empirical Mode Decomposition

The EMD is an iterative algorithm that removes the highest frequency oscillation from the analyzed data in each iteration. After each repetition, a lower frequency information residue remains that is further decomposed until only a trend remains. The resulting components of this adaptive decomposition are the intrinsic mode functions (IMFs) and represent the intrinsic oscillations of the signal so when summed they should result in the original signal. These IMFs are defined as functions with equal number of extrema and zero crossings (at most differed by one)

with zero average between their upper and lower envelopes. As they represent a simple oscillatory mode they can be seen as the equivalent to a spectral line in the Fourier estimated spectrum with the difference that IMFs may be frequency-modulated. The EMD is a fully data-driven mechanism and doesn't require previous knowledge about the signal, contrary to filtering. Norden Huang developed it [Huang et al 1998] for the analysis of nonlinear, non stationary geophysical time series, which might have been the motivation for the data-driven character of this method, and has spread to different applications from biomedical to financial fields. Despite the stated good results of this relatively new signal analysis tool, it is a completely empirical and non-linear method without any mathematical basis until now. IMFs are restricted to have the same number of zero crossings as extrema at most differed by one so that they will have well behaved Hilbert transforms, which will allow the calculation of the signals instantaneous frequencies in a time frequency distribution (Hilbert Spectrum). Because the EEG is a non-stationary signal, with its frequency content quickly changing across time, the Hilbert Spectrum tends to be more satisfying than the classical spectral analysis like Fourier or wavelet transform providing, at least, more frequency resolution [Liang et al. 2005]. The application of this method for the EEG data analysis is also promising as it allows the detection, and separation of a wide variety of EEG recording artifacts as power line and EMG interference while preserving important characteristics of the original signal [Sweeney-Reed and Nasuto 2007]. Nevertheless, some limitations have been reported with this method. Patrick Flandrin and associates applied the EMD to a specific type of noise and observed that it behaves like a dyadic filter [Flandrin et al. 2004]. However, this can be due to the nature of the data applied and even if such problem exists, EMD would not suffer from certain limitations of band pass filtering like spurious harmonics or negative frequencies [Sweeney-Reed and Nasuto 2007]. Other limitation demonstrated by Liang and associates is that EMD fails to decompose correctly a signal composed of two intersecting chirps [Liang et al. 2005]. Again, this poses no threat to the actual paradigm of EEG aided cognitive analysis that focuses on particular frequency ranges and cognition aspects [Sweeney-Reed and Nasuto 2007].

The first step is taking the time series $m_1(n)$ and, identifying all its maxima and minima. Secondly, the identified maxima are connected by a cubic spline curve forming the upper envelope $x_u(n)$. The same process is repeated for all the identified minima originating the lower envelope $x_l(n)$. The mean between these envelopes is then calculated as a new series $m_1(n)$ where each point is valued after:

$$m_1(n) = \frac{x_u(n) - x_l(n)}{2} \quad (2)$$

This mean is subtracted from the original series resulting in the first attempt of an IMF:

$$h_1(n) = x(n) - m_1(n) \quad (3)$$

The above procedures constitute the sifting process which is used to remove riding waves from the series. Until an IMF is extracted, $h_1(n)$ will have to undergo the

sifting process until a certain stoppage criterion is met. Now the new IMF approximation will be calculated by:

$$h_{11}(n) = h_1 - m_{11}(n) \tag{4}$$

where $m_{11}(n)$ is the series with the mean values of the upper and lower envelopes of $h_1(n)$. This process is repeated k times until the stopping criteria is met, leading to:

$$h_{1k}(n) = h_{1(k-1)} - m_{1k}(n) \tag{5}$$

The sifting process ends and $h_{1k}(n)$ is the first IMF denoted as $c_1(n) = h_{1k}(n)$ supposedly contains the finest scale component of the signal. These oscillations are then separated from the rest of the data forming the residue:

$$r_1(n) = x(n) - c_1(n) \tag{6}$$

Because the residue might contain more oscillations (lower than the ones extracted in the previous IMFs), it is treated as a new signal to which all the above procedures are repeated M times, until the M -th residual is seen as constant. By now, M IMFs have been extracted. The first and last IMFs extracted correspond to the fine and coarse scales of the signal respectively. When all the IMFs and the M -th residue are added, the signal should be reconstructed almost identically:

$$x(n) = \sum_{i=1}^N c_i(n) + r_M(n) \tag{7}$$

The stoppage criterion for the sifting iterations serves two purposes: Most importantly, as a way to guarantee that the resulting IMF has a well behaved Hilbert transform and secondly to guarantee that the extracted IMF corresponds exactly to the highest oscillation component of the processed signal by eliminating riding waves in the approximation to the IMF. The first is to guarantee that the IMF has a well behaved Hilbert transform the mean of its envelope should be zero at all points and the number of zero-crossings and extrema must differ at most by one. The second is achieved when the approximations start to saturate which can be expressed as a very low difference between two consecutive sifting results:

$$SD = \sum_{n=0}^N \frac{|h_{k-1}(n) - h_k(n)|^2}{h_{k-1}^2(n)} \tag{8}$$

To guarantee that the IMF has a well behaved Hilbert transform the mean of its envelope should be zero at all points and the number of zero-crossings and extrema must differ at most by one.

4.2 Instantaneous Frequency

After all the IMFs are determined, the instantaneous frequency (IF) of each IMF at each time point can be calculated. For a given time series $z(t)$ this calculation uses its analytical signal $z(t)$ defined as [Liang et al. 2005]:

$$z(t) = x(t) + iH[x(t)] = a(t)e^{i\theta(t)} \quad (9)$$

where $a(t)$ and $\theta(t)$ are the instantaneous amplitude and phase, respectively, of the analytical signal and $H[x(t)]$ is a signal orthogonal to $x(t)$, its Hilbert transform. Now, the instantaneous frequency can be obtained by differentiating the instantaneous phase:

$$IF(t) = \frac{1}{2\pi} \frac{d\theta(t)}{dt} \quad (10)$$

For a discrete signal the instantaneous frequency can be calculated by:

$$IF(n) = \frac{1}{4\pi} [\theta(n+1) - \theta(n-1)] \quad (11)$$

However, if the IMF is not a pure sinusoidal waveform, the instantaneous frequency at a certain point may not reflect the frequency of the current oscillation, but its intrawave frequency modulation. Two oscillations with the same frequency can have different instantaneous frequencies if their waveform is not the same. Huang presents two examples where this frequency modulation occurs [Huang et al. 1998]: non sinusoidal wave (Stokes wave) and an amplitude modulated sinusoid with its envelope decaying exponentially. This has the same cause as the occurrence of harmonic distortion when using a linear system to approximate a nonlinear one, like the harmonic components in the Fourier Transform of the EEG.

4.3 Instantaneous Oscillation Frequency

To avoid intrawave modulation, the frequency of each oscillation is calculated using the time value between extremes in $\theta(t)$. These represent the phase shifts in $z(t)$ and so, each oscillation period T . In the scope of EEG analysis, the extra information in the IFs could be useful to search for a correspondence between different signatures in for IF in different recording locations thus, assigning signatures to possible macrocolumn sources. For the scope of this work the most important information is each oscillations frequency and amplitude. Therefore, if a complete oscillation is present in an IMF there is no need to calculate its instant frequency for every point. This calculation is left for the IMFs extremities where complete oscillations are not present. For the rest of the signal the instantaneous oscillation frequency (IOF) is calculated.

This approach was not seen in the reviewed literature and consequently needs further testing for its validity in the analysis of non-linear and non-stationary signals such as the EEG. However, our findings suggest it can be very promising not only for signal analysis but also for EEG Biofeedback.

In Figure 6 an EEG recording in Cz corresponding to an eyes closed state is decomposed into its IMFs. The real time region of one second is used for frequency analysis using the spectrum estimation using the FFT (Figure 7) and the IOF as described above (Figure 8). In Figure 6 it is possible to identify a phenomenon similar to the explained above of amplitude modulated sinusoids with increasing and decreasing envelope in the IMFs 1 and 2.

