Proceedings of the 3rd Annual South-East European Doctoral Student Conference

Proceedings of the 3rd Annual South-East European Doctoral Student Conference

Edited by
Iraklis Paraskakis
Andrej Luneski

ISBN 978-960-89629-7-2
ISSN 1791-3578

17 Mitropoleos str. 54624 Thessaloniki, Greece
Tel: +30 2310 253 477
Fax: +30 2310 253 478 ext. 115
http://www.seerc.org

Preface

It is my pleasure to welcome you to the 3rd Annual South East European Doctoral Student Conference (DSC2008) taking place June 26th and 27th in Thessaloniki, Greece, and organised by the South East European Research Centre (SEERC).

This volume contains the papers from the programme of the 3rd Annual South East European conference. The aim of the conference is to initiate an exchange of knowledge between young researchers and to help establish a network of scholars currently undertaking research in South-East Europe. Having identified academic isolation as a problem that many doctoral students face today, SEERC aims to bring researchers together for establishing collaborative links between disciplines, for testing the ground for innovative ideas and for engaging the wider academic community.

Building on the success of the two previous conferences, this year's conference attracted a large number of submissions resulting in around 85 presentations of both papers and posters. The audience of the conference expanded beyond the boundaries of South East Europe and we had presentations from throughout Europe confirming the need for Doctoral Students to come together, discuss their experiences and gain external feedback to their work as well as listen to the progress and methodology of fellow PhD candidates.

Additionally, this year a workshop entitled: ***Building Research Skills*** was organised offering to delegates hands-on tips regarding how to write a research paper as well as how to present a paper in a conference. This workshop was organised in response to a recognised need for developing and mastering these research skills. It suffices to say that even some experienced researchers find still difficult to master such skills. The workshop was conducted by Dr. Petros Kefalas from CITY College.

Moreover, this year the Risk, Well Being and Cognition Track were joined by students from the European Society of Cognitive Psychology (ESCoP) Summer School on Neuroscience of Attention. Doctoral students from the Summer School joined the conference on its second day and presented part of their work from the Summer School to the delegates and received feedback.

The keynote speech this year was given by *Prof. Eleftherios Iakovou*, from the Aristotle University of Thessaloniki. The title of the keynote speech is "Global Trends and Challenges: Impact on Logistics and Supply Chain Management Research".

The scope of the conference was, again, multi-disciplinary spanning throughout the areas in which SEERC is doing active research and therefore it was divided into four parallel sessions:

- Enterprise and Regional Development
- Information and Communication Technologies
- Governance Politics and Society
- Risk Well Being and Cognition

There were 125 submissions and of these 62 were accepted as full papers and 18 for poster presentations. The full papers were divided as follows:

- 14 for the Enterprise and Regional Development Track
- 22 for the Information and Communication Technologies Track
- 17 for the Governance Politics and Society Track
- 9 for the Risk Well Being and Cognition Track

One of the main objectives of the conference has been to provide an opportunity for PhD students to receive advice from experts in their chosen field of research. This would not have been accomplished without the hard work of the members of the Programme Committee as well as that of the reviewers, who's reading and commenting of the submissions has been pivotal. Also the participation of the invited discussants and their feedback to the presentations made by the participants has been greatly appreciated. The list of the discussants according to the research track is as follows:

Enterprise, Innovation and Development

- Prof. Siniša Zarić (Director, Centre of International Studies, Faculty of Economics, University of Belgrade, Republic of Serbia),
- Prof Elias Carayannis, (School of Business, The George Washington University, USA and CITY College, Greece)
- Dr. Konstantinos Priporas, (Department of Marketing & Operations Management, University of Macedonia, Greece)

Information and Communication Technologies

- Dr Anca Vasilescu, (Dept of Informatics, University of Brasov, Romania),
- Dr. Petros Kefalas, (Dept of Computer Science, CITY College, Greece)
- Dr Anthony Simons, (Dept of Computer Science, University of Sheffield, UK)
- Dr. Panagiotis Bamidis, (Dept of Medical School, Aristotle University of Thessaloniki, Greece)
- Prof. Demos Stamatis, (Dept of Informatics, TEI of Thessaloniki, Greece)
- Dr Ilias Sakellariou, (Dept of Applied Informatics, University of Macedonia, Greece)

Governance Politics and Society

- Dr Nikos Sifakis, (School of Humanities, Hellenic Open University, Greece)

- Dr Nicola Rehling (School of English, Aristotle University of Thessaloniki, Greece)
- Dr Alkmene Fotiadou, (Division of Humanities, CITY College, Greece)
- Ass. Prof Gregory Paschalidis, (Dept of Journalism and Mass Communication, Aristotle University of Thessaloniki, Greece)
- Dr Christos Frangonikopopoulos, (Dept of Journalism and Mass Communication, Aristotle University of Thessaloniki, Greece)

Risk Well Being and Cognition

- Dr Elvira Masoura (Department of Psychology Aristotle, University of Thessaloniki, Greece)
- Dr. Lambros Lazouras (Department of Psychology, CITY College, Greece)
- Dr. Susie Savvidou (Department of Psychology, CITY College, Greece)

SEERC would like to thank all the above named discussants for accepting our invitation and providing their valuable feedback to the PhD students that made their presentations.

Finally, we would like to also thank all the presenters and participants as well as our colleagues at SEERC and CITY that contributed in making DSC2008 a successful event. We are looking forward to the 4th conference in June 2009.

May 2008

Iraklis Paraskakis
Andrej Luneski

Programme Committee:

Prof Hamideh Afsarmanesh, *University of Amsterdam, Netherlands*
Prof Galia Angelova, *Bulgarian Academy of Science, Bulgaria*
Prof Elena Antonacopoulou, *University of Liverpool, UK*
Prof Tudor Balanescu, *University of Pitesti, Romania*
Prof Nick Bassiliades, *Aristotle University of Thessaloniki, Greece*
Dr Panagiotis Bamidis, *Aristotle University of Thessaloniki, Greece*
Prof Betim Cico, *Polytechnic University of Tirana, Albania*
Prof Dimo Dimov, *Bulgarian Academy of Science, Bulgaria*
Prof Andrew Geedes, *University of Sheffield, UK*
Prof Dimitris Georgarakos, *University of Frankfurt, Germany*
Dr Maria Georgiadi, *University of Crete, Greece*
Dr Stylianos Hatzipanagos, *Kings College, University of London, UK*
Prof Antonis Makredemetres, *University of Athens, Greece*
Prof Yannis Manolopoulos, *Aristotle University of Thessaloniki, Greece*
Dr Elvira Masoura, *Aristotle University of Thessaloniki, Greece*
Prof Luis Camarinha-Matos, *New University of Lisbon (Uninova), Portugal*
Prof Roumen Nikolov, *University of Sofia, Bulgaria*
Dr Miguel Nunes, *University of Sheffield, UK*
Prof Demetrios Sampson, *University of Piraeus, Greece*
Prof Eftichios Sartzetakis, *University of Macedonia, Greece*
Dr Klio Semoglou, *Aristotle University of Thessaloniki, Greece*
Prof George Sideridis, *University of Crete, Greece*
Prof Dimitris Sotiropoulos, *University of Athens, Greece*
Dr Paul Stubbs, *Institute of Economics, Croatia*
Prof Davidkov Tsvetan, *University of Sofia, Bulgaria*
Dr Ana Vasconcelos, *University of Sheffield, UK*
Dr Anca Vasilescu, *University of Brasov, Romania*
Prof Dimitris Vrontis, *University of Nicosia, Cyprus*
Prof Asteris Xouliarias, *Harrokopeio University of Athens, Greece*
Prof Sinisa Zaric, *University of Belgrade, Serbia*
Dr Alexandros Zangelidis, *University of Aberdeen, UK*
Prof Katerina Zdravkova, *University of Ss Cyril & Methodius, FYROM*

Organising Committee:

Dr George Eleftherakis
Ms Sara Hannam
Dr Efrosyni Kalyva
Ms Erika Melonashi
Dr Alexandros Psychogios
Mr Martin Zlatev

Additional Reviewers:

Table of Contents

INFORMATION AND COMMUNICATION TECHNOLOGIES

RISK AND WELLBEING

INFORMATION AND COMMUNICATION TECHNOLOGIES

Analysis of the Justification of Vented Disk Installation onto the Front Axle in a Passenger Motor Vehicle

Raicevic Mile[1], Demic Miroslav[2], Milenkovic Predrag[3]

[1,3]Zastava cars, Kragujevac,Serbia, mile23@verat.net[1]

[2]Mechanical faculty, Kragujevac, demic@kg.ac.yu

The process of mastering production of the new car parts is a very expensive one, and the result of such investments should result in improved vehicle performance. In earlier years, „Zastava" car factory did not use to equip its vehicles with vented disk brakes. Since the efficient cooling is the key issue for braking efficiency, it can be expected that the application of the vented disk would lead to the improvement of the vehicle's braking parameters. The aim of this paper is to compare braking parameters of the vehicle in cases when the vehicle is equipped either with solid or with vented disk brake (mounted onto the vehicle's front axle). For the purpose of collecting necessary data, the experimental researches were conducted. According to the comparison of the experimentally obtained data, it is concluded that for a vehicle with similar constructional characteristics as the vehicle used in the experiment, the installation of the vented disk is not justified.

Keywords

brake, experiment, solid disk, vented disk.

1. Introduction

The need for increased cooling of disk brakes led to the development of vented disks, anyhow the advantages of vented disks over solid disks is the subject of some conjecture. The primary advantage of vented rotors is increased heat dissipation from internal pumping of air, however, under slow speeds the pumping action of the vanes is minimal and only becomes pronounced as the rotor speed increases [10]. At higher speeds the airflow flowing around the disk as a result of the forward movement of the vehicle, tends to prevent effective pumping of air through the vanes [1, 10].

Friction coefficient between the braking pads and the disk was calculated according to the results of the experimental research and is described by the expression [4, 5, 6] that shows the immense influence of the temperature:

$$\mu_p = 0.512 pp^{-0.047} v^{0.012} \theta^{-0.046} \quad \textbf{(1)}$$

Where:

- *pp* – pressure given in bars measured in the front braking line [bar]
- *v* – speed of the vehicle [km/h]
- θ - temperature [^{o}C].

The continuous race among the car manufacturers concerning the offer of car models that are equipped with more advanced and more improved braking features is also run in the field of competitive prices. The new requirement which car manufacturers are currently facing is the transition on the new type of disk brakes that consequently causes the rise in costs for all the activities that are included in the production chain: from the start of production to the final control. Thus it is important to test how significant the new improvement of braking characteristics is and which particular regime shows the greatest improvement of braking performances. However, if the improvement of the braking characteristics is not considerable and if it is present only within the certain regimes that the vehicle rarely exploits, than the application of vented disk seems highly improbable. This paper presents the results of experimentally obtained data which justify the decision to equip Florida vehicle (a vehicle belonging to „Zastava automobili" product range) with vented disk system.

2. Experimental Research

In order to collect the data necessary for the comparison of the braking performances, the experimental research was conducted [3,7,8]. For the purpose of the experiment, the vehicle Zastava Florida 1.3 equipped with Bosch ABS was used. The following measurements were taken: time measured from the beginning of the braking action, the speed of the vehicle, brake pedal force, the pressure of the working fluid within the brake lines directed to the left front wheel, the pressure of the working fluid within the brake lines directed to the rear wheels in front of the braking corrector and finally, the angular speed of each wheel and FADE tests. For the purpose of illustration only in figure 1 is shown a vehicle on the runway with velocity sensor corevit L on vehicle's side panel.

The experiment was conducted on runway near the town of Kraljevo. The runway in question has the asphalt concrete surface that is characteristic for its good adhesion. For the purpose of compliance with the existing regulations, the breaking corrector was installed. The measurements were taken for the vehicle fitted with either solid disk or vented disk brakes onto the front axle and in the following circumstances:

- initial speed 80 km/h till halt for a vehicle carrying 2 passengers and 20 kg luggage
- initial speed 80 km/h till halt for a vehicle carrying 5 passengers and 50 kg luggage

- initial speed 140 km/h till halt for a vehicle carrying 2 passengers and 20 kg luggage
- initial speed 140 km/h till halt for a vehicle carrying 5 passengers and 50 kg luggage.
- Two drivers took part in the experiment and they were instructed to brake as if it had been panic braking. The fulfilment of their task was confirmed by consulting the paper [9] which thoroughly deals with the issue of panic braking as well as with the driver's reaction.

Figure 1 Vehicle on the runway

As the signal for the beginning of measuring, the signal emittted by the rear stop light switch mounted on the brake pedal was used. Thus the possibility of late reaction of the driver was excluded etc. This issue was thoroughly examined in previously published paper [9].

FADE test was implemented according to ZA technical documentation, i.e. Automobile Research Institute [2,11] that is more strict than ECE 13 rule book. This program includes two tests, each containing 25 consecutive braking cycles, starting at the initial velocity of 100 km/h and ending with end velocity of 50 km/h with deceleration of 5 m/s2. A single braking cycle is actually the cycle from the moment of braking to the vehicle re-speeding to the initial speed; this cycle lasts for 45 seconds. After the first test is conducted, brakes are given 10 minutes to cool off, and than the second test is carried out- again consisting of 25 braking cycles. Here, the vehicle is loaded to a maximum loaded weight, i.e. the total weight is 1350 kg. The temperature is measured by thermocouple on the friction lining of the brake pad (2 mm beneath the contact surface).

3. Experimentally Measured Data

In order to get a general idea, only the partial results of the experimental research for four cases mentioned in item 2 are given. The vehicle has very good braking characteristics even without ABS and achieves deceleration in compliance with the regulations of 5.8 m/s^2 and with the brake pedal force of 174 N continuosly, when the vehicle is ready for driving and with 1 driver i.e. total weight is 1020 kg. For the fully loaded vehicle i.e. 1350 kg total weight, the brake pedal force should be 232 N [2]. However since the time till full halt in an ABS equipped vehicle depends on the initial reaction of the driver (more detailed information in [9]) and that during the experiment the magnitude of the initial speed was not strictly taken as the signal for the beginning of braking, thus for the sake of comparison of the vehicle's achieved speeds, the values shown in the diagrams are slightly smaller than the values of initial speeds.

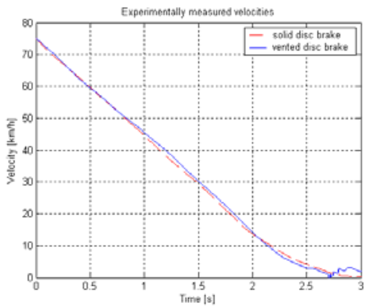
In order to make an adequate comparison of the vehicle's braking parameters, the disk brakes characterized by the following features were used: the working cylinder diameter of the brake, outer brake disk diameter and active radius of the braking effect were identical. Identical braking pads were also applied. The clamp had to be additionally modified when the vented disk was used (disk width 19 mm, the width of the paddle slot 6 mm) because of the difference in width between the vented and the solid disk brakes.

Experimentally measured data for driver M and initial speed 80 km/h till halt for a vehicle carrying 5 passengers and 50 kg luggage are shown on figures 2-5.

Experimentally measured data for driver Z and initial speed 140 km/h till halt for a vehicle carrying 2 passengers and 20 kg luggage are shown on figures 6-9.

Experimentally measured data for driver M and initial speed 140 km/h till halt for a vehicle carrying 5 passengers and 50 kg luggage are shown on figures 10-13.

Experimentally measured data for driver Z and initial speed 140 km/h till halt for a vehicle carrying 5 passengers and 50 kg luggage are shown on figures 14-17.



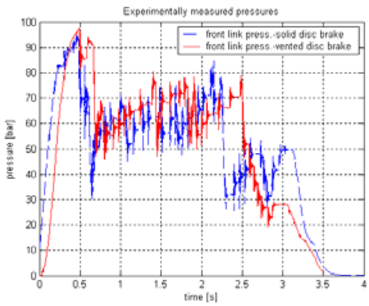


Figure 2 Speeds of the vehicle at the initial speed speed 80 km/h with 5 passengers and 50 kg luggage

Figure 3 Front link pressures at the init. 80 km/h with 5 passengers and 50 kg luggage

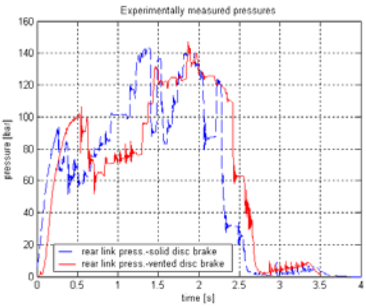


Figure 4 Rear link pressures at the initial speed initial 80 km/h with 5 passengers and 50 kg luggage

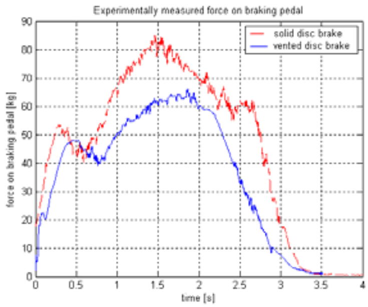


Figure 5 Measured forces on pedal at the speed 80 km/h with 5 passengers and 50 kg lugg.

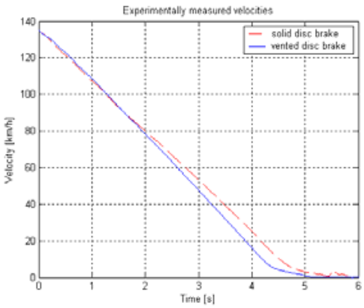


Figure 6 Measured velocities at the initial speed speed 140 km/h with 2 passengers and 20 kg luggage

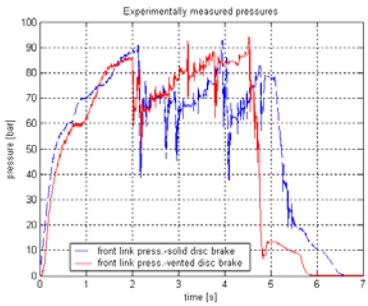


Figure 7 Front link pressures at the initial 140 km/h with 2 passengers and 20 kg lugg.

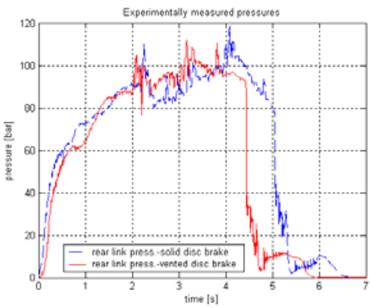


Figure 8 Rear link pressures at the initial speed initial 140 km/h with 2 passengers and 20 kg lugg.

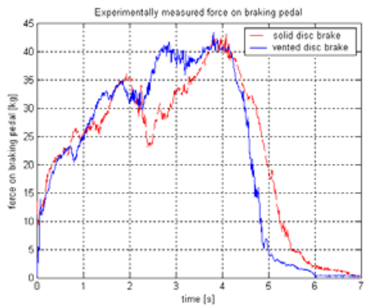


Figure 9 Measured forces on pedal at the speed 140 km/h with 2 passengers and 20 kg lugg.

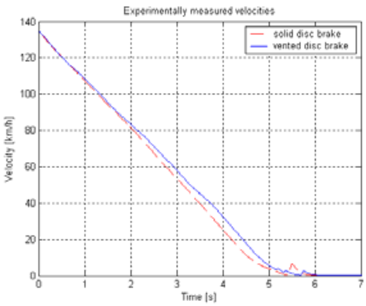


Figure 10 Measured velocities at the initial speed 140 km/h with 5 passengers and 50 kg lugg. driver M

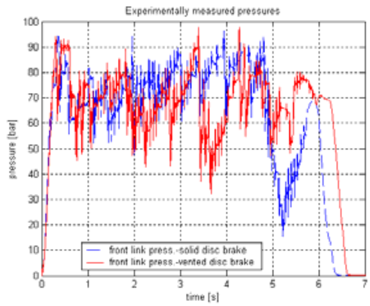


Figure 11 Front link press. at the initial speed 140 km/h with 5 passengers and 50 kg lugg. driver M

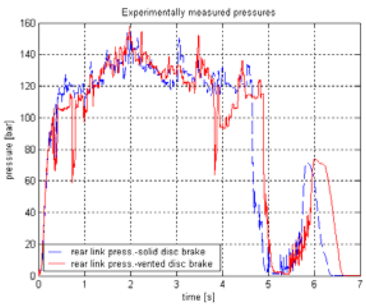


Figure 12 Rear link press. at the initial speed 140 km/h with 5 passengers and 50 kg lugg. driver M

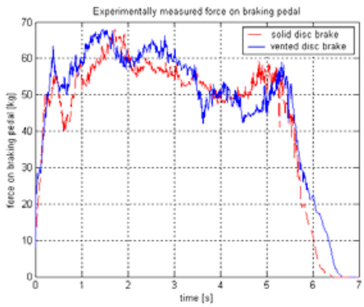


Figure 13 Measured forces on pedal at the initial speed 140 km/h with 5 passengers and 50 kg luggage driver M

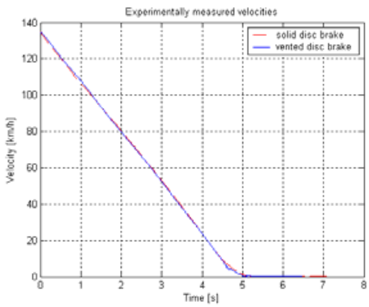


Figure 14 Measured velocities at the initial speed 140 km/h with 5 passengers and 50 kg lugg. driver Z

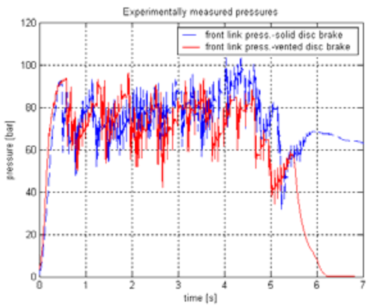


Figure 15 Front link press. at the initial speed 140 km/h with 5 passengers and 50 kg lugg. driver Z

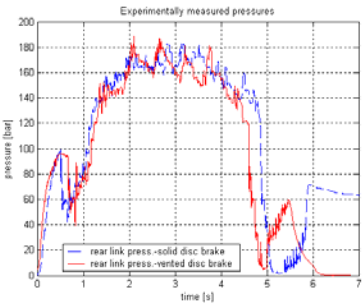


Figure 16 Rear link press. at the initial speed 140 km/h with 5 passengers and 50 kg lugg. driver Z

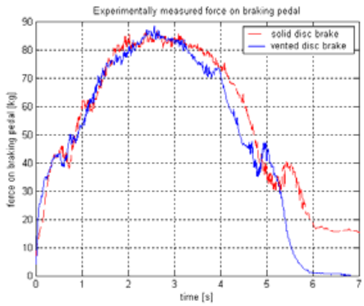


Figure 17 Measured forces on pedal at the initial speed 140 km/h with 5 passengers and 50 kg luggage driver Z

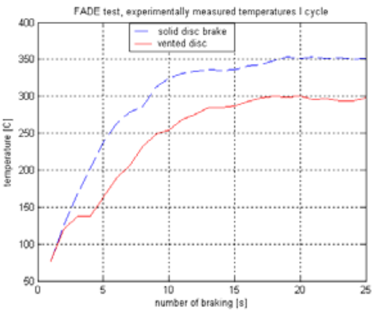


Figure 18 FADE test – I cycle

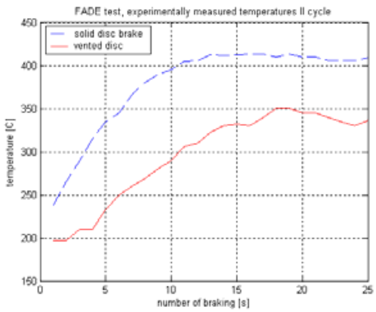


Figure 19 FADE test – II cycle

Figures 18 and 19 show contiguous diagrams for the achieved temperatures during the braking process implemented within the FADE test framework, for the vehicles equipped with vented or solid disk on their front axle, while drum brakes are mounted on the vehicles' rear axle in both cases.

It is evident from the figures 18 and 19, that the temperature of the brake actuator is significantly higher for the solid disk. In order to determine the degree of decline the vehicle's braking performances, which occurred due to the increase in temperature in brake actuators, additional measurements were taken. After the 25 braking cycles were completed during the second test, the vehicle was sped up to the velocity of approximately 140 km/h (initial braking speed deviation is within permitted limits). As it has been previously mentioned in this paper, paragraph 2, the implemented research procedure is much stricter than the one prescribed by the ECE regulations. Again, the braking action was done as if it were panic braking. The results of these measurements are given in Figures 20-23. According to the speeds measured during the braking action, the declarations were computed for this 26th braking that belongs to the second FADE cycle and are given in the Figure 24.

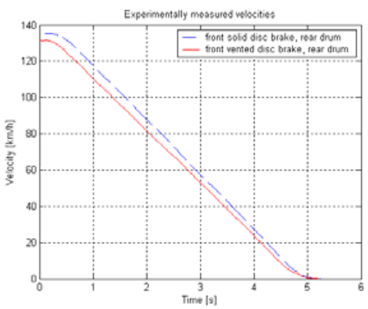


Figure 20 Measured velocities at the initial speed 140 km/h with 5 passengers and 50 kg lugg. driver Z

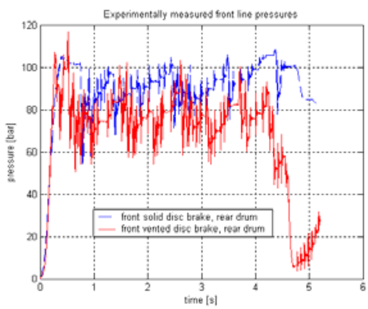


Figure 21 Front link press. at the initial speed 140 km/h with 5 passengers and 50 kg lugg. driver Z

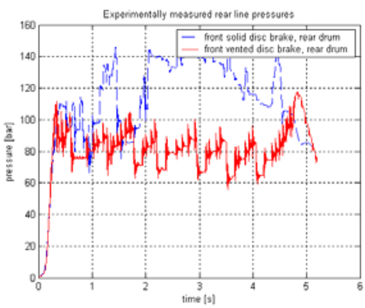


Figure 22 Rear link press. at the initial speed 140 km/h with 5 passengers and 50 kg lugg. driver Z

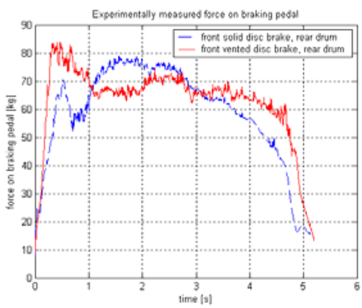


Figure 23 Measured forces on pedal at the initial speed 140 km/h with 5 passengers and 50 kg lugg. driver Z

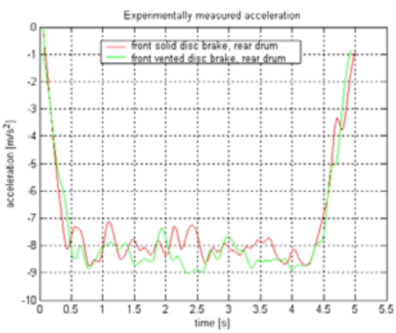


Figure 24

4. Analyses of the Experimentally Measured Data

One panic braking was performed under conditions previously described in this paper. Brake pads characterized with very satisfactory heat resistance features were used and they were in compliance with Zastava Vehicles technical documentation [11]. Under the above mentioned conditions, the difference between the times necessary for the full halt of the vehicle are not great for one panic braking, because the difference in vehicle's breaking performances depends on the change of the adhesion coefficient which in turn depends upon the temperature.

As far as the change of the pressure in the line that leads to the front wheels is concerned, by applying vented brakes we achieved lower average values for the pressure at the initial speed of 140 km/h. The mean pressure values in the front installation were calculated by using Matlab mean function computer program in the following cases – for the first 3 seconds of braking at the initial speed of 80 km/h as well as for the first 5 seconds of braking at the initial speed of 140 km/h.

When the initial speed was 80 km/h, the mean values of pressures in the front installation were slightly grater in the vehicle that had vented disk brakes installed. The mean pressure value shown in the figure 3 for the solid disk is msd=57.8407 bars while the mean pressure value for the vented disk brakes is msvd=59.5951 bars.

However, this ratio changes when the initial braking speed is increased. Namely, in the cases when the initial speed is 140 km/h, the mean pressure value is greater in the vehicles that use solid disk brakes. Thus the mean pressure value in the case shown in figure 11 for the solid disk brakes is msd=73.4033 bars, and for the vented disk brakes it is msd=67.8508 bars.

The mean pressure values in the line that leads to the rear wheels have similar values.

From the figures 5, 9, 13 and 17 one can conclude that the drivers fulfilled the recommendation to brake as if it were panic breaking.

From the Figures 18 and 19 one can see that the temperature of the vented disk brake pad is much smaller than the temperature of the solid disk pad when both are found under the same experimental conditions. During the experiment a 10 minute pause between the first and the second measurement phase was made. It is interesting that both types of brakes cooled off for approximately 100°C. This is due to the fact that the thermocouple was placed on the braking pads that are made of identical material and possess identical characteristic, also when the vehicle is not in motion these brake pads do not touch brake disks. However, one can see from the Figure 19 that the solid brakes showed more rapid increase in temperature in the first 5 seconds of the test.

As seen in the Figure 24, the vehicle that is equipped with vented disk mounted onto its front axle, did achieve slightly better mean deceleration. However, the vehicle that is equipped with solid disk on its front axle also possesses excellent braking characteristics, even after great number of braking actions. The achieved decelerations that are given in the figure 24 have high values because the braking actions were performed on the dry asphalt-concrete surface that is characterized by high adhesion coefficient. These mentioned conditions are considered as a critical case; they are also the conditions in which vented disk should demonstrate its advantages.

5. Conclusions

The application of vented disks mounted on the front axle improved vehicle's braking parameters.

According to the assumption expressed at the beginning of this paper, the advantages of ventilated disks brakes were clearly distinguished when the initial speeds at the beginning of braking were greater.

Braking pads that were applied in this experiment had very satisfactory features in terms of small change of the friction coefficient for the temperature increase. The data related to the fact that at higher initial speeds, the smaller pressure was measured in the lines leading to the front wheels when the vented disks brakes were installed, speaks for itself and supports the claim that braking performances were definitely improved.

The advantages of the vented disks are not significantly distinguishable during a single panic braking. The substantial advantages of vented disk brakes are clearly distinguishable during FADE test. The achieved results point to a fact that in the conditions of repeated braking actions, vented disks achieve better results.

However, according to the conducted research, it could be concluded that the vehicle that was used in the experiment, has excellent braking characteristics, even when the solid disk is mounted onto its front axle. This is due to the fact that the constructional characteristics of the vehicle (maximum speed, weight, applied friction pair materials, brake actuator size,…etc.) for which the application of the vented disk has been investigated, are such that the achieved temperatures can not significantly endanger braking effectiveness.

References

1 Fowkes F G R, Fulton P M. Critical appraisal of published research: introductory guidelines. BMJ 1991; 302: 1136-40.

2 Technical documentation of ZA a.d., Research and Development Center 2007.

3 Demić M.,Raičević M. Modelling of suspension system behaviour for a vehicle equipped with ABS during the straight line braking, International congress motor vehicles and motors, paper MVM20060050, Kragujevac, Serbia, October 4th-6th, 2006.

4 Todorović J.,Duboka Č.,Arsenić Ž. Modelling of non-asbestos disk brakes friction, journal MVM br.81/82, The Mechanical Faculty of Kragujevac 1988.

5 Todorović J.,Duboka Č., Arsenić Ž.,Mandić D. Laboratory tests of pads FIAZ 2102/07 for front disk brake of YUGO Florida A103.00 according to ZCZ TU7. H2000, Report No. 09-1145/97, The Mechanical Faculty of Belgrade, 1997.

6 Technical documentation of Zastava cars a.d. regarding the disk brake research with materials N890, textar T46866, Galfer 3000 GG, Kragujevac, Serbia, 1991.

7 Raicevic M. ABS algorithm influence on vehicle's braking parameters, Master of Science thesis, Mechanical Faculty, Kragujevac 23.01.2007.

8 Demić M.,Diligenski Dj,Raičević M. Simulation model of a vehicle fitted with ABS during the straight line braking, MVT 2006., Timisoara, Romania, November 2006.

9 Demić M., Raičević M., Neagu E. Comparison of braking parameters and drivers' reactions in case of vehicles with and without ABS, Scientific bulletin N^{o}16, Automotive series XII pages 50-59, University of Pitesti, Faculty of mechanics and technology, 2006. Pitesti, Romania

10 Stephens A. Aerodynamic Cooling of Automotive Disk Brakes, thesis submitted for the degree of Master of Engineering, School of Aerospace, Mechanical & Manufacturing Engineering RMIT University, March 2006.

11 Zastava automobili technical protocol 9.02122, 9.02122/64 supplement 25/64 and 26/64, Zastava cars a.d. Research and Development Center, May 1989.

A New Approach to Software Pattern Detection

Mocanu Mihai Stefan

University of Pitesti, 1st Targu din Vale Street, Arges, Romania,
mocanushtefan@yahoo.com

This article aims at introducing a new approach in detecting complex software constructs, which arise from daily needs to solve recurring problems, constructs known as software patterns. Traditional software detection approaches usually either tend to detect only but simple patterns, or tend to have a very complicated theoretical approach. In this article, the theoretical base is simple, easy to understand and to extend, while having the potential to detect complex patterns.