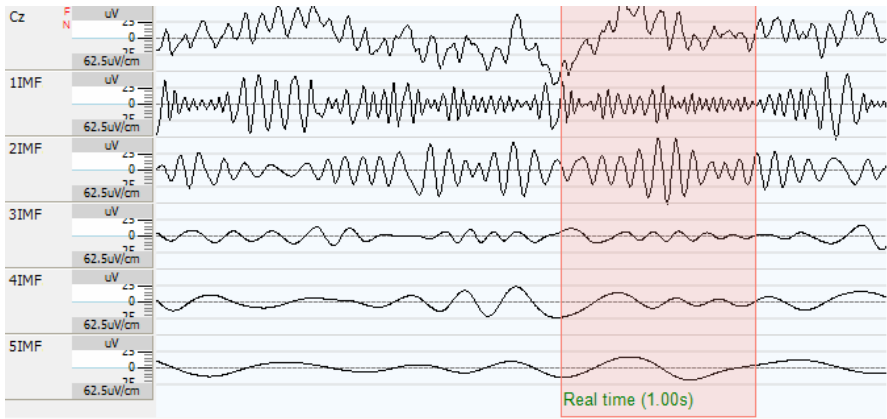


Fig. 6 EEG signal being decomposed into its constituting IMFs in the *Somnium* software application. The real time region is set to an arbitrary position but it can be set to acquire the signal as it is received from the driver and use it in real time for EEG biofeedback. Signal in Cz is notch filtered for 50Hz and band passed for 0.5Hz to 30Hz just for display. EMD uses the unfiltered signal and the first IMF – 0IMF – is omitted as it contains the mains hum

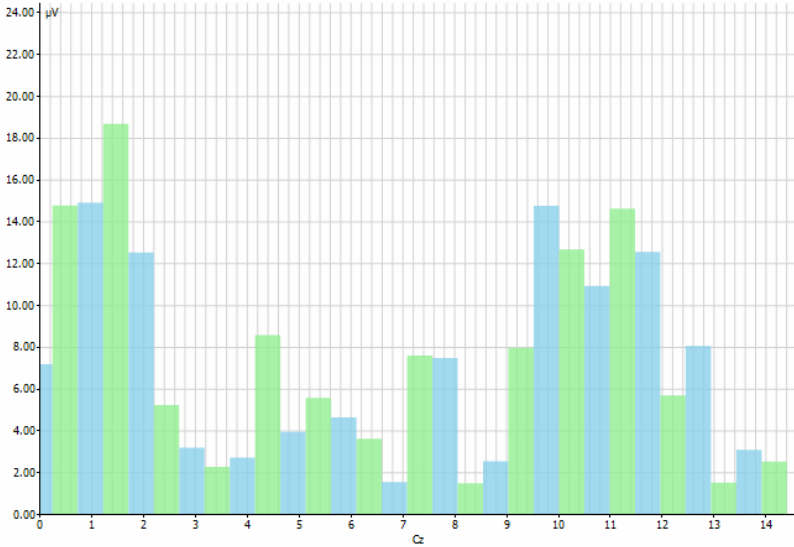


Fig. 7 Spectrum estimation using a 512 point FFT for the signal Cz in the 1 second real time region in Figure 6. Frequency resolution is 0.5Hz due to the 512 points FFT and sample rate of 256 samples per second

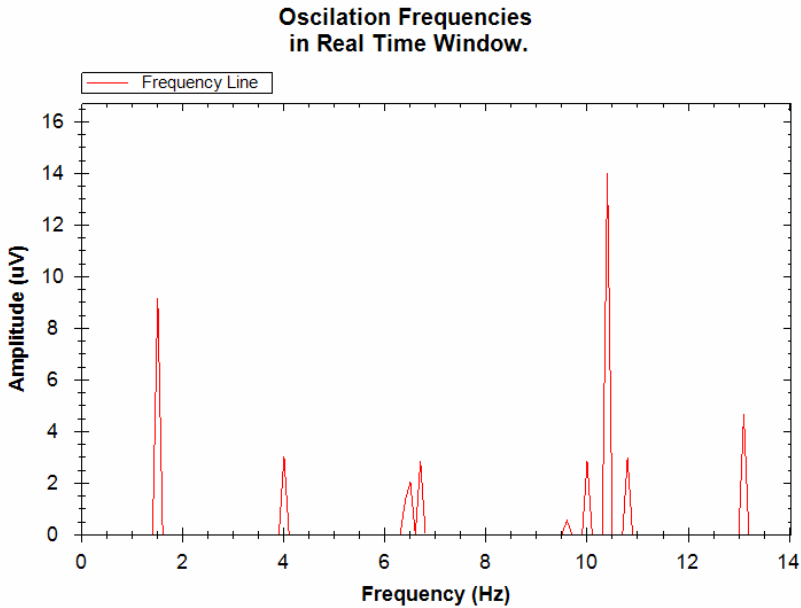


Fig. 8 IOF with frequency resolution of 0.1Hz for the signal Cz in the 1 second real time region in Figure 6

The frequency resolution for the IOF was set to 0.1Hz still, any other value could be chosen as it does not depend on the sampling rate. Only the frequency band depends on the sample rate as stated by the Nyquist Theorem. With a one second window it is also not possible to detect frequencies below 1Hz with the IOF as the complete oscillation will not be captured by the window. For that purpose, the window length can be increased.

4.4 Concluding Remarks about IOF and EEG Biofeedback

Concerning the application of this technique for EEG biofeedback it is important to assure small calculation times due to the real demand. In an Intel Core i7 920 @2.67Ghz the IOF for the example above was calculated in 66 milliseconds and for a 2 seconds window it took 240 milliseconds. For a one second signal composed of two sinusoids with different frequencies it took 5 milliseconds to calculate their IOFs. Unfortunately, it is not easy to predict the time necessary for this calculation because besides depending on the length of the signal it also depends on its complexity as it influences the number of IMFs to be extracted. Yet, there is the possibility of parallelizing the process and start calculating the IOF in a separate process immediately after its respective IMF is available.

Interestingly, the highest oscillation present in Figure 8 has the same frequency as the PAF for that subject and is present in Cz in the entire period of the eyes closed condition.

5 Future Directions for EEG Biofeedback

This chapter provided a brief introduction to different types of EEG biofeedback and a study focused on a specific subtype guided by the IAF band. Albeit the developed platform followed some advices present in the literature that are essential for the efficacy of any biofeedback protocol aiming to produce lasting changes in the EEG, it separated from some pre-established concepts such as standard frequency bands or fixed trials and sessions. In fact, providing a set of initial sessions that guide the user and advices about the best cognitive strategies to achieve EEG self regulation was essential. By the end of these sessions the subject that achieved control and lasting changes over the IAF already knew what cognitive strategies to apply and in the following sessions could be focused on the training rather than trying different strategies. Therefore, it is legit to propose this step in any of the biofeedback protocols discussed in the beginning.

Other details like the use of immersive and intuitive environments for the biofeedback display are also possible candidates for the requirements of any protocol.

Further support for the benefits of individualized frequency bands can also be concluded by this study. However, this only affects brainwave guided biofeedback protocols as the others rely entirely on QEEG databases or knowledge about the brain functional connectivity.

Finally, the use of a precise measure for EEG activity should be considered. As explained previously, the use of FFT based methods can present several drawbacks in this area despite their simplicity and massive use. We presented a modification to an already existing method, focusing on a single aspect of the EEG rather than accounting for all the signals information. It tries to ignore intrawave modulation and measure only the period of each oscillation, sinusoid or not, that composes the time series. The rationale under this decision is that most studies about EEG whilst using FFT based techniques focus more on the oscillatory nature of the EEG and less on other characteristics as the shape of the oscillations. Additionally, very few or none information exist about the correlates of intrawave modulation and cognitive processes.

References

- [Flandrin et al. 2004] Flandrin, P., Rilling, G., Goncalves, P.: Empirical mode decomposition as a filter bank. *IEEE Signal Processing Letters*, 111–114 (2004)
- [Gruzelier et al. 2006] Gruzelier, J., Egner, T., Vernon, D.: Validating the efficacy of neurofeedback for optimising performance. *Event-Related Dynamics of Brain Oscillation* 159, 421–431 (2006)

- [Huang et al. 1998] Huang, N.E., Shen, Z., Long, S.R., Wu, M.L.C., Shih, H.H., Zheng, Q.N., Yen, N.C., Tung, C.C., Liu, H.H.: The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proc. of the Royal Society of London Series A-Mathematical Physical and Engineering Sciences* 454(1971), 903–995 (1998)
- [Klimesch et al. 2008] Klimesch, W., Freunberger, R., Sauseng, P., Gruber, W.: A short review of slow phase synchronization and memory: Evidence for control processes in different memory systems? *Brain Research* 1235, 31–44 (2008)
- [Liang et al.] Liang, H.L., Bressler, S.L., Buffalo, E.A., Desimone, R., Fries, P.: Empirical mode decomposition of field potentials from macaque V4 in visual spatial attention. *Biological Cybernetics* 92(6), 380–392 (2005)
- [Rodrigues et al. 2010] Rodrigues, J.P., Migotina, D.G., da Rosa, A.C.: EEG training platform: Improving Brain-Computer Interaction and cognitive skills. In: *Proc IEEE 3rd Conf. on Human System Interaction*, Rzeszow, Poland, pp. 425–429 (2010)
- [Serman and Egner 2006] Serman, M.B., Egner, T.: Foundation and practice of neurofeedback for the treatment of epilepsy. *Applied Psychophysiology and Biofeedback* 31(1), 21–35 (2006)
- [Sweeney-Reed and Nasuto 2007] Sweeney-Reed, C.M., Nasuto, S.J.: A novel approach to the detection of synchronisation in EEG based on empirical mode decomposition. *J. of Computational Neuroscience* 23(1), 79–111 (2007)
- [Vernon 2005] Vernon, D.J.: Can neurofeedback training enhance performance? An evaluation of the evidence with implications for future research. *Applied Psychophysiology and Biofeedback* 30(4), 347–364 (2005)
- [Walker et al. 2007] Walker, J.E., Kozlowski, G.P., Lawson, R.: A modular activation/coherence approach to evaluating clinical/QEEG correlations and for guiding neurofeedback training: modular insufficiencies, modular excesses, disconnections, and hyperconnections. *J. of Neurotherapy* 11(1), 25–44 (2007)

EEG Signal Processing for BCI Applications

A. Roman-Gonzalez

Department of Electronics Engineering, Universidad Nacional San Antonio
Abad del Cusco, Peru
a.roman@ieee.org

Abstract. In this article we offer a communication system to people who undergo a severe loss of motor function as a result of various accidents and/or diseases so that they can control and interact better with the environment, for which a brain-computer interface has been implemented through the acquisition of EEG signals by electrodes and implementation of algorithms to extract characteristics and execute a method of classification that would interpret these signals and execute corresponding actions. The first objective is to design and construct a system of communication and control based on the thought, able to catch and measure EEG signals. The second objective is to implement the system of data acquisition including a digital filter in real time that allows us to eliminate the noise. The third objective is to analyze the variation of the EEG signals in front of the different tasks under study and of implementing an algorithm of extraction of characteristics. The fourth objective is to work on the basis of the characteristics of the EEG signals, to implement a classification system that can discriminate between the two tasks under study on the basis of the corresponding battles.

1 Introduction

The work presented in this paper is based on [Roman-Gonzalez 2010a; Roman-Gonzalez 2010b]. There are a significant number of people suffering from severe motor disabilities due to various causes, high cervical injuries, cerebral palsy, multiple sclerosis or muscular dystrophy. In these cases the communication systems based on brain activity play an important role and provide a new form of communication and control, either to increase the integration into the society or to provide to these people a tools for interaction with their environment without a continued assistance. There are various techniques and paradigms in the implementation of brain-computer interfaces (BCI). A brain-computer interface is a communication system for generating a control signal from brain signals such as EEG and evoked potentials. The Communication between the two essential parts of BCI (brain and computer), is governed by the fact that the brain generates the command and the

computer must to interpret [Roman-Gonzalez 2010a]. The amyotrophic lateral sclerosis (ALS) is a progressive neurodegenerative disease and is characterized by the death of motor neurons, which turns in a loss of control over voluntary muscles [Wolpaw et al. 2002; Kuo-Kai et al. 2010]. A stroke or other accident can lead to degeneration of parts of the brain, which makes people unable to communicate more with the environment, they have the same cognitive abilities, this is what is known as Syndrome "Locked-In" in France there is approximately 500 patients with this syndrome and about 8000 and 9000 patients with ALS, data published in [Lecocp and Cabestaing 2008]. To measure and study the brain activity signals, there are different methods such as: magnetic resonance imaging (MRI), computed tomography (CT), the ECOG scale, single photon emission computed tomography (SPECT), positron emission tomography (PET), magnetoencephalography (MEG), functional MRI (fMRI), but these signals are not practical to implement a human-machine interface, because some are only anatomical information, other techniques are very invasive, others are a lot of exposure to radiation and another are very expensive [Kirby]. To work with electroencephalographic (EEG) is the most convenient and therefore the BCI is based on detecting the EEG signals associated with certain mental states.

The paper is structured as follow: Section 2 presents an overview on the theory. Section 3 shows practical application. Finally Section 4 reports our conclusions.

2 Theoretical Background

2.1 The Electroencephalogram (EEG)

The electroencephalogram (EEG) is a study of brain function that reflects the brain's electrical activity. To collect brain electrical signal using electrodes placed on the scalp, which is added a conductive paste to enable the brain electrical signal, which is of a scale of microvolts, can be recorded and analyzed. EEG signals have different rhythms within the frequency band with the following characteristics: [Roman-Gonzalez 2010a; Kirby 2004].

Rhythm Alfa or Mu: It is characteristic of the state of consciousness and physical and mental rest with the eyes closed.

- Low voltage (20-60 μv /3-4mm) with variable morphology.
- High frequency (8-13 Hz).
- Zones of origin: later.
- Visual blockade before palpebral opening and stimuli (reactivity).

Rhythm Beta: It is characteristic of the state of consciousness in states of cortical activation (replace of α).

- Low voltage (10-15 μv /1-1.5 mm) with variable morphology.
- High frequency (13-25 δ + Hz) to greater predominant frequency in anxious and unstable subjects.
- Zones of origin: central frontals.

Rhythm Theta: It is characteristic of the state of deep and normal sleep in the childhood (10 years), abnormal during the state of consciousness.

- Appearance in specific physiological conditions (hyperventilation and deep sleep).
- High voltage (50 μv /7mm).
- Low frequency (4-8 Hz).
- Zones of origin: thalamic zones, parietotemporal region.

Rhythm Delta: It is characteristic of indicative pathological states of neuronal difficulty (comma) and occurs during deep sleep.

- High voltage (70–100 μv /9-14 mm) with variable morphology.
- Low frequency (4 - δ Hz).
- Subcortical origin (not defined).

In the EEG signals, can be observed what is called evoked potentials, these evoked potentials is a neurophysiologic examination that assesses the role of acoustic sensory system, visual, and somatosensory pathways through evoked responses to a stimulus known and standardized. There are several types of event-related evoked potentials (ERP) and visual evoked potential (VEP) evoked potentials acoustic (PEA), motor evoked potentials (MRP), Steady State Visual Evoked Responses (SSVEP), etc. which are discussed in articles [Wolpaw et al. 2002 ; Lecocq and Cabestaing 2008].

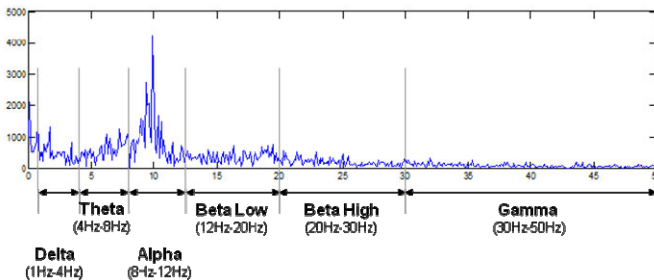


Fig. 1 EEG rhythms in time and frequency domain [Kuo-Kai 2010; Kirby 2004]

2.2 International System of Positioning Electrodes 10/20

Although, there are several different systems (Illinois, Montreal, Aird, Cohn, Lennox, Merlis, Oastaut, Schwab, Marshall, etc.), the 10/20 international system is the most widely used at present. To place the electrodes according to this system proceeds as follows:

The inactive or common electrode is placed remote of the skull (earlobe, nose, or chin). It is counted on data points such as: nasion and inion. Ten percent of the data points are the prefrontal and occipital planes. The rest is divided in four equal parts of 20% each.

Five cross-sectional planes exist: Prefrontal (Fpz), frontal (Fz), vertex (Cz), parietal (Pz) and occipital (Oz).