The approach consists in establishing sets of jointly sufficient conditions that need to be fulfilled by a software construct to be classified as a known software pattern. While one such set of conditions can detect only part of the implementations, it is possible to verify several such sets of conditions for a given software construct in order to classify it as a know pattern. The approach yields good results as it closely follows the normal steps a human operator would employ in the attempt to discover such patterns.

Keywords
application logic , reverse engineering , software patters detection.

1. Introduction

Software patterns are special software constructs that represent time-proven solutions to well-known problems in various areas in software development. They come to solve various problems, related to object creation, concurrency, object accessing, etc.

A cornerstone work in the area is the book "*Design Patterns: elements of reusable object oriented software*" [1], which defines what are now known as classical software patterns.

While important, these are not the only software constructs that are useful or well-known, but they are important information in the quest of understanding an application's logic. Detection of various software patterns in applications results in proper understanding of the source code of the application and easier evaluation of quality standards. It can provide valuable metrics, as well as insights on various levels of functionality isolation, with impact on application maintenance and migration. Also, by understanding various regions of the application, tests can be better designed.

Throughout the article I will use Java examples, although the principles in this article can be extended for other programming languages.

2. Software Pattern Example: Observer-Observable

2.1 Background

The Observer pattern is a design pattern that is used in software development to communicate information regarding the occurrence of an event.

In this pattern, also known as Publisher-Subscriber, one or more objects – known as Observers – are registered to observe an event that might occur on another object – the Observable.

This software construct offers a decoupling of the entity that generates the event and the entities that are interested in being notified by the occurrence of the event, offering flexibility and a standard mean of adding functionality that is not anticipated at the moment of development of the Observable.

In a class diagram view, an example describing the concept can be seen in Figure 1.

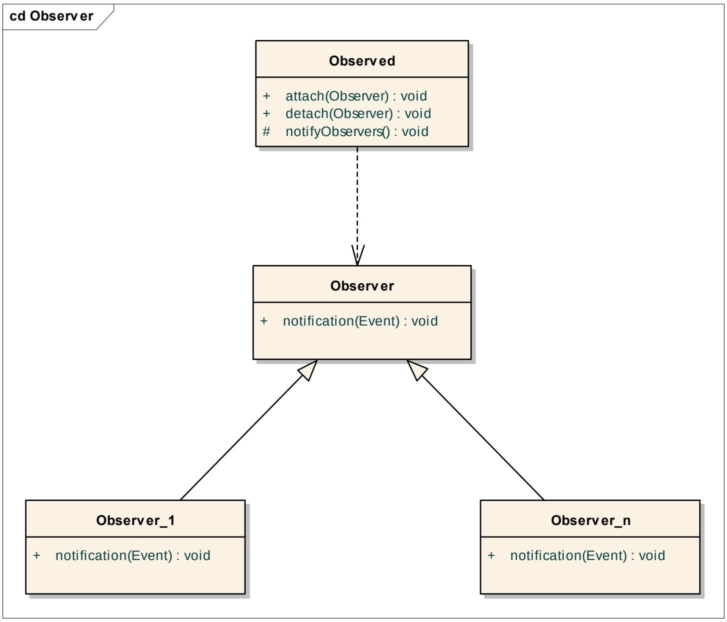


Figure 1 Class view for the Observer Pattern.

2.2 Approach Description

I will start by defining a set of necessary and sufficient conditions a set of classes needs to fulfil in order to be classified as an implementation of the Observer pattern:

Condition set 1

1a) A class (Observable) holds Multiple References to objects from another distinct class (the Observer)

2a) The Multiple References to the Observer class can be dynamically added other references through a publicly available method in the Observable class

3a)A method (Notification Method) is invoked on a part (or all) the Multiple Reference

A couple formed by (Observable, Observer) is an implementation of the Observer pattern.

Condition 1a) arises from the need to store references to the objects that need to be notified about the occurrence of the event. It is necessary that the Observable holds references to the observer in order to notify them about the event's occurrence.

Condition 2a) is a general requirement that objects that are interested in receiving events can be registered (or register themselves) at runtime.

Condition 3a) is necessary in order for the Observable to notify the occurrence of the event.

All the above conditions are independently necessary. Conversely, if a set of classes fulfil these conditions, it is sufficient for them to be an implementation of the Observer pattern.

Therefore, the above set is a necessary and sufficient set of conditions that need to be fulfilled for a set of classes to be an implementation of the Observer pattern. We can conclude that:

$$1a) \wedge 1b) \wedge 1c) \Leftrightarrow O \qquad (1)$$

where O denotes that the classes identified by the above conditions represent an implementation of the Observer pattern.

In practice, some of the conditions might be hard to verify completely, therefore I will define weaker conditions that will be tested. On the downside, they will not verify all the possible implementations of the pattern, but on the other side, the new conditions are easier to verify while test results show that they will detect the vast majority of the implementations.

Condition set 2

A class (Observable) contains an Iterable property (it fulfils the java.lang.Iterable contract, either by implementing the interface, one of its subinterfaces or acting as a delegate of a class that fulfils this contract)

The Iterable property defined previously is publicly writable

The Iterable property defined previously is iterated and on each iteration a method is invoked on the object retrieved in the iteration step.

Condition 1b) is a sufficient condition for Condition 1a), although it is not necessary: in Java, there is a standard way to define a multiple reference, by using collection-type classes: lists, maps, arrays, vectors, etc. However, this is not the only way to hold a multiple reference. For example, one could use exotic language constructs, through external code (like JNI or remote invocation), or additional software constructs that are hard to verify that they are collections (for instance, using reflection). We have

1b) => 1a) (2)

By restraining the condition 1a) to 1b), the condition 2a) needs to restrict the applicability area to the one defined at 1b), so instead of verifying the multiple references defined at 1a), we just need to verify the condition that it is possible to insert additional elements in the collection defined at 1b). We have

2b) => 2a) (3)

Similarly, by restricting 3a) to the Java collections, we can now limit to the methods available in those classes (iterators, array index, specific collection operations, etc.). We have that

3b) => 3a) (4)

So far, we have restricted the first set of conditions to another set that can be more easily verified. By applying these restrictions, we have obtained a set of conditions that are *jointly sufficient*, without being independently necessary or independently sufficient. That is, from (1), (2), (3) and (4) we obtain that:

1b) ∧ 2b) ∧ 3b) => O

3. Testing

In order to test the previous conditions, I have designed 3 test categories, and used samples from each test category as an input to various test applications that intend to detect software patterns, as well as an application I have written a Java that will attempt to verify the above conditions.

3.1 Test Applications

Similarity Scoring [2] uses the Similarity Matrix and the Similarity Score [3] to identify software patterns [4].

Pinot [5] analyses the Abstract Syntax Tree of the source code in order to identify the patterns [6].

FUJABA [7] analyses the source code, creating graphs that are later analyzed using specific approaches to detect the software patterns [8].

CASA (Computer Aided Software Analysis) is an application written by the author for the purpose of analyzing the Abstract Syntax Tree of the input source code in order to detect software patterns. It is in the experimental phase and is likely to be further modified.

3.2 Test Cases

I have divided the tests into 3 categories, depending on the degree of compliance with the Observer pattern.

3.2.1 Minimal Implementations

The minimal implementations are implementations of the pattern that contain only the necessary language elements in order to correctly and completely implement the desired pattern.

In this implementation, we do not use additional language elements, or intermediary constructs, or any other language elements to implement the pattern. However, the implementation is complete, meaning that the classes are an implementation of the Observer pattern. We can see this as the classes that contain the minimal code for an implementation of the pattern.

A minimal implementation of the Observer is defined in the following code:

```
public interface Listener {
    public void onNotification();
}
```

This interface defines a class of objects interested in receiving notifications regarding the occurrence of an event. In this simplistic approach, only the occurrence of the event is notified, with no details regarding the event. A class that is able to register such listeners can be defined as follows:

```
public class Observable {
    protected ArrayList listenersList = new ArrayList();
    public void registerListener(Listener l) {
            listenersList.add(l);
    }
    public void test() {
            Iterator iterator = listenersList.iterator();
            while(iterator.hasNext()) {

    ((Listener)iterator.next()).onNotification();
            }
    }
```

```
}
```

This class contains the minimal code necessary to implement the Observable paradigm in the Observer pattern. It contains a basic collection-type property, which is publicly writable by the method `registerListener`, and inside the method `test` all the registered listeners will have the notification method called. The next code sample shows how to use the above classes:

```
public class ObserverTest {

   public static void main(String[] args) {

            Observable obs = new Observable();

            Listener l = new Listener() {

                     public void onNotification() {

                              System.out.println("Notification
received");

                     }

            };

            obs.registerListener(l);

            obs.test();

   }

}
```

Using the previous classes as input for the tested applications, I have obtained the following results:

Application tested	Result
Similarity Scoring	1
Pinot	1
Fujaba	0
CASA	1

Table 1 Result of running the applications against a simple implementation.

Results marked with "1" denote that the application has correctly identified the classes as an implementation of the Observer pattern, while the ones marked with "0" failed to detect them.

As we can see, the majority of the applications tested will recognize the very simple implementation. Of course, such simple implementations are not common case; in practice it's more likely to have implementations from the following categories.

3.2.2 Complex Implementations

These implementation use additional language constructs to implement the design pattern. While not strictly necessary for a complete implementation of the pattern, this is used when there is need for other functionality besides the simple Observer pattern.

As the pattern is usually part of a more complex system, it is common that additional functionality is required, and the possibilities are countless: we might require logging, selection of the observers to be notified, persistence for the observer references, just to name a few.

One of the techniques used when we need additional information attached to an object is *wrapping*. This is the encapsulation of the original object and the required information into a single class, as shown in Figure 2.

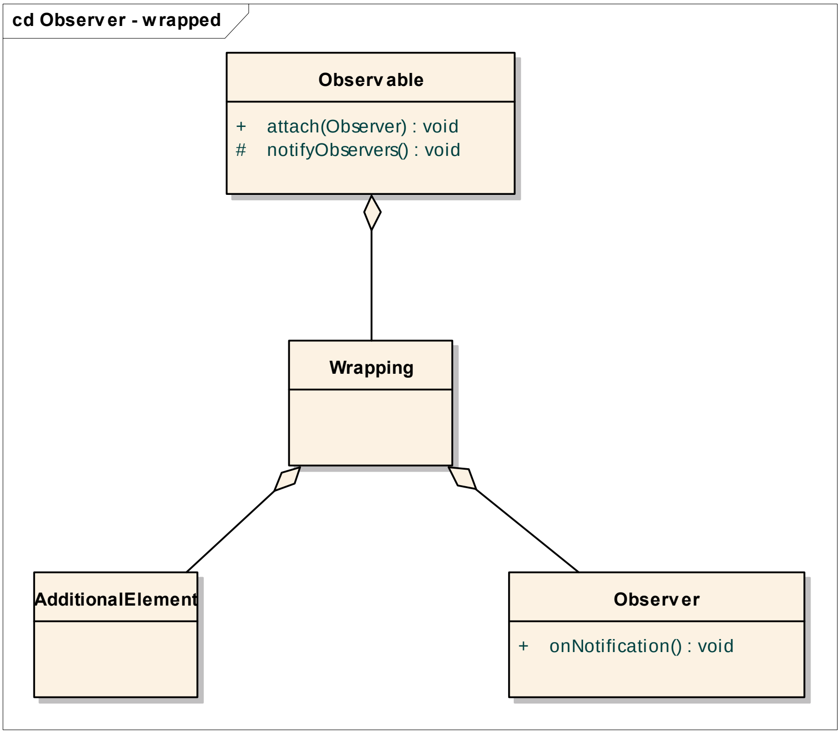


Figure 2 Wrapping the Observer.

For instance, we would like to remember the last time the Observer has been notified. For this purpose, we will wrap the Observer together with the last notification time, like in the following code sample:

```
class StampedListener{
    protected Date lastNotification;
    protected Listener actualListener;
    public StampedListener() {
    }
```

```
    public StampedListener(Listener actualListener) {
            this.actualListener = actualListener;
    }
    public Date getLastNotification() {
            return lastNotification;
    }
    public void setLastNotification(Date lastNotification) {
            this.lastNotification = lastNotification;
    }
    public Listener getActualListener() {
            return actualListener;
    }
    public void setActualListener(Listener actualListener) {
            this.actualListener = actualListener;
    }
}
```

The Observer does not change, but the Observable can be implemented as follows:

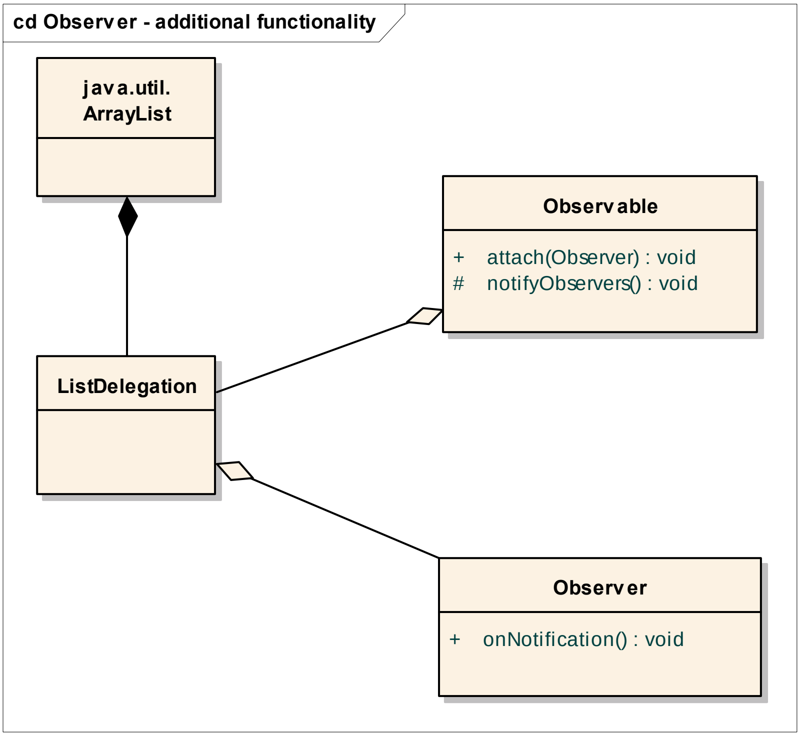
```
public class Observable {
    protected ArrayList<StampedListener> events = new
ArrayList<StampedListener>();

    public void registerListener(Listener l) {
            events.add(new StampedListener(l));
    }
    public void test() {
            for(int i=0 ; i<events.size() ; i++) {

    events.get(i).getActualListener().onNotification();
            }
    }
}
```

Running the applications against this input provides the following results:

Application tested	Result
Similarity Scoring	1
Pinot	0
Fujaba	0
CASA	1

Table 2 Result of running the applications against the wrapped Observer.

One other sample exemplifies the situation in which a simple list does not fit the application needs. For instance, we would like additional functionality for the collection that holds the observers, like the need to selectively notify the observers, let's say according to privileges. For this purpose, we would like to extend the functionality of the list. A popular choice for extending the functionality is delegation. For this example, I will aggregate a collection and expose part of its functionality. This is represented in Figure 3.



b

Figure 3 Delegating the contract of an ArrayList.

The following sample exemplifies a delegation:

```
public class ImprovedList {
    protected List list = new ArrayList();
    protected ThreadLocal threadUsername = new ThreadLocal();
    public Object tryObtain(int index) {
            if("admin".equals(""+threadUsername.get())){
                    return list.get(index);
            }else {
                    return null;
            }
    }
    public void register(Object o) {
            list.add(o);
    }
    public int numberOfElements() {
            return list.size();
    }
}
```

An implementation of an Observer that makes use of the above list is described in the following code:

```
public class Observable {
    protected ImprovedList listeners = new ImprovedList();
    public void registerListener(Listener l) {
            listeners.register(l);
    }
    public void test() {
            for(int i=0 ; i<listeners.numberOfElements() ; i++)
{

    ((Listener)listeners.tryObtain(i)).onNotification();
            }
    }
}
```

Running the applications against this input provides the following results:

Application tested	Result
Similarity Scoring	0
Pinot	0
Fujaba	0
CASA	1

Table 3 Result of running the applications against the Observers using delegated list.

As a conclusion, running the complex implementation of the patterns as input for various pattern detection applications tend to generate incorrect results of the type *False Negative*, meaning that the application will not detect the pattern, although it is a valid implementation.

3.2.3 Incomplete Implementations

These implementations attempt to implement a valid Observer pattern, but however, the implementation is incomplete, failing to verify one or more of the necessary and sufficient conditions 1a), 2a) or 3a).

The reasoning of this test category is that in practice, there are other software constructs that are similar (in principle and/or implementation) with the target pattern.

For instance, if an Observable does not add Observers to the existing references, it would not qualify as an implementation of the pattern.

Let's take the example of the simple implementation of the Observer pattern and modify it so that the method `registerListener` does not add the listener to the collection:

```
public class Observable {
    protected ArrayList listenersList = new ArrayList();
    public void registerListener(Listener l) {
            //listenersList.add(l);
    }
    public void test() {
            Iterator iterator = listenersList.iterator();
            while(iterator.hasNext()) {

    ((Listener)iterator.next()).onNotification();
            }
    }
}
```

Running the applications against this input provides the following results:

Application tested	Result
Similarity Scoring	0
Pinot	1
Fujaba	1
CASA	1

Table 4 Result of running the applications against the first incomplete implementation.

Results marked with “1” denote that applications have not identified the classes as an implementation of the Observer pattern, while results marked with “0” detected such an implementation, even though it is incomplete.

A more subtle example would be to modify the wrapped example in the previous section so that the Observer is not passed to the wrapped object. In this way, we break the rule 2a), so the following code is not a valid implementation of the Observer pattern:

```
public class Observable {

   protected ArrayList<StampedListener> events = new
ArrayList<StampedListener>();


   public void registerListener(Listener l) {

           //events.add(new StampedListener(l));

           events.add(new StampedListener());

   }

   public void test() {

           for(int i=0 ; i<events.size() ; i++) {

   events.get(i).getActualListener().onNotification();

           }

   }

}
```

Running the applications against this input provides the following results:

Application tested	Result
Similarity Scoring	0
Pinot	1
Fujaba	1
CASA	1

Table 5 Result of running the applications against the second incomplete implementation.

As a conclusion, running the incomplete implementations as input for the various pattern detection tools tend to generate incorrect answers of the type *False Positive*, meaning the application has detected an implementation of the pattern, although the implementation is incomplete.

4. Conclusions

The myriad of existing software applications and the number of developers makes the task of detecting a software construct a fuzzy domain, even for simple and well known software patterns, like the one presented here, the Observer.

Testing using artificially constructed cases show that applications in general tend to detect simple implementations of the patterns, while more complex implementations are hard to detect.

Additional testing on various open source applications complements the testing based on categories defined earlier, showing that this approach has potential to be extended to other more complicated software constructs, like the Object Pool pattern, or architectural patterns, like MVC, SOA, Domain Objects, etc.

References

1 E. Gamma, R. Helm, R. Johnson, and J. Vlissides, Design Patterns: elements of reusable object oriented software, Addison Wesley, Reading MA, USA, 1994.

2 http://java.uom.gr/~nikos/pattern-detection.html.

3 V.D. Blondel, A. Gajardo, M. Heymans, P. Senellart, and P. Van Dooren, "A Measure of Similarity between Graph Vertices: Applications to Synonym Extraction and Web Searching," SIAM Rev., vol. 46, no. 4, pp. 647-666, 2004.

4 N. Tsantalis, A. Chatzigeorgiou, G. Stephanides, S. T. Halkidis, "Design Pattern Detection Using Similarity Scoring", IEEE Transactions on Software Engineering, vol. 32, no. 11, pp. 896-909, November, 2006.

5 http://www.cs.ucdavis.edu/~shini/research/pinot/

6 ***Reverse Engineering of Design Patterns from Java Source Code*** Nija Shi and Ron Olsson Appeared in the 21st IEEE/ACM International Conference on Automated Software Engineering (ASE'06), Tokyo, Japan, September 2006.

7 http://www.fujaba.de

8 J. Niere, W. Schäfer, J. P. Wadsack, L. Wendehals, and J. Welsh, “Towards Pattern-Based Design Recovery”, Proceedings of the 24th International Conference on Software Engineering, Orlando, Florida, USA, 2002, pp. 338-348.

Appendix

List of Acronyms

MVC	Model-View-Controller
SOA	Service-Oriented Architecture

The Importance of Information and Communication Technologies in the Learning Process: The Case Study of Mathematics in Secondary Education in Greece

Keramida Konstantina

University of Macedonia, Greece, keramida@uom.gr

Purpose-**This case study evaluates the rate at which the integration of ICT in the learning process has improved the teaching of Mathematics, emphasizing on the extraction of useful information on teachers' conduct with regard to their age, sex, special qualifications, skills and the type of secondary education they work in, after their attendance of the European Program "Information Society".**

The purpose of this case study is to:

Examine whether the mathematicians who took part in this European program, actually use the knowledge acquired.

Evaluate their opinion on the effectiveness of the use of ICT and also of the use of the knowledge they acquired, in class.

Design/methodology/approach-**The methodology that was followed includes questionnaires and personal interviews. The questionnaire was filled in by 101 teachers who teach Mathematics in secondary schools, attended the educational program "Information Society", have access to computer laboratories during classes and also have access to the internet from home.**

Findings-

1.There is a differentiation between male and female teachers as well as, between teaching in secondary education and college. According to the findings of this study, male teachers use ICT in school and at home more frequently in greater length compared to their female colleges. Furthermore, secondary teachers of

maths science are more experienced ICT users compared to their college colleges. This differentiation is attributed to the fact that secondary teachers have less pressure during their lesson and more spare time for ICT. On the contrary, college teachers have more time constrains to cover and conclude the subjects taught and take on work at home, limiting their spare time to be used on other purposes, including ICT.

2. The "Information Society" European Program did not succeed in enabling the maths teachers to effectively use ICT in class or even use it at all. They could not assimilate the material taught due to the short time frame of the program and the diversity of the participants with regard to sex, subject specialization, age and ICT background.

***Originality/value*-The findings of this study contribute to the improvement of this European Training Program in ICT and support the integration of ICT in the learning process.**

Keywords

ICT and secondary teaching, high school and college, "information society", secondary education, mathematics and ICT.

1. Introduction

This case study evaluates the rate at which the integration of ICT in the learning process has improved the teaching of Mathematics, emphasizing on the extraction of useful information on teachers' conduct with regard to their age, sex, special qualifications, skills and the type of secondary education they work in, after their attendance of the European Program "Information Society". The research based on the following 2 hypotheses:

1st hypothesis:

The type of school in which the Mathematics teacher works in determines the degree of use of the internet, during classes as well as in his or her spare time, for informative reasons with regard to the maths science. High school teachers make better use of the internet compared to college teachers.

2nd hypothesis:

The teacher's gender influences the extend of use of the internet for the above mentioned reasons. Male teachers make better use of the internet and are more experienced users compared to their female colleagues.

2. Methodology

2.1 Data Analysis

The participants' answers to the survey, after being coded, were input into the statistics programme SPSS 11.5 for Windows and analysed. Initially, the data were statistically descripted and the summary statistics were presented. Secondly, the variation of the variables E1 and E2 was carried out with respect to the rest of the variables to establish whether any correlation is statistically significant. This analysis was conducted using ANOVA (Analysis Of Variance). It was considered the best method for our Loglinear model, Logit model in particular, taking into account that our data is qualitative and stratified. Furthermore, we used OLS method for the regression analysis of our data. We examined the correlation indexes between the type of school and the gender and the rest of the variables. For this purpose, we used Crosstabs, Correlation Coefficient Spearman and Pearson ChiSquare.

2.2 Sample

In the present research, the method of non-random sampling was used. The statistic population of this research is all mathematics teachers of secondary and college education of the Municipality of Thessalonica. Out of the 742 Mathematics teachers of schools of Secondary Education at the Municipality of Thessalonica (380 teach in high schools and 362 teach in colleges) (Table 1), we selected for the purpose of our research 101, that is 13.61 per cent of the total statistic population. Our sample is made up of 101 maths teachers of which, 52 teach in high schools and the other 49 are college teachers (Table 2), 72 of those teachers are men and 29 are women (Table 3). The vast majority of the participants are aged 46 to 50 years old. The data collection was accomplished through the use of a questionnaire and personal interviews. The questionnaire was sent to the teachers via email.

Type of School	Number of Teachers of Mathematics	Teachers (%)
High School	380	51
Colleges	362	49
Sum	742	100

Table 1 Teachers of Statistic Polpulation

Type of School	Number of Teachers of Mathematics	Teachers (%)
High School	52	51
Colleges	49	49
Sum	101	100

Table 2 Teachers of Sample

Type of School	Male Teachers of Mathematics	Female Teachers of Mathematics	Male Teachers (%)	Female Teachers (%)
High School	32	19	45	66
Colleges	39	10	55	34
Sum	72	29	100	100

Table 3 Teachers of Sample by Gender and Type of school

3. Data-Questionnaire Variables

The questions that were presented to the teachers are shown bellow. Every question is represented by a variable.

VARIABLE	QUESTION
E1	In which type of secondary education school do you teach?
E2	Gender
E3	Age
E4	Do you have internet access at home?
E5	In which user category would you classify yourself before taking part in the ICT programmes?
E6	Was your reason for taking part in the ICT programmes educational?
E7	To what extent do you use ICT in the teaching process as a tool?
E8	To what extent do you use ICT in preparing for your lessons?
E9	To what extent do you use the internet in the teaching process as well as at home for informative reasons with respect to Maths science?
E10	To what extent are you experienced users in Word Processor and Excel?
E11	To what extent are you experienced in teaching Word Processor and Excel?
E12	What are the skills that you acquired after participating in ICT?

Table 4 List of Variables

4. ANOVA (ANalysis Of VAriance)

4.1 Analysis of Variance E1

It is deduced from our research that the extent of use of ICT in the teaching process of Mathematics in Secondary Education is 5 per cent statistically significant. There is, thus, statistical significance between high school and college teachers in the use of ICT in the teaching process of Mathematics (E1-E7).

In addition it is concluded from our research that the extent of use of ICT in the preparing the lessons of Mathematics in Secondary Education is 5 per cent statistically significant. There is, thus, statistical significance between high school and college teachers in the use of ICT in preparing for the lessons of Mathematics (E1-E8).

It is important to refer that the extent of use of ICT and internet in the teaching process as well as at home for informative reasons with respect to Maths science is 5 per cent statistically significant. There is, thus, statistical significance between high school and college teachers in the use of ICT for searching information relative to Maths (E1-E9).

4.2 Analysis of Variance E2

It is deduced from our research that the extent of use of ICT in the teaching process of Mathematics between males and females teachers is 5 per cent statistically significant. There is, thus, statistical significance between men and women teachers in the use of ICT in the teaching process of Mathematics (E2-E7).

In addition it is concluded from our research that the extent of use of ICT in the preparing the lessons of Mathematics by gender is 5 per cent statistically significant. There is, thus, statistical significance between male and female teachers in the use of ICT in preparing for the lessons of Mathematics (E2-E8).

It is important to refer that the extent of use of ICT and internet in the teaching process as well as at home for informative reasons with respect to Maths science is 5 per cent statistically significant. There is, thus, statistical significance between male and female teachers in the use of ICT for searching information relative to Maths (E2-E9).

5. Findings

According to the results of regression analysis, we have proven that there correlation between variables, this means that we accept the 1st hypothesis as well as the 2st hypothesis. Based on the above findings, we conclude that:

1. There is a differentiation between male and female teachers as well as, between teaching in secondary education and college. According to the findings of this study, male teachers use ICT in school and at home more frequently in greater length compared to their female colleges. Furthermore, secondary teachers of maths science are more experienced ICT users compared to their college colleges. This differentiation

is attributed to the fact that secondary teachers have less pressure during their lesson and more spare time for ICT. On the contrary, college teachers have more time constrains to cover and conclude the subjects taught and take on work at home, limiting their spare time to be used on other purposes, including ICT.

2. The “Information Society” European Program did not succeed in enabling the maths teachers to effectively use ICT in class or even use it at all. They could not assimilate the material taught due to the short time frame of the program and the diversity of the participants with regard to sex, subject specialization, age and ICT background.

References

1 Collis, B 1989. Using Information Technology to Create New Educational Situations. Higher Education Policy. vol 2, no 4, pp. 63-70.

2 Bell, A 1993. Some experiments in diagnostic teaching. Educational Studies in Mathematics. vol. 24, pp.115-137.

3 Chambers, M 2005. Problem Set 7: Serial correlation and dynamic models. University of Essex. Department of Economics.

4 Christensen, R 1997. Effect of technology integration education on the attitudes of teachers and their students. PhD thesis. University of North Texas. Denton.

5 Chronis, K 2001. New Practices with New Tools in the Classroom: Educating Teacher Trainers in Greece to Generate a “School Community Use of New Technologies. Themes in Education. 2:4. pp. 381-399.

6 Colley, A & Comber, C 2003. Age and gender differences in computer use and attitudes among secondary school students: what has changed?. Educational Research. vol. 45. no. 2. pp.155-165.

7 Fuchs, T & Wossmann L, Computers and student learning: Bivariate and Multivariate Evidence on the Availability and use of computers at home and at school. viewed October 2004. http://www.cesifo.de/link/fuchs_t.htm.

8 Gipps, C & James, M 1996. Assessment matched to learning. paper presented in a symposium of BERA Assessment Policy Task Group. 12 September 1996. University of Lancaster.

9 Gujarati, D 1988. Basic Econometrics. McGraw-Hill Book Company.

10 Gunilla, J & Jorden, N 2004. ICT in the Classroom: Is Doing More Important than Knowing?. Education and Information Technologies. 9:1. pp. 37-45.

11 Javris, P Holford, J & Griffin, C 1998. The Theory and Practice of Learning.

12 John, P & Gravani M 2005. Evaluating a “New” in-service professional development programme in Greece: the experiences of tutors and teachers. Journal of In-service Education. vol. 31. no.1. pp. 105-129.

13 Knezek, G & Christensen, R 2002. Impact of new Information Technologies on Teachers and Students. Education and Information Technologies. 7:4. pp. 369-376.

14 Kozma, R B 2003. Technology and Classroom Practices: An International Study. Journal of Research on Technology in Education. vol. 36. no. 1. pp. 1-14.

15 Li, Q 2003. Would we teach without technology? A professors experience of teaching mathematics education incorporating the internet. Educational Research. vol. 45. no. 1. pp. 61-77.

16 Mccormick, R & Scrimshaw, P 2001. Information and Communications Technology. Knowledge and Pedagogy. Education. Communication and Information. vol. 1. no 1. pp. 37-57.

17 Nason, R & Woodruff, E 2002. New ways of learning Mathematics: Are you ready for it?. proceedings of the International Conference on Computers in Education. IEEE.

18 Papert, S 1979. Computers and Learning. The Computer age: A twenty-year View. Cambridge. MA: The MIT Press. pp.73-86

19 Russell, A L 1995. Stages in learning new technology: naïve adult email users. Computers Education. vol. 25. no. 4. pp 173-178.

20 Sandholtz, J Haymore, R Cathy, D & David, C 1997. Teaching with Technology: Creating Student-Centered Classrooms. Teachers College Columbia University. New York. NY 10027.

21 Shashaani, L 1994. Gender-differences in computer experience and its influence on computer attitudes. J. Educational Computing Research. vol 11. no. 4. pp. 347-367.

22 Turner, J. R & Thayer, F. J 2001. Introduction to analysis of variance. Sage Publications.

23 Underwood, J & Underwood, G 1990. Computers and Learning. Helping children acquire thinking skills. Blackwell Oxford UK & Cambridge USA.

24 Wang, P Cheng, W Wang, W & Hung, P 2002. An elementary school Mathematics dynamic learning system and its effects. proceedings of the International Conference on Computers in Education. IEEE.

25 Webb, M E 2002. Pedagogical Reasoning: Issues and solutions for the teaching and learning of ICT in secondary schools. Education and Information Technologies. 7:3. pp. 237-255.

26 Wendy, M 2001. Not Just Tools: the Role of E-technologies in Culture of Learning. Education. Communication & Information. vol. 1. no. 2. pp.229-235.

Web Page Classification Based on Artificial Neural Network

Zlatica Korkaric

University of Novi Sad, Technical Faculty "Mihajlo Pupin", Dj. Djakovica bb, Zrenjanin, Serbia, zlata@tf.zr.ac.yu

Nowadays, the focus of research is on the automatic classification of web page information. Application of artificial neural networks is suitable for many complex problems such as multi-class pattern recognition. For the automatic classification of the web pages advantages of neural networks is used. In this research neural network is designed and trained to classify web pages with educational content in defined classes. Some preliminary experimental findings show that this approach is successful in solving problem of classification of unseen web pages. Suggested approach can be useful in solving tasks related to the quality web page classification.

Keywords
artificial, classification, network, neural, page, Web.

1. Introduction

In the information processing areas automatic classification and ranking of web pages is very popular [1], [2]. Considering the problem of the automated classification of web pages, web pages' text information, structure information and hyperlink information had to be taken in to the consideration. Many sources outline usage of then artificial neural networks in solving the problems of web page classification and ranking. Nowadays, the focus of the research is set on the automatic classification of web page information. Due to the diversity of web pages' content, complexity of structure and other characteristics, it was very difficult to improve the accuracy of automatic classifications. Although researchers have designed all kinds of classifiers for automatic classification such as Naïve Bayes classifier, K-neighbor Clustering, Self-Organizing Mapping Neural Network, Support Vector Machine, these classifiers have their own characteristics and conditions of application [3].

In this paper one approach to the automatic classification of web pages is suggested. Web pages that are used in this experiment are the web sites that were made by the students of Technical faculty in Zrenjanin as assignments during the one of study courses [4]. Web sites as a set of web pages were designed on different themes. This experiment set of web pages had 2063 web pages. Feature of web page was based on the crucial keywords. And specially designed SOM neural network was trained to classify web pages from the experiment set of web pages.

2. Web Page's Feature Extraction

Two aspects of web page's feature extraction based on keywords were taken in to consideration. One aspect is keyword appearance in a document of a certain class; and second aspect is frequency of the keyword. Software for comparison has been designed and implemented in to process of extraction of web page features. At first keyword dictionary is set up. Then this dictionary was used to scan the web pages and get the frequency of keywords. Keyword dictionary consists of many words, whereas a large part of them does not contribute to the classification. The selection of useful keywords is very important. Feature compression is one of the methods to select keywords. Today we have several feature compression algorithms and some of them are: Principal Component Analysis (PCA), Information Entropy Loss, and word frequency threshold [5], etc.

We have selected Vector Space Model (VSM), nowadays, one of the popular ways to dentate text features. In this model, the document space is regarded as a vector space composed by a group of orthogonal word vector, and every document d represent a specific vector in it: $V(d)=(t_1, x_1(\mathrm{d}); \ldots; t_\mathrm{i}, x_\mathrm{i}(\mathrm{d}); \ldots; t_\mathrm{n}, x_\mathrm{n}(\mathrm{d}))$, with t_i representing the word and $x_\mathrm{i}(\mathrm{d})$ representing the weight of t_i in document d. The frequency of t_i in d is defined as function of $tf_i(d)$, notated as this: $x_1(\mathrm{d})=\varphi(tf_i(d))$. Functions φ are as follows.

Boolean function: $\varphi = \begin{cases} 1, tf_i(d) \geq 1 \\ 0, tf_i(d) = 0 \end{cases}$ (1)

Square Root function: $\varphi = \sqrt{tf_i(d)}$ (2)

Logarithm function: $\varphi = \log(tf_i(d) + 1)$ (3)

TFIDF function: $\varphi = tf_i(d) \times \log\left(\frac{N}{n_i}\right)$ (4)

Forms to denote our feature vectors are as follows:

1) Simply word frequency feature. Every record is corresponding to a vector $V_j=(x_{j1}, x_{j2}, \ldots, x_{jn})$, where x_{ji} is the number of appearing times of the i-th keyword in the j-th document.

2) Feature of division of word frequency by document length. Every record is corresponding to a vector $V_j=(x_{j1}/L_j, x_{j2}/L_j, \ldots, x_{jn}/L_j)$, where x_{ji} is the number of appearing times of the i-th keyword in the j-th document and L_j is the length of the j-th document.

3) ID/TIF feature set. Every record is corresponding to a vector $V_j=(x_{j1}/v_1, x_{j2}/v_{j2}, \ldots, x_{jn}/v_n)$, where v_i is the number of documents in which the i-th word appears.

4) Normalized frequency feature set. Every record is corresponding to a vector $V_j=(x_{j1}/S_j, x_{j2}/S_j, \ldots, x_{jn}/S_j)$, where $S_j= \sqrt{\sum_{i=1}^{n} x_{ji}^2}$.

In order to filter useless keyword feature, we introduce the concept of Frequency Covering Rate.

$$x_{ij} = \frac{f_{ij}}{\sum_{j=1}^{n} f_{ij}} \cdot \frac{d_{ij}}{N_j} \quad (5)$$

Where f_{ij} is the frequency of the i-th keyword in the j-th document, d_{ij} is the number of the j-th class documents in which the i-th keyword appears, and N_j. is the total number of the j-th class documents. x_{ij} reflects the correlating degree between the i-th keyword and the j-th class documents. Size of x_{ij} denotes the characteristics of the j-th class. So we have set a threshold T to filter keyword features.

$$\overline{x_i} = \max_j \left\{ x_{ij} \right\} \quad (6)$$

If $\overline{x_i}$ >T, then we keep the i-th keyword and use it in dictionary.

3. Explanation of the SOM Classifier

One of the SOM neural network models is a Kohonen model. A Kohonen model is composed of two layers neurons, the input layer and the output layer, which also called competition layer. We can denote input vector as $\mathbf{x} = (x_1, x_2, ..., x_n)^T$, then the connecting weight between the neurons in the input layer and those in yhe competing layer is $w_j=(w_{1j}, w_{2j}, \ldots, w_{nj})^T$, j=1, 2, …, H, and the output of competition layer neurons is

$$y_j = \sum w_{ij} x_i = w_j^T \mathbf{x}, \ j=1, 2, \ldots, H \quad (7)$$

Presented in the vector form looks like following:

$$y = Wx \quad (8)$$

By self-organizing competition, SOM neural network can change the disordered input set $X = \{x^{(k)}\}_{k=1}^{K}$ into ordered topology connection in the competition layer, such as the distribution of clustering centers. Among all the objects in the input set, clustering centers are those connection weights corresponding to the winner unit in the competition layer's output. So self-organizing is the process of looking for connection weights that best match the input vector. We use the following criterion to find the winner neuron.

$$h(\mathrm{x}) = \arg \min_j \left\| x - w_j \right\| , \ j=1, 2, \ldots, H \quad (9)$$

In the stated equation, $\left\| \ \right\|$ is the Euclid norm, it describes a sort of mapping, which is a continuous input vector space $R^n \supset X$ to the discrete neuron space $\{j=1, 2, \ldots, H\}$ by competitive actions among competing units, with the restriction $h(x) \in \{1, 2, \ldots, H\}$. Thus, we have achieved ordered division of the object X in the vector input space R^n..

4. Explanation of the Learning Algorithm

SOM neural network's learning can be done by Hebb learning rule [3]. However, considering the irreversible transferring of information, we have to introduce the oblivious factor $g(y_i)\, w_j$, with w_j be the connecting weight corresponding to the j-th neuron in the competition layer, and $g(y_i)$ be the responding function of y_i.

If we let $g(y_j) = \eta\, y_j$, as η is the learning step length, then we get the expression to modify the weights.

$\Delta w_j = \eta\, y_j x - g(y_j)\, w_j$ (10)

If we let $y_j = h_{i,j(x)}$, then the above written equation can be rewritten as following:

$\Delta w_j = \eta\, h_{i,j(x)}(x - w_j)$ (11)

And weights' iteration formula is

$w_j(t+1) = w_j(t) + \eta(t)\, h_{i,j(x)}(t)\, (x - w_j(t))$ (12)

Where, respectively, $w_j(t)$, $\eta(t)$ and $h_{i,j(x)}(t)$ are the weight, learning step length and neighbor area function after the t-th iteration.

The neighbor area function is defined as following:

$$h_{i,j(x)} = \exp\left(-\frac{d_{i,j}^2}{2\sigma^2(t)} \right) \quad (13)$$

Where, $\sigma(t) = \exp(-\frac{t}{\tau})$, with τ be a constant designating the maximum number of iterative times.

The SOM is a kind of unsupervised learning algorithm. If we have known the labels of the samples, we can go farther using LVQ (Learning Vector Quantization) algorithm to modify the clustering centers. LVQ algorithm is described simply as follows.

For all the input vectors $x^{(k)} \in X$, repeat the following process. Assuming weight vector w^c was the most alike vector with input vector x_i, w^c and x_i belong to the C_{w^c} class and $C_{x(i)}$ class, respectively, then the modifying rule of weight vector w^c is as follows.

If $C_{w^c} = C_{x(i)}$, then

$w^c(t+1) = w^c(t) + \alpha_t[x^{(k)} - w^c(t)]$, $\alpha_t \in (0,1)$.

If $C_{w^c} \neq C_{x(i)}$, then

$w^c(t+1) = w^c(t) - \alpha_t[x^{(k)} - w^c(t)]$,

Otherwise weight vectors remain unchanged..

5. Conclusions

Described neural network was created in MatLab software and trained. These students' web pages in total of 2096 web pages, was divided in to tree classes. In Entertainment class was 1012, in Community/Society class was 52 and in Education/Science class was 936 web pages.

Initial dictionary collected 236891 words, and after filtering the keywords using the class covering rate threshold, dictionary was resized to 5236 keywords. Afterwards corresponding feature vector was established by counting the numbers of times these keywords appear in certain web page.

In experiment of automatic classification of 1012 web pages, only 851 were classified in Entertainment class, of 52, only 38 were classified in Community/Society class and of 936, only 894 were classified in Education/Science class. That makes the percentage precision 84.09% Entertainment class, 73.07% in Community/Society class and 95.51% in Education/Science class.

Further research would be focused on testing the neural network on greater set of web pages with comparison to some representative other kinds of neural networks. Some hypotheses will be based on larger number of classes and different approaches to creation of keyword dictionary.

References

1 Scarselli F, Young S L, Gori M, Hagenbuchner M, Tsoi A C, Maggini M, Graph Neural Networks for Ranking Web Pages. Proceedings of the 2005 IEEE/WIE/ACM International Conference on Web Intelligence, 2005.

2 Liang J, Chinese Web Page Classification based on Self-Organizing Mapping Neural Networks, Proceedings of the Fifth ICCIMA'03, 2003.

3 Demuth H, Beale M, Neural Network Toolbox For Use with MATLAB, 1998.

4 Korkaric Z, Methodology for the Evaluation of the Educational Web Presentations, Thesis on master studies, Authors' reprint (in serbian), Zrenjanin 2004.

5 Sebastiani F, Machine Learning in Automated Text Categorisation, ACM Computing Surveys, Vol. 34, March 2002.

Appendix

List of Acronyms

SOM	Self-Organizing Mapping neural network
VSM	Vector Space Model
LVQ	Learning Vector Quantization

Comparisons of Verifications in VDM and the Z Notation Formal Methods

Cozma Pompiliea[1], Tudor Balanescu[2]

[1,2]University of Pitesti,Faculty of Mathematics and Computer Scienc.Department of Computer Science, 110040,Pitesti.Str. Targu din Vale, nr.1, Arges, Romania, pcosma99@hotmail.com[1], tudor.balanescu@upit.ro[2]

The aim of this paper is to present, from a comparative point of view, those verifications which can be performed within the specification of software system by means of Vienna Development Method (VDM) and Z Notation. The frame of formal systems is considered from the mathematics perspective of the sets theory. Firstly, the general context of this paper will be briefly mentioned - formal specifications methods, properties of formal systems (consistency, completeness, decidability). Secondly, the main verification of VDM and Z Notation will be mentioned in the subsection 2.2. Refinement and Proof Obligation in VDM and Z Notation section will compare different approaches to VDM and Z Notation. Thus, data refinement - initial state validation, domain obligation, result obligation - will be analyzed and we will discuss the conditions which have to be met theoretically in the two VDM and Z Notation formal systems for the conclusion – the data refinement theory for VDM and Z Notation specifications are one and the same. In perspective it is interesting to study this conclusion with the formal perspectives on software evolution from refinement to retrenchment.

Keywords

formal specifications methods, verification of formal systems, Vienna development method, Z notation.

1. Introduction

The mathematics theory - algebra, logic, sets theory represent the basis of formal specifications methods [1]. This confers the main properties of formal systems: consistency (all provable statements are semantically true), completeness (all semantically true statements are provable) and two important consequences of completeness (compactness and decidability) which are very important for the verifications that can be performed within the specification of software systems [2], [3].

2. VDM and Z Notation

VDM and Z Notation with a common base of the mathematics theory - sets theory, are included in the class of model-oriented formal specifications methods and they are described with the help of representational abstraction and operational abstraction [3], [4].

2.1 Comparisons of Specifications in VDM and Z Notation

In VDM the representational abstraction is used in building the data types and the details of structure, while in Z Notation it is used to describe the syntactic structures - type definition, global constants and state space declaration.

The operational abstraction in VDM describes the abstract algorithmic manipulation of the data introduced in the representational abstraction as first - order predicate logic formulas, while in Z Notation it formalizes operation and function within an abstract state.

The different approaches of the abstraction in VDM and Z notation determine the different basis structures: a VDM specification uses various blocks identified with a distinct keyword, while the kernel of the Z notation is 'the scheme' notation.

The blocks used in the VDM specification are: types, values, functions, states, operations, each definition corresponds to a notion.

The scheme notation, used in the Z Notation, can be viewed as an encapsulated structure which is commonly used to represent types, state spaces and operations.

A comparative case study in VDM and Z Notation is analyzed in [5].

2.2 The Role of Proof in VDM and Z Notation

A general view on the proof is a mechanism to prove that the claim Q is a logical consequence of some premises P1 and Pk. A formal proof consists of the steps which show that all implicit assumptions are made explicit, with reference to the sources used in deriving each step [6].

Therewith the development of software requires the kind of role of proof which should provide de necessary accuracy. In traditional programming languages, the consistency checking for program is performed at run time. But formal specifications are not executable in general.

Thus, the role of proof has more perspectives - the check of specification for consistency, the refinement proofs or the program proving.

These notions are studied separately in more papers as [6], [7], but it is difficult to establish if proprieties are respected. Also, for two formal systems, the way for checking these notions is different for consistency and the program proving.

Due to this, the common perspective in VDM and Z Notation of role of proof - refinement and proof obligation is analyzed in follow-on.

3. Refinement and Proof Obligation in VDM and Z Notation

In VDM and Z Notation there are two ways by which a specification can be refined - operation refinement and data refinement.

These notions are studied in VDM or Z Notation in more papers as [9], proposing the generating proof obligation to verify object –Z specification or [10], as proof in VDM – study case.

The common perspective of operation refinement and data refinement in VDM and Z Notation are analyzed below.

3.1 Operation Refinement

The different definition of the operation refinement in VDM: one level operations are refined to one or more operations in which details are explicit, while in Z Notation it focuses on mapping an operation in the abstract state space to one or more operations in the concrete state space, which determines a different way of implementing these notions.

3.2 Data Refinement

A proof obligation for the data refinement in VDM establishes that for every abstract data type there exists at least one concrete type that implements it, while in Z Notation data refinement a data type is an abstract state space which is refined into another data type in the concrete state space.

VDM proof obligation for a data refinement requires showing that no data are lost and no new data are introduced; for every operation that modifies data in the abstract level, the data in the refined data type is modified to yield the same effect. Formalizing this consideration (by defining a retrieve function that maps a concrete state space into its abstracte space) five components in a proof obligation [8] are acquired – signature verification, adequacy obligation, initial state validation, domain obligation, result obligation.

In Z Notation a data refinement process must satisfy three conditions [6] which will be compared in the following sections with initial state validation, domain obligation, result obligation of VDM.

This comparison focuses on mapping of concrete state spaces to abstract state spaces, defined as a function (retrieve function) in VDM, whereas it is defined by a scheme in Z Notation.

From this perspective these conditions will be analyzed in the following.

3.3 Initial State Validation

Taking into account the following two definitions, from the previous perspective, one can notice that these two are identical.

This requirement in VDM consists in: the retrieve function should match the initial concrete state to the initial abstract state as defined by the initialization functions in both state spaces.

The correspondent requirement in Z Notation is: every initial state for the concrete state space corresponds to a valid initial state for the abstract state space.

3.4 Domain Obligation

For every concrete operation Opc, in VDM, the precondition of the corresponding abstract operation Opa in conjunction with the retrieve function ensures the precondition of Opc. Informally, domain obligation ensures that the precondition of the concrete operation is weaker than that of its corresponding abstract operation.

In Z Notation, the corresponding condition is: a concrete operation terminates whenever the corresponding abstract operation is guaranteed to terminate. That is to say, the precondition of the concrete operation must be weaker than that of the abstract operation.

One can notice again that the two conditions coincide.

3.5 Result Obligation

Finally, result obligation in VDM is defined as follows: for every abstract operation Opa the postcondition of its corresponding concrete operation Opc, in conjunction with the retrieve function, ensures the postcondition of Opa.

Corresponding condition in Z Notation is: every state in which a concrete operation terminates corresponds to a member of those abstract states in which the corresponding abstract operation could terminate.

Again, the previous perspective leads to the same result.

3.6 Refinement Theory

It is interesting to notice that by including the abstract and concrete states in the declaration part of the refinement scheme, the latter satisfies the signature and adequacy obligations warranted by data refinement.

4. Conclusions

Thus, the data refinement theory for VDM and Z Notation specifications are one and the same.

In perspective, it is interesting to study this conclusion with the formal perspectives on software evolution from refinement to retrenchment, which are approached in [11].

References

1 Cazanescu V. Curs de bazele informaticii. Tipografia Universitatii Bucuresti. 1983.

2 Kroger F. Temporal logic of programs. EATCS Monographs on Theoretical Computer Science. New - York Springer -Verlag. 1987.

3 Norcliffe A, Slater G. Mathematics of software construction. Ellis Horwood. 1991

4 Slonnereger K, Kurtz B L. Formal syntax and semantics of programming languages. Addison-Wesley Publishing Company. 1995.

5 Hayes I J. VDM and Z: A comparative case study. Formal aspects of computing. 4(1). 76-99. 1992.

6 Alagar V S, Periyasamy K. Specification of software systems. Springer. Berlin. 1998.

7 Woodcook J C P, Davis J. Using Z: Specification, Refinement and Proof. Prentice Hall. 1996.

8 Jones C B, Jones K D , Lindsay P A, Moore R. mural: AQ Formal Development Support System. Springer-Verlag. London. 1991.

9 Zhicheng W, Huaikou M, Hongwei Z. Generating Proof Obligation toVerify Object-Z Specification. Software Engineering Advances. International Conference on Volume Issue Oct 2006. Page(s):36-38.

10 Bicarregui J C (Ed). Proof in VDM : Case Studies. Springer. Berlin. 1998.

11 Poppleton M, Groves L. Formal perspectives on software evolution from refinement to retrenchment. http://eprints.ecs.soton.ac.uk/12347/

Three Case Studies of Cities that have been Nominated as Intelligent Communities by the Intelligent Community Forum

Angelidou Margarita[1], Komninos Nicos[2]

Urban and Regional Innovation Research Unit (URENIO), Faculty of Engineering, Aristotle University of Thessaloniki, A.Papanastasiou 138, 54249 Harilaou, Thessaloniki, Greece, mangelidou@gmail.com[1], komninos@urenio.org[2]

Issy-les-Moulineaux (France), Waterloo (Ontario) and Singapore are three cities which have been nominated as Top Intelligent Communities by the Intelligent Community Forum in the period 2005-2007. This paper presents key aspects of their policy towards creating a broadband economy and highlights similarities and differences amongst them. Whereas each city has followed its own path towards becoming an intelligent community, long-range planning, strong vision and insightful approach stand out as the keys of success.

Keywords

information and communication technology, intelligent community, issy-les-moulineaux, Waterloo, Singapore.

1. Introduction

Over the past decades, Information and Communication Technologies (ICT) are undergoing radical advance and experience the continuous emergence of new applications. The diffusion of the Internet, as well as of wireless technologies lead to what is well known as *technological convergence* and *mobile and interactive computing*, thus forerunning a very different future of the urban life. In the context of this change, cities strive to climb up the urban hierarchy, by using the broadband economy to enhance security, health, education and culture services, and also by gathering around them complexes of highly specialized activities, based on services, innovation and commerce.

On the other hand, the Intelligent Community Forum (ICF) [1] is a nonprofit think tank, which focuses on job creation and economic development in the broadband economy. It has established a series of *Intelligent Community Indicators*, which are used as the criteria for the annual nomination of the world's *Top 7 Intelligent Communities of the Year*. The ICF is a highly respected organization, and these cities are considered as pioneers and role models for the development of vibrant Digital Age communities in the 21st Century.

The purpose of this paper is to study and compare three case studies of cities that have been designated as intelligent communities by the ICF over the past few years, thus reaching to conclusions about the ICT policies that these cities have followed. Those cities are Issy-les-Moulineaux (France), Waterloo (Ontario, Canada) and Singapore. The selection of the case studies for the purpose of this paper has been done on the basis of their different history, culture and geographical dispersion (one european, one american, and a developing eastern country), thus highlighting the emerging contradictions in matters of policy, public administration and strategic planning.

Research is mainly based on Internet sources. Sources on the Internet are quite sufficient on an initial level, including some official documents, abstracts of interviews and promotion videos of those cities. This would be rather presumable, since Issy-les-Moulineaux, Waterloo and Singapore have been characterized as intelligent communities partly due to their infrastructure in electronic governance and electronic services. It should be however mentioned, that the nature of the study is such, that is restrictive in terms of in situ observation (due to geographical distance) and printed bibliographical resources are extremely poor.

The paper consists of two parts. In the first part (paragraphs 2-4), each of the cities is analyzed separately, in terms of general economical and social facts and focusing on their policies regarding the integration of the broadband economy. The analysis is conducted in a bullet-like way, with the purpose of facilitating comparison amongst them. The second part (paragraph 5) comprises a comparative analysis of the characteristics of each city, referring specifically to critical points of their policy. The paper closes with critical conclusions which emerge throughout the analysis (paragraph 6).

2. Issy-les-Moulineaux, France

2.1 Facts about Issy-les-Moulineaux

Issy-les-Moulineaux is a commune in the southwestern suburbs of Paris, France, located along the river Seine, at a distance 6.6 km from the centre of Paris. With a population of over 63.000 inhabitants and a land area of 4.25 ha, it is today one of the most densely inhabited communes in Europe, with a solid economical base which counts 70.000 jobs [2], [3].

The town of Issy has a history of almost 2.000 years. In the 19^{th} Century, the community became an industrial zone of the Paris region and later a major site for the first aviation attempts, due to the construction of an airfield. Following the World War II, Issy underwent the same de-industrialization as other communities in the 1970s and 1980s. Today, having successfully gotten over its industrial past, it lies within the heart of the Val de Seine business district, the largest cluster of telecommunication and media businesses in France [3], [4].

Characteristic for Issy has been the stability of its political leadership- since 1980 it has had the same mayor, André Santini. This fact allowed the full implementation of a series of plans for local development, which focus on environmental sustainability, high quality of urban space and urban regeneration. Most importantly, the political leadership focused already since 1980 on the attraction of companies in the field of ICT and the continuous upgrade and employment of ICT services not only for entrepreneurs, but also for the inhabitants. Today 60% of the companies based in Issy are in information and communications technology, including Cisco Systems Europe, France Telecom, Hewlett Packard, Orange Internet, Sybase, Canal+, Canal Satellite, Eurosport, France 5 and France 24. Apart from that, 70% of households are connected to the Internet [2], [4].

Issy-les-Moulineaux was chosen in 2005 and 2007 as one of the top7 Intelligent Communities of the Year by the Intelligent Community Forum.

2.2 E-Services

In 1996 the Local Information Plan of Issy was launched, leading to the inauguration or a series of institutions for technological development and innovation, and also set the vision for digital democracy, digital government and the rendering of Issy-les-Moulineaux a Digital City. The most important initiatives that came as a result of the Local Information Plan were:

- Broadband society: free Wi-Fi access in public places and 170 computers or terminals at the public's disposal. There has been an organized effort to raise the public's awareness in matters of ICT and to bridge the digital divide. 100% of schools have now broadband Internet connections, while there is one computer available for every 13 primary school pupils [2]
- Cyber Kindergarten: parents follow the early-learning games of their children via a relative website [5]
- The Cube: the first multimedia cultural centre, completely dedicated to digital art recreation [5]
- Municipal Call Center: Call Center for public services to residents [5]
- Issy Mobile initiative: request and receipt of information (from weather alerts to cultural events) directly via SMS and paying for parking places by mobile phone [3]
- Initiatives for the elderly: a specialized training program has been implemented, helping elderly people to get acquainted with the use of Internet and e-mail. A club of "cyber seniors" has been created to allow these elderly "net-surfers" to meet, exchange their virtual experiences and share know-how with others [5]

2.3 E-Democracy – e-Voting

E-democracy and e-Voting has been an important issue for Issy:

- 900 inhabitants are members of the “Citizen Panel”, a tool for e-Democracy, through which they give their opinion and help in decisions every 3 months[6]
- Registration on the Citizen Panel is free to all inhabitants through the city’s website [2]
- Interactive City Council: since 1997 live broadcasting of the City Council’s sessions on cable TV and Internet. People can ask questions to elected representatives during breaks. Web users can participate through the city website [6]
- The Participative Budget Making Platform: since 2000 allows the inhabitants to contribute to the formation of the city’s primary goals (such as investments and developments works) [2]
- Since 2000 the “World eGov Forum” on e-democracy is conducted every year in Issy [6]
- The European Cybervote Project allows E-Voting for district councils since 2000
- Since 2007 all electoral sites are equipped with electronic voting machines [6]

2.4 E- Government

The city’s website, launched 10 years ago (www.issy.com) is the main source of information for citizens, associations, businesses, parents, youth and seniors [2]. Through the website, access is granted to:

- Issy.com and Issy.TV, the first ever local Web TV channel, which allow access to practical information, local information and news, and use of the website to present personal pages or question local politicians or get informed about the municipal budget by looking up the on-line financial report. This can be combined with computer animation for children between 9 and 14 years of age [2].
- IRIS (Information and Receipt of ISséens), a new administrative services centre which can be contacted via visit, phone call, mail, fax and email. It has it own Knowledge Base.
- Online services [5]:
    - Ordering birth, marriage and death certificates, ID cards and passports
    - Paying taxes
    - Registration on electoral rolls
    - Indicating a change of domestic situation for school files
    - Hotel room, parking spot reservations
    - Reservation of books, discs and DVDs at the Media Library
    - Reservation of games and CD-ROMs at the Game Library
    - Access to City Publications (guides and newspapers)

3. Waterloo, Ontario, Canada

3.1 Facts about Waterloo

Waterloo is a small city in Ontario, Canada. Its permanent population is about 100.000 residents, plus 20.000 temporary residents (mainly students), in an area of 6.400ha. It is adjacent to the larger city of Kitchener, although they are two distinct cities. It was first inhabited around 1800 by immigrants from Pennsylvania, and later, during the 18^{th} century, by immigrants from Germany [7].

Waterloo has a service-oriented economy, of which prominent are health insurance and high-tech sectors, as well as two major universities. The city is part of Canada's Technology Triangle (CTT), a joint economic development initiative of Waterloo, Kitchener, Cambridge and the Region of Waterloo that markets the region internationally [7]. The CTT is home to 334 technology companies and another 404 providing related services that employ about 10% of the labor force, but account for 45% of job growth. Waterloo is home to 40% of the high-tech firms, such as Sun Life Financial/Clarica, Manulife Financial, Research In Motion (maker of the Blackberry), Sybase, Google, Oracle, Adobe, McAfee, NCR Corporation and Agfa. Furthermore, 76% of businesses and 47% of households are connected to the Internet by broadband connections [4]

Apart from these, there are two major universities with 50 research institutes, which play a very important role for the local development, thanks to their various activities, ranging from innovation technology, technology transfer, spin-offs creation and the compilation of plans for local development, to the production of skilled executives for local enterprises.

Waterloo was awarded in 2006 and in 2007 as one of the top7 Intelligent Communities of the Year by the Intelligent Community Forum.

3.2 E-Services

Numerous e-services are available in the city [8]:

- Wi-Fi: Free wireless access for the public is available at the public library and within the central business district of the City
- Real time broadcasting of the City Council meeting via the Internet: viewers may submit online questions to the Council for discussion during question and answer periods
- Virtual e-Learning resources for some of the most commonly used software applications.
- Electronic payment services allow individuals to conduct business transactions with the city via the Internet and by telephone (register and pay online for lessons or programs, tax payment of taxes, parking tickets etc)

- Online surveys/polls about municipal happenings and issues
- E-library Strategy: Began in 2003 by the Waterloo Public Library, focusing on the familiarization of the public with the use of new technologies and accessibility for all. Free access to 7.000 e-books, laptop rental
- Community Learning Spaces: portal which provides information related to business, tourism, arts and culture to the communities of the Waterloo Region.
- MyWaterlooRegion.com: online forum for the discussion of local issues and politics (traffic studies, development proposals, education and local elections)
- Grand River Hospital (GRH): comprehensive patient portal that provides individual patients with access to their treatment plan, their schedule (and the ability to change their schedule), the tracking of symptoms and side affects, their care team, their diary and specific educational resources concerning their illness. In the future, the portal will also hold the patient's electronic records.

3.3 E- Government

The city's website (http://www.city.waterloo.on.ca) is the main source of information for residents, administration, businesses, visitors. Some of the available online services are [4], [9]:

- Minutes of city council meetings and city program registration
- Tax assessment tools
- Interactive GIS maps
- Marriage license registration
- Constitution of building permits
- Overnight parking spot reservation, payment of parking tickets
- Youth programs

3.4 Educational and Business Initiatives

Waterloo's economy is solidly based on the numerous universities, research centres and institutions of the area. The most important ones are [8]:

- University of Waterloo: computer science, mathematics and engineering programs. Features 53 research institutes, including 15 in information and communications technology and an Innovation Accelerator
- Wilfrid Laurier University: business and economics programs. Designs and delivers management programs to organizations in and around Canada's Technology Triangle

- Institute for Quantum Computing: leading centre for quantum and nanotech research
- Centre for New Oxford English Dictionary and Text Research: research through for the development of application-driven text management software. Responsible for the computerization of the Oxford English Dictionary
- Waterloo Institute for Health Informatics Research
- Centre for International Governance Innovation: research and policy development functions relating to multilateral government. Provides training to academics and practitioners through workshops and seminars. In collaboration with the World Economic Forum, the United Nations and other research institutes world-wide.
- Canada's Technology Triangle: not-for-profit, private-public sector economic development organization, which markets Waterloo and the other cities in the region to the world.
- Innovate Inc: company to support the innovative spirit of students, faculty, staff and alumni, and to help bring their ideas to the marketplace
- Communitech: capacity-building association focusing on technology
- Waterloo Technology StartUp Network: website through which budding entrepreneurs can connect with more experienced peers and financial backers.
- WREPNET (Waterloo Regional Education and Public Network): fibre-optic data network which covers 247 sites (schools, libraries, municipalities, colleges, Grand River Hospital)
- ORION (Ontario Research and Innovation Optical Network): high-speed optical network that brings broadband and connectivity to research and advanced educational institutions

3.5. Planning on ICT

For more than three decades the City of Waterloo has followed a strategy of supporting and encouraging innovation.

In 1990 the City of Waterloo commissioned a five-year Strategic Information Resource Plan, of which key-themes were data-sharing and data integration. In 1993 the City completed a Land Related Information System Project (LRIS), which related business applications to Geographic Information Systems (GIS). Later on, in 1998, the City piloted the Internet-based Waterloo Information Network (WIN), while in 2000 City of Waterloo undertook a significant, city-wide public consultation process by the name Imagine!Waterloo to discern the best possible future for Waterloo for planning and development purposes. Waterloo's strategic priorities focused on a variety of community challenges, ranging from environmental protection to transportation and culture to city communications [8].

In its 2005 Strategic Plan, the City has revised and updated its focus on intelligent and sustainable development for the community. The new Strategic Plan provides a focus for the evolution and sustainability of the Intelligent Community concept through the expansion of both infrastructure and local service delivery initiatives. Five key strategies frame the new plan:

- Planning for Growth and Change
- Addressing Service Needs
- Safe and Caring Community
- Building Partnerships
- Pursuing Operational Excellence

4. Singapore

4.1 Facts about Singapore

Singapore is an island nation located at the southern tip of the Malay Peninsula. It covers an area of 70.000ha and has a population of 4,7 million. It was inhabited by indigenous populations and became a British colony in the 19th century. Singapore became an independent republic in 1965 and since then it has evolved to a modern economy based on electronics manufacturing, petrochemicals, tourism and financial services alongside the traditional duty-free trade center.

Its political leadership stability and the setting of consistent goals for the entire period from 1965 until today, favored the constitution of a successful and transparent market economy and the overcoming of a series of crises of the past decade (Asian economic crisis, SARS virus, terrorists attacks). Its multicultural identity (75% Chinese, the rest are Malays, Indians and Eurasians) favored the opening of the local economy to the foreign markets.

Singapore has a highly developed market-based economy, which historically revolves around extended entrepot trade. Tourism is also one of the largest industries in Singapore, and especially gambling tourism (casinos) and medical tourism contribute significantly to the local economy. Furthermore, around 38,000 people work in the media in Singapore, including publishing, printing, broadcasting, film, music, digital and IT media sectors [10]. The government has also constituted the Infocomm Development Authority (IDA), the authority which is responsible for all matters and actions regarding ICT in Singapore. IDA continuously implements plans for the investment and use of ICT in public administration, services, and businesses for almost 30 years now, and today Singapore is considered one of the pioneers on e-Government, e-Services and benchmarking. There were 2.000 vacancies in IT jobs in 2005 [11].

Singapore was awarded in 1999, 2002 and 2005 as one of the top7 Intelligent Communities of the Year by the Intelligent Community Forum.