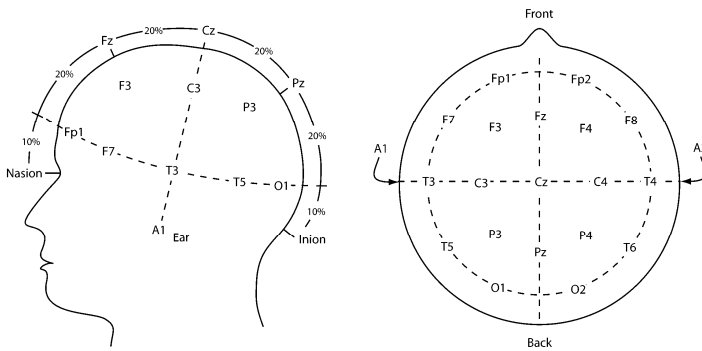


Fig. 2 Positioning of the electrodes

The number of electrodes used and the position, depends on the particular signal that we want to analyze. The oscillation of the sensorimotor cortex, changes dynamically the execution of the movement of a member:

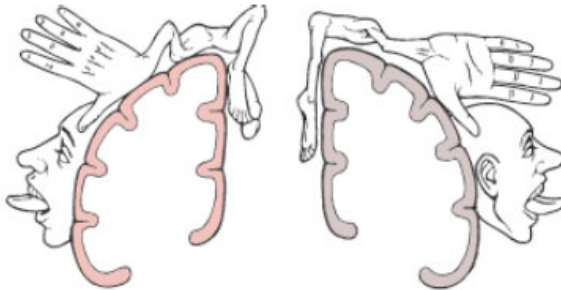


Fig. 3 Sensory and Motor Homunculus [Solis-Escalante and Pfurtscheller 2009]

2.3 The Brain Computer Interfaces

A brain computer interface is a communication system that can generate control signals from brain signals, i.e. a BCI is a system that translates brain activity into commands for a computer or other device. A BCI allows users to interact with their environment using just brain activity, without using nerves and muscles.

A general block diagram for a brain-computer interface is shown below:

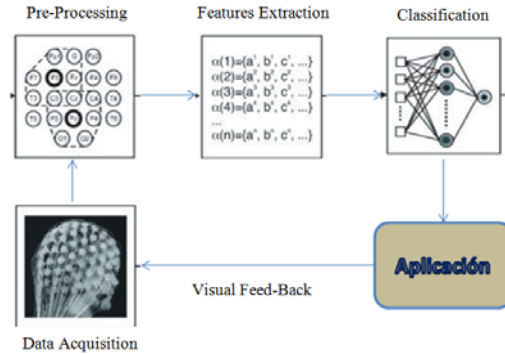


Fig. 4 General block diagram for a BCI

2.4 Asynchronous Interfaces

This kind of interface analyzes the user's voluntary activity; this analysis maintains a continuous communication link with the system. In this case, the system continuously analyzes the signals from the user's brain activity and classifies mental status periodically. In other cases, the interface can measure temporal variations in the rates associated with motor activity of the user; such amplitude variations can be detected and then transformed into commands. The analysis of motor activity requires lengthy training.

Spontaneous brain activity produces the following types of signals that are used in interfaces [Lecocq and Cabestaing 2008]:

1. Slow Cortical Potential Shifts (SCPS).
2. Oscillatory activity sensorimotor.
3. Spontaneous EEG signals.

2.5 Synchronous Interfaces

This type of interfaces analyzes EEG signals evoked by potential stimuli received by the user from the system (can be visual, auditory or tactile), in this case is the

system that performs the task of communication, the user simply react or not to a series of stimuli. In this case do not work with spontaneous brain activity, if not rather with the brain's response to stimuli and then transform this response commands. For such interfaces requires a limited learning.

The main types of signals that are used in these synchronous interfaces are [Lecocq and Cabestaing 2008]:

1. Steady State Visual Evoked Responses (SSVERs).
2. Event Related Potentials (ERPs).

2.6 Invasive or Noninvasive Interfaces

The signals of brain activity that can be measured can be signs at the scalp as the electroencephalogram (EEG) can be at the level of the cerebral cortex as the electrocorticogram (ECoG) or the need for implanting electrodes into the brain. Then we distinguish the invasive methods such as those that require the installation of electrodes inside the skull. Noninvasive methods are those that can measure signals only from the surface of the scalp [Lecocq and Cabestaing 2008]. In the invasive methods, when an electrode is connected directly to a neuron, it measures its post-synaptic electrical activity and / or the potential cast for its axon [Lecocq and Cabestaing 2008].

The most used non-invasive technique is to work with the EEG signals collected from electrodes placed on the scalp.

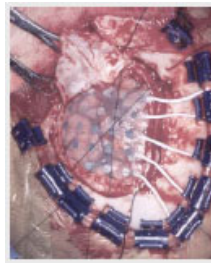


Fig. 5 Invasive method for measuring brain activity [Lecuyer 2007]

2.7 BCI P300 Speller

This kind of BCI was originally proposed by Farwell and Donchin in 1988 and is also studied in [Lecocq and Cabestaing 2008], is a non-invasive communication

interface based on event-related evoked potentials ERPs P300 type. This interface allows the user to write a text on the computer, is a 6x6 matrix that is displayed on the screen and is made up of 26 letters of the alphabet, nine numbers and a symbol that enables the cancellation of the previous selection.

The P300 speller is based on a paradigm which consists of presenting stimuli in the form of lighting in each row or column. The user's task is to take attention to the character to select and count the times that is affected by lighting. The illuminations are done in a random and repeated several times for each character.



A	B	C	D	E	F
G	H	I	J	K	L
M	N	O	P	Q	R
S	T	U	V	W	X
Y	Z	1	2	3	4
5	6	7	8	9	

Fig. 6 P300 Speller matrix

2.8 Wheelchair Control with BCI

Currently there are several research teams working to develop and improve the control system of a wheelchair based on measurements of the EEG signals of brain activity in patients with severe loss of motor activity. In this area, one of the first to submit a rough prototype wheelchair controlled by EEG signals was by Tanaka in [Tanaka et al. 2005] and is also studied in [Lecocq and Cabestaing 2008]. Tanaka used a noninvasive BCI asynchronous analyzing EEG signals between 0.5 and 30 Hz, in the training phase of the system the user must imagine the movement left and right for 20 seconds for each move, the acquisition is made at 1024 Hz and based on these signals the system learns to discriminate between both types of movement.

One of the latest studies in relation to control a wheelchair with EEG signals was introduced by Toyota [Toyota 2009]. This system has the capacity to analyze the EEG wave signal every 125 milliseconds and decide whether to turn left, turn right or forward. The analyzed waves are shown in real time on the computer screen to give visual feedback. This system uses a cheek movement to slow or stop the wheelchair; this movement can be made by an accumulation of air in that area.



Fig. 7 Toyota wheelchair controlled by BCI [Toyota 2009]

Another work with wheelchair control based on EEG is done by the project OpenViBE [Lecuyer 2007]. OpenViBE is a free platform to develop BCI applications, within these different applications was a control of a wheelchair, for which uses electrodes at positions C3 and C4 of the international position of electrodes 10/20 to capture the signals of intention to move left or right hand and thus represent the rotation the wheelchair to the right or left respectively, for EEG signals representing the movement of feet, an electrode is placed in the front and thus represents the advancement of the wheelchair. In a first moment is perceived to be very difficult to handle the wheelchair with these premises, so in a second experiment using the signal from the feet to select from several target destinations, so once you select your destination, as Wheelchair uses other algorithms to get to your chosen destination and progress.

3 Development of the Work

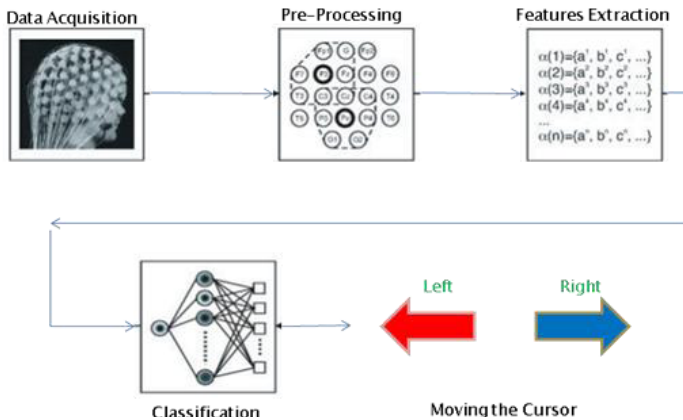


Fig. 8 Block Diagram

3.1 Data Acquisition

For the data acquisition, electrodes of 8 mm of Ag/AgCl fixed on C3 and C4 of the international system of positioning 10/20 were used. Signal amplification was made through amplifier EEG of 8-channel model Procomp Infinity. The sampling frequency is of 256 Hz. A digital band-pass filter has been implemented between 0.5 and 30 Hz in real time to especially eliminate the noise originating from the mains and other sources.



Fig. 9 Electrodes



Fig. 10 Amplifier Procomp Infinity

3.2 Features Extraction

The stage of extraction of characteristics is probably the most critical step in the processing of signal EEG, with a view to maximizing the potential success of the classification stage as well as the global yield of the system. A second objective of the stage is to compress the data without loss of excellent information through the process of classification so that it can operate in real-time. The rhythm μ , which corresponds to an oscillation of signal EEG between the 8 and 13 Hz, is caught in the sensorimotor zone located in the central hairy region. This rhythm, present in most adults, has particularity to present attenuation in its amplitude when some types of movement are performed, or what is more important when the intention is had to realize some movement, or simply imagining movements of the extremities, as shown in Figures 11 to 14.