4.2. E-Services

ICT technology usage in Singapore households ranks among the highest in the world, with at least one computer in 74% of Singapore homes, more than 40% dial-up Internet penetration, household broadband penetration at 42%, 99% island-wide broadband coverage and mobile penetration rates at 92% [12], [13]. More than 600 wireless hotspots across the island (2003). Special e-Services regard:

- Logistics: the government has committed to invest up to S$50 million over five years to develop an integrated IT platform to facilitate exchange of trade documents. The country also invests on Radio Frequency Identification (RFID) across sectors like retail, logistics, manufacturing, and aviation, thus facilitating technology-driven transactions.
- Healthcare: through the Healthcare.NET initiative, Alexandra Hospital, IDA, Microsoft and various industry partners coming together to pilot and test the use of ICT in developing a hassle-free hospital, with improved efficiency in service provision and greater convenience for patients [11]
- Education: the BackPack.NET Centre, established through a strategic collaboration between IDA, Microsoft and National Institute of Education, prepares teachers on how technology is expected to influence pedagogical methods and improve the education environment for both teachers and future generations of students [11]. Also, schools are now a test bed for digital textbooks, Tablet PCs and other innovative technology, as well as for in-depth technology training for both teachers and students [12]
- Security: IDA collaborates with the Ministry of Home Affairs to develop standards for biometric passports. The Fully Automated Seamless Travel system (FAST) is already in operation at the Changi airport and allows full check-in of two minutes' durations through biometric cards [13].

4.3 E-Government

The city's website (http://www.gov.sg/) [14] is the main source of information for the government, citizens, businesses and non-residents. It is all-inclusive and it is built in 4 parts. Some of the quick links are:

- Pay Fines, Fees, Taxes and Licenses
- Buy Government Publications, Reports, Statistics etc
- Mobile services: information and alerts through SMS

4.4 Planning on ICT

The Government set in the early 80's with the aim of transforming the Singapore Government into a world-class user of ICT. The process was realized by IDA, who prepared a series of *National Infocomm Plans* (for the national strategy on ICT in matters of society and industry) and their correspondent Government Infocomm Plans (for the use of ICT in government and administration). Briefly, the National Infocomm Plans are [15]:

- The National Computerization Plan (1980-85): work functions automation and deployment of IT in the Public Service
- The National IT Plan (1986-91): provision of one-stop services through cross-agency linkages and automation and integration of public services.
- IT2000 (1992-99): launched to position Singapore as a global IT hub, by improving quality of life, boosting the economic engine and linking communities locally and globally
- Infocomm 21 (2000-03): Spurred by the convergence of telecoms and IT, it was launched with the vision of a global and prosperous e-economy and e-society. This plan placed all public services online.
- Connected Singapore (2003-6): use of ICTs as a means for the creation of new ideas, especially in new business opportunities, consumer value and cultural experiences.
- iN2015 (now in action): Intelligent Nation 2015 is a 10-year masterplan to realize the promises of ICT over the next decade. iN2015 reaffirms the strategic role that ICT will play in achieving the vision of An Intelligent Nation, a Global City, leveraging ICT for innovation, integration and internationalization. It focuses on 7 areas: Digital Media&Entertainment, Education&Learning, Financial Services, Government Services, Healthcare& Biomedical Services, Logistics&Hi-Tech Manufacturing, Tourism-Hospitality& Retail [13]

5. Comparative Analysis

This paragraph summarizes the characteristics mentioned in the previous paragraphs and demonstrates the similar or different ways in which each of the three cities have exploited the opportunities of the broadband economy. In particular, Table 1 refers to general socio-economical characteristics of Issy, Waterloo and Singapore, giving the frame within which the city has evolved, whereas Table 2 mentions explicitly the aspects of Planning on ICT, as they were followed by each city.

	Issy-les-Moulineaux	Waterloo	Singapore
population/ density	63.000 / 147 residents/ha	120.000 / 15 residents/ha	4.680.000 / 64 residents/ha
political stability	same mayor since 1980	the Waterloo Regional Council has had the same chair since 1985	governed by the same party since 1959
Strategy for ICT since	1980	1990	1980

existence of a special organization for ICT	no	Canada's Technology Triangle (CTT)	Infocomm Development Authority (IDA)
top industries	information technology, communications, broadcasting, financial services, pharmaceuticals, publishing	education, research, insurance & financial services, information technology	electronics manufacturing, petrochemicals, tourism, financial services, duty-free trade center.

Table 1 General characteristics of Issy, Waterloo and Singapore

	Issy-les-Moulineaux	Waterloo	Singapore
vision	strong and focused, due to the Local Information Plan since 1996	strong, within the general frame of creating a service-oriented economy	strong, within the general frame of creating a service-oriented economy
objectives	becoming a Cyber-City;	not explicitly expressed	becoming a world-class user of ICT
long-range planning	Series of projects implemented within The Local Information Plan running since1996	within the general planning for development	a series of consequent national ICT plans have been implemented. The latest is iN2015
interference of universities and research institutions	medium; planning is mainly a job of administrative authorities	very strong; they contribute actively to planning processes	medium; planning is mainly a job of administrative authorities
public participation	very strong; the public can express their views through elected representatives, the Participative Budget Making Platform etc	strong; Imagine!Waterloo (2000) was a city-wide public consultation aimed to determine the best possible future of the city	medium; planning is mainly a job of administrative authorities
focus on hard/soft infrastructure	focus on soft infranstructure	focus on hard infrastructure	focus on hard infrastructure

existence of target groups/ specialized actions	Separate target groups: elected representatives, associations, residents, businesses, parents, young, seniors. Specialized ICT initiatives (i.e. Issy Mobile, Issy TV, e-voting, e-democracy Forum, Cyber Kindergarten, cyberseniors)	Target Groups: residents, visitors, businesses.	Target Groups: residents, businesses, non-residents

Table 2 Characteristics of Planning Policy on ICT for Issy, Waterloo and Singapore

6. Conclusions

Towards the closure of this paper, this paragraph seeks to describe in a comprehensive way the similarities and differences in the ways the cities of the three case studies have made their way towards the creation of a broadband economy. Once the comparative analysis is complete, it is obvious that a series of conclusions can be pointed out. Conclusions refer to the past, the present and the future of these cities, demonstrating their long-range endurance and adaptability.

First of all, all three cities foresaw the oncoming era of ICT prevalence in public life from an early stage. They adopted strategies for the exploitation and incorporation of new opportunities offered by the ICT sector even since the 1980s, in an innovative and experimental spirit, seeking ways to improve public services, administrative procedures and community development. This means that they have been innovative for more than 25 years in a row. They managed to evolve with flexibility in all situations, constantly monitoring the public realm and revising strategies for local development according to emerging needs.

Persistence, a strong vision and a vibrant spirit have also been the keys of success in every case, whereas political stability has undoubtedly been helpful in setting and realizing long-term goals. However, not all three case studies have celebrated the inauguration of a special organization which occupies itself with ICT exclusively; only Singapore has developed its own Infocomm Development Authority, whereas Issy doesn't act within such a special organization and Waterloo has been acting within the general direction of the Canada's Technology Triangle, which includes other cities as well. This can be also justified due to their differences in sizes and characteristics.

Top industries also vary among the three case studies. Issy focuses on information technology, communications, broadcasting, financial services, pharmaceuticals and publishing; Waterloo on education, research, insurance, financial & health services and information technology; Singapore on electronics manufacturing, petrochemicals, tourism, financial services and trade. Their diversity, however, has not kept them back from becoming intelligent communities but instead each one of them was creative enough to use ICT applications to their best interest; i.e. Singapore uses it for the facilitation of trade, whereas Waterloo uses it in health services. Although in diverse ways, they all evolved into service-oriented economies, enhanced by ICTs.

Furthermore, population and density appear as less important factors. Issy, Waterloo and Singapore are indeed very different when it comes to population data. For instance, Issy is a very densely inhabited area (147 residents/ha), whereas Waterloo is almost 10 times more scarcely inhabited (15 residents/ha), although they fall in the same classification rank in matters of population, as small to medium sized cities. Singapore falls somewhere in between in matters of population density (64 residents/ha) and has of course a far bigger population, but the Singapore case demonstrates that ICT inclusion is possible in various scales with proportionate results.

As expected, all three cities have had a strong vision and long-range planning for growth and development, regardless of their focus on intelligent technology or not. It is clear, for instance, that Issy has used ICT as the very object of its own development, investing in e-government, e-services and e-democracy, whereas Waterloo has merely used ICT as a tool for development. This justifies the fact the Issy and Singapore have explicitly expressed the objective of becoming a Cyber-City or a world-class user of ICT, in contrast to Waterloo, which aimed high but did not make ICT its primary goal. The same applies to long-range planning; Issy and Singapore have implemented development plans exclusively dedicated on ICT use and growth, whereas Waterloo provisioned ICT use within the general planning for development.

Perhaps the most interesting part of this analysis has to do with the nature of ICT utilization within the context of the intelligent community. Further explaining, this is related to the degree of interference of universities and research institutions in planning for ICT development, participative processes of the public, customization, existence of target groups and/or specialized programs/actions. Issy, for instance, focuses primarily on human capital, the establishment of soft infrastructure and institutional density based on a broadband economy. This *human-centric* approach justifies the existence of special target and age groups, as well as of specialized and customizable programs available for all, whilst citizens can participate actively in governing and reforming their city through ICT usage. Waterloo, on the other hand, aims on hard infrastructure, knowledge creation, innovation and spillovers, assisted by numerous universities and research centers. It is what one could call a *business-driven* intelligent economy. Public participation and services for residents do exist; but aren't the epicenter of the ICT policy. The Singapore case lies somewhere in between, as it could be also characterized as a business-driven economy, but besides financial services it also focuses on tourism and trade.

Whereas each city has followed its own path towards becoming an intelligent community, long-range planning, strong vision and insightful approach stand out as the keys of success. Three different cases; three successful intelligent communities, each in its own unique way.

References

1 Intelligent Community Forum, The Top Seven Intelligent Communities by Year, <URL: www.intelligentcommunity.org>, December 2007

2 Town of Issy-les-Moulineaux official website, <URL: http://issy.com>, March 2008
3 Wikipedia, Issy-les-Mounlineaux. <URL: http://en.wikipedia.org/wiki/Issy_les_ Moulineaux>, February 2008
4 The Intelligent Community Forum, The Top 7 Intelligent Communities of 2007, <URL: http://www.intelligentcommunity.org/>, December 2007
5 Legale E., van Lerberghe D., Issy-Les-Moulineaux: E-Government and E-Democracy in Action, 2003, <URL: http://unpan1.un.org/ intradoc/groups/public/documents/Other/UNPAN022096.pdf>, February 2008
6 Legale E., The Example of Issy-les-Moulineaux*,* Global Forum 2007, 2007, <URL:http://www.items.fr/spip.php?article296&var_ recherche=legale>, February 2008
7 Wikipedia, Waterloo, <URL:http://en.wikipedia.org/wiki/Waterloo%2C_ Ontario>, February 2008
8 Intelligent Waterloo, <URL: http://www.intelligentwaterloo.ca> February 2008
9 City of Waterloo official website, <URL:http://www.city.waterloo. on.ca>, March 2008
10 Wikipedia, Singapore, <URL: http://en.wikipedia.org/wiki/Singapore>, February 2008
11 Sadasivan B., Kwok Cheong L. (2005), Be Ready To Seize Opportunities, <URL:http://www.ida.gov.sg/News%20and%20Events/ 20050717153719.aspx?getPagetype=21>, February 2008
12 The Intelligent Community Forum, The Top 7 Intelligent Communities of 2005, <URL: http://www.intelligentcommunity.org/>, December 2007
13 iN2015 Masterplan & iN2015 Steering Committee, Fact Sheet, <URL:www.ida.gov.sg/doc/News%20and%20Events/News_and_Events_Level2/2 0050508/in2015_&_SC.doc>, February 2008
14 Singapore Government official website, <URL: http://www.gov.sg/>, March 2008
15 The Government's Infocomm Journey, <URL: http://www.igov.gov.sg/ Strategic_Plans/Our_Journey.htm>, February 2008

Appendix

List of Acronyms

CTT	Canada's Technology Triangle
GIS	Geographic Information Systems
ICF	Intelligent Community Forum
ICT	Information and Communication Technology
IDA	Infocomm Development Authority
iN2015	Intelligent Nation 2015
IRIS	Information and Receipt of ISséens
IT	Information Technology

LRIS	Land Related Information System Project
ORION	Ontario Research and Innovation Optical Network
RFID	Radio Frequency Identification
WIN	Waterloo Information Network
WREPNET	Waterloo Regional Education and Public Network

The Independent Component Analysis of Clusterized Data on Principal Directions – Basis

Doru Constantin, Luminita State

The University of Pitesti, Dept. of Computer Science, Str. Targu din Vale, No.1, Romania

Since in the case of the nonlinear independent components analysis the pattern estimation presents a series of difficulties, in this article we propose an estimation method based on the use of the principal components and on the use of the algorithms of the linear ICA pattern. Thus, for the estimation of the nonlinear ICA pattern we propose a local analysis pattern that involves the use of the principal directions for the classification of the input data and the application of the linear ICA pattern to the level of each class of elements. We use a classification part that corresponds to the nonlinear representation of the mixed data and also a part of local application of the linear ICA pattern in order to describe the independent characteristics of the data. The purpose is to obtain a better data representation than by applying the methods corresponding to the linear ICA pattern at a global level. The proposed algorithm has been tested in signals separation applications and the obtained results prove good recognition performances.

Keywords
independent component analysis, blind source separation, principal component analysis.

1. Introduction

A central task in neural-network research, as well as in statistics and signal processing, is to find a suitable representation or transformation of the data. For computational and conceptual simplicity, the representation is often sought as a linear transformation of the original data. Many methods and principles can be accomplished for suitable linear transformation. The principal component analysis (PCA) can be a widely used method. PCA is defined by the eigenvectors of the covariance matrix of the input data. Such a representation is adequate for Gaussian data. However, the input data can contain a non-Gaussian data, for example, in various applications of communications, signal and image processing. Recently, independent component analysis (ICA) has been proposed as an alternative method for PCA. It is a signal processing technique whose goal is to express a set of random variables as linear combinations of statistically independent variables. The main applications are the blind source separation and the feature extraction. Basically PCA considers the second order moment only and it dose not correlate data, while ICA accounts for higher order statistics and it identifies the independent source components from their linear mixtures. ICA of a random vector x consists of finding a linear transformation, $s = Wx$, so that the components, s_i, are as independent as possible, in the sense of maximizing some function F(s) that measures independence. ICA provides a more powerful data representation than PCA. The measures of non-Gaussianity, such as fourth order cumulant (or kurtosis) and negentropy, have been used for estimating the ICA model. The minimization of mutual information and the maximum likelihood estimation are also used for ICA estimation.

2. The Derivation of the PDLICA Algorithm

Independent Component Analysis is an independent characteristics extraction technique and according to the previous definition it usually uses a linear pattern. Through an extension of the ICA linear pattern that supposes that \$x\$ input data nonlinearly depend on certain independent components, we obtain the ICA nonlinear pattern. In this case the input data (mixtures) can be represented through $x = T(s)$ where s is the vector of the independent sources, and $T : R^m \to R^m$ is an unknown nonlinear mixing function. So as to keep it simple we consider that the number of mixtures is equal to that of the independent components. Under these conditions the nonlinear ICA problem consists of determining the reverse of the mixing function, T, that estimates the independent components, respectively a function $I : R^m \to R^m$ under the form of $y = I(x)$ where y designates the estimations of the independent sources. For the nonlinear ICA pattern we proposed several estimation options, but their application presents a series of difficulties, as it comes out from the studies made in the field. First of all there is a problem related to the nonunicity of the nonlinear ICA pattern solution. Although we proposed certain reverse function estimation constraints, in certain applications we cannot clearly say which constraints are appropriate for the pattern. And we also stressed the fact that in case we determine the I function by imposing the constraint that the y elements be mutually independent, we do not ensure the covering of the pattern original sources. A second problem met when estimating the nonlinear ICA pattern refers to the calculation difficulties since once with the increase of the problem dimension we also increase the calculation complexity. As an alternative method for estimating the nonlinear ICA pattern we propose a local analysis pattern that supposes the use of the principal directions for classifying the input data and the application of the linear ICA pattern at the level of each class of elements. Thus we use a classification part that corresponds to the nonlinear representation of the mixed data and also a part of local application of the linear ICA pattern in order to describe the independent characteristics of the data. The purpose is to obtain a better data representation than through applying the methods corresponding to the linear ICA pattern at a global level. The independent components analysis pattern (PDLICA model) implies two stages, respectively:

- the classification of the observation data through the use of the principal directions;

- the estimation of the independent directions corresponding to each class set at step 1.

The first stage of the pattern implies the input data representation through their classification into M classes, each class having a definite structure (skeleton) rendered by the associated principals directions. Solving this stage is carried out through the use of the classification algorithm based on principal directions [2]. There is two variants of the algorithm, one of non-adaptive kind and the other is adaptive and which proves good generalization performances. The second stage of the pattern implies the use of the estimation methods and principles of the linear ICA pattern in order to determine the independent components at the level of each class obtained at the first step. For the estimation we can use specific methods based on the nongaussianity maximization, the obtaining of the independent components for each cluster being made through the application of the FastICA algorithm.

3. Experimental Results

The analysis pattern (PDLICA) proposed in this article has been tested in blind signal separation applications. The performed tests contain input signals recorded based on two original signals and thus in the first stage when we determine the classes of signals using the principal directions we obtain two clusters. Each cluster shall be submitted to an analysis in the independent components and we shall obtain independent signals corresponding to the original data. We have accomplished tests in which the test data are the input data, but also tests in which test signals are new examples of the two classes of signals taken into consideration.

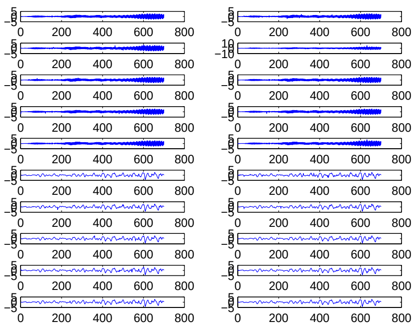


Figure 1 The initial clustered signals of the algorithm

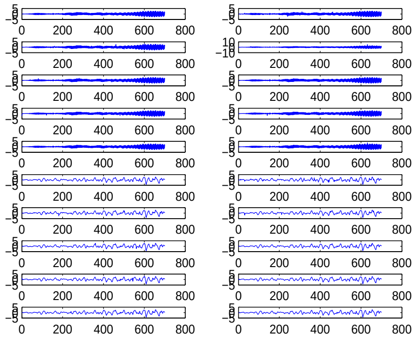


Figure 2 The clusters of signals discovered by the algorithm.

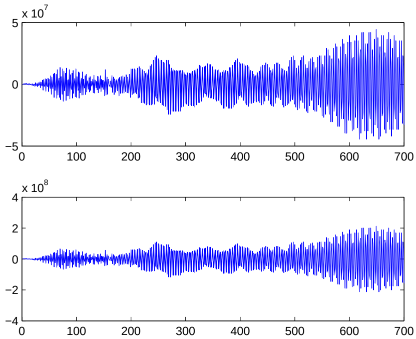


Figure 3 Independent components of the cluster 1.

Thus the performances of the proposed algorithms were set both for the version of the algorithm that classify initial sample data used at step one of the proposed model and for the new sample data for classification that differ of the training data and thus we have two tests. For the first case we not consider new examples, the algorithm working with the initial data. The obtained results revealed a classification with maxim percentage of trustworthiness. The initial signals and the obtained clusters are shown in figures 1 and 2. Cluster 1 is represented by the first five signals from each two columns and cluster 2 is represented by the rest of the signals. The independent components corresponding to the two clusters are shown in figures 3 and 4. The corresponding original signals of the two classes are shown in figure 5. By analyzing the components revealed by the algorithm in the two clusters we notice a good identification of the original sources.

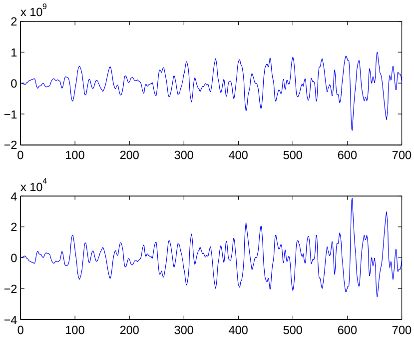


Figure 4 Independent components of the cluster 2.

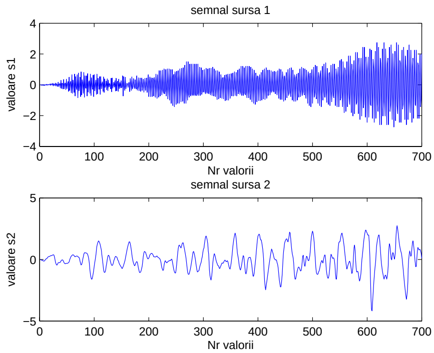


Figure 5 Original independent components.

In the second case we took into consideration new examples which have been classified and afterwards we analyzed the independent components for each cluster. As an average value, the classification percentage was placed around 95%. In this case, too the results show a good performance in determining the original sources. Thus the algorithm corresponding to the proposed pattern proved to have a good recognition and generalization capacity in applications specific to the blind source separation problems.

References

1 Hyvarinen ., Karhunen J, Oja E. Independent Component Analysis, John Wiley & Sons, 2001.

2 State L, Cocianu C, Panayiotis V, Constantin D. PRINCIPAL DIRECTIONS - BASED ALGORTHM FOR CLASSIFICATION TASKS, Proceedings of the Ninth International Symposium on Symbolic and Numeric Algorithms for

Scientific Computing (SYNASC 2007) 26-29 September 2007 in Timisoara, IEEE Computer Society.

3 Cho Y H, Hong S J, Park S M. Face Recognition by Using Combined Method of Centroid Shift and Fixed-Point ICA Based on Secant Method, Proceedings of the Sixth International Conference on Intelligent Systems Design and Applications (ISDA'06), IEEE Computer Society, 2006.

4 Sener S, Unel M. A New Affine Invariant Curve Normalization Technique Using Independent Component Analysis, The 18th International Conference on Pattern Recognition (ICPR'06), 2006.

5 Stone J V. Independent Component Analysis, The MIT Press, Cambridge, Massachusetts London, England, 2004.

6 Bingham E.Advances in Independent Component Analysis with Applications to Data Mining, Helsinki University of Technology, Espoo 2003.

Drina River Basin Hydro Information System: Simulation Model Concept

Vladimir Milivojevic[1], Dejan Divac[2], Zoran Simic[3]

[1]*Faculty of Mechanical Engineering,University of Kragujevac, Sestre Janjic Street, 6, Kragujevac, Serbia, vladimir.milivojevic@gmail.com*

[2,3]*"Jaroslav Černi" Institute for the Development of Water Resources Jaroslav Černi Street, 80, Belgrade, Serbia, ddivac@eunet.yu[2], simzoran@gmail.com[3]*

The Drina River Basin simulation model, named as Drina Hydro Information System, is constantly being developed at the Jaroslav Černi Institute, in collaboration with the Centre for Information Technology at Faculty of Mechanical Engineering of University of Kragujevac. The model development project is supported by the Serbian ministry responsible for water management and the Electric Power Industry of Serbia. The strategic objective of the Drina HIS is to create environment for optimum water resources management and to address and resolve existing and potential conflicts of interest in the region relating to multi-purpose use of water, and the misalignment of interests of the various stakeholders in the river basin. The functional objective of the simulation model is to support water management decision making (i.e. to aid users in their assessments of the consequences of various management scenarios and to support planning within various hydrologic, climatic, economic, regulatory and political constraints). Combined use of rainfall/runoff model and DEVS based simulation for system elements makes a new approach when it comes to water and hydropower management. The model and simulation core make a solid base for application of optimization algorithms, in order to provide more automated decision support system.

Keywords

discrete event systems, hydro information system, rainfall, runoff, simulation.

1. Introduction

The Drina River Basin (Drina RB) simulation model, the Drina Hydro Information System (HIS), is being developed at the "Jaroslav Černi" Institute, in collaboration with the University of Kragujevac/Faculty of Mechanical Engineering. The model development project was supported by the Serbian ministry responsible for water management and the Electric Power Industry of Serbia. The strategic objective of the Drina HIS was to create environment for optimum water resources management and to address and resolve existing and potential conflicts of interest in the region relating to multi-purpose use of water, and the misalignment of interests of the various stakeholders in the river basin. The functional objective of the simulation model was to support water management decision making.

2. Overview of the Drina River Basin

The Drina river basin (RB) represents the most significant hydro potential in the Balkans which is not being fully utilized. The surface area of the Drina RB is some 19,570 km2 (Serbia 30.5 %, Montenegro 31.5%, and Bosnia and Herzegovina 37%). The average altitude of the Drina RB is 934 m (altitudes range from 75 m at the mouth of the Drina to more than 2500 m in the highest mountains). The multi-annual average precipitation level in the Drina RB is about 1100 mm, ranging from 700 mm in the northern and eastern portions of the river basin to 3000 mm in the source area of the Lim River in the Prokletije Mountains. The average discharge of the Drina at its mouth is slightly above 400 m^3/s. The southern portions of the river basin are usually much richer with water than the central and northern portions. Specific runoff from the mountainous areas in the southern portions of the river basin is at times higher than 15 l/s per km^2. Specific runoff in the central portion of the river basin ranges from 10 to 15 l/s per km^2, while specific runoff in the northernmost, lowland portion of the river basin can be as low as 2 l/s per km^2.

To date, 9 hydro power plants (HPP), the Uvac HPP, the Kokin Brod HPP, the Bistrica HPP, the Potpeć HPP, the Piva HPP, the Višegrad HPP, the Bajina Bašta HPP, the Bajina Bašta PS-HPP, and the Zvornik HPP, have been built in the Drina RB; their total installed power is 1932 MW and their average annual output is 6350 GWh.

3. Scope and Objectives of the Drina Hydro Information System

The Drina RB can accommodate a number of other major hydropower facilities, which would provide an additional annual power output in excess of 7000 GWh. Such hydropower projects would have to include the formation of large reservoirs, which would provide: irrigation of several tens of thousand hectares of farmland in Serbia (Mačva and Srem) and Bosnia and Herzegovina (Semberia); the provision of water supply for several million people and numerous industries in Serbia, Bosnia and Herzegovina, and Montenegro; flood risk attenuation over the entire Drina RB and a portion of the Sava RB; and major water quality improvements.

However, even after protracted efforts aimed at better utilization of the Drina RB hydro potential, the future development of the river basin has not yet been comprehensively defined due to a mis-alignment of various stakeholder interests, including those of the governments of Serbia, Montenegro, and Bosnia and Herzegovina (Republika Srpska and the B&H Federation); electric power companies which generate electricity utilizing the Drina RB hydro potential and deliver electricity to different regions; local governments; public utilities; industries; various nature conservation organizations; and the like. As such, the only proper approach is to address the entire basin as a unique water management unit.

The Drina Hydro Information System (the Drina HIS) is a distributed information system which supports water management in the Drina RB and is comprised of several interactive components: integration software for distributed measurement, data acquisition and data archiving systems; simulation model; optimization software; prognostic model; database; user interface; and river basin stakeholder access and communication software. The simulation model is the basic component of the complex software and constitutes the core of the distributed system for Drina RB integrated water management support.

4. Spatial Decomposition of the River Basin, Theoretical Background and General Logic

The model addresses water flow and water use over a large and complex area, which encompasses the entire Drina RB (ca. 20,000 km^2). In general, it is important to note the difference between two types of water flow: controllable water flow, or water flow which can be controlled by artificial structures (some of which have already been erected, while others still have to be built), and inexorable water flow which cannot be affected by management decisions. Water enters the system in the form of atmospheric precipitation and is subjected to the user demand (power generation as a function of time or abstraction of specific volumes of water as a function of time). As such, the model includes the generation of runoff, taking into account the effects of snow, topography and soil, as well as all relevant types of linear flow: morphology-based flow along natural streams and flow through structures (dam spillways and outlets, hydro power plants, tunnels, channels, pipelines, etc.).

Additionally and very importantly, modelling includes the variation in flow conditions as a function of time, as a result of management decisions (deliveries, priorities and constraints, synchronized with pre-defined power and water demand, as a function of system status parameters) [1],[2],[3].

The model was developed for calculations with daily or hourly time steps.

In view of the spatial and functional complexity of the system, the river basin was broken down into various elements which can simulate different types of water flow (natural and artificially created, uncontrolled and controllable, through existing or potential future structures), based on the following concepts and descriptions [20],[23].

The hydroprofile is a model element assigned to each site which holds an existing or will hold a (planned, potential) future dam, water-gauging station, water intake regardless of the type of water use, used water outlet, and river mouth. A hydroprofile is situated solely on a river (natural stream) and its existence determines the control profile of the associated sub-catchment. A modeled hydroprofile can exist in one of the following four options (states): reservoir hydroprofile, run-of-river hydroprofile, inactive hydroprofile, and input hydroprofile.

The reservoir hydroprofile is a type of hydroprofile which, as an option, is assigned to each site of the existing and planned dams in the Drina RB. The reservoir hydroprofile is used to model the operation of the reservoir and dam facilities, with due regard being given to all natural and artificially created phenomena encountered during the flow of water, which are described by suitable mathematical equations: transformations within the reservoir, controllable and uncontrolled flow over dam spillways, controllable discharge of water via dam foundation outlets, uncontrolled seepage through the dam and the dam site, uncontrolled evaporation from the water table (water exiting the system), and formation of summary flow downstream from the dam, including setting of biological minimum flow requirements downstream from the hydroprofile.

With regard to increasing demand, the reservoir hydroprofile can function in several ways: the reservoir meets the demand as far as it can, and transfers the unserved demand to upstream assets; the reservoir sets the demand and requests charging to the normal water level; the reservoir meets the demand as far as it can and does not transfer the unserved demand to upstream assets, and the like. Water deliveries are prioritized by means of a demand serving order of facilities which draw water from the reservoir.

The run-of-river hydroprofile is a type of hydroprofile which, as an option, is assigned to each reservoir hydroprofile, but also to each site which holds a water-gauging station, water intake regardless of the type of water use, used water outlet, and river mouth. A run-of-river hydroprofile is used to readily model flow continuity.

The inactive hydroprofile is an inactive-state-type of hydroprofile. The associated sub-catchment of the inactive hydroprofile is added to the first downstream active hydroprofile and it represents the meeting point of two river stretches in which it is the initial and ultimate hydroprofile.

The input hydroprofile is a type of hydroprofile by which the catchment upstream from that profile is replaced with appropriate inflow. This allows for the observation of any portion of the Drina RB, without the need to configure upstream elements.

The sub-catchment is a spatial element determined by upstream and downstream hydroprofiles and river network sub-catchment boundaries. When the downstream hydroprofile is inactive, the sub-catchment of the inactive hydroprofile is added to the first downstream reservoir or run-of-river hydroprofile. Integration of several sub-catchments results in catchments to a particular hydroprofile. The sub-catchment-type element can be used to model: the creation of precipitation in drainage areas (entry of water into the system), the transformation of precipitation into surface runoff and groundwater flow, and the loss of water along drained surfaces (exit of water from the system) [4],[5],[6].

A physically-based hydrodynamic model was adopted. Each sub-catchment was divided into a network of elementary surfaces, or hydrologic response units (HRUs). The HRU is a basic unit used to model the formation of runoff, taking into account the influence of the topography, vegetation and soil.

The first layer simulates water retention by the vegetation cover when precipitation is in the form of rainfall. The second layer, in addition to entrapment by vegetation, simulates water retention within the snow cover. The output from this layer is snowmelt which enters the next layer. The third layer represents the unsaturated layer of soil. It simulates surface runoff and seepage into deeper layers of the soil. Following saturation of the soil and inflow from the previous layer, a portion of the water flows to the aquifer from which groundwater flow or base flow (fourth layer) originates. The fifth layer is in effect the retention capacity of the surface and the topsoil. The rate of flow between layers is determined by the characteristics of the vegetation cover, the topsoil and the hydrogeological strata.

Closed flow is an element used for linear modeling of water flow through a tunnel or pipeline, which results in a certain loss of potential energy while retaining equality between the input and output hydrographs. This element creates a link between reservoir/run-of-river hydroprofiles and hydropower elements (HPP, pumping station, and pumped-storage HPP), as well as a link between the reservoir hydroprofiles themselves.

Open flow is an element used for linear modelling of water flow in rivers, based on river channel morphology, including the transformation of the input hydrograph into an output hydrograph based on Muskingum/Muskingum-Cunge model equations. This element creates a link between two active hydroprofiles. In the case of an inactive hydroprofile, open flow links the upstream and first downstream active hydroprofile. The open-flow-type unit is determined (generated-regenerated) automatically, based on the active hydroprofile status.

The hydro power plant is an element used to model the control of power generation and associated water flow. The modelled HPP can be of the run-of-river type (water is drawn and returned within the same reservoir hydroprofile) or diversion type (water is drawn from one and returned to another hydroprofile). The HPP tailwater can be the tailrace of the reservoir hydroprofile, a run-of-river hydroprofile or reservoir. The HPP operation model is based on the use of turbine hill charts (power – net head – discharge), taking into account losses within the HPP's inlet/outlet tract.