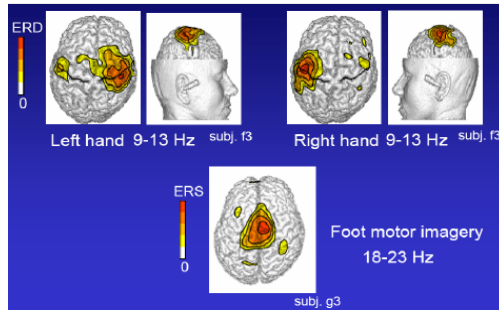


Fig. 11 Cerebral activity during the imagination of movements of the right hand and left hand

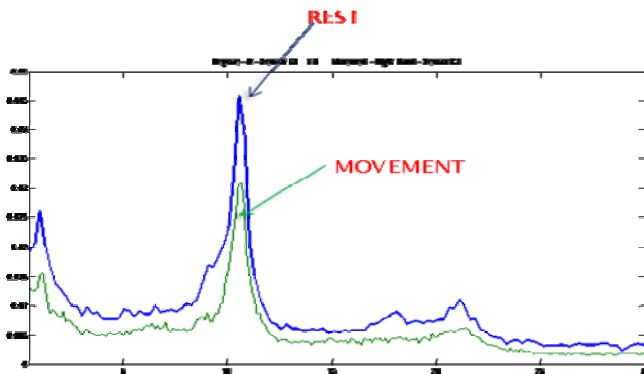


Fig. 12 Difference in the frequency band alpha between movement and rest

It is possible to stress that movements of the right hand produce a variation in the activity of the left part of the brain and vice versa.

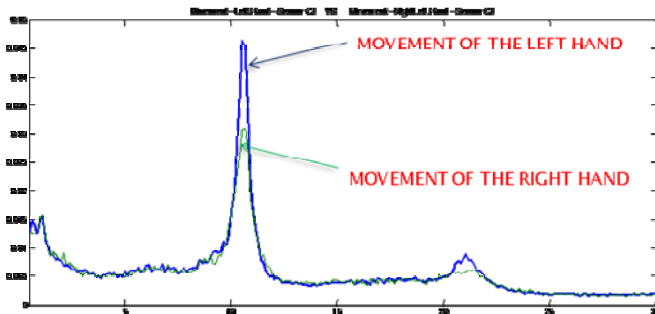


Fig. 13 Difference in the frequency band alpha between movement of the right hand and left hand in the electrode of the position C3

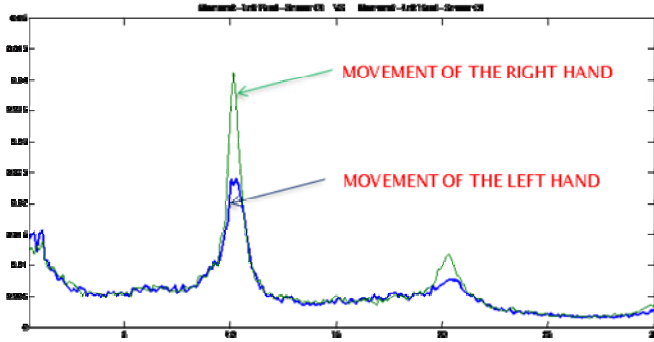


Fig. 14 Difference in the frequency band alpha between movement of the right hand and left hand in the electrode of the position C4

To quantify these characteristics (visibly observable) and then apply the classification methods, we use two different sets of characteristics: The first is a set of autoregressive parameters that represent spectral analysis, and the second will be to obtain spectral energy in Mu and Beta band for each electrode.

3.3 Autoregressive Adaptive Parameters (AAR)

To represent the characteristics previously described in numbers that allow us to implement a sort algorithm because we used autoregressive adaptive parameters that allows us to represent the frequency response of the signal as shown in Figure 15. A model AAR of order p is written as follows:

$$\begin{aligned}
 y(t) &= a_1(t) * y(t-1) + \dots + a_p(t) * y(t-p) + x(t) \\
 &= a(t)^T * Y(t-1) + x(t)
 \end{aligned}
 \tag{1}$$

The difference with the stationary molding autoregressive (AR) is that parameters AAR vary with them; the prediction of the error is calculated as follows:

$$e(t) = y(t) - \hat{a}(t-1)^T * Y(t-1)
 \tag{2}$$

For parameter calculation, we used the method of least mean squares (LMS), which is given by the following equation:

$$\hat{a}(t) = \hat{a}(t-1) + (UC / MSY) * e(t) * Y(t-1)
 \tag{3}$$

where:

$$UC \rightarrow \text{Update Coefficient} = 0.0055$$

$$MSY \rightarrow \text{Signal Variance} = \frac{1}{N} \sum_{t=1}^N Y_t^2$$

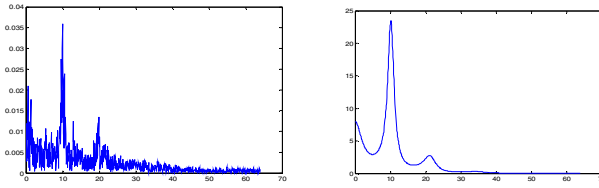


Fig. 15 Comparison of frequency response with the FFT and parameter AAR

With 6 parameters AAR at each electrode, there are a total of 12 characteristics.

3.4 Spectral Energy in Mu and Beta Band (PST)

In this case we will calculate the spectral energy in the Mu band (8-13 Hz) and Beta band (17-24 Hz) for each electrode (positions C3 and C4) which is why we will have in total a set of 4 features. The analysis is performed continuously in each moment of time as shown in Figure 16, we take a window of 1 second, this window will move in every moment of time in each window the work being done is to filter the signal first with a bandpass filter of 8-13 Hz and then calculate the spectral energy in the band Mu using the equation (4):

$$PST = \frac{1}{N} \sum_{t=1}^N Y_t^2 \tag{4}$$

where:

PST = Spectral energy.

N = 256, as the windows is 1 sec. and de sampling frequency is 256 Hz

Then we filter the window with a bandpass filter between 17 and 24 Hz to calculate the energy in the Beta band.

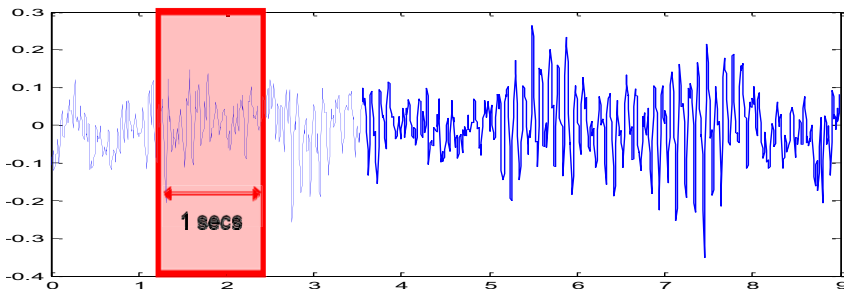


Fig. 16 Window signal to obtain the spectral energy in the Mu and Beta Band Leer fonéticamente

3.5 Classification

The phase of classification is the final task of the process. The entrance to the sort algorithm is the set of characteristics extracted in the previous stage, and the exit is an indication of the mental state of the user. In this case, we are working with two states: left and right.

For the present work, two methods of classification were developed: linear discriminating analysis and neuronal network. Both methods give similar results of a constant weight vectors; this way, the activation function would be as follows:

$$AC = \sum x_i * w_i + cte \tag{4}$$

$$AC = X * W + cte$$

4 Test Process

4.1 Fixation of Electrodes

Bipolar electrodes are used; each electrode is placed at 2.5 centimeters toward the back and at positions C3 and C4, as shown in Figures 17 and 18.

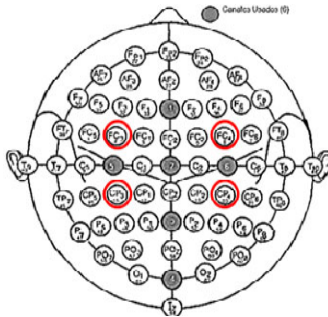


Fig. 17 Fixation of the bipolar electrodes in C3 and C4



Fig. 18 Photographs with the fixed electrodes

To fix the electrodes, gel and conductive grease were used.



Fig. 19 Gel and conductive grease

4.2 Acquisition of the Signal and Training

Each of the tests lasts for only 9 seconds, and during the training process, we performed 60 tests. The test begins at rest, and after 3 seconds, the system randomly chooses a value to send to the right or left signal. This is why the person will have the 6-second rest to imagine the movement specified (for better understanding, please refer to Figure 20).

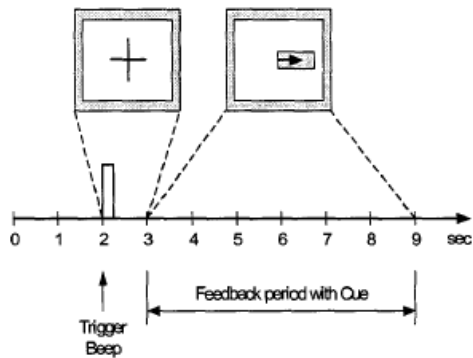


Fig. 20 Composition of the 9 seconds of the test [Schlogl et al. 2003]



Fig. 21 Photographs during the process of acquisition of EEG signals

For the training, two stages were performed. First is offline training, where there is no feedback; this serves to register and keep the data for analysis. Second is online training where feedback regarding function to the preliminary results of the offline analysis exists; this training serves so that the user can learn to control the cerebral activity more effectively. In Figures 22 and 23, we can observe the forms implemented for each of the trainings.

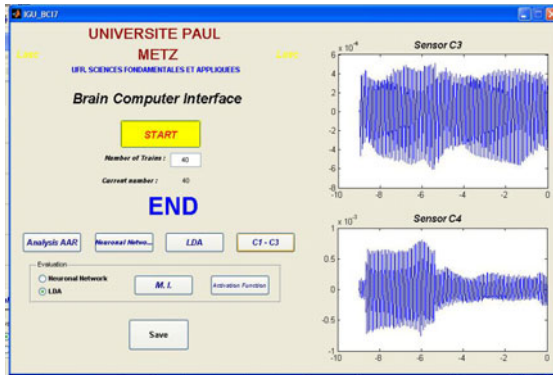


Fig. 22 Interface for offline training

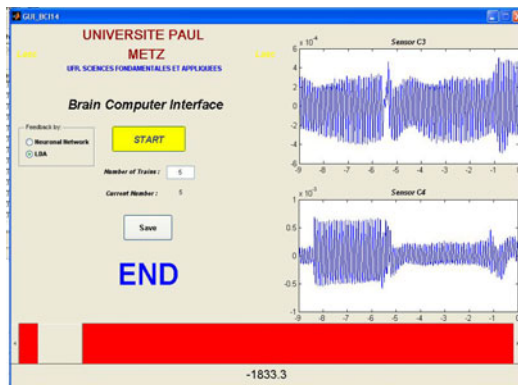


Fig. 23 Interface for online training with feedback

5 Results

To be able to evaluate the obtained results, 2 methods were taken into account:

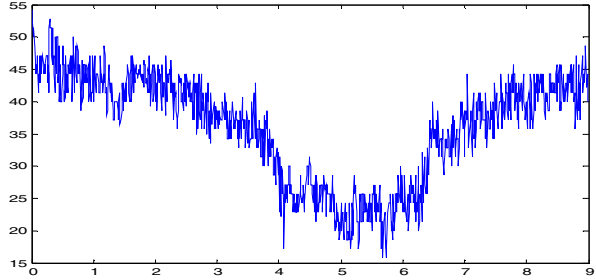
Error Rate. This is the error that takes place when trying to classify the produced signals both enters types of tasks under study (movement of the right hand or the left hand).

Mutual Information. This is the amount of information that can be recovered through classification and the extracted characteristics.

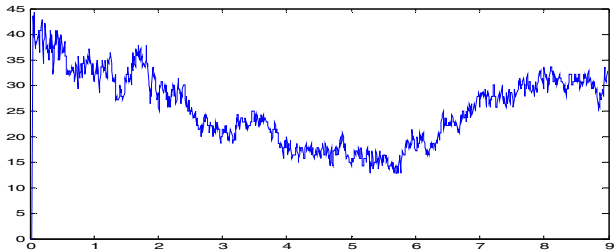
We have worked with 2 data bases: Graz Data Base and Metz Data Base. As the analysis was performed in a continuous manner because the evaluation was realized in every moment of time with each sample of collected, we can observe the values of the error and mutual information based on time below:

Results for Graz Data:

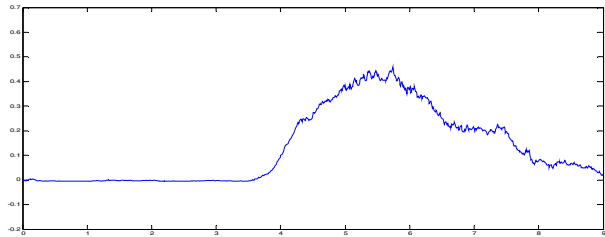
AAR, p = 6, Method LMS, ERROR RATE
UC = 0.0055
Neural Network
Min = 15.7143 %
Time = 5.6797 sec



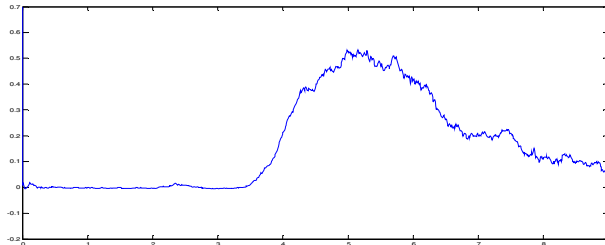
AAR, p = 6, Method LMS, ERROR RATE
UC = 0.0055
LDA
Min = 12.8571 %
Time = 5.2969 sec



AAR, p = 6, Method LMS, MI
UC = 0.0055
Neural Network
Max = 0.4583
Time = 5.7422 sec



AAR, p = 6, Method LMS, MI
UC = 0.0055
LDA
Max = 0.5328
Time = 5.1563 sec

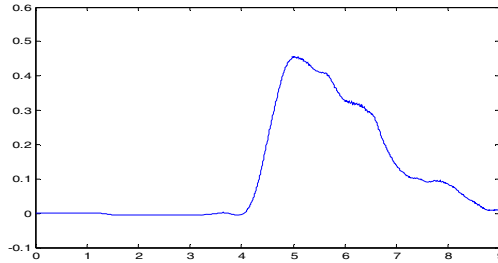


The difference between the spectral energy of the band can be graphical alpha (8 – 13 Hz) of C3 and C4 when types of movement both take place. Furthermore, we can graph the function of the resulting activation of the classification method that is obtained when the movements is either left or right. Both graphs are based on time because of the continuous analysis that I am realized in every moment of time, as shown in Figures 24 and 25.

Using the spectral energy (PST) we have the following results:

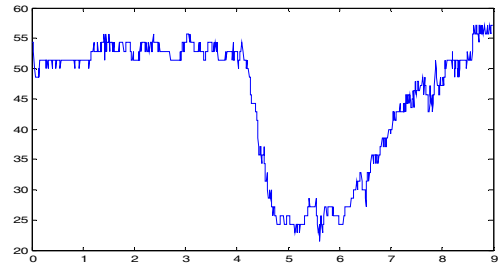
Method PST, **MUTUAL INFORMATION**

Neural Network
Máx = 0.4567
Time = 5.0469 sec
Files TRAIN



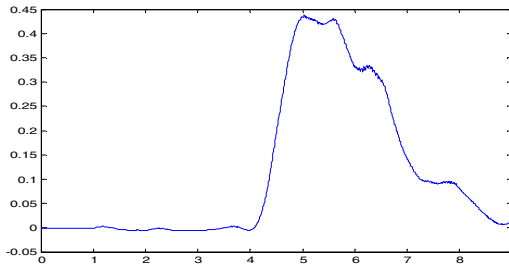
Method PST, **ERROR RATE**

Neural Network
Min = 21.4286 %
Time = 5.6250 sec
Files TRAIN



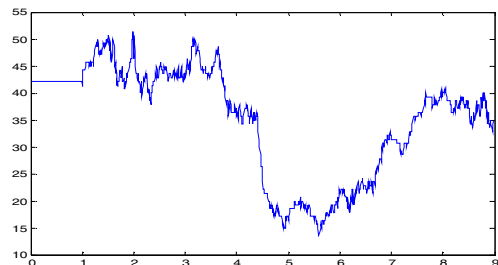
Method PST, **MUTUAL INFORMATION**

LDA
Max = 0.4369
Time = 5.0469 sec
Files TRAIN



Method PST, **ERROR RATE**

LDA
Min = 13.5714 %
Time = 5.5938 sec
Files TRAIN



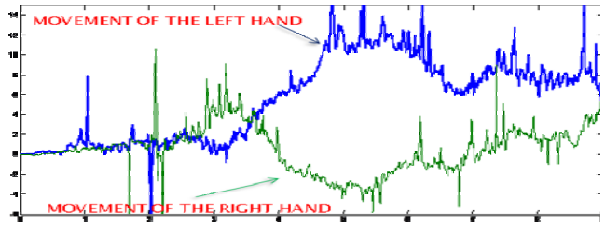


Fig. 24 Difference of the spectral energy between the electrodes C3 and C4

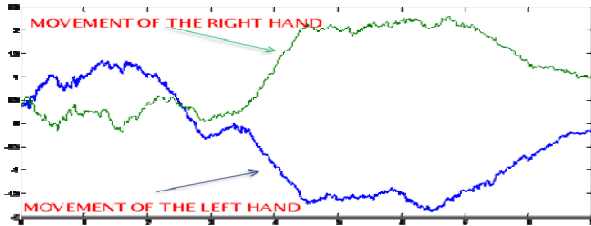


Fig. 25 Graph of the function of the activation for movements of the right hand and left hand