The model provides options for several HPP operating modes (depending on the time step and the type of problem being solved). In general, HPP operation can be modelled with a pre-set power or energy demand as a function of time, or based on available inflow, including a wide range of possibilities for the modelling of different management scenarios relating to the distribution of power and discharge among power generating units.

The pumping station is an element used to model the management of energy consumption and flow (transfer) of water from the lower to the upper reservoir. The model is founded upon water transfer estimation, based on pumping station requirements and characteristics (net head is used). It is possible to model pumping station operation in several ways; in general, pumping station requirements are based either on energy demand or water level in the lower reservoir.

The pumped-storage HPP is an element used to model the management of energy consumption and flow (transfer) of water from the lower to the upper reservoir in the pumping mode, and to model the management of power generation and water flow from the upper to the lower reservoir in the turbine mode. The Drina HIS provides two operating scenarios of a pumped-storage HPP: pre-set energy level for both operating modes (turbine or pump) or pre-set power for the turbine mode and pre-defined program for the pump mode.

The user is an element used to model controllable water consumption from reservoir or run-of-river hydroprofiles by users (water supply, irrigation), including partial return of water into the downstream hydroprofiles of the system (except when water is routed away from the Drina RB). The user defines the water demand in the form of a hydrograph. The demand is served according to priorities pre-defined during the configuration procedure. Examples of water uses include: municipal and industrial water supply, irrigation of farmland, cooling of thermal power plant facilities, and the like.

The most complex configuration of the Drina HIS reflects potentially full asset availability. A complete breakdown of the entire Drina RB results in a system configuration comprised of: 127 hydroprofiles, 127 subcatchments, 127 open flows, 27 closed flows, 64 HPPs, 2 pumping stations, 2 pumped-storage HPPs, and 43 users (water supply, irrigation).

5. Numerical Aspects

Since the simulation of the system involves discrete changes in the system or its environment, which is not continuous over time, a method was developed based on the Discrete Event System Specification (DEVS) approach. The DEVS allows for the representation of all systems whose input/output behaviour can be described by sequences of events, provided system states have a finite number of changes during any time interval. The DEVS model [15] was developed by *Bernard Zeigler* in the mid 1970s and has, since then, been the most extensively used approach in computer system and network simulations [16], while it is still being researched as a simulation method for continuous physical systems [17],[18]. In addition to the high level of generality, since it integrates continuous and discrete, or hybrid, models, the DEVS provides a suitable environment for the implementation of artificial intelligence [19], which can be used for data based models (experimental, monitoring, historic, etc.).

The formal description of the DEVS atomic model is formally defined as: $\mathbf{M} = \left(\mathbf{X}, \mathbf{Y}, \mathbf{S}, \delta_{\text{int}}, \delta_{ext}, \lambda, ta\right)$, where **X** is a set of input events (e.g., in the case of a reservoir, this can be a change in the inflow rate, a change in spillway or foundation outlet control, etc.), **Y** is a set of output events (in the case of a reservoir, this includes variation in the overflow rate, discharge through the foundation outlet, water level, and the like), **S** is a set of system status variables (variables relevant to asset status definition, which can be basic or derived, e.g., in the case of a reservoir, the basic status variable is the current volume, while all other quantities, such as the current water level, surface area, etc. are derived from the volume), and $\delta_{\text{int}}, \delta_{ext}, \lambda$ and *ta* are functions which define system dynamics.

Every possible state *s* $\left(s \in S\right)$ has its associated *time advance* which is calculated by means of the *time advance function* $ta\left(s\right)\left(ta\left(s\right): S \rightarrow \Re_0^+\right)$ (e.g., if the difference between the flow to and from the reservoir is *Q*, then the time advance is obtained as $ta = Q / V_{quant}$, that is, it is the time during which the volume will change by V_{quant} if *Q* is unchanged). The output of the time advance function is a non-negative real number which indicates for how long the status of the system will remain unchanged, in the absence of any external influence.

If the state of the system is s_1 at time t_1, after $ta(s_1)$ time units (or at time $ta(s_1)+t_1$)), the system will undergo an internal transition and will change its state to s_2. The new status is obtained as $s_2 = \delta_{\text{int}}\left(s_1\right)$. The function $\delta_{\text{int}}\left(\delta_{\text{int}} : S \to S\right)$ is referred to as an *internal transition function* (one example is the change in reservoir volume due to the move to the next point in time $V_{t2} = V_{t1} + \left(t_2 - t_1\right)Q$ at an unchanged Q).

When system state changes from s_1 to s_2, the output event $y_1 = \lambda\left(s_1\right)$ is generated. The function $\lambda\left(\lambda : S \to Y\right)$ is referred to as an *output function* (its task is to generate output messages, e.g. on request, the reservoir generates output messages relevant to the operation of other assets – discharge, headwater level, total flow to the reservoir, and the like). Functions ta, δ_{int} and λ define the autonomous behavior of the DEVS model.

If an input event occurs at any time, the system status is changed instantaneously. The new system status does not depend only on input events, but also on the previous status and the time elapsed since the last transition. If the state of the system changes to s_3 at time t_3, and then an input event occurs at time t_3+e, whose value is x_1, the new state is obtained as $s_4 = \delta_{ext}\left(s_3, e, x_1\right)$, where $ta\left(s_3\right) > e$ is implied. In such a case, we can say that an external transition has occurred. The function $\delta_{ext}\left(\delta_{ext} : Sx\Re_0^+ xX \to S\right)$ is referred to as an *external transition function* (if the reservoir's input port receives a message that the rate of inflow has changed, the reservoir must update its state variables and duration in order to continue to participate in the simulation). External transition does not generate output events.

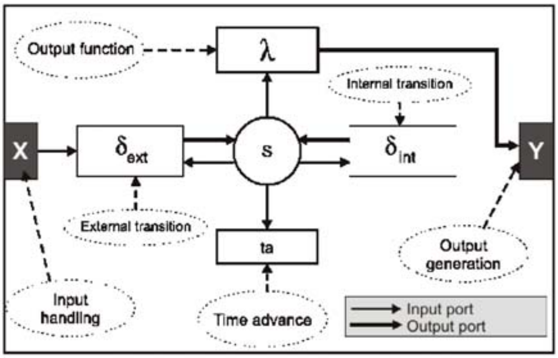


Figure 1 Schematic representation of the DEVS atomic model.

As mentioned above, the DEVS is a model formulated in general terms and it can be used to describe highly complex systems. However, the representation of complex systems, based on stringent principles of physics, using only transition and time advance functions, can become an overly complex procedure. Difficulties arise because it is necessary to predict and describe all possible cases which can be encountered during a simulation, using only a few functions. Of course, complex systems can also be viewed as a number of coupled simple elements. Following coupling, the output events of a sub-system become input events of another sub-system, to which the former is coupled (e.g. if the reservoir and its HPP are bonded into a single bonded system, then the reservoir's output event – which includes information about the tailwater level, current water balance, discharge to the HPP, etc., becomes an input event for the HPP, based on which it computes its status variables, duration, and the like). The theoretical set-up ensures that the coupled system will behave like an atomic model with respect to its environment, such that complex models can be created hierarchically; this is an appropriate basis for the development of object-oriented simulation software [14].

6. Database

Database content and structure. All data used by the Drina HIS are classified and stored in a database, whose primary task to service models. It contains information about the following: system configuration, performance of existing and potential future facilities (reservoirs, spillways, outlets, HPPs, pumping stations, and the like); catchment areas (topography, vegetation, soil, etc.), the hydrographic network, watercourses, hydrometeorological stations, hydrology, weather, users, and the like.

In addition to data about modeled assets (127 hydroprofiles, 64 HPPs, etc.), the Drina HIS database contains information about 23654 HRUs with 118270 runoff functions, as well as information about 10 types of vegetation, 8 types of soil, and 6 hydrogeological formations within the Drina RB. The hydrographic network is comprised of 1957 nodes and 1955 river segments.

The Drina HIS database is comprised of 98 tables linked with 87 relations, thus ensuring data consistency in a form adaptable to system changes [21].

Application of GIS technology. The database relies on the Geographical Information System (GIS), which allows for association with specific spatial and geographical features.

Data within the Drina HIS database are arranged in a manner which is most similar to the ArcHydro model; this model is a widely accepted standard for the management of GIS data relevant to water resources management [12]. The arrangement of data into layers, and their inter-linkage, leads to a level at which it is possible to use standard models and procedures for the definition and analysis of river networks and catchment areas.

The accepted and applied ArcHydro data model standard for database management also allows the application of standard GIS software, such as ArcView or Autodesk Map [13].

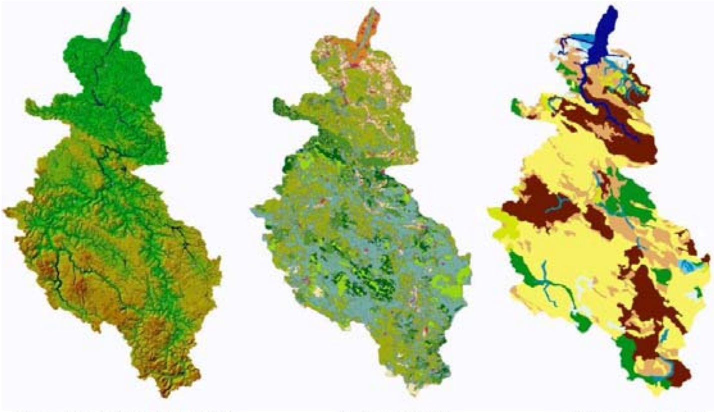


Figure 2 Portion of the model's GIS-content layer.

7. Software Platform

The aim of system architecture design and selection of suitable software technologies was to create an open, scalable platform which can equally support a distributed environment, which is currently most often the case. Since it is a complex system subjected to upgrading and increasing complexity, application scalability is very important and the ability of a large number of users to access the system was kept in mind from the very start. Contemporary information systems handle enormous amounts of data and operating principles established only several years ago are already obsolete. The Drina HIS was not developed on the basis of a conventional singlelayer system, in which an application directly accesses the data, but on the basis of a three-layer model [7]. A three-layer model makes a clear distinction between three functional units: the presentation layer, the business logic layer, and the data layer. The presentation layer is the part of the application which is visible to the user. It is implemented via Windows tools which are made available to the user. The business logic layer can be implemented in two ways: in the form of codes within the applications or in the form of an independent web service with which applications communicate via SOAP messages. The first approach was used for the current software version. The data layer represents any database supported by a .NET environment (in this case the Microsoft SQL Server), and communication with the central layer is provided via the ADO.NET environment[8].

The Scalable Vector Graphics (SVG) language is used to describe two-dimensional graphics in XML [9]. SVG specifies the use of three types of graphical objects: vector objects, figures, and text. The objects may be grouped, their styles may be changed, they can be transformed, etc. In the Drina HIS user interface, the SVG is used for visualization of the simulation model and GIS content, since it is able to handle vector displays and raster data equally, and to thereby ensure full interaction[22].

8. Parameter Estimation

In addition to experimental data about the performance of all elements of the system (e.g., reservoir volume curves or turbine hill charts) or information about catchment areas (topography, vegetation, soil, etc.), the database contains „model parameters“ which can be determined by observation or measurement of flow or catchment area characteristics. One example is the Muskingum model parameter of open flow, which has no direct physical meaning and cannot be measured. It is a weight coefficient which is an indicator of the relative importance of downstream and upstream discharges during calculations. As such, model parameters include: 12 parameters for each sub-catchment, 4 parameters for each type of vegetation, soil and hydrogeology; and 2 parameters for each open flow.

These parameters were estimated through optimization (application of evolutionary algorithms), with the goal of achieving the best possible match between computed and measured discharges at a particular hydroprofile where a representative hydrologic station with reliable instrumentation is available. Using known precipitation levels and temperatures, the computed values are obtained through an iterative process (simulation, assessment, comparison, correction, and repeated simulation).

The parameters were estimated over one period (2-3 years), while validation was assessed over a different, independent period [24]. Measured and simulated values for two selected profiles, following calibration, are shown in Fig. 3.

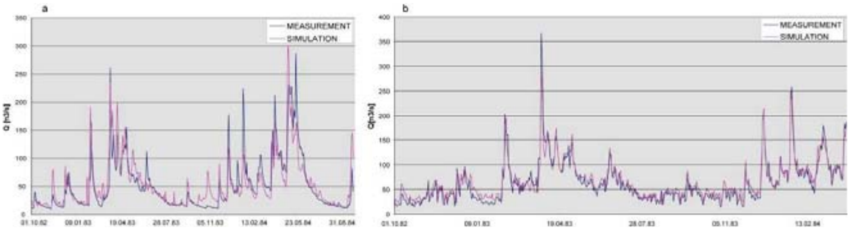


Figure 3 Measured vs. simulated values (a - Prijepolje Hydrologic Station hydrograph, and b - Potpeć Dam hydrograph).

9. Simulation Outputs

Simulation outputs include hydrographs and water level diagrams for dam sites, hydrologic station sites and other sites (i.e. all hydroprofiles), hydrographs for dam evacuation facilities, generated electricity, number of operating power units, specific energy, HPP discharge (as well as power output, turbine efficiency and turbine discharge for all active HPP power units), and power consumption for transfer pumping. The outputs are comprised of suitably discretized time series; both are graphical and numerical and can be exported by means of the copy/paste function [22] (Fig. 4 illustrates several simulation outputs).

Figure 4 Simulation outputs (a - Potpeć HPP power generation, b - Water level of the Kokin Brod Reservoir).

10. Conclusion

The HIS, which supports water management within the Drina River Basin, is a tool by which a more dynamic and more efficient dialog can be established between all river basin stakeholders, at all decision-making stages (spanning from strategic investment planning to operational management) and at all levels of involvement (ranging from measurement and information gathering to the provision of complex evidence in legal procedures).

The unique coupling of time-continuous rainfall/runoff model and DEVS simulation model provides opportunities for further research, in areas of parallel processing and optimization algorithms for automatic recalibration of model parameters.

References

1 Goodman Alvin S.: Principles of Water Resources Planning, Prentice-Hall, 1984.

2 Wurbs, R.A.: Reservoir Simulation and Optimization Models, Global Climate Change Response Program, Denver, Colorado, August 1991.

3 Divac, D., Grujović, N., Milovanović, M., (in Serbian) A New Simulation Model for Water Balance Assessments of Water Systems: Methodology, Software, and

Applications, Monograph: Management of Serbian Water Resources, ’99, Jaroslav Černi Institute for the Development of Water Resources, Belgrade, 1999.

4 J.G. Arnold, S.L. Neitschn, J.R. Kiniry, J.R. Williams: SWAT - Soil and Water Assessment Tool Users Manual, Texas 2001.

5 Prohaska, S., Ristić, V.: (in Serbian) Hydrology, Part 1: Hydro-Meteorology, Hydrometry, and the Water Regime, Belgrade, 2003.

6 Prohaska, S.: (in Serbian) Hydrology, Part 2: Hydrologic Forecasting, Modeling, and Applications, Belgrade, 2006.

7 Prosise, J.: Programming Microsoft® .NET Microsoft Publishing, Redmond, 2003.

8 Chand, M., Talbot, D.: Applied ADO.NET: Building Data-Driven Solutions, Apress Publishing, Berkeley, 2003.

9 Campesato, O.: Fundamentals of SVG Programming: Concepts to Source Code, Charles River Media, 2003.

10 Blind, M.W., Adrichem, B., Groenendijk, P.: Generic Framework Water: An open modeling system for efficient model linking in integrated water management - current status, EuroSim 2001, Delft, 2001.

11 Blind, M.W., Adrichem, B., Groenendijk, P.: Generic Framework for hydro-environmental modelling, HydroInformatics 2000, Cedar Rapids, 2000.

12 Maidment, D. R.: ArcHydro GIS for Water Resources, ESRI Press, Redlands, 2002.

13 Struve, J., Westen, S., Millard, K., Fortune, D.: Harmonit – State-of-the-Art Review, London, 2002

14 Bernard P. Zeigler: Objects and Systems: Principled Design with Implementations in C++ and Java, Springer, 1997.

15 Bernard P. Zeigler, Tag Gon Kim, Herbert Praehofer: Theory of Modeling and Simulation, Academic Press, 2000.

16 Jerry Banks, John Carson, Barry L. Nelson, David Nicol: Discrete-Event System Simulation, Prentice Hall, 2004.

17 James Nutaro: Parallel Discrete Event Simulation with Application to Continuous Systems, PhD thesis, University of Arizona, Tuscon, Arizona, 2003.

18 Ernesto Kofman: Discrete event simulation of hybrid systems, SIAM Journal on Scientific Computing, 2004.

19 Hessam S. Sarjoughian, Francois E. Cellier: Discrete Event Modeling and Simulation Technologies: A Tapestry of Systems and AI-Based Theories and Methodologies, Springer, 2001.

20 The Drina Hydro Information System, Simulation Model, Version 2, Vol. 1: Concept and Mathematical Model (in Serbian), Jaroslav Černi Institute for the Development of Water Resources, Belgrade, 2005.

21 The Drina Hydro Information System, Simulation Model, Version 2, Vol. 2: Database (in Serbian), Jaroslav Černi Institute for the Development of Water Resources, Belgrade, 2005.

22 The Drina Hydro Information System, Simulation Model, Version 2, Vol. 3: Operating Instructions and Verification Examples (in Serbian), Jaroslav Černi Institute for the Development of Water Resources, Belgrade, 2005.

23 The Drina Hydro Information System, Simulation Model, Version 2.1, Vol. 1: Corrections, Data, Parameters, and Recommendations (in Serbian), Jaroslav Černi Institute for the Development of Water Resources, Belgrade, 2006.

24 The Drina Hydro Information System, Simulation Model, Version 2.1, Vol. 2: Model Application: Pilot Studies (in Serbian), Jaroslav Černi Institute for the Development of Water Resources, Belgrade, 2006.

Appendix

List of Acronyms

DEVS	Discrete events system
HIS	Hydro information system
HPP	Hydro power plant
HRU	Hydrologic response unit
RB	River basin
SVG	Scalable vector graphics
XML	Extensible markup language

A Subliminal Channel in the Framework of P Systems

Luciana Maria Morogan[1], Adrian Atanasiu[2]

[1]Spiru Haret University, Faculty of Mathematics-Informatics, 13 Ion Ghica Str, Bucharest, Romania, l.morogan.mi@spiruharet.ro

[2]University of Bucharest, Dept. of Theoretical Computer Science, 14 Academiei Str, Bucharest, Romania, adrian@galaxyng.com

Differently of the majority of the cryptography models, the molecular cryptography systems based on membrane computing have to their origins both biology elements, inspired by the membrane computing (P systems), and subliminal channels witch are cryptography elements. Gustavo's J. Simmons showed (at Crypto `83) that a transmitter and a chosen receiver - that are exchanging secret information – in the absence of a secrecy channel to allow them to convert a portion of the authentication information to a hidden communications channel, could pervert an authentication. Under some conditions, even the detection of the existence of this hidden channel can be made as difficult as the authentication algorithm was "cryptosecure". In view of this open existence (yet undetectable), such a hidden channel was called a "subliminal" channel. This paper reviews briefly the essential features of a tissue-like P system and then discuses the protocol implementation of a one bit subliminal channel in the framework of membrane computing.

Keywords

membrane computing, P systems, subliminal channel.

1. Introduction

Molecular computing (DNA computing, ciliates based computing and membrane computing) refrazed in terms of *soft computing*, created an important crossroad of many researchers` interests in this biology inspired domain. Differently of the sudy of biomolecular cryptography models based on DNA computing or ciliates based computing, that we can find in a quite big numer of already published papers, in the new field of membrane based cryptography a small number of papers are already known. The present paper proposes the modelling of a one bit subliminal channel as an absolute novelty found at the frontier of two domains: cryptography and membrane computing.

Membrane computing is a branch of natural computing which investigates computing models abstracted from the structure and functioning of living cells and from the interactions in tissues or higher order biological structures. Briefly, a membrane system, that we will describe above, is a distributed computing model processing multisets of objects either in the compartments of a cell-like hierarchical arrangement of membranes (hence a structure of compartments which corresponds to a rooted tree), or in a tissue-like structure consisting of cells placed in the nodes of an arbitrary graph. Both the objects of the membranes, the membranes, and the links among them evolve according to some rules. For instance, the multisets of objects evolve mainly by means of rewriting rules, which have the form of usual chemical equations (several objects react and get transformed into some product objects). A *crucial aspect* of this processing is the resulting *communication* of objects through membranes, between regions of the same cell, between cells, and between cells and their environment. We have used this aspect for creating a one bit *subliminal channel* in the framework of a tissue-like P system.

Gustavo's J. Simmons showed (at Crypto `83) that a transmitter and a chosen receiver - that are exchanging secret information – in the absence of a secrecy channel to allow them to convert a portion of the authentication information to a hidden communications channel, could pervert an authentication. Under some conditions, even the detection of the existence of this hidden channel can be made as difficult as the authentication algorithm was "cryptosecure". In view of this open existence (yet undetectable), such a hidden channel was called a "*subliminal*" channel.

Let's imagine our P system as a tissue-like structure - placed in a common environment - consisting of two membranes, witch are models that process multisets of objects in the compartments of a cell-like hierarchical arrangement of cells, and another elementary membrane placed between them and who plays the role of the host of the communication channel. The three membranes, that from now on we will refer to as cells, are placed in the nodes of a connex graph, corresponding to a communication network established among adjacent cells by making their protein channels cooperate, moving molecules directly from one cell to another. Supposing we want to communicate *m* bits of information from the transmitter-cell to the chosen receiver-cell. This means that we need at least *m*+*r* bits for the information transfer, because *r* bits represent the redundant information, being used only by the receiver-cell to divide the set of all possible messages, the multiset denoted with *M*, into two disjoint subsets of messages: acceptable messages, meaning the messages that contain redundant information, represented by a multiset M' of objects, and unacceptable messages, meaning the messages that are altered or fraudulent, represented as $M \setminus M'$. For preventing the fact that the host-cell may sneak through information containing the authentication and be able to alter the message, in general, this information must be secured from outsiders by encryption. This must be either a block cipher (if *m*+*r* is small enough), either a block chain or a feedback cipher, so as to produce the desired "spreading" of symbol dependence [3]. The essentials points to an authentication without secrecy channel are:

- the receiver-cell authenticates a message through the presence of *r* bits of redundant, i.e., expected, information in the decrypted cipher;
- the host-cell to the communication channel verifies that nothing has been concealed by decrypting the ciphers and verifying that the resulting message is precisely what he expected on an a priori knowledge of the message.

2. Tissue-like P systems

In order to model our system, we need to review the class of P systems based on a *tissue-like* structure, with membranes placed in the nodes of an arbitrary graph witch, in the basic variant, is a total one. What matters is the communication graph, dynamically defined during computations. Several (elementary) membranes – also called cells – are freely placed in a common environment and can communicate either with each other or with the environment by *symport/antiport rules*, witch are rules that governs the communication through a specific membrane (no object can be changed during the trans-membranes movement). Those rules corresponds to some processes by witch two molecules pass together across a membrane and in this case the process is called *symport*, or the two molecules pass simultaneously through a protein channel, but in opposite directions, in witch case, the process is called *antiport*. For more information, see [11]. In the case of tissue-like P systems, we consider antiport rules of the form (i, x/y, j), where i, j are labels of cells or at most one is zero, identifying the environment, and x, y are multisets of objects. This means that the multiset x is moved from i to j at the same time as the multiset y is moved from j to i. If one of the multisets x, y is empty, then we have a symport rule. The communication among cells is done either directly (in one step), either indirectly (one cell throws some objects in the environment and other cells can grab these objects in the next step or later). The environment contains a specified set of objects in arbitrary many copies. A standard computation starts from the initial configuration, using the rules in the nondeterministic maximally parallel manner, and halts in a specified cell. The result of the computation is the number of objects from that specified cell.

Tissue-like P systems with channel-states are constructs of the form

$$\Pi = \left(O, T, K, w_1, ..., w_m, E, syn, \left(s_{(i,j)}\right)_{(i,j)\in syn}, \left(R_{(i,j)}\right)_{(i,j)\in syn}, i_0 \right),$$

where:

- O is the alphabet of *objects*;
- $T \subseteq O$ is the alphabet of *terminal* objects;
- K is the alphabet of *states* (not necessarily disjoint of O);
- $w_1, ..., w_m$ are strings over O representing the initial multisets of objects present in the cells of the system (we assume that we have m cells, labelled with 1, 2, …, m);
- $E \subseteq O$ is the set of objects placed in arbitrary many copies in the environment;
- $syn \subseteq \{(i,j) \mid i,j \in \{1,2,...,m\}, i \neq j\}$ is the set of links among cells, called *synapses*; 0 indicates the environment such that for $(i,j) \in \{1,2,...,m\}$ at most one of (i, j), (j, i) is present in *syn*;
- $s_{(i,j)}$ is the *initial state* of the synapse $(i,j) \in syn$;

- $R_{(i,j)}$ is a finite set of rules of the form $(s, x/y, s')$, for some $s, s' \in K$ and $x, y \in O^*$, associated with $(i,j) \in syn$;
- $i_0 \in \{1,...,m\}$ is the *output* cell.

There is the restriction that there is at most one synapse among two given cells, and the synapse is given as an ordered pair (*i*, *j*) with witch a state from *K* is associated. A rule of the form $(s, x/y, s')$ is an antiport rule, acting only if the synapse (*i*, *j*) has the state *s*, and the application of the rule means (1) moving the objects specified by *x* from *i* to *j* (if *i* = 0, it means the environment), at the same time with the move of the objects specified by *y* in the opposite direction, and (2) changing the state of the synapse from *s* to s'. The computation starts with the multisets specifed by $w_1,...,w_m$ in the *m* cells: in each time unit, a rule is used on each synapse for which a rule can be used (if no rule is applicable for a synapse, then no object passes over it and its state remains unchanged). The use of rules is sequential at the level of each synapse, but it is parallel at the level of the system: all synapses which can use a rule must do so (the system evolves synchronously). The computation is successful if and only if it halts and the result of a halting computation is the number of objects from *T* present in the output cell in the halting configuration. The set of all numbers computed in this way by the system is denoted by *N*($\prod$). If we can compute vectors, we consider the multiplicity of objects from *T* present in the output cell in the halting configuration.

3. Setting up the One Bit Subliminal Channel in the Framework of a Tissue-Like P System with Channel States

In a common environment that contains a specified set of objects in arbitrary many copies, we place our tissue-like structure of a P system. The structure consists of one elementary membrane (called the host-cell of the communication channel) and two non-elementary membranes (called the transmitter-cell and the chosen receiver-cell as we have already mentioned in the first chapter of this paper) placed in the nodes of a connex graph corresponding to a communication network established among adjacent cells by making their protein channels cooperate, moving molecules directly from one cell to another. This communication is governed by making use of inter-cellular symport/antiport rules of a form that we will discuss later in this chapter (no object can be changed during the trans-membranes movement). As our system presents the structure of a connex graph, for all cells labelled *i*, *j* in the graph, $i \neq j$, there is at least a chain of cells, starting with *i* and ending with *j* (by chain we mean a succession of cells in witch all two neighbour cells are adjacents). This allows us to make a restriction in the system that will help us setting the subliminal channel: all the information-objects that are exchanged between the transmitter-cell and the receiver-cell, must across the host-cell of the communication channel. Let us suppose that we may have $\Lambda = \{T, H, R, T', R', En\}$ a finite set of cell labels. For the intra-cellular communication of the multisets of objects placed in the transmitter-cell (denoted by *T*)/receiver-cell (denoted by *R*) respectively, across the inner cell (labelled with T', R' respectively) we use symport/antiport rules of the form (*objects*, *tar*), where $tar \in \{in, out\}$. The two non-elementary cells process multisets of objects, in the compartments of a cell-like hierarchical arrangement of cells, by using a set of cooperative rewriting rules of the form $u \to v$ where *u*, *v* are objects (with *u* arbitrary), from a finite alphabet of objects.

Formally, *SC Tissue-like P system with channel-states* (*SC* coming from the subliminal channel abbreviation) is a construction of the form

$$\Pi = \left(O, Ter, K, \mu, w_T, w_H, w_R, V, syn, \left(s_{(i,j)} \right)_{(i,j)\in syn}, \left(R_{(i,j)} \right)_{(i,j)\in syn}, R_{ic}, R_{rew}, \rho, R' \right),$$

where:

- O is the alphabet of *objects*; $M \subset O$ is the multiset of all possible messages and $M' \subset M$ the multiset of messages that contain redundant information for the receiver-cell;
- $Ter = \{c\} \subseteq O, c \in O^+$ is the alphabet of *terminal* objects (in this system we need only one object);
- K is the alphabet of *states*;
- $\mu = [\mu_T, \mu_H, \mu_R]$ is the *tissue-like structure*, where μ_T, μ_H, μ_R are cell structures corresponding to the transmitter-cell, host-cell and receiver-cell witch have the forms: $\mu_T = [_T[_{T'}]_{T'}]_T, \mu_H = [_H]_H, \mu_R = [_R[_{R'}]_{R'}]_R$; *T*, *H*, *R*, T', R' are elements from a finite set of cell labels;
- w_T, w_H, w_R are strings over *O* representing the initial multisets of objects present in the cells labelled with *T*, *H* and, *R* respectively; we assume that M, the multiset of all possible messages, is initially placed in each of these strings;
- $V \subseteq O$ is the set of objects placed in arbitrary many copies in the environment;
- $syn \subseteq \{(i,j) \mid i,j \in \{T,H,R,En\}, i \neq j\} \setminus \{(T,R),(R,T)\}$ is the set of links among cells, called *synapses*; *En* indicates the environment such that at most one of (*i*, *j*), (*j*, *i*) is present in *syn*;
- $s_{(i,j)}$ is the *initial state* of the synapse $(i,j) \in syn$;
- $R_{(i,j)}$ is a finite set of rules, called *inter-cellular communication rules*, of the form $(s, x/y, s')$, for some $s, s' \in K$ and $x, y \in O^*$, associated with $(i,j) \in syn$. A rule of the form $(s, x/y, s')$ is an antiport rule, acting only if the synapse (*i*, *j*) has the state *s*, and the application of the rule means (1) moving the objects specified by *x* from *i* to *j* (if *i* = *En* or *j* = *En* it means the environment), at the same time with the move of the objects specified by *y* in the opposite direction, and (2) changing the state of the synapse from *s* to *s′*. As given one synapse among two given cells,

the synapse is given as an ordered pair (i, j) with witch a state from K is associated (s - before the communication and s' - after the communication). By convention, these rules will be considered of the form $(i, x/ y, j)$, where i, j are the labels of the cells (if $i = En$ or $j = En$ then we have the environment) and x, y are multisets of objects. As in our model, we will have only rules of the form $(i, x/ 0, j)$ then the final representation of such rules is (i, x, j).

- R_{ic} is a finite set of rules, called *intra-cellular communication rules*, of the form (*object*, *tar*), where $tar \in \{in, out\}$. We have $R_T \cup R_R = R_{ic}$, where R_T and R_R are sets of rules associated with the transmitter-cell and the receiver-cell: (1) if we have a rule of the form (*object*, *in*), acting in the T cell, then the *object* will be sent in the inner cell, T' (respectively, if we have a rule of the form (*object*, *in*), acting in the R cell, then the *object* will be sent in the inner cell, R'); (2) if we have a rule of the form (*object*, *out*), acting in the T' cell, then the *object* will be sent out of the inner cell, to T (respectively, if we have a rule of the form (*object*, *out*), acting in the R' cell, then the *object* will be sent out of the inner cell, to R);
- R_{rew} is a finite set of rules, called *cooperative rewriting rules*, of the form $u \to v$ where u, v are objects (with u arbitrary) over O.
- ρ is a partial order relation over the rules of $\prod$ specifying the priority in the same cell: *cooperative rewriting rules* > *intra-cellular communication rules* > *inter-cellular communication rules*;
- R' is the *output* cell.

A standard computation starts from the initial configuration, using the rules in the nondeterministic maximally parallel manner, and halts in the output cell. The result of the computation is the number of objects from the output cell, objects belonging to *Ter*.

So, how do we set up a subliminal channel in such a framework?

Considering M the multiset of all possible messages and $M' \subset M$ the multiset of messages containing the redundant information, messages that will be accepted as authentic by the receiver-cell, the existence of M is essential, representing the means of the receiver-cell in detecting and avoiding deception. We must recall that initially the multiset M is placed in all the transmitter-cell, the host-cell and the receiver-cell. The essential idea of setting up a subliminal channel is the implementation, in the communication channel, of a crypto-algorithm based on two secret encryption keys, in witch the host-cell has the decryption key in advance, for verifying that this open channel is not misused. This is why into the multiset of objects placed in arbitrary many copies in the environment we throw, in arbitrary many copies, the decryption key d, witch is public. Generally, the encryption algorithm that encrypts a message m, using a secret encryption key e obtaining a cipher c, $E(m, e) = c$, is implemented as a rewriting rule $E \in R_{rew}$ of the form $E : me \to c$; and the decryption algorithm that decrypts a cipher c, using the public decryption key d and obtaining the original message m, $D(c, d) = m$, is implemented as a rewriting rule $D \in R_{rew}$ of the form $D : cd \to m$, where $m \in M', e, c \in O, d \in V$. We suppose that there are different messages and, for each of such messages, there are two ciphers such that by decryption with the public key one obtain the same correct message. The host-cell, that has only one of the two encryption systems, can decrypt the cipher using his decryption key, introduced into the cell by the rule (*Em*, *d*, *H*), being convinced that nothing was concealed in the message, witch is actually true. The decryption key was taken, in advance, from the environment (in arbitrary many copies). The receiver-cell decrypts the cipher for authentication and recovers the information, being capable to learn one bit of information, sufficiently enough to suggest him witch of the two encryption systems was used for the message encryption. This feature of the channel is called the *subliminal channel*.