Through these graphs, it can be observed that the classes under study are separable.

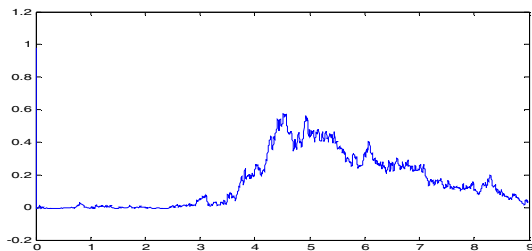
After making this first step OFFLINE training and have analyzed the data, we turn to step ONLINE training using the vector and the constant found in the previous step based on the results shown. As a result we obtained the following confusion matrix

Confusion Matrix

	Classe	Right	Left	TOTAL
	Right	59	11	70 (*)
	Left	8	62	70 (+)
%	Right	84.29	15.71	100
	Left	11.43	88.57	100

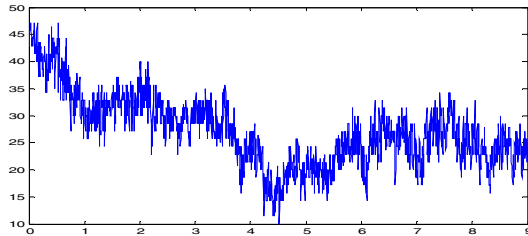
Results for Metz Data:

Method AAR,
MUTUAL INFORMATION
 Neural Network
 Máx = 4.5234
 Time = 0.5751 sec
 Files TRAIN



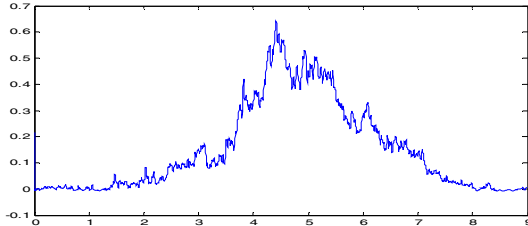
Method AAR, **ERROR RATE**

Neural Network
 Min = 10 %
 Time = 4.5156 sec
 Files TRAIN



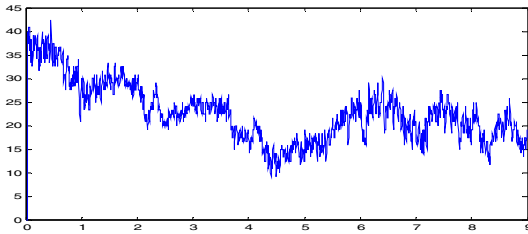
Method AAR, **MUTUAL INFORMATION**
LDA

Max = 0.6433
 Time = 4.4141 sec
 Files TRAIN



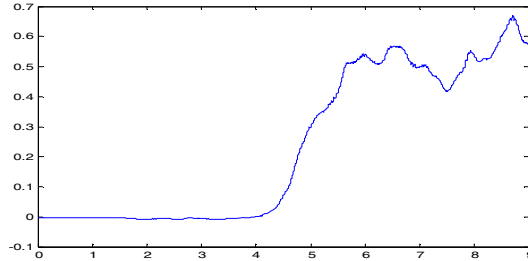
Method AAR, **ERROR RATE LDA**

Min = 9.1667 %
 Time = 4.4023 sec
 Files TRAIN



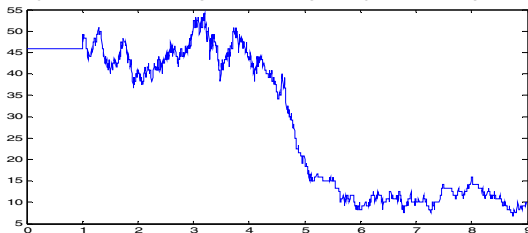
Method PST, **MUTUAL INFORMATION**
LDA

Max = 0.6697
 Time = 8.7031 sec
 Files TRAIN



Method PST, **ERROR RATE LDA**

Min = 6.6667 %
 Time = 8.7500 sec
 Files TRAIN



Confusion Matrix

	Classe	<i>Right</i>	<i>Left</i>	<i>TOTAL</i>
	<i>Right</i>	26	3	29 (*)
	<i>Left</i>	1	30	31 (+)
%	<i>Right</i>	89.66	10.34	100
	<i>Left</i>	3.23	96.77	100

6 Discussion and Conclusions

We can observe that in the results obtained on data published by the University of Graz, the use of autoregressive parameters provides better results than the spectral energies, whereas the reverse is true for our Metz data. This may be due to the fact that data published by the University of Graz match better filtered signals, and therefore the AAR model which reflects all the spectrum is more significant than the energies of the Mu and Beta band. Moreover, one can observe in the data provided from the University of Graz and in our database, there is a smaller error when trying to classify a signal representing a movement of left hand that represents a movement of the right hand this may be caused likely that the system assumes a state of rest as a movement to the left hand and thereby the left hand classification would be more decisive. We conducted trials with much hand movement and with only the imagination of movements, the best results are obtained when the user only imagine the movement, this may be because there is a greater concentration only when we imagine the movement while it is possible to achieve movements automatically without thinking.

Through this article we provide the basis and foundation for developing a brain-computer interface, showing the different steps to implement a BCI, the different stages of processing and analyzing the different techniques currently used.

The most important aspects to be taken into account in order to have good results: A good fixation of the electrodes on the scalp, which required a measure of the impedance of the electrodes on the scalp, which should be less than 5 K ohms. It is always necessary prior training stage. However there are investigations that seek to perform discrimination tasks without training, but the results are not encouraging. Each person has a different way of managing their brain activity. To ensure good training, each individual or user needs to perform at least 60 tests.

The analysis has been performed in a continuous manner during the 9 seconds of each test, and the best results—with minimum error and maximum value for mutual information—are found between the fifth and sixth seconds

The results so far are very encouraging, in some cases reaching rates of 93% effective, but even more must be done about it because it is necessary to increase the number of free degrees, a better definition of states, speed in the interpretation, to be able to have more complex applications

References

- [Kirby 2004] Kirby, M.: Some mathematical ideas for attacking the brain computer interface problem. Department of Mathematics, Colorado State University (2004)
- [Kuo-Kai et al. 2010] Kuo-Kai, S., Po-Lei, L., Ming-Huan, L., Ming-Hong, L., Ren-Jie, L., Yun-Jen, C.: Development of a low-cost FPGA-based SSVEP BCI multimedia control system. *IEEE Trans. on Biomedical Circuits and Systems* 4(2), 125–132 (2010)
- [Lecocq and Cabestaing 2008] Lecocq, C., Cabestaing, F.: Les interfaces cerveau-machine pour la palliation du handicap motor severe. LAGIS – Laboratoire d’Automatique, Génie Informatique & Signal, Université des Sciences et Technologie de Lille (2008)

- [Lecuyer 2007] Lecuyer, A.: Interfaces cerveau-machine: Avances recentes et perspectives a travers le projet open-vibe. Journée IrisaTech. (2007)
- [Roman-Gonzalez 2010a] Roman Gonzalez, A.: System of communication and control based on the thought. In: Proc. 3rd International Conference on Human System Interaction, Poland, pp. 275–280 (2010)
- [Roman-Gonzalez 2010b] Roman Gonzalez, A.: Communication technologies based on brain activity. In: World Congress in Computer Science, Computer Engineering and Applied Computing – WORLDCOMP 2010, Las Vegas, Nevada, USA (2010)
- [Solis-Escalante and Pfurtscheller 2009] Solis Escalante, T., Pfurtscheller, G.: Brain switch asincrónico basado en ritmos sensorimotors. Seminario de Bioingeniería Elche (2009)
- [Tanaka et al. 2005] Tanaka, K., Matsunaga, K., Wang, H.: Electroencephalogram-based control of an electric wheelchair. IEEE Trans. on Robotics 21(4), 762–766 (2005)
- [Toyota 2009] Toyota Motor Corporation, Real-time control of wheelchairs with brain waves. RIKEN (2009)
- [Wolpaw et al. 2002] Wolpaw, J.R., Birbaumer, N., McFarland, D.J., Pfurtscheller, G.: VaughanTM Brain-computer interfaces for communication and control. Clinical Neurophysiology 113, 767–791 (2002)

Author Index

- Alcaraz, J.J. 165
Andrushevich, A. 183
Augustyniak, P. 523
- Bajorek, M. 453
Bartczak, T. 253
Borchers, J. 503
Bujnowski, A. 453
Burda, A. 3
- Cavalieri, S. 201
Chandramouli, A. 343
Choraś, R.S. 403
Chugo, D. 471
Costa-Montenegro, E. 165
- Dattolo, A. 311
Djouani, K. 433
- Eno, J. 343
- Fercu, M. 183
Ferrara, F. 311
- Gauch, S. 343
González-Castaño, F.J. 165
Grabska, E.J. 135
Grzech, A. 103
- Hajder, M. 253
Hamam, Y. 433
Hareźlak, K. 49
Hippe, Z.S. 3
Hoffmann, A. 503
- Hopf, J. 183
Hurtmanns, J. 503
- Józwiak, I. 359
- Kapuściński, T. 539
Klapproth, A. 183
Kluska-Nawarecka, S. 85
Kolbusz, J. 299, 327
Korniak, J. 299, 327
Kościuk, M. 149
Kraus, P.H. 503
Kronenbuerger, M. 503
Krzemiński, K. 359
Kulikowski, J.L. 67
- Lewicki, A. 271
López-Matencio, P. 165
Luhandjula, T. 433
- Mączka, T. 13
Malinowski, A. 239
Marnik, J. 539
Mertens, A. 503
- Nawarecki, E. 85
- Oszust, M. 539
Ozaki, H. 471
- Paradowski, M. 387
Pham, N. 239
Popa, M. 487
Porta, M. 417

Portmann, E. 183
Prusiewicz, A. 103

Ravarelli, A. 417
Regulski, K. 85
Rejer, I. 33
Rodrigues, J.P. 555
Roman-Gonzalez, A. 571
Rosa, A. 555
Rozycki, P. 299, 327
Ruminska, J. 453
Ruminski, J. 453

Samolej, S. 539
Schlick, C. 503
Sieiro-Lomba, J.L. 165
Sieklicki, S. 149
Sieklicki, W. 149
Skabek, K. 287
Śluzek, A. 387
Sochan, A. 287
Strumillo, P. 373

Tadeusiewicz, R. 271
Takase, K. 471
Tasso, C. 311

Vales-Alonso, J. 165
van Wyk, B.J. 433

Wacharamanatham, C. 503
Werner, A. 49
Wilamowski, B.M. 239
Williams, Q. 433
Winiarczyk, R. 287
Wiszniewski, B. 225
Wtorek, J. 453
Wysocki, M. 539

Yokota, S. 471

Żabiński, T. 13
Zaitseva, E. 119
Zięba, M. 103

Subject Index

A

- ACO heuristic strategy 271
- Ambient intelligence platforms 165
- Ant colony optimization algorithm 271
- Artificial
 - intelligence concepts 183
 - neural networks 3
- Atomic services granularity 103
- Autonomous
 - agents 387

B

- BCI
 - applications 571
- Blind
 - people 387
 - volunteers 373
- Brain
 - computer interface 571
- Brute force extrapolation algorithm 433

C

- CAD tools 239
- Car navigation systems 487
- Castings defects 85
- Client/server
 - model 239
 - exchange of information 201
- Cloud computing model 239
- Cognitive
 - mapping 373
 - processes 555
- Collaborative computing 225
- Color
 - recognition 453
 - vision deficiency 453
 - vision problem 453
- Communication
 - cost 103
 - quality improvement 253
 - system 571

- Compound personal 523
- Computer
 - graphics 539
 - networks 239
- Computing utilization 239
- Control
 - plane architecture 299
 - scheme 471
 - system 149, 471

D

- Data
 - flows processing 103
 - navigation systems 487
 - transfer minimizing 523
 - centric computing 225
- Database access and management 49
- Decision
 - making method 165
 - making task 271
- Deontological knowledgebases 67
- Design process 135
- Designing floor layouts 135
- Diagnostic-advisory system 85
- Dichromacy 453
- Direction recognition 433
- Distributed
 - systems 253
 - virtual museum 287
- Document-centric computing 225
- Dynamic
 - programming approach 165

E

- Educational tool 539
- EEG (electroencephalography)
 - biofeedback 555
 - signal processing 571
- E-engineering 359
- Electroencephalogram 555
- Environment for teaching 49

Evaluating overheads 201
 Expert
 model 33
 modeling process 33
 Eye tracking 417

F

File
 system 343
 timestamps 343
 Folksonomy 311
 Formal model 135
 Future internet 287
 Fuzzy
 information retrieval 183
 model 33

G

Gabor wavelets 403
 Gesture recognition 539
 Gmpls network 299
 Graph-based data structure 135
 Grid computing model 239

H

Hand
 motion detection 433
 posture recognition 403
 HCI systems 13, 239, 403
 Health monitoring 523
 Healthcare
 system 119
 Heart rate 165
 Heterogeneous training programs 165
 Hilbert huang transform 555
 Human
 activity supporting 67
 system interface 13, 239
 machine communication 253
 machine interface 149, 433

I

Image
 simulation 453
 Information society 359
 Intellectual capital 271
 Intelligent
 manufacturing system 13

Interactive agents 225
 Internet of things services 183
 IT system 359

K

Kantorovitsch space 67
 Kinetic tremor 503

L

Labelling method 453
 Learn-and-test paradigm 3
 Location data 183
 Logic models 135

M

Malicious traffic 327
 Man-computer relationship 85
 Markov decision processes 165
 Mesh topologies 299
 Mobile
 agents 225
 health care applications 523
 interactive document (MIND) 225
 navigation 487
 Moodle platform 49
 Motor
 disabilities 417
 function 571
 Multi-step training scenario 165
 MySQL 49

N

Navigation 373
 Neural
 networks 3
 Non-linear
 classification methods 433
 Non-visual presentation 373

O

On-off model 327
 Ontology
 of conceptual visual design 135
 backed search queries 183
 OPC UA 201
 Oracle DBMS 49
 OSPF-TE 299
 Out-door mobility 373

P

Pedestrian navigation systems pns 487
PHP 49
Progressive mesh 287
Prometheus framework 183

Q

Quality
 of services improvement 299
 pages 343
Queue validation 3

R

Radon transform 403
Real-time
 services 253
Recommender systems 311
Recommending tags and resources 311
Rehabilitation robotic walker 471
Reliability
 analysis 119
 engineering 119
Remote
 control system 149
Residential infrastructure 523
Resubstitution 3
Ring topologies 299
Robotic walker system 471
RSVP-TE 299
Running track 165

S

Seating motion 471
Self
 similar traffic model 327
 similarity feature 327
Semantic
 approaches 311
 spaces 183
Sensory substitution methods 373
Service oriented paradigm 103
Services
 merging, partitioning and
 execution 103
 quality 103
Sitting motion assistance 471
Small and medium enterprises 3
Social
 semantic relations 311

 tagging 311
 web 311

SQL
 server 49
Storage management of agricultural
 products 149
Storehouse local controller 149
Surveillance systems 523
Swabbing movements 503
Swarm algorithm 271
Synergic
 intranet 359

T

Teleassistance system 373
Texture features 403
Therapeutic tool 539
Therapy of children 539
Touch screens 503
Traffic generator 327
Tremor symptoms 503

U

Ubiquitous health supervision 523
Uncertain data 3
Uniform resource locator (URL) 343
Unknown environment 387
Unobtrusive devices 417
URL ordering 343

V

Virtual museum 287
Vision-based understanding 387
Visual design aided by computer 135
Visually impaired 373, 387

W

Wayfinding 373
Web
 crawlers 343
 log 343
Wheelchair motion 433
Wiping movements 503
Wireless
 data transfer 149
 sensor network 165
WWW images 453