What is actually taking place, can be seen in Figure 1.

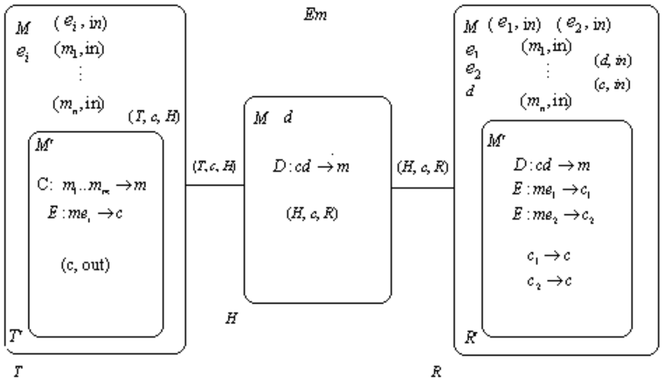


Figure 1 One bit Subliminal Channel in the framework of a SC Tissue-like P system.

In place of the secret encryption key *e*, there are two secret encryption keys $e_1, e_2 \in O$ witch encrypt the multiset of acceptable messages, M', into two disjoint multisets of ciphers (corresponding to the encryptions keys). This fact is unknown to the host-cell. So, we consider the choice of *E* and *D*, with $E, D \in R_{rew}$, as follows: there are two secret encryption keys $e_1, e_2 \in O$ and there is a public decryption key $d \in V$ such that

- for $E: me_1 \to c_1, E: me_2 \to c_2$, the two ciphers are different, and
- for $D: c_1 d \to m, D: c_2 d \to m$ the message obtained is the same in the both cases, for $c_1, c_2 \in O, m \in M'$.

Between the transmitter-cell and the receiver-cell there is the convention that if the encryption is made by using e_1 then the bit subliminally transmitted is a 0, and if the encryption is made by using e_1 then the bit subliminally transmitted is a 1. For avoiding the indeterminism induced by introduction in the transmitter-cell of both secret key in the same time (because of the parallel manner of treating the rules, it may happen that both secret keys will be applied in the same time), we will make the assumption that we will introduce into the transmitter-cell the proper key for the desired bit (only one key will be introduced for the transmition of one bit). This is not the case of the receiver-cell witch needs both secret keys to be introduced into the cell for being able to recover the subliminal bit. Now we have in the transmitter-cell the multiset of all possible messages *M* and the secret key object e_i, with *i* being either 1, either 2, but never both. In the host cell there is already the decryption public key object *d* (introduced into the system by a rule as above) and *M;* and in the receiver-cell both $e_1, e_2 \in O$ and *M*.

In the *transmitter-cell* the fallowing actions are taking place:

- the multiset M' of all the redundant messages, in arbitrary many copies, along with the secret key object are passing into T' following the intra-cellular communication rules from R_T, rules of the form (*object*, *in*), that are applied in parallel for all *m* in M' and $e_i \in O, i = 1$ or $i = 2$;
- supposing $M' = \{m_1, ..., m_n\}$ is the multiset of objects in arbitrary many copies in T', we apply a "choice rule" to all messages from M' that will decide a message *m* that will be chosen to be transmitted – this rule has the form C: $m_1 ... m_n \to m$ with $C \in R_{rew}$;
- *m* is encrypted fallowing the rule $E : me_i \to c, c \in O, i = 1$ or $i = 2$;
- *c* in arbitrary many copies is sent out, from T' to *T*, by this intra-cellular communication rule from R_T : (c, out);
- the transmitter cell sends *c* to the cell host, using an inter-cellular communication rule (*T*, *c*, *H*).

The *host-cell* receives and decrypts *c* using the public decryption key *d*, following the rule $D : cd \to m.$ For *H*, as *m* belongs to *M* and verifying the message, obtains that nothing was concealed in the information he just received, witch is actually true, and so the cipher *c* will be forwarded to the receiver-cell: (*H*, *c*, *R*).

The *receiver-cell* receives *c*:

- for all *m* in the multiset M' of all the redundant messages, found in arbitrary many copies in *M*, along with the secret key-objects $e_1, e_2 \in O$, *R* passes into R' the multiset M' and the keys in the same time with similar parallel operations taking place in the transmitter-cell;
- in order to decrypt the message, the public decryption key object *d* is brought into the receiver-cell, in arbitrary many copies, from the environment using the synapse (*Em*, *d*, *R*);
- the decryption key *d* and the cipher *c* are passed into R' following the intra-cellular communication rules from R_R, rules of the form (*object*, *in*), applied in parallel for $d \in V$ and for the received cipher *c*;

- in the inner cell R', the cipher c is being decrypted following the rule $D: cd \to m$; if m contains no redundant information, meaning that the message is fraudulent or altered, the cell is rejecting it and no hidden information will be found; else, *m* is an authentic message and remains in the cell;
- the subliminal bit is obtained. Two encryptions are taking place in parallel (by using the two secret keys objects e_1, e_2): $E: me_1 \to c_1$ and $E: me_2 \to c_2$. As only one of c_1 or c_2 is already in R' as *c*, the other one will be rejected by the cell. The remaining cipher, wich will be either c_1 or c_2, will be rewritten as *c*. This means that in R' there are another two rewritting rules belonging to R_{rew}: $c_1 \to c$ and $c_2 \to c$. These copies of *c*, witch are identical copies of the ones received from the host-cell, gives the secret key used for encrypting the original message object and, in this way, the subliminal bit. If an e_1 was used in the encryption process, then the subliminal bit received is a 0; else, if an e_1 was used in the encryption process, then the subliminal bit received is a 1.

For *example*, we make the supposition that the subliminal bit desired to be transmitted from the transmitter-cell to the receiver-cell is a 0. This means that in the transmitter-cell we introduce the encryption key e_1 witch, following the convention above, corresponds to the proper transmition of the bit 0. In the initial state of the system, in all three of the cells involved in our processes, we find an arbitrary number of copies of all possible messages from the set *M* and : 1) in the transmitter-cell, *T*, we also introduce the encryption key e_1 and 2) in the receiver-cell, *R*, both the encryption keys e_1 and e_2 are being introduced. The steps of the computation are as fallows:

Step1. In both transmitter and receiver cells, the messages containing the redundant information for *R* (messages that will be accepted as authentic) are being introduced in the inner cells T' and R' respectively. If we make the assumption that the set of messages containing the redundancy, M', contains messages of the form $m_1, ..., m_n$ in arbitrary many copies, then all $m_i, i = \overline{1, n}$ will pass into T', R' respectively, following rules of the form $(m_i, in), i = \overline{1, n}$. In the same time also the encryptions keys are being passed: 1) in T' passes e_1 by (e_1, in) and 2) in R' passes e_1 and e_2 by (e_1, in) and (e_2, in).

In parallel, an arbitrary number of copies of the decryption key, *d*, are brought into the host cell, *H*, from the environment using the following rule: (*Em*, *d*, *H*).

Step2. A "choice rule" is applied for choosing (in arbitrary many copies of course) the message *m*, from the set M', to be transmitted: $C: m_1 ... m_n \to m$.

Step3. In order to be transmitted, we first encrypt *m*, obtaining the encrypted message *c* in an arbitrary number of copies that is equal to the arbitrary number of copies of *m* ($E: me_1 \to c$).

Step4. All the encrypted messages *c* are passing through the cell membrane of T' into the upper cell, *T*: (*c*, *out*).

Step5. Following the rule (*T*, *c*, *H*), all copies of *c* are being sent to *H* through the synapse (*T*, *H*), witch changes its state from s_T to s_H.

Step6. The host-cell, *H*, receives *c* and using the decryption key *d*, decrypts *c* into *m* ($D: cd \rightarrow m$).

Step7. As nothing was concealed in the received information, *H* sends the encrypted message *c* to the receiver-cell following a rule of the form (*H*, *c*, *R*), through the synapse (*H*, *R*) changing its state from s_H to s_R.

Step8. *R* receives *c* and in order to decrypt it, making use of the rule (*Em*, *d*, *R*) it brings an arbitrary number of copies of the decryption key *d,* from the environment into the system.

Step9. In addition, for the message to be decoded, *d* and *c* are passing into the inner cell R'. This is done by the means of the two rules applied in parallel for all copies of *d* and *c*: (*d*, *in*) and (*c*, *in*).

Step10. The decryption of *c* is taking place using this rule: $D: cd \rightarrow m.$ As *m* is not an altered or fraudulent message, meaning $m \in M'$, it is not rejected by the cell R' and the computation goes to the next step.

Step11. The message *m,* obtained in *Step10*, is encrypted in parallel for both the encryption key found in R', in order to obtain the subliminal bit: $E: me_1 \rightarrow c_1$ and $E: me_2 \rightarrow c_2$. As only c_1 can be found in R' having the form of *c*, the encrypted message c_2, along with its copies, is rejected by the cell.

Step12. The computation stops after applying the last step of rewriting c_1 into *c*. As in the original encryption of the message *m* in the transmitter cell, the key that was used is found to be e_1, this leads us to the subliminal bit transmitted from the transmitter-cell to the receiver-cell. That is a 0.

4. Conclusions

The paper belongs to the cryptography model research domain with orientation to the biology inspired architectures. A new perspective of communication via subliminal channels is created in the membrane computing manner and presented by the implementation of such a channel in the framework of a tissue-like P system. The role of information exchanged between a transmitter and a chosen receiver is now played by objects, two secret encryption keys and a public decryption key, placed, initially, in the environment. These objects are evolving according to some rules representing processes, like encryption, decryption or communication, used in the basic protocols of communication through membranes, between regions of the same cell, between cells, and between cells and their environment. Although recently introduced, the study of molecular computing based on membrane computing, called the attention of a large number of researchers. But, in the new field of membrane based cryptography a small number of papers are already known. This paper, in witch we propose the modelling of a subliminal channel in the framework of membrane computing, represents an absolute novelty that leads us in new research over this domain, starting with an analyze of the channel security in this environment.

References

1 G. J. Simmons. The Prisoners' Problem and the Subliminal Channel. Proceedings of Crypto 83, Santa Barbara, CA, August 21-24, 1983, published by Plenum Press.

2 G. J. Simmons. Message Authentication without Secrecy. Secure Communications and Asymmetric Cryptosystems, ed. by G. J. Simmons, AAAS Selected Symposia Series, West View Press, Boulder, CO (1982), pp. 105-139.

3 G. J. Simmons. The Subliminal Channel and Digital Signatures. Eurocrypt '84, Paris, France, April 9-11, 1984, Advances in Cryptology, Ed. by T. Beth et al.., Springer Verlag, Berlin, 1985, pp. 364-378.

4 C. J. Simmons. A Secure Subliminal Channel (?). Crypto '85, Santa Barbara, CA. August 18-22, 1985, Advances in Cryptology, Ed. by H. C. Williams, Springer-Verlag, Berlin, 1986, pp. 33-41.

5 C. J. Simmons. Subliminal Communications is Easy Using the DSA. Advances in Cryptology - Eurocrypt '93, LNCS 765, pp. 218-232, Springer-Verlag, Berlin, Heidelberg 1994.

6 G. J. Simmons. Subliminal channels: Past and present. Eur. Trans. Telecommunication, vol. 5, no. 4, pp. 459-473, July / Aug, 1994.

7 G. J. Simmons. The History of Subliminal Channels. IEEE Journal on Selected Areas in Communication, Vol. 16, No. 4, May 1998.

8 Gh. Paun, G. Rozenberg, A. Salomaa. DNA Computing. New Computing Paradigms. Ed. Springer-Verlag, Berlin Heidelberg, 1998.

9 Gh. Paun. Computing with membranes (P systems): Twenty six research topics. It can be found in C. S. Calude, M.J. Dinneen, Gh. Paun. Pre-proceedings of the workshop on multiset processing (WMP-CdeA 2000). Pages 203-217, CDMTCS-140, August 2000.

10 C. Calude, Gh. Paun. Computing with cells and atoms. Taylor and Francis, London, 2000 (Chapter 3: Computing with membranes).

11 G. Ciobanu, Gh. Paun, M. J. Perez-Jimenez. Applications of membrane computing. 2005, elecronic format. (Chapter 1: Gh. Paun. Introduction to membrane computing)

12 G. Ciobanu, Gh. Paun, M. J. Perez-Jimenez. Applications of membrane computing. 2005, elecronic format. (Chapter 10: O. Michel, F. Jacquemard. An analysis of a public-key protocol with membranes).

13 L. M. Morogan. Coding information in new computing models. DNA Computing and Membrane Computing. In printing at Ed. Fundatia Romania de Maine, Bucharest.http://psystems.disco.unimib.it/

Mathematical Model of Serbian-Romanian Hydropower System “Iron Gates”

Nikola Milivojevic[1], Nenad Grujovic[2], Boban Stojanovic[3]

[1,2]Faculty of Mechanical Engineering,University of Kragujevac, Sestre Janjic Street, 6, Kragujevac, Serbia, {nikolam[1], gruja[2]}@kg.ac.yu

[3]Faculty of Science, University of Kragujevac, Radoja Domanovića, 6, Kragujevac, Serbia, boban@kg.ac.yu

Mathematical model for hydropower estimation and operational management is developed at the Centre for Information Technology of the University of Kragujevac. The objective of the mathematical model is to ensure efficient utilization of the Danube’s hydropower potential, to address the demand of Serbian and Romanian electrical power systems which differs in terms of power and time, and to comply with a number of constraints, which are defined in bilateral agreements. In this paper it has been proven possible to simulate integrated system of cascading power plants, unlike old uncoupled model, with concurrent simulation of both hydraulic and electric processes. In a view of the complexity of this problem, the model employs a genetic algorithm mechanism described using an example involving a problem related to the optimum operating regime of all facilities of the system. It is expected to provide daily management support and is a means by which the outcomes of operational planning within different hydrologic, economic, legal and other frameworks can be assessed, and to obtain conditions for optimum water resource management and the resolution of existing and potential conflicts in the region with regard to any mis-alignment of stakeholder interests.

Keywords
optimal hydropower generation, numerical simulation, decision support system, adaptive genetic algorithms.

1. Introduction

A mathematical model for hydropower estimation and operational management of the Iron Gate 1 hydro power plant (HPP) and Iron Gate 2 HPP is being developed at the “Jaroslav Černi” Institute, Belgrade, in collaboration with the Centre for Information Technology, of University of Kragujevac/Faculty of Mechanical Engineering. It was commissioned by PD Djerdap, a company which operates within the scope of the Electric Power Industry of Serbia. The objective of the mathematical model is to ensure efficient utilization of the Danube's hydropower potential, to address the demand of Serbian and Romanian electrical power systems which differs in terms of power and time, and to comply with a number of constraints at various control profiles of the Danube, which are defined in bilateral agreements.

2. Description of the Iron Gate 1 and Iron Gate 2 HPP System

The Iron Gate 1 system was built on a stretch of the Danube (rkm 943+000) which is shared by Serbia and Romania. The main structure (dam) is 1280 m long and is symmetrically divided into its Serbian and Romanian portions, each of which is comprised of: a navigation lock; non-overflow earth dam, a hydro power plant with 6 power units, and the respective portion of a gravity concrete dam with 14 spillways (25 m clearance).

The major characteristics of the dam are: net head 15.4 - 31 m, total installed discharge 9,800 m^3/s, and total installed power 2,165 MW. The Iron Gate 1 Reservoir was formed in a complex river system comprised of the Danube and its tributaries: the Tisa, Sava, Velika Morava, Tamiš, Nera, Mlava, Pek and Poreč rivers. An important characteristic of the Iron Gate 1 Reservoir is its variable length, extent of backwater and volume, as a function of the flow rate and HPP operating modes. The volume of the reservoir under average hydrologic conditions is 3,500 million m^3.

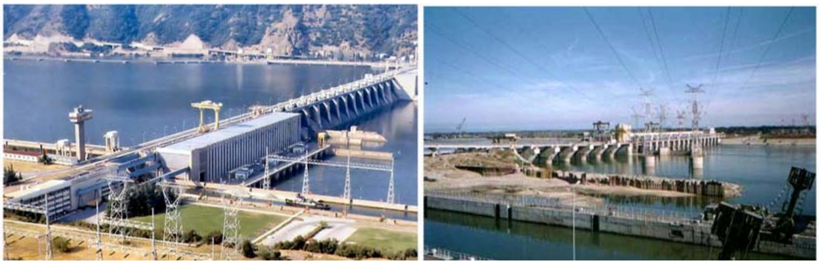


Figure 1 Iron Gate 1 and Iron Gate 2 dams.

The Iron Gate 2 system is the downstream part of a cascade and its operation is coupled with that of the Iron Gate 1 system. Two HPPs, with 8 power generating units each, are located on the dam at rkm 862+800 of the Danube's main stream, along its left bank. At mid-section, there is an overflow dam with 7 spillway fields, and closer to the right bank there is a Serbian navigation lock and an additional HPP with two power generating units. The Romanian navigation lock is located in a channel running through the Island of Mare. The dam on the Gogoš Arm of the Danube (rkm 875) has a spillway at mid-section with 7 spillway fields, and an additional Romanian HPP with 2 power generating units, adjacent to the right bank of the river arm. The HPPs are equipped with horizontal encapsulated power generating units whose installed power is 27 MW each, making up a total of 540 MW. The total installed discharge is 8,500 m^3/s. Gross head varies from 2.5 m to 12.75 m, depending on the flow of the Danube. The Iron Gate 2 Reservoir is an 80 km long stretch of the Danube's channel. At maximum water levels, the volume of the reservoir is 820 million m^3. No significant tributaries empty into this reservoir.

3. Model Scope and Objectives: Optimum Management of the System of HPPs

The mathematical model can simulate and optimize the operation of the complex Iron Gate 1 and Iron Gate 2 hydropower system based on pre-defined individual facility performance levels, Serbian and Romanian electrical power demand, water level and discharge constraints at control profiles, and initial and boundary conditions, whereby it can address various scenarios of initial data and required outputs. This allows for efficient daily management decision-making with regard to appropriate operating modes.

Namely, water is evacuated by both Romanian and Serbian facilities (each drawing on its portion of the potential), via respective power units and dam spillway fields. The system is managed in such a way as to ensure optimum utilization of the Danube's hydropower potential, provide for unhindered navigation, and ensure that riparian lands are not threatened [1], [2]. It is especially noteworthy that the upstream portion of the Iron Gate 1 Reservoir stretches over lowland and that the riparian lands of the Iron Gate 2 Reservoir also lie relatively low.

As a result, bilateral agreements impose water level restrictions at characteristic flow profiles, implicitly determining HPP operating modes. In compliance with Iron Gate 1 and Iron Gate 2 Operation Regulations [18], daily production plans for the current day and following days are generated and synchronized by the Romanian and Serbian parties. The daily plan defines: daily average water levels, daily average discharges, water evacuation modes for the dams, amounts of water to be evacuated, total available energy, maximum/minimum HPP power, and overflow magnitude expressed by means of overflow energy. The daily plan is adjusted if daily plan reviews, previous day analyses, and updated forecasts of the Danube's flow rate indicate that water levels will exceed set constraints, which cannot be tolerated.

Therefore, the basic task of the mathematical model is to determine the sequence and dynamics of power unit engagement (and of spillways, as needed), relative to predicted inflow rates and pre-defined hourly output plans, or pre-defined hourly production priorities for each Serbian and Romanian HPP, while complying with predefined constraints and striving to minimize any departure from plan and minimize water consumption (i.e. maximize hydropotential utilization).

4. Spatial Decomposition of the System, Theoretical Background and General Logic of the Mathematical Model

The model addresses water flow and power generation in the entire Iron Gate 1 and Iron Gate 2 system. The entry of water into the system is represented by the flow rates at river system profiles upstream from the reservoir (the main stream of the Danube and all of its tributaries). On the other hand, it has to cope with user demand (Serbian and Romanian electrical power demand as a function of time) and prescribed constraints. As a result, the model includes all relevant types of linear flow: morphology-based flow in natural watercourses and flow through facilities (HPPs, dam spillways, dam outlets, navigation locks, and the like). Additionally and very importantly, modeling includes the variation in flow conditions as a function of time, due to management decisions. The model has been developed for an hourly time-step environment.

In view of the spatial and functional complexity of the system, the modeled area has been broken down into various elements which can be used to simulate different types of water flow, both natural and artificial (Fig. 3)[3].

The basic element of the river network is a river reach which is used to model a portion of open flow between the junction, bifurcation and man-made hydropower assets. The model which describes the complex river system is obtained by joining river reaches which define open flow, introduce tributaries, create river islands, locate hydropower assets, and hydraulically link such assets within an integrated model [4], [16], [17]. The complexity of the system, from both the modeling and numerical solving perspective, results from a large number of bifurcations, of which the most important is the Gogos-Iron Gate 2 loop. In this portion of the system there are two HPP dams in two parallel branches, so that internal conditions dictated by hydropower system management have to be met, in addition to the necessary compliance with constraints at the initial bifurcation and the point where the river arm ultimately re-joins and the main stream.

The dam structure (which includes both the HPP and spillway fields) has been modeled using internal boundary conditions. The flow through the HPPs and over the spillways is a function of the headwater and tailwater, as well as other parameters which define the operating modes of these facilities. Headwear and tailwater levels are recorded at each time step, as are other parameters which affect the operation of an HPP or spillway. The discharge is determined based on the obtained values and functions which describe the operation of an HPP or spillway. This discharge is set at one or both boundary nodes of the dam structure. Such an approach to HPP and spillway modeling allows for the definition of non-analytical forms, such as a number of decision-making methods and the inclusion of various parameters which might affect HPP or spillway operation [10].

In the above context, it should be noted that an HPP asset constitutes a set of individual power generating units which are engaged in accordance with the criterion which requires minimization of the total flow through the HPP. The transformation of gross head and discharge into energy, or the definition of the required discharge for a certain level of power relative to the current net head, is performed at every step, for each power unit based on its characteristics, or turbine hill charts (power – net head – discharge), taking into account losses in the inlet/outlet tract (whereby losses are time-dependent parameters) [5]. Consequently, the number of engaged power units is defined for each time step of the simulation, based on the minimum water consumption criterion and with the aim of achieving the required electrical power output. The spillway facility constitutes a set of individual spillway fields with their respective characteristics (overflow curves: water level – clearance – discharge), which are engaged if water evacuation through the HPPs cannot respond to all the prescribed constrains.

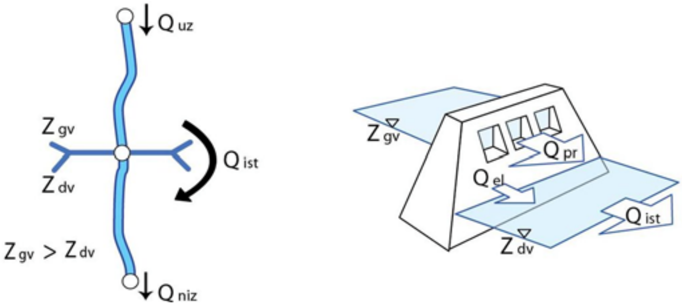


Figure 2 Representation of the dam, including HPPs and spillway, as an internal boundary condition.

In general, flow simulation is based on a one-dimensional model of unsteady flow which is used to solve basic equations of mass conservation and momentum conservation laws [6], [7]. Approximate numerical integration of 1D unsteady-unconfined flow equations was performed by implementing the "four point" method. Four points in the x-t plane are used to define the area in which these equations are (approximately) integrated, in order to obtain a system of algebraic equations for the reach. The weighted trapezoidal rule [8] was applied in the model.

The defined equations describe the laws of discharge and water level variation within an open flow network. Once the system of equations is formed, all equations must be solved simultaneously. Unknown variables of the system include discharges and water levels for a given simulation time step. Upon completion of calculations for a time step, computed values become initial values for the next time step. Once the characteristics of the entire model for a given time step have been determined, the system of non-linear equations is solved iteratively, applying the Newtonian method. Since there are active hydropower facilities (dam with HPP and spillway), whose functioning depends on current headwater and tailwater levels, it is also necessary to adjust discharges through the given facilities. Using such adjusted values, the entire equation solving procedure is repeated until the convergence criterion has been achieved.

5. Optimization within the Mathematical Model

A very important aspect of the mathematical model developed for the Iron Gate 1 and Iron Gate 2 HPP system is the solving of operation optimization problems, whereby different objectives may be encountered in practice.

For example, the objective may be to achieve maximum utilization of the hydropotential based on pre-defined hourly production priorities. The objective may also be minimal variation from the pre-defined production plan. In both cases, the number of engaged power generating units is also optimized based on the minimal water consumption criterion. All of these optimizations have to adhere to prescribed constrains (generally water level restrictions at characteristic flow profiles).

In most cases, this type of problem is solved by evolutionary algorithms [11], [12], dynamic programming [13] and the augmented Lagrange multiplier method [14]. In the present model, satisfying imposed constrains is a complex problem since there can be a time lag between the cause and the actual violation of a constraint. The time interval between the occurrence of the cause and its effect has not been uniquely defined, since it depends on a large number of other system parameters and on the flow of the simulation itself. In a view of the complexity of this problem, the model employs a genetic algorithm mechanism which will be described using an example involving a problem related to the optimum operating regime of all facilities of the system, based on pre-defined hourly production priorities.

The priority plan is given for two complete systems (Serbian and Romanian), in tabular form and such that preference is defined by priority level instead of the weight coefficient for every hour. The priority structure is entered into target programming and the assumption is made that constraints are primary criteria which have to be met and they are, therefore, given the highest priority. The task is addressed by the weight coefficient method, which is the most frequently applied method in multicriteria optimization. This method introduces weight coefficients w_i for all criteria functions f_i^* $i = 1,\ldots,n$, and the vectorial optimization problem is reduced to scalar optimization $\max \sum w_i P_i \Delta t$ $i = 1,\ldots,24$, where $P_i \Delta t$ is the generated energy E during time interval Δt.

The target function for production optimization in terms of pre-defined hourly priorities has the form of $\max \left\{ \sum \sum w_{i,j} P_{i,j} \right\}$ $i = 1,\ldots,n,\ j = 1,\ldots m$, where $w_{i,j}$ are priority levels for the j^{th} HPP facility in the i^{th} step, and $P_{i,j}$ is the power achieved by the j^{th} facility in the i^{th} step.

In order to improve algorithm efficiency, solutions which violate constraints, or lie outside of the feasible space, are also addressed and the proposed plan assessed with regard to the intensity of the potential constraint violation. This is achieved by an internal addition of a penalty term to the objective function $\sum \alpha_k \sum \left| g_l\left(z_k\right) \right| = 0$ $k = 1,\ldots,q,\ l = 1,\ldots,r$. Functions $g_l\left(z_k\right)$ represent a numerical value which describes the number and intensity of violations of the l^{th} constraint at the k^{th} flow profile. To achieve better convergence of the algorithm, a separate weight coefficient is added to each constraint. It should be noted that a solution is acceptable if, and only if, the sum $\sum \alpha_k \sum \left| g_l\left(z_k\right) \right| = 0$, since in that case there is no violation of any system constraint. However, even though the basic objective is to comply with all constraints, the introduction of this factor allows for the evaluation of solutions which violate any of the constraints, but in such a way that the next step favors the proposed adjustment which was closest to satisfying the imposed system constraints. Weight coefficients α_i allow for preference to be given to particular internal terms of the fitness function.

Unit commitment is implemented on an hourly basis, during which time a particular HPP facility is engaged based on the proposed HPP output and the power generating units are engaged based on minimum consumption.

The solution, in the form of hourly production plans for individual power generating units, is coded into the binary gene. The process results in one or more genes which, in effect, represent the optimum hourly plan for a particular power generating unit and the extent of any achieved optional overflow. Genetic algorithm performance improvements, in terms of maintaining a favourable exploration/exploitation ratio during the entire optimization process, were achieved by introducing fuzzy logic controllers which were used to adapt genetic algorithm parameters [15]. The adaptive genetic algorithm approach is such that at every n generations, applying the proposed fuzzy rules, the mutation probability (p_m) is determined on the basis of its value in the previous generations and the achieved best individual fitness (f_n) improvement.

6. Software Structure

Major software modules are: a user interface, a module which simulates unsteady flow in the open-flow network and optimizes operation of hydropower facilities, and a database [21].

User interface. A user-friendly, modern, graphically oriented interface has been developed, which interactively guides the user through all simulation model application stages: database search, handling of input data and model object parameters, initiation of the simulation/optimization process, handling of output data, and creation of reports.

Simulation/optimization module. This module activates implemented numerical analysis methods and optimization algorithms which were described earlier in the text. The module provides two-way communication with the user interface (problem definition and output review/analysis).

Database. The database integrates required data with an appropriate record structure, and is based on existing hardware platforms and types of databases used within the system. The database contains diverse data: comprehensive system configuration information (e.g., hydrographic network, facilities, monitoring sites, and riverbed morphology), comprehensive facility performance data (e.g., turbine hill charts, spillway-field discharge characteristics, etc.), constraints, history of measured reservoir water levels, history of measured HPP electrical and non-electrical parameters (e.g., power, discharge, net and gross head, power output, spillway field discharge), and the like.

The software has been designed for a Windows platform and developed using a three-layer model, which makes a clear distinction between functional units: a presentation layer, a business logic layer, and a data layer. The portion of the application which interacts with the user is referred to as the presentation layer; it is implemented via Windows forms. The business logic layer has been implemented as a code within the forms. The data layer represents any database which is supported by the .NET environment (in this case the Microsoft SQL Server), and communicates with the business logic central layer via ADO.NET objects [4].

7. Parameter Estimation and Accuracy Verification

In addition to experimental data relating to the performance of all system components (e.g., overflow curves or turbine hill charts) or riverbed morphology, the database includes „model parameters“ which cannot be determined by observation or direct measurement of flow characteristics [20]. These include Manning coefficients of roughness, which vary as a function of the physical position along the flow and the flow rate.

Model parameters have been estimated through optimization, with the goal of achieving the best possible match between computed and measured water levels at control profiles. Based on known inflow and outflow data, the computed value is obtained through an iterative process (simulation, assessment, comparison, correction, and repeated simulation). The previously-described evolutionary algorithms are also included in the estimation procedure. The target function of the evolutionary algorithms is minimum deviation of computed water levels from corresponding measured water levels.

The roughness coefficient estimation process encompasses a wide range of total flow rates to the reservoir, from 2500 to 10000 m^3/s, including both quasi-steady flow periods and periods of sudden flow rate variation. Figure 3 is a graphical representation of a comparison between measured and simulated values for the dam and the most important control profile, reflecting an arbitrary historic period of 7 days.

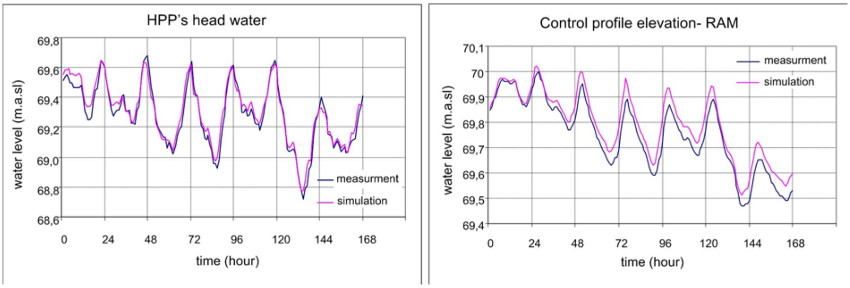


Figure 3 Measured vs. simulated values: Dam profile and control profile.

Even though historic data contained a certain degree of inaccuracy (e.g., dam discharges and, particularly, estimates of natural flow to the reservoir), their careful interpretation and the application of the described parameter estimation procedure resulted in sound mathematical model calibration, and it is, therefore, possible to re-compute an episode from history to a desired level of accuracy.

8. Application of the Mathematical Model

When applying the mathematical model, the first step is to define its spatial and temporal framework, model performance levels and parameters, constraints which have to be complied with, input flow time series, and the functional mode (which determines the simulation/optimization procedure for solving a specific problem), along with relevant energy requirements [19]. The next step is pre-processing, or numerical computation of the initial status of the system. Then, based on the initial status and given incoming flow rates and electrical power demand, hydraulic/hydropower simulations and system operation optimization are conducted based on predefined management criteria.

Management criteria are primarily defined by the model's functional mode, as well as by selection of relevant attributes depending on the type of analysis being conducted. In the widest sense, there are three mathematical model functional modes:

Mode 1: Review and adjustment of a specified production plan

This functional mode of the mathematical model is used to check and modify the existing (daily) production plan and to define the needed spillway field discharge regime (as required), with the goal of complying with constraints and minimizing any departure from the proposed plan. This, of course, implies that system inflow forecasts, initial water levels and the production plan are available.

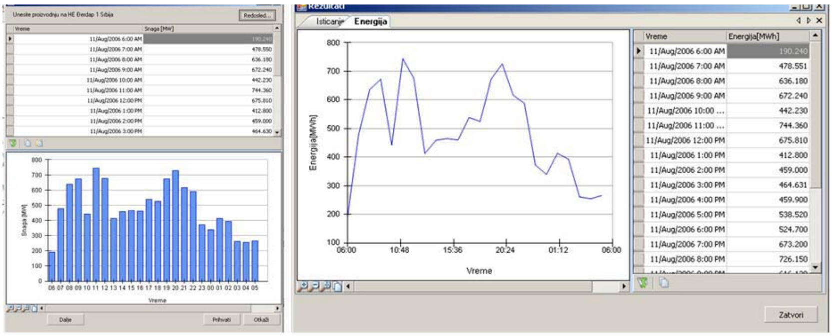


Figure 4 Energy at the Iron Gate 1 HPP: Serbian side (a: Input data – given production plan; b: Simulation output, adjusted production plan).Mode 2: Optimum operation in the absence of a specified production plan

The second mode is used to solve problems which do not include a pre-defined hourly production plan. The hourly production plan is replaced with set hourly priorities, which give preference to particular periods during the day. Priority plans are specified for all HPPs included in the configuration, in tabular form, with preference defined by the priority level for each hour. The priority level is represented by an integer which, in the general case, can be from 1 to 24 in a 24-hour time step sequence. The goal of this mode is to arrive at an hourly plan of power unit engagement (and an overflow plan, as needed), based on specified hourly priorities which do not violate constraints and maximize power output. This, of course, implies that system inflow forecasts and initial water levels are available.

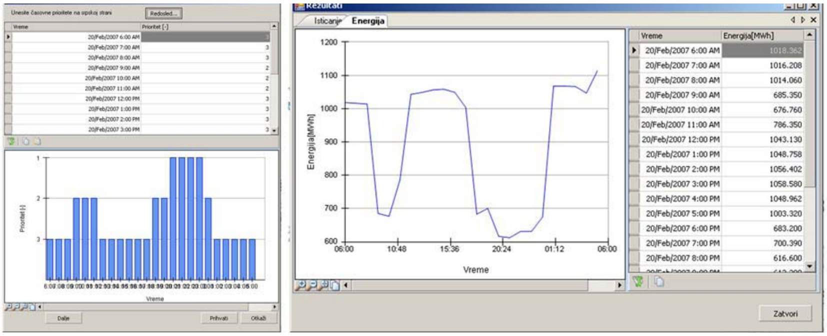


Figure 5 Energy at the Iron Gate 1 HPP: Serbian side (a: Specified priority plan - input data, b: Simulation output).Mode 3: Explicit setting of power unit/spillway field operational parameters

This mode, which explicitly specifies operational parameters for the power units and spillway fields (e.g., power unit discharge, spillway field discharge, or individual power unit output and gate clearance of individual spillway fields), is used to repeat historic periods and to check the operation of HPP and spillway facilities, as well as to perform supplemental estimations of mathematical model parameters.

9. Conclusion

In concluding, it is worth stressing some relevant aspects concerning the proposed modelling and software development methodology. The planning of the electric power utilities has been one of the most important areas of application of the operational research methods, and this is due not only to the nature of the problems in this area, in general well structured, always challenging and linked to the minimization of costs associated with improvements to operational procedures, but also to the technical guidance provided by the managers and the ample availability of data and information.

Now, when power generation firms move towards a competitive environment, the planning methodology and models must be adapted or even changed to confront a wider range of objectives, that are not only strategic, but also financial and behavioural, especially in such cases when there is more than one entity (firm, state etc.) operate at the same objects, competing in usage of the common resource. The possibilities of management failure, previously nonexistent now tend to accentuate with the increase in competition.

The mathematical model for hydropower estimation and operational management of the Iron Gate 1 and Iron Gate 2 HPP system is a complex software which has been designed to simulate and optimize operation of cascade HPPs, based on pre-defined facility performance levels, initial/boundary conditions, electrical power system demands, and prescribed constraints at control profiles and facilities.

This software is expected to provide daily management support and is a means by which the outcomes of operational planning within different hydrologic, economic, legal and other frameworks can be assessed. The development and application of this software is a step toward the strategic goal: the creation of conditions for optimum water resource management and the resolution of existing and potential conflicts in the region with regard to any misalignment of stakeholder interests.

References

1 Methodology for Hydro Power Plant Operating Mode Planning under Unsteady Flow Conditions (in Serbian), 1984., Jaroslav Černi Institute

2 Study of a Mathematical Model for Optimum Operation of the Djerdap 1 and Djerdap 2, 1985, Jaroslav Černi Institute

3 Yeh W. W. G., Reservoir management and operations models: A state-of-the-art review, water resources research , agu, vol. 21, no. 21, December 1985.

4 Divac, D., Grujović, N., Milanović, M., (in Serbian) A New Simulation Model for Water Balance Assessments of Water Systems: Methodology, Software, and Applications, Monograph: Management of Serbian Water Resources, '99, Jaroslav Černi Institute, 1999.

5 Branislav V. Djordjević, Utilization of Water for Hydropower Generation (in Serbian), Belgrade, 2001

6 Cunge, J.A., Holly, F.M., Jr., and Verwey, A., Practical aspects of computational river hydraulics: London, Pitman Publishing Limited (reprinted 1986, Iowa City, by the Iowa Institute of Hydraulic Research), 1980.

7 Abbott, M.B., and Basco, D.R., Computational fluid dynamics: New York, Longman Scientific and Technical with John Wiley and Sons, 1989.

8 Preissmann, A: Propagation des intumescences dans les canaux et rivieres: First Congress of the French Association for Computation, Grenoble, France, 1961.

9 HEC-RAS Hydraulic Reference Manual, US Army Corps of Engineers, USA, 2002.

10 Technical report TR12/06, Center for Informational Technology, Kragujevac, Serbia, 2006.

11 Sheble, G.B., Fahd, G.N., Unit commitment literature synopsis, IEEE Trans. Power Systems 9, 1994.

12 Santos, E.F. Ohishi, T., A hydro unit commitment model using genetic algorithm, IEEE Evolutionary Computation, CEC2004, 2004.

13 Jaeeung Yi, Steven Stitt, Dynamic Optimal Unit Commitment and Loading in Hydropower Systems, Journal of Water Resources Planning and Management, Vol. 129, No. 5, 2003.

14 Sharma, V, Jha, R., Naresh, R., An augmented Lagrange programming optimization neural network for short-term hydroelectric generation scheduling, Engineering Optimization, Volume 37, Number 5, 2005.

15 Herrera F., Lozano M., Fuzzy adaptive genetic algorithms: design, taxonomy, and future directions, Soft Computing - A Fusion of Foundations, Methodologies and Applications, Volume 7, Number 8, Springer, 2003.

16 The Trebišnjica Hydro Information System (in Serbian), Jaroslav Černi Institute, 2002.

17 The Drina HIS Simulation Model, Phase 1 (in Serbian), Jaroslav Černi Institute, 2002.

18 Regulations Governing the Organization and Operation of the Joint Djerdap HPP/FE Portile de Fier Electric Power Dispatch Service

19 The Serbian-Romanian Hydropower System „Djerdap“: Mathematical Model, Vol. 1, Hydraulic reference manuals (in Serbian), Jaroslav Černi Institute, 2007.
20 The Serbian-Romanian Hydropower System „Djerdap“: Mathematical Model , Vol. 2, Database reference manuals (in Serbian), Jaroslav Černi Institute, 2007.
21 The Serbian-Romanian Hydropower System „Djerdap“: Mathematical Model , Vol. 3, User manuals (in Serbian), Jaroslav Černi Institute, 2007.

Appendix

List of Acronyms

HPP	Hydro power plant

Special Tailored Boosting Method for Face Recognition

Maria Miroiu[1]*, Ion Vaduva*[2]

[1]*University of Pitesti, Faculty of Mathematics and Computer Science, 1 Targu din Vale Str, 110040 Pitesti, Arges, Romania, maria.miroiu@upit.ro*

[2]*University of Bucharest, Faculty of Mathematics and Computer Science, 14 Academiei Str, 70109 Bucharest, Romania, vaduva@fmi.unibuc.ro.uk*

The seminal work of Viola and Jones [6] managed to impose a special class of developments by applying boosting to robust and fast object detection tasks, therefore proposing a new area in computer vision. The most significant application area refers to face detection, face recognition and pedestrian detection. Inspired also by the idea of age estimation using boosting based regression [8], we propose a new image recognition approach to young face detection that uses an extended haar-like features [3] instead of the basic haar-like features proposed by Viola and Jones. We trained our system on images from FGnet database [1] and we saw on testing that the system is capable of detecting successfully upright frontal young faces observed in reasonable lighting conditions.

Keywords

boosting, face recognition, haar-like features.

1. Introduction

Face recognition is a challenging topic in pattern recognition mainly because the variations of head poses and different facial expressions such as smiling/non smiling or open eyes/closed eyes as well as shadings and other variations in lighting imply serious difficulties in the identification of individuals. Also aging modeling is an important step for face analysis and recognition.

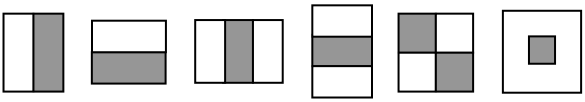
During the recent years, the boosting method has become a popular and robust ensemble method in improving the accuracy of a given learning algorithm. The most representative boosting algorithm is AdaBoost (Freund and Schapire, 1995). Viola and Jones [VJ01] introduced a new and effective face detection algorithm based on simple features trained by the AdaBoost algorithm. The detection could be made in real time and yet is very flexible in the sense that it can be trained for different levels of computational complexity, speed and detection rate suitable for specific applications.

We focused here on young faces detection. We were able to successfully construct and apply the a robust detection system mainly based on the algorithm proposed by Viola and Jones, but we have used instead an extended set of haar-like features. Although the extended feature set usually complicates the learning, it was more paid of by the added domain knowledge.

2. Image Representation and Feature Extraction

For image representation we used the values of features instead of pixel intensities directly, mainly because the features encode domain knowledge that is difficult to learn using a finite quantity of training data better than pixels. Other reason is that a feature-based system can be much faster than a pixel-based system. In our work we used an extended Haar-like features [3] as an alternative to simple Haar-like features [6], [7]. face analysis and recognition.

The *simple Haar-like features* used by Viola and Jones [6] are of three kinds. The value of a *two-rectangle feature* is the difference between the sums of the pixels within two rectangular regions. The regions have the same size and shape and are horizontally or vertically adjacent (see figure 1). A *three-rectangle feature* computes the sum within two outside rectangles subtracted from the sum in a center rectangle. Finally, a *four-rectangle feature* computes the difference between diagonal pairs of rectangles. The position and size of the feature can vary over the detection sub-window of the image that is used.



(1a) (1b) (1c) (1d) (1e) (1f)

Figure 1 Prototypes of simple Haar-like features

The number of features derived from each prototype is quite large and differs from prototype to prototype. For example, the number of features of type (1a) of high h and width of white region w equal with the width of black region b inside of a $W \times H$ detection window is

$$\sum_{w=1}^{[W/2]}(W+1-2w)\cdot\sum_{h=1}^{H}(H+1-h)$$

Table 1 lists the number of features for a detection window size of 24×24.

Feature type	Minimum sizes of the feature	Maximum sizes of the feature	#
1a ; 1b	2×1 ; 1×2	24×24 ; 24×24	43,200
1c ; 1d	3×1 ; 1×3	24×24 ; 24×24	27,600
1e	2×2	24×24	20,736
1f	3×3	24×24	8,464
Sum			100,000

Table 1 Number of features inside of a 24×24 window

Upright rectangle features can be computed very quickly using the integral image method which creates a new image *ii* (called *integral image*) for each test image *i* such that the integral image at location (x, y) is the sum of all pixels values above and to the left of (x, y), inclusive:

$$ii(x, y) = \sum_{x'<x, y'<y} i(x', y')$$

For this, each test image can be scanned once and the corresponding integral image is found using the following pair of recurrences:

$$s(x, y) = s(x, y\text{-}1) + i(x, y)$$

$$ii(x, y) = ii(x\text{-}1, y) + s(x, y) ,$$

where $s(x, y)$ is the cumulative row sum with $s(x, -1) = 0$ and $ii(-1, y) = 0$.

Using this method is no longer necessary to add or subtract individual pixels. The brilliance in using an integral image to speed up a feature extraction lies in the fact that any rectangular sum can be calculated from the corresponding integral image, by indexing the integral image only four times. Given a rectangle specified as four coordinates (x_1, y_1) upper-left corner and (x_2, y_2) lower-right corner, evaluating the area of a rectangle is done in four integral image references:

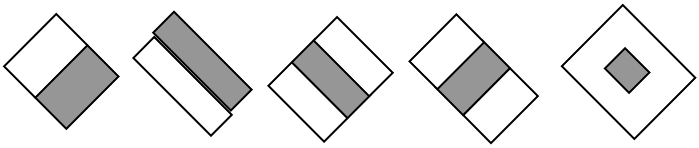
$$rs(x_1, y_1, x_2, y_2) = ii(x_1, y_1) + ii(x_2, y_2) - ii(x_1, y_2) - ii(x_2, y_1)$$

Since the two-rectangle features from Figure 1 involve adjacent rectangular sums, they can be computed using six pixels from the integral image, eight in the case of the three-rectangle features and nine for four-rectangle features.

For a given set of training images, we can extract a large collection of features very fast using the idea above. The hypothesis of Viola and Jones is that a very small number of these features can be combined to form an effective classifier.

This rectangle features are sensitive to the presence of edges, lines, bars, and other simple image structure. But the only orientations available are vertical, horizontal, and diagonal.

Lienhart [3] introduced an additional set of rotated Haar-like features, which significantly enrich the basic Haar-like features from Figure 1 and which can also be computed rapidly at all scales in constant time. This set contain 45^0 rotated features so they add additional domain-knowledge to learning framework.



(2a) **(2b)** **(2c)** **(2d)**

Figure 2 Examples of rotated Haar-like features

The number of features from Figure 2 inside of a 24×24 window is

8,464 + 20,736 + 4,356 + 3,600 + 1,521 = 38,677.

Also, rotated rectangle features can be computed very fast using an auxiliary image, which this time is the *Rotated summed area table iii*. $iii(x, y)$ gives the sum of the pixels of the rectangle rotated by 45^0 with the right most corner at (x, y) and extending till the boundaries of the images:

$$iii(x, y) = \sum_{x'<x, x'<x-|y-y'|} i(x', y')$$

It can be calculated with two passes over all pixels. The first pass from the left to right and top to bottom determines

$$iii(x, y) = iii(x\text{-}1, y\text{-}1) + iii(x\text{-}1, y) - i(x, y) - iii\ (x\text{-}2, y\text{-}1)$$

with $iii(\text{-}1, y)= iii(\text{-}2, y)= iii(x, \text{-}1)=0$, whereas the second pass from the right to left and bottom to top calculates

$$iii(x, y) = i(x, y) + iii(x\text{-}1, y\text{+}1) - iii\ (x\text{-}2, y).$$

From this, the pixel sum of any rotated rectangle like the one from Figure 3 can be determined by four image references:

$$\text{iii(x+w, y+w) + iii(x-h, y+h) – iii(x, y) – iii(x+w-h, y+w+h).}$$

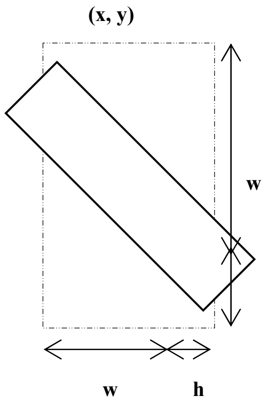


Figure 3 Example of a 450 rotated rectangle

3. AdaBoost Training and Feature Selection

Each Haar-like feature has its own attributes: type, position, and size. Given a moderate size of image, one can generate a huge number of Haar-like features by varying the rectangle feature attributes. Even though each feature can be computed very efficiently, computing the complete set is prohibitively expensive. Also some of the features could be slightly scaled versions of other features. That's why we need now a method that can reduce this feature set and give us the best features that discriminate the young faces from the other and also complement each other.

In this paper we use a variant of AdaBoost algorithm [2] both to select a small set of relevant (best) features and train the classifier. In general, *boosting methods* convert (boost) a weak learning algorithm into a strong one. Boosting has been analyzed carefully and tested empirically by many researchers. We focus on the discrete AdaBoost (adaptive boosting) algorithm introduced by Freund and Schapire [2]. It adaptively re-weights the training examples instead of resampling them. The basic algorithm has as input a training set with positive and negative labeled examples and an initial uniform distributed weights over the examples. Based on them, a weak classifier is trained. A weak classifier is called a classifier that performs slightly better than random guessing, i.e., for a binary decision task, the error rate is less than 50%. The classifier is obtained by applying a learning algorithm (e.g. applying statistical learning for a decision stump). Then the example weights are updated in order to emphasize those which were incorrectly classified by previous weak classifier. Therefore, the algorithm focuses on the difficult examples. The process is repeated, and a new weak classifier is added on each boosting iteration, until a certain stopping condition is met (e.g. a given number of weak classifiers are trained). Finally, a strong classifier is computed as weighted combination of weak classifiers followed by a threshold. Freund and Schapire [2] proved strong bounds on the training and generalization error of AdaBoost. For the case of binary classification the training error drops exponentially fast with respect to the number of boosting rounds T (i.e. number of weak classifiers). Schapire et al. [4] showed that boosting algorithms maximizes the margin and proved that larger margins for the training set are translated to superior upper bounds on the generalization error.

Boosting for feature selection was introduced by Tieu and Viola [7]. Feature selection from a large set of feature is done by AdaBoost. The main idea is that each feature corresponds to a single weak classifier and the boosting algorithm selects an informative subset from these features. Training proceeds similar to the above described boosting algorithm. Given a set of possible features, the algorithm builds a weak hypothesis based on the weighted training samples on each iteration step t. The best one forms the weak hypothesis h_t which corresponds to the selected feature f_t. With respect to the error of the chosen hypotheses the weights of the training samples are updated. Finally, a strong classifier H is computed as a weighted linear combination of weak classifiers, where the weights are inversely proportional to the training errors of the corresponding weak classifiers.

Following the spirit of the Viola and Jones [6], we used one-dimensional *decision stumps* as primitives to construct the weak function set H. The advantages of using decision stumps includes that they are robust to appearance variation, that they are local features, that they are fast to evaluate using integral image, and, most importantly, that they allows an incremental feature selection scheme that will be addressed later.

A one-dimensional (1D) decision stump $h(x, f, p, \theta)$ is associated with a Haar-like feature f, a decision threshold θ, and a parity direction indicator p that takes the value of either +1 or -1.

$$h(x, f, p, \theta) = \begin{cases} 1 & \text{if } pf(x) < p\theta \\ 0 & \text{otherwise} \end{cases}$$

Here x is a 24×24 pixel detection sub-window of a face/non-face image and the output of our weak classifier is 1 if the image x is classified as a face and 0 if it is classified as a non-face. Thus a set of features define a set of weak classifiers. Finding the threshold of each weak classifier is an optimization step and any optimization algorithm that minimizes the error of classification can be used.

As an observation, the weak classifier h is not a true classifier in our final algorithm. In fact it is a weighted histogram of the outputs of one feature applied to all our training data along with an optimum threshold that separates the young-face output from the other outputs and hence provides an error of classification associated with a particular feature.

If we would use an intuitive procedure for training our features and finding the set of best features that constructs the weak classifier for each feature and rank the features in order of small error of classification, we'll obtain the first best feature and a set of features that look very similar to the first but are slightly scale or shifted. These features are all basically the same and fail on the test images. The AdaBoost algorithm fixes this problem by changing the weights used in computing the classification error of weak classifier. A small error is now weighted more and this ensures that our fist best feature and any other feature similar to it will not be chosen as our second best feature. This means that our best second feature is no longer similar to our first best feature and a whole different feature is selected as our next best feature. This second best feature ideally compliments our first best feature in the sense that it is successful at classifying faces that the first best feature failed on. This process is repeated to find as many best feature as desired. The detailed AdaBoost algorithm for feature selection is presented below.

```
Given labeled examples (x1,y1),…,(xN,yN), where yi=1(0) for
young (no young) face images.

Initialize weights associated to each labeled example:
w1,i=(2m)-1 for yi=-1 and w1,i=(2n)-1 for yi=1, i=1,…,N, where
m and n are the number of non-young face images and young face
images respectively.

For t=1,…,T

      Normalize the weights
```

$$w_{t,i} = \frac{w_{t,i}}{\sum_{j=1}^{N} w_{t,j}}$$

```
, 1,…,N.
   Select the best weak classifier with respect to the weight
   error
```

$$\varepsilon j = \min_{f,p,\theta} \sum_i w_{t,i} \left| h_j(x_i, f, p, \theta) - y_i \right|$$

```
   Choose the classifier ht as the hj with the lowest error εj
   Update the weights
```

$$w_{t+1,i} = w_{t,i}\beta_t^{1-e_i}$$

where ei=0 if an example xi is correctly classified, ei=1 otherwise, and βt = εt /(1-εt).

The final strong classifier is

$$H(x) = \begin{cases} 1, \sum_{t=1}^{T} \alpha_t h_t(x) \geq \frac{1}{2} \sum_{t=1}^{T} \alpha_t \\ 0, \text{otherwise} \end{cases}$$

where αt = log(1/βt).

Once a number of best features have been found, they can be used to detect faces in a test image. Each feature votes on weather it thinks the test image is a young face or not. Each feature is weighted in log-inverse proportion to the error of that feature. So a feature with a smaller error gets a heavier weighted vote.

4. Experiments

For training our detection system we used the Fgnet aging database [1]. There are 1002 facial images in the database with age ranges from 0 to 69. The face images involve all possible variations including illumination, pose, expression, beards, moustache, spectacles, etc. We used 326 images containing frontal young faces and 267 images containing frontal adult faces and with reasonable lighting. Figure 4 shows a few of these training faces. The training faces from the images were cropped, converted to grayscale and then resized to 24×24 pixels. After that we normalized them and the integral image of each training image is created for use them by our algorithm.

Figure 4 Example of cropped young and adult faces used for trainng

For a 24×24 detection window we obtain a set of close to 65,000 features of minimum size 8. Because of the huge number of features, the training process lasts for at most two days on a 2.66 GHz processor and we obtained ten best features. Although training takes much time, the detection algorithm is fast and can be used to scan large images quickly. The best ten features obtained are distributed around eyes especially eyeballs, the region under and above mouth, eyebraws. Two of them are ilustrated in figure 5. These features also make intuitive sense as they represent the lighter and dark areas of eyes, eyebraws, mouth and region around on a typical young face.

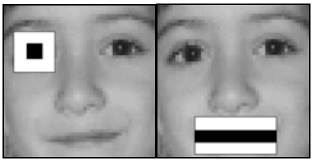
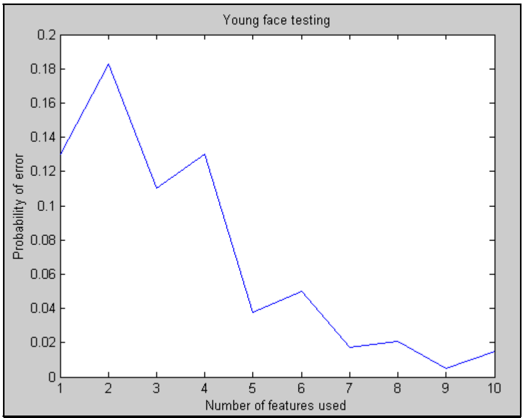


Figure 5 Two examples of best features selected after training

These best features were used to find probability of error plots. Figure 6 shows these plots. It is seen that the number of young face images misclassified decreases as the number of features used increases and the same is true for adult faces. We used the same training set as a test set for measuring the accuracy and correctness of the obtained best features. We also saw that the probability of error is very low over the training set so these features are very successful at actually doing what they were trained to do.



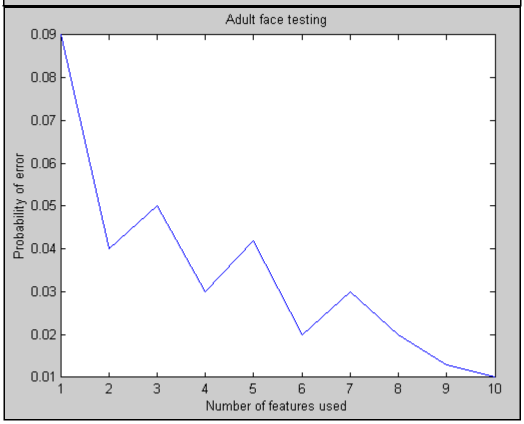


Figure 6 Probability of error varying with the number of best feature used

We also tested our detection system on some random images other then the training ones. The program scans the image at multiple levels and crops out parts of the image, resizes the parts to 24×24 pixels, normalizes and classifies it. In general, our detector is capable of detecting upright frontal faces observed in reasonable lighting conditions. It recognizes babies and typical young and adult faces, but it fails on girl and boy faces with beetle-browed or who look like an adult.

A single young face within an image usually had multiple detection windows that decided that there is a young face inside, since it is classified as a young face at multiple levels and by different scans. In order to limit the number of false positives within images, a lower bound on the size of the cropped out parts from images was imposed on our detection window.

On the random test images we saw also that the system doesn't recognize them at all if the image has tiny faces. Of course we can change the minimum features size to less than 8 for fine tuning the detection.

5. Conclusions

We were able to successfully construct and apply a detection system on young face recognition. The system is an adaptation of the Viola and Jones proposal using instead an extended Haar-like feature set.

References

1 FGnet Aging Database, www.fgnet.rsunit.com

2 Freund Y, Schapire R. A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences 1997, 55:1, 119–139.

3 Lienhart R, Maydt J. An extended set of Haar-like features for object detection. Proceedings International Conference on Image Processing 2002; 900-903.

4 Schapire R, Freund Y, Bartlett P, Lee W S. Boosting the margin: A new explanation for the effectiveness of voting methods. In Proceedings International Conference on Machine Learning 1997; 322–330.

5 Tieu K, Viola P. Boosting image retrieval. In Proceedings IEEE Conference on Computer Vision and Pattern Recognition 2000; 228-235.

3 Viola P, Jones M. Robust real time object detection. Proceedings IEEE Conference Computer Vision and Pattern Recognition 2001; 1, 511-518.

4 Viola P, Jones M. Robust real time face detection. International Jornal of Computer Vision 2004; 57(2), 137-154.

5 Zhou S K, Georgescu B, Zhou XS, Comaniciu D. Image based regression using boosting method. 10th International Conference on Computer Vision 2005; 1:541-548.

Optimal Luminance-Chrominance Downsampling through Fast K-Means Quantization

Nicolaie Popescu-Bodori[1], Luminita State[2]

Dept. of Mathematics and Computer Science, Spiru Haret University, Bucharest, Romania, bodorin@ieee.org[1], radus@sunu.rnc.ro[2]

Abstract – This paper presents an algorithm within the k-means paradigm optimized for fast image processing and its applications to the downsampling of the luminance-chrominance channels of continuous-tone still images and to the iris segmentation.

Entropy reduction of any given image is achieved here by applying three main processing operations on each of the luminance / chrominance channels: a preliminary scalar quantization with 256 equally spaced reconstruction levels; a k-means computation of a new set of reconstruction levels containing fewer, non-uniformly spaced but much more significant values; the final quantization step using this second set of reconstruction values. No explicit median filter is applied.

This approach enables us to find the appropriate reconstruction values according to the image's chromatic features (gradient and edges) without explicitly computing them, while preserving local visual significance and the quality of the processed picture. Our results shows that severe downsampling rate (for example 1:1/16:1/16:1 downsampling corresponding to 256:16:16 YCbCr reconstruction levels) can be achieved with no significant loss in the visual quality of the processed image by applying the proposed Fast-K-Means Quantization algorithm. Finding 16 reconstruction values in each of the CbCr channels through Fast K-Means algorithm is achieved here with a computational cost comparable with that of the RGB to YCbCr conversion.

Keywords

iris segmentation, k-means, run-length filtering, fast k-means quantization.

1. Problem Statement

To avoid any confusion we will remind that the terms downsampling / downsizing / subsampling (and sometime resampling) are used in image processing to identify *a* *procedure by which the spatial resolution of an image is reduced* [1]. At first sight, there is no connection between downsampling and quantization defined as being the process in which *the amplitude of a signal is compared to a set of decision levels* [2]. But in fact, downsampling decreases the storage of an image creating (and ignoring) gaps in the original image, gaps that will be further reconstructed by some median filter. Consequently, downsampling/upsampling procedures retain in the image storage other values than those originally being stored by initial quantization of the image. This fact that implicitly links together the two discussed concepts enables us to formulate the following problem: *for a given image is it possible to identify a set of reconstruction values and a set of neighborhoods, both of the same length, ensuring that the mean of each neighborhood is the corresponding reconstruction value?* Theoretically, if it is possible, it means that we have found both a quantization and a median filter whose results are perfectly matching each other. More than that, from a practical point of view, we can expect that some morphological properties of the original image will become accessible (visible, at least) in the quantized image.

Further in this paper we will prove that k-means algorithm gives affirmative answer to the above question. Also, as a practical outcome, we will present two applications of Fast K-Means Quantization Algorithm to Luminance-Chrominance Downsampling and to Iris Segmentation.

2. Preliminaries

The classical approach to downsampling [11] is to consider the minimum coded units as being blocks, or much more generally neighborhoods, sometimes mutually exclusive, sometimes not, but small enough to minimize the visual effects of some median filter applied to them. In this way, the initial image is viewed as being covered by a uniformly spaced grid defining the blocks or by a uniformly shaped collection of neighborhoods. A lucky guess of such a cover ensures that chromatic variation across these neighborhoods is smaller than a desired threshold. Otherwise, an important (noise-like) local chromatic variation corresponding to an essential detail may produce a strong unwanted visual effect.

Also, downsampling can be done adaptively to amplitude of the high frequency components [3] and also by varying the threshold and the type of neighborhoods according to local chromatic variation. By doing so, we negotiate between making a more accurate numerical representation of any particular processed image and preserving enough regularity in the neighborhoods shape to ensure computational efficiency. Even so, while remaining an explicit median filtering on imposed uniformly shaped neighborhoods, adaptive downsampling doesn’t solve the following issues:

- The encoding of one fixed chromatic value depends on the local chromatic medians computed across those neighborhoods containing that value instead of depending on the value itself. In this way, we are reducing the entropy in each of the minimum coded units but we are introducing in the encoded image an entropy that wasn’t really there at the beginning (an intrinsic noise of the median filter that has been applied).

- By varying the threshold or the shape, dimensions and extent of the neighborhoods, we generate an encoding rule that is depending on the pixel position. Consequently, coding the encoding rule in a compressed standardized format becomes a challenge in itself.
- No matter how strong the visual effects are, median filtering on imposed uniformly shaped neighborhoods will always produce new edges placed on the each border of the filtered neighborhood. Surprisingly, we hope to preserve significant edges in the original image by introducing some new edges in the filtered image. This is also another noise, intrinsic to median filters, whose visibility is increasing with the regularity of the neighborhoods collection covering the image and also with the extent of these neighborhoods.

Of course, when speed is the main issue in processing, we don't bother about the fact that we will eventually see, in the downsampled image, things that weren't really there in the original image [9,10].

On the other hand, the simplicity level of the classical downsampling methods based on median filtering is difficult to achieve in other downsampling approaches. Consequently, a good negotiation between precision and simplicity becomes our target.

3. Principles of K-means Optimal Downsampling

This paper is an outcome of a few experimental studies that have been done by us following some new proposed principles of k-means optimal downsampling, principles suggested step by step by our practice and set out as follows:

- Any image is a self-described package of information, each chromatic layer being an array of pixels with random chromatic values for which the location is not important (when we are downsampling an image, in fact we are truncating the histogram to get more redundancy in order to reduce the entropy);
- During the downsampling process it is mandatory to maintain a functional correspondence between the chromatic values within original image and those within processed image. Doing otherwise is against the first principle from above and also, from a practical point of view, we implicitly add in the downsampled image a description of some facts that never existed in the original image.
- In the k-means context, the set of optimal reconstruction values (of a quantization at any given number of chromatic levels) and also the shape, dimensions and extent of the elements within the neighborhoods set covering the image are all intrinsic properties of the given image and consequently need not be imposed during the computation (we will let the image to talk about itself).
- Any robust efficient implementation of k-means algorithm designed to process at least a subclass of 8-bit/channel images must support uint8[1] acceleration over that subclass, or in other words, the normalized double representation of the processed

[1]Matlab naming convention for 8-bit unsigned integer data type

image must be a stability point of computational mechanism tolerating at least round-off errors perturbations, meaning that for a given image and for a given choice of initial set of centroids, if we do substitute the uint8 representation for the double normalized one, the algorithm must converges nearly to the same solution (to a nearly optimal solution).

These principles had enabled us to formulate the Fast K-Means Quantization algorithm, an efficient iterative procedure in which:

- The classified clusters are the optimal choice of neighborhoods set computed without imposing restrictions on their shape, dimension and extent;
- The centroids of the clusters are the *optimal* choice of reconstruction levels that *is minimizing the within-cluster sums of point-to-centroid distances*.

4. Generic K-Means Algorithm

The problem formulation had suggested from the beginning that k-means algorithm is the first candidate for answering our question formulated in the first section of this paper. A succinct presentation of k-means algorithm is the following:

```
Generic K-Means Procedure:
      INPUT: dataset to clusterize, desired number of clusters,
               eventually other custom data (such initial choice
      of centroids);
      While the termination condition isn't satisfied:
          For each element in the dataset:
              Find the closest centroid (ambiguity must be
          treated also);
              Mark the appartenence of the element to the
          corresponding cluster;
          EndFor;
          If none of the elements in the dataset changes its
          appartenence then
              Fulfil the termination condition;
          EndIf;
          For each of the computed clusters:
              Recompute cluster's centroid as being the mean of
              contained elements;
          EndFor;
      EndWhile;
END.
```

The results obtained by applying this algorithm on luma-chroma representation of 24-bit images normalized in double precision were very good, almost indistinguishable from the original images in many cases. The big problem is that the generic algorithm converges very slowly due the fact that very fine tuning of the values of the centroids is done in a large number of iterations through a large dataset. When we were working with normalized double-precision images of dimension 512x512, our first m-script implementation of the generic k-means algorithm has spent over 20 seconds to find 16 to 32 clusters in one chroma channel. Therefore, the next step was to move the computations in uint8 domain, as much as possible.

5. Fast K-Means Image Quantization Algorithm

The main algorithm presented in this paper is an adaptation of the generic k-means algorithm, tuned for fast chromatic clustering in Matlab. It takes all computational advantage of working with uint8 representation of 24-bit images:

Collapsing memory storage by avoiding chromatic values redundancy: as we can see in the generic k-means algorithm, there is no need to keep trace of the pixel position in the image. This enables us to exploit the big redundancy of chromatic values which will make the image's storage to collapse. In the k-means context, there are only two important data to retain from the each luma-chroma plane of the input image: what unique chromatic values it contains and how many times they appear. Consequently, the exact histogram of the given chromatic plane is the proper storage to work with (is the shortest container for the two required pieces of data). All the information we need from the original chromatic plane could be contained in a maximal 2x256 uint8 array, no matter how big the original image. Consequently, the initial storage is read only once prior to k-means iterations.

Collapsing memory storage by avoiding appartenence index redundancy: initially, the index memorizing the appartenence of each pixel to one specific cluster is an two dimensional array as wide as the original image is, but again, its values depends on the chromatic values of the corresponding pixels instead of depending on the pixels position in the image. Therefore, the initial appartenence index can also be efficiently stored in a maximal 2x256 uint8 array containing duplicate data of the first part of the exact histogram (maximum 256 unique values) and corresponding appartenence index of each chromatic value.

The algorithm also make the most of Matlab acceleration mechanisms such as vectorized [6] optimized loops and efficient handling of large spreaded array subsets (loops elimination) through the use of logical indexing [7]. At first sight, these optimizations seems to add unnecessary (or even unacceptable) particularization of the algorithm, but all of these mechanisms are replicable, even with better time performances, in other programming languages.

Taking into account that, latter in this paper, we will be interested in estimating the algorithm's performance, isolating the true computational kernel from input, output and collapsing operations is the proper way to formulate the algorithm:

```
function RCP = resample_kmeans_uint8(CP, NC):
  [EH, OC] = exact_histogram(CP);
  % deflates chroma plane in its exact histogram;
  [C, CI] = kernel_kmeans_uint8(EH, OC, NC);
  RCP = parse_output(CP, CI, EH, C);
  % inflates EH and CI to reconstruct chroma plane;
function [C, CI] = kernel_kmeans_uint8(EH, OC, NC);
  CI = uint8(zeros(size(EH'))); C = uint8([]);CC = uint8([1:NC]);
  T = false;
  while ~T
```

```
    C = get_new_centroids(C, EH,OC, CI, NC, CC);
    [T, CI, CC] = recompute_clusters(EH,OC, C, CI, T);
end;
```

where the meaning of each variable is as follows:

- RCP – reconstructed chroma plane array;
- CP – chroma plane array;
- NC – desired number of clusters;
- EH – array of values appearing in exact histogram of CP;
- OC - array of occurrences accounted for each of the values of EH;
- C – array of computed centroids;
- CI – collapsed cluster index;
- T – termination flag;
- CC – index of changed clusters

The main improvement against the generic k-means algorithm is the fact that, no matter how big the initial image, the computational kernel here is working with only a maximal set of 1024 values for each luma-chroma plane: maximum 3x256 uint8 values for EH, CI and C (but usually NC is small and consequently C is also shorter than EH), and maximum 256 unsigned integer values (uint32 for very large images such as 65536x65536 images) for OC. Therefore, the computational complexity of the computational kernel *does not depend on the size of the processed image*.

Taking into account that, in the reconstructed chroma plane, each chromatic value is replaced by the centroid of the cluster containing it, the fast k-means algorithm formulated above requantize the original chroma plane using the centroids of computed clusters as reconstruction values, as is shown in the following figure:

Being a quantization, k-means algorithm is implicitly an approximative reading operation on the original data, something that usually median filters aren't.

It is generally accepted [2] that in some ideal conditions (almost never satisfied by k-mean clustering mechanism) the optimal choice for the reconstruction values of a uniform scalar quantization is to place them in the middle of each decision interval. Our k-means requantization here certainly doesn't do that at all. The difference between the k-means computed centroids and those "ideally placed" is a measure of how ideal the real images aren't, and also for how un-natural are some ideal statistic criteria to human eye and to the image itself. A telling example can be seen in figure 2.

Another important aspect is that the Fast K-Means Image Quantization Algorithm tends to 'recognize' clusters that are also meaningful to human understanding of the picture. This is because the algorithm tends to ignore relatively constant chromatic areas in direct relation with their extent and with their position in their cluster: a relatively small area with relatively constant chromatic usually doesn't have an important contribution to the mean of the cluster containing it, especially when the values in that area are very different from the cluster's centroid.

Initial bandwidth of the signal:	First quantization at 256-levels during acquisition process:	Requantization through the fast k-means algorithm:

$$
\left.\begin{array}{lcl}
d_{256} & & \\
\vdots & \rightarrow & r_{256} = 255 \\
d_{255} & & \\
\vdots & \rightarrow & r_{255} = 254 \\
d_{254} & & \\
\vdots & & \vdots
\end{array}\right\} \rightarrow \begin{array}{c} C_{CN} \\ \vdots \end{array}
$$

$$
\begin{array}{c} \vdots \end{array}
$$

$$
\left.\begin{array}{lcl}
\vdots & \rightarrow & \vdots
\end{array}\right\} \rightarrow \begin{array}{c} C_k \\ \vdots \end{array}
$$

$$
\left.\begin{array}{lcl}
\vdots & & \vdots \\
d_2 & & \\
\vdots & \rightarrow & r_2 = 1 \\
d_1 & & \\
\vdots & \rightarrow & r_1 = 0 \\
d_0 & &
\end{array}\right\} \rightarrow \begin{array}{c} \vdots \\ C_1 \end{array}
$$

Figure 1 Image requantization through fast k-means algorithm.

All of the above observations and also the simplicity of the Fast K-Means Image Quantization Algorithm suggest that, at least in chromatic clustering, a good strategy is to assume that an image is a collection of pixels with no specific properties and to keep considering that until the moment when the image itself will tell us something else. Taking other way means to overwrite the initially available information within the original image hoping to obtain better results from ideal hypothesis than from approximative reading.

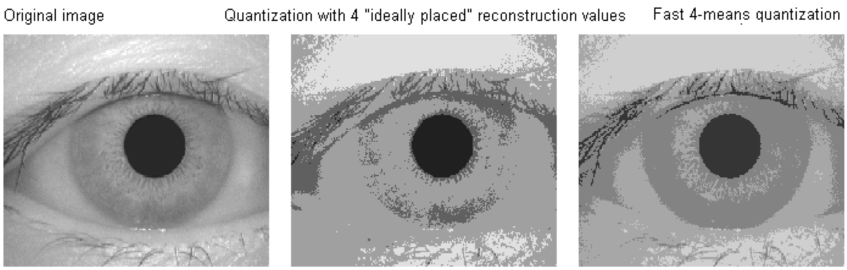


Figure 2 A 4-means quantization example (original from CASIA V1 iris database).

On the other hand, the distance from ideal case to real image is not very big suggesting that the number of k-means iterations will be smaller enough when idealized reconstruction values will be used as initial choice for centroids. But when timing is critical, even a uniform partition from minimum to maximum of the histogram is a good choice for the initial values of the centroids. Our practice shown that few additional k-means iterations run incomparably faster than fine tuning in search of some decision intervals on which to ensure a constant probability density of the chromatic values [2].

6. Application of the Fast K-Means Image Quantization Algorithm to the Luminance-Chrominance Downsampling

The Fast K-Means Image Quantization Algorithm was initially designed for studying the entropy of 24-bit RGB images while we were searching for any kind of transform that could decrease image's entropy and could preserve visual quality of the processed picture also, especially for such kind of transform that could homogeneously spread the error (the difference between the original and the reconstructed image) over an irregular set of neighborhoods instead of accumulating it near a rectangular grid (part of the blocking effect in DCT[2]-coded images).

Visual results of the Fast K-Means Image Quantization Algorithm are usually very good, as it can be seen in the figure 3.

Figure 3 Fast 16-Means Chroma Quantization of Lena image (256 luma levels, 16 chroma blue/red levels).

[2]Discrete Cosine Transform.

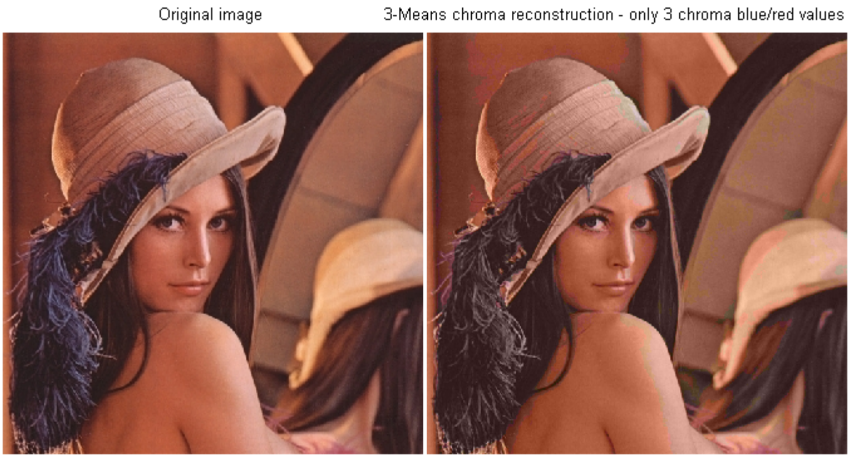


Figure 4 Fast 3-Means Chroma Quantization of Lena image (256 luma levels, 3 chroma blue/red levels).

Decreasing of the visual quality becomes evident for severe reduction of chroma levels number (figure 4) but, as expected, the visual quality is more sensitive to luma requantization (figure 5).

Figure 5 Fast 16-Means Luma-Chroma Quantization of Lena image (16 luma levels, 16 chroma blue/red levels).

All the benchmark tests that we have done so far includes many usually standard RGB-to-YCC and YCC-to-RGB conversions such as HDTV-709, PC-709, SDTV-601, PC-601, VIDEO-601, JPEG-601, PALTV, SECAM, Matlab implicit conversion, and other custom experimental conversions. A comprehensive m-script including the most of them will be available on the internet [8] as soon as possible, in uint8 version.

There is a sufficiently large class of images which are indistinguishable from their 256/16/16 luma-chroma reconstructions. For this kind of images, all the information in the chroma blue/red channels can be stored in one single channel (8 bit/pixel without chroma downsizing). Therefore, a first possible application is to use one 8-bit channel to store something else such an infrared image of the same subject, or the scaled response of the surroundings to a radar signal, or prediction values for the next frame, or any other data that can be fitted in the uint8 domain.

7. Application of the Fast K-Means Quantization Algorithm to the Iris Segmentation

The most important application of the Fast K-Means Image Quantization Algorithm that we have been done so far is the iris segmentation.

This section of the present paper uses CASIA V1-3 iris databases [12]. We express our gratitude to Chinese Academy of Sciences, Institute of Automation, for giving us permission to work with these databases. There are also few images from the old CASIA V1 database (in which the pupil isn't filtered) that we are working on.

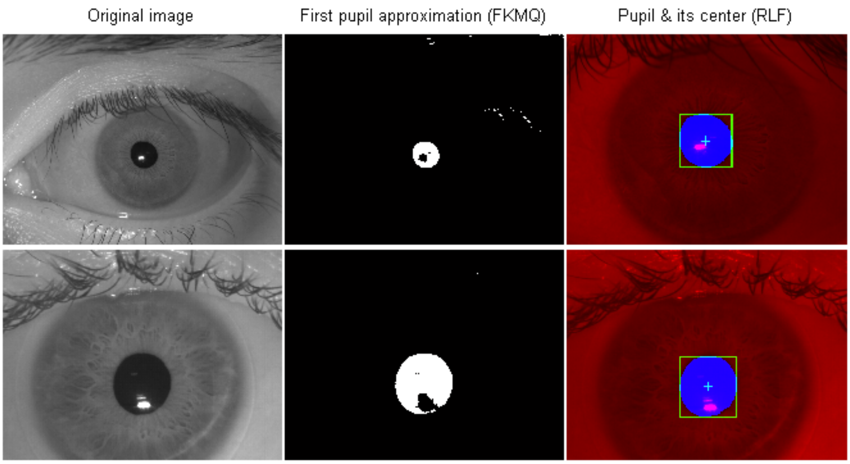


Figure 6 Fast 16-Means Iris Segmentation: finding the pupil and its center through Fast-K-Means Quantization and Run-Length Filtering (two images from old CASIA V1).

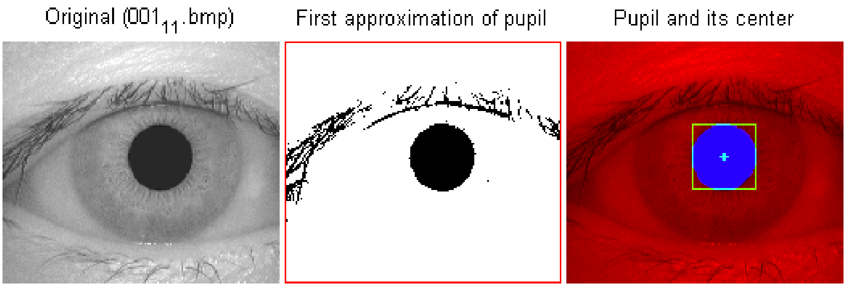


Figure 7 Fast 3-Means Iris Segmentation: finding the pupil and its center through Fast K-Means Quantization and Run-Length Filtering (image from CASIA V1).

We consider the generic process of iris recognition [4] as being divided into the following parts:

- acquire the image;
- crop the acquired image to significant area;
- locate the pupil and its center;
- do the iris segmentation;
- detect and reject cases of visual exposure inconsistency;
- compute the Key Features Array from extracted iris area;
- compare the Key Features Array against those stored in some ID Keys Database;

This paper investigates especially localization of pupil and its center, and also the next two steps.

Initially the pupil localization has been done for some images from old CASIA V1 database (like those in the figure 6) and this has been proved to be simple enough to achieve using the proposed Fast K-means Quantization Algorithm which converged in a very few iterations to a very good approximation of the pupil area (the target signal) slightly perturbed only by some specular light and by some eyelashes points heaving chromatic values among the values of the pupil cluster. Here both specular light and eyelashes have the meaning of a noise. The signal to noise ratio being sufficiently good there was no problem to filter the unwanted noise. We have used Run-Length Encoding to filter the noise, deleting pixels around the pupil (erosion) and dilating the remaining cluster to fit its interior (or to close its interior when the specular light perturb the pupil's border, or both of the cases). Such a filter can be easily derived from any usual Run-Length Encoding procedure.

In the context of finding the pupil location through k-means algorithm, the present CASIA V1 database has been proven to be more challenging. This is because of the fact that the images are pre-processed, meaning that the pupil is filtered and the specular light on the pupil isn't present. There are two kind of data losses involved in this operation: the specular light can't be further used as an estimation of pupil location (whenever it exists, the specular light on the pupil is the easier area to locate in the entire image), and the second one - and the more important thing, the chromatic values in the pupil area becomes closer to those of the eyelashes. Therefore, the first approximation of the pupil computed through Fast K-Means Quantization Algorithm becomes noisier (figure 7) and all the segmentation procedure had to be recalibrated. All in all, the present V1 database has been proven to be more demanding when it came to test robustness of our segmentation algorithm. For all of these reasons from above, we have chosen the present CASIA V1 database to work with in this paper.

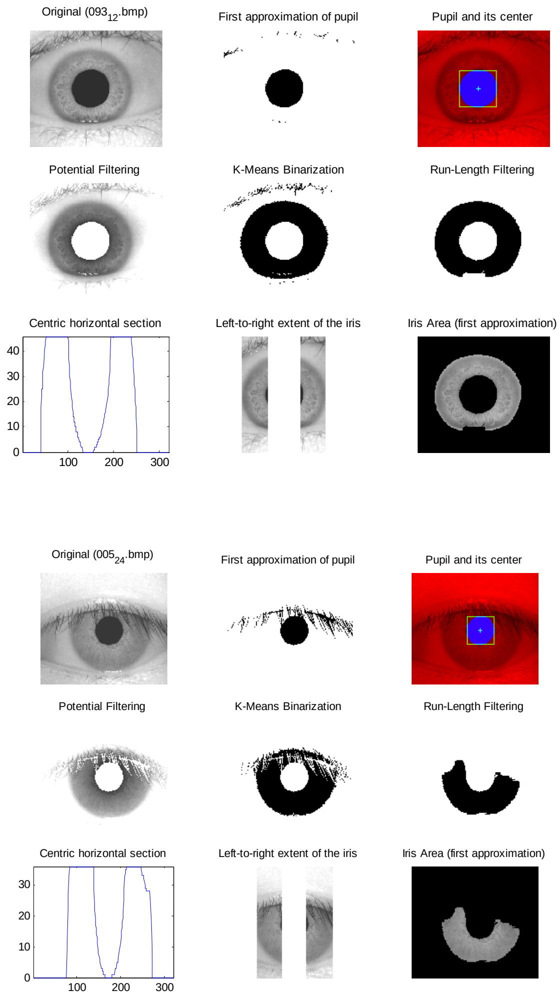


Figure 8 Iris Segmentation: an ideal example / a noisy example.

Our iris segmentation algorithm is the following:

```
Generic Iris Segmentation Procedure;
1. INPUT: current iris image and eventually other custom
     data (calibration variables);
2. Run the Fast 4-Means Quantization adaptively to the first
     cluster consistency to obtain the first approximation
     of pupil;
3. Use Run-Length Filtering to denoise (eliminate
     surrounding noise and fill the gaps in the pupil, if
     any);
4. Identify the pupil with all remaining pixels in the
     pupil’s cluster;
5. Compute the center of the pupil and the pupil radius;
6. Filter entire image with a central potential field of
     energy originating in the center of the pupil (the
     iris will become darker and the ‘non-iris’ whiter).
7. Extract the entire cluster obtained in (2) from the
     filtered image obtained in (6);
8. Binarize the result of (7) through Fast 2-Means
     Quantization;
9. Use Run-Length Filtering to denoise (if necessary) the
     result of (8);
10. If the result from (9) satisfy some consistency criteria
     then go to (11), else go to (1);
11. OUTPUT: result of (9) is a binary index of a noisy iris
     segment;
```

The algorithm obtains very good results but as a whole is still experimental. The pupil localization is fully calibrated for CASIA V1 database and its success rate is 100% on this database. It went wrong (3-pixel error for the pupil center) on a single case which has also being rejected for severe inconsistency of the iris segment (sleepy eye, very noisy eyelashes, pupil severely obstructed by the superior eyelid, all at once).

Few examples of iris segmentation results can be seen in the figure 8.

8. Fast K-Means Image Quantization – an Algorithm for Image Processing Supercomputing

This section summarizes some of the most important properties of the main proposed algorithm, Fast K-Means Image Quantization:

First of all, the computational kernel of the algorithm is independent from image size, making the algorithm suitable for very large images.

The algorithm can be used successfully as a component in other applications such as segmentation algorithms.

At the last but not the least we will show that the computational kernel of the proposed algorithm truly does have an incredible speed. The next test has been done over a database of 24-bit RGB usual images containing 84 pieces (from landscape to macro pictures) of dimension 512x512, and is designed to illustrate the time performance of the proposed algorithm. During the test, algorithm (re)quantizes each chroma blue/red plane at the given numbers of chromatic levels mentioned in the Table 1.

	Fast 128-Means Chroma Quantization			Fast 64-Means Chroma Quantization			Fast 32-Means Chroma Quantization		
Function Name	Time(sec.)	Calls	Time/call (sec)	Time(sec.)	Calls	Time/call (sec)	Time(sec.)	Calls	Time/call (sec)
resample_kmeans_uint8	73.90500000	84	0.87982142857141	54.43800000	84	0.64807142857141	36.85300000	84	0.4387261904762
parse_output	59.71900000	168	0.35547023809520	40.11000000	168	0.23875000000028	22.18300000	84	0.2946547619050
rgb2ycc_uint8	24.51900000	84	0.29189285714286	24.48200000	84	0.29145238095260	24.75100000	84	0.2810238095239
ycc2rgb_uint8	23.40900000	84	0.27867857142878	23.34200000	84	0.27788095238070	23.60600000	168	0.1320416666667
exact_histogram	11.05200000	168	0.06578571428565	11.20100000	168	0.06667261904736	11.22000000	168	0.0667857142857
imread	4.41800000	84	0.05259523809504	4.47200000	84	0.05323809523816	4.56900000	84	0.0543928571425
kernel_kmeans_uint8	2.31300000	168	0.01376785714285	2.30000000	168	0.01369047619049	2.75400000	168	0.0163928571428
recompute_clusters	1.68600000	336	0.00501785714287	1.76800000	627	0.00281977671450	2.17300000	1109	0.0019594229035
get_new_centroids	0.56300000	336	0.00167559523808	0.45300000	627	0.00072248803830	0.53400000	1109	0.0004815148783

	Fast 16-Means Chroma Quantization			Fast 8-Means Chroma Quantization			Fast 4-Means Chroma Quantization		
Function Name	Time(sec.)	Calls	Time/call (sec)	Time(sec.)	Calls	Time/call (sec)	Time(sec.)	Calls	Time/call (sec)
resample_kmeans_uint8	27.41000000	84	0.32630952380970	22.47100000	84	0.26751190476183	17.59400000	84	0.2094523809525
parse_output	11.86300000	168	0.07061309523817	6.55200000	168	0.03900000000026	3.32000000	168	0.0197619047620
rgb2ycc_uint8	24.60800000	84	0.29295238095217	25.57800000	84	0.30450000000023	25.23400000	84	0.3004047619050
ycc2rgb_uint8	23.38700000	84	0.27841666666652	24.31400000	84	0.28945238095254	23.91900000	84	0.2847499999998
exact_histogram	11.13500000	168	0.06627976190464	11.29500000	168	0.06723214285696	11.40900000	168	0.0679107142856
imread	4.70500000	84	0.05601190476210	4.49900000	84	0.05355952380917	4.72200000	84	0.0562142857140
kernel_kmeans_uint8	3.71200000	168	0.02209523809532	3.91900000	168	0.02332738095226	2.07800000	168	0.0123690476191

Table 1 Time performance of the Fast K-Means Image Quantization Algorithm (Processor: P IV Prescott 2.8 GHz

In the mentioned table we are tracing the trend of the execution times, especially for the computational kernel of the algorithm (kernel_kmeans_uint8), for I/O functions (exact_histogram and parse_output) , and also for 3 'witness' functions: the imread Matlab implicit function, and the RGB-YCC conversion functions (see section 5 - Fast K-Means Image Quantization Algorithm).

It can be seen in the Table 1 that time complexity of the computational kernel is bounded in a narrow band, far away from any polynomial or superpolynomial time [5].

References

1 International Telecommunication Union (ITU), The International Telegraph and Telephone Consultative Committee (CCITT), International Standard ISO/IEC, 10918-1: 1993(E), CCITT Recommendation T.81, Digital compression and coding of continuous still images – Requirements and Guidelines, Section 3 (Definitions, abbreviations and symbols), 1993.

2 William K. Pratt, Digital Image Processing, Chapter 6 - Image Quantization, John Wiley & Sons, 2001, third edition.

3 Weisi Lin, Li Dong, Adaptive Downsampling to Improve Image Compression at Low Bit Rates, IEEE Transactions on Image Processing, vol. 15, no. 9, September 2006 2513.

4 Li Ma, Tieniu Tan, Yunhong Wang, Dexin Zhang, Efficient Iris Recognition by Characterizing Key Local Variations, IEEE Transactions on image processing, vol. 13, no. 6, June 2004, 739.

5 David Arthur, Sergei Vassilvitskii, How Slow is the k-Means Method?, Stanford University, http://www.cs.duke.edu/courses/spring07/cps296.2/papers/kMeans-socg.pdf

6 The MathWorks Inc. http://www.mathworks.com/support/tech-notes/1100/1109.html

7 Steve Eddins, the MathWorks Inc., http://blogs.mathworks.com/steve/2008/01/28/logical-indexing/

8 http://www.mathworks.com/matlabcentral/fileexchange/

9 http://glennchan.info/articles/technical/chroma/chroma1.htm, figures 6-10,

10 http://www.impulseadventure.com/photo/chroma-subsampling.html

11 http://en.wikipedia.org/wiki/YUV_4:4:4#Types_of_subsampling

12 http://www.sinobiometrics.com/

Third Party Copyrights

Designing e-Testing Systems in Service Oriented Architecture

Goce Armenski[1], Marjan Gusev[2]

Institute of Informatics, Faculty of Natural Sciences and Mathematics, University "Ss Cyril and Methodius", Arhimedova bb, PO Box 162, Skopje, R. Macedonia, armenski@ii.edu.mk[1], marjan@on.net.mk[2]

Assessment is an important integral part of every learning process. The nature of the assessment process has deep impact on the ways students learn, defining the contents they assume important and the ways they organize their learning process [1]. The extensive use of technology in learning and working, is forcing its use in the assessment process. A lot of software packages exist in the market to realize automated assessment. Several of them are very comprehensive, but most of them are stand alone applications without possibilities for interoperability, adaptability according to learner characteristics and possibilities for content reuse.

In this paper we describe the purposes and the process of designing interoperable E-Testing Framework by remodelling an existing eTesting system and introducing new structured Service Oriented Architecture, based on encapsulating existing business functions as loosely coupled, reusable, platform-independent services which collectively realize required business objective. This common framework should provide greater interoperability between different systems, greater flexibility, access and sharing of common information, greater pedagogical flexibility and adaptability towards the learner characteristics.

Keywords

eAssessment, eLearning, eTesting, SOA, interoperability.

1. Introduction

Knowledge has always been a main driver for economic growth and social development. The ability to innovate and create new knowledge has always been a main tool for creating well-being.

In the past several decades there has been fundamental reshaping of the global economy influenced by the advances of modern information and communication technologies. In an increasingly global economy where the capacity to use information in the right time and on right place gives advances on the market, knowledge has become the key resource. Knowledge has value, and creating value is about creating new knowledge and capturing its value.

The emerging knowledge society signifies a new era for education and training. Knowledge and skills of citizens are becoming increasingly important for the quality of citizens' life. Workers in the 21st century knowledge society will need to be lifelong learners, adapting continuously to changed opportunities, work practices, business models and forms of economic and social organisation.

The characteristics of this society, brings new challenges for higher education. Higher education institutions have main role in the process of redefining the models for acquiring knowledge and skills. Technology is more often used in learning as a tool for lectures, delivery of materials, and assessment of student knowledge.

In the past several years lots of reviews and analyses, predicted that eLearning will drastically change the way people acquire knowledge, giving optimistic numbers about its acceptance. Still, nowadays there are lots of reports that eLearning failed to fulfil these expectations. Variety of reasons are discussed and analyzed by experts. They all agree that the institutional and pedagogical reservations from one side and technical issues such as interoperability and security on the other are main barriers for broader uptake of eLearning systems [2].

Pedagogical considerations and business processes to facilitate learning, however defined, are of paramount concern in developing e-learning infrastructure [3]. Student-centered learning and constructivist approaches, are just some of the paradigms which emerged, and are being supported by technological advances.

2. eAssessment

Assessment takes central place in the learning process. The assessment nature have deep impact on the way people learn, defining the contents they will assume as important and defining the way they will spend their time [1].

Assessment is a process in which examples of person's attitude are taken at particular time and they are evaluated. According to the evaluation of these examples, conclusions are made for the person's achievement, potential, intelligence, attitude or motivation. Different forms of assessment exist and each of them has different use. Besides the traditional summative and formative assessment, in the past several years newer types of assessment are becoming more popular, such as competence assessment, performance assessment, portfolio assessment and peer assessment. Compared to the traditional ones, they are more integrated and embedded in the learning context which requires higher level of student involvement in the assessment process. These types of assessments try to give an adequate answer to the ideas of a learning process where teaching, learning and assessment interact.

The broadest term which is used in literature when discussing assessment automation is computer assisted assessment. This term cover any use of computers in the process of assessing knowledge, skills and abilities of individuals [4].

There are several systems for automatic assessment on the market, mainly as part of distance learning systems. However, there are independent software packages for computer based assessment, web based assessment or electronic assessment. Many of these systems are very comprehensive but most of them are stand alone applications without possibilities for interoperability, adaptability according to learner characteristics and possibilities for content reuse [5] [2] [6].

The system for electronic testing at the University “Ss Cyril and Methodious” - eTest, is a result of continuous development of concepts and software, which are used for conducting frequent assessments, on which more than 500 students take part. The original idea was to create a system that can help realization of exams for cases where number of students is very big (several hundreds on each exam), and in cases when each student is allowed to apply for an exam each month. Afterwards, this idea was expanded to realize an independent system of testing with a lot of intelligence, applicable both for conventional and distance learning.

3. eTest Architecture

3.1 System Architecture

The system is realized as WEB application with three layer architecture. Main reason to realize this system not as classic client server application is the characteristic of Web applications to be installed on one computer and used on any computer on Internet through common web browser. This is very convenient for students since they are not obliged to come at university to use the system, and can efficiently use e-business paradigm anytime, anywhere.

The three layered architecture of the system is shown on Fig.1, realized by separate database layer, application layer for basic system modules and user interface layer.

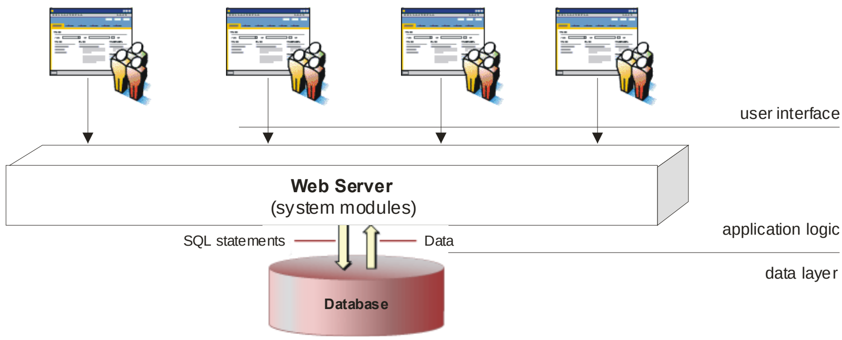


Figure 1 Three layered architecture of the system

3.2 Course Organization

The basic structure of the system consists of courses which material is divided in lectures. A tree like organization of lectures is implemented. Each lecture consists of smaller parts and each part consists of different sets and finally of learning objectives, as shown in Fig.2 [7]. The course material in the lesson is divided in at least three parts (in Fig.2 marked as A, B and C).

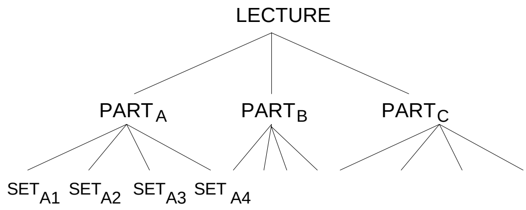


Figure 2 Tree like organization of the lectures

The system for eTesting can have unlimited number of courses. The courses are independent from each other and have their own structure and question bank.

3.3 Test Creation Algorithm

The electronic testing system realized as computer-based testing is realized in such a way that each test generated will measures verbal, quantitative and analytical skills related to a specific field of course study. A different time constraint and score mark is associated to each area. The area consists of a set of questions defining one concept or one knowledge skill. We differ three classes of questions: verbal, quantative and analytical questions, similar to the description in [7].

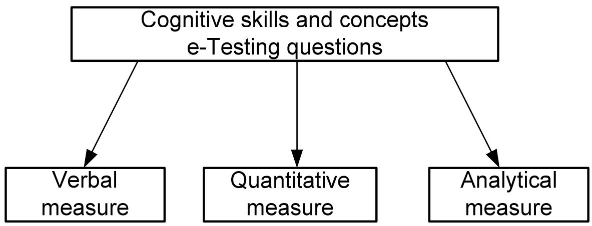


Figure 3 Knowledge based cognitive skills and concepts can implement e-Testing models by three types of questions with: verbal, quantitative and analytical measure

The system for eTesting has implemented the following types of questions:

- single choice questions;
- multichoice questions;
- short answer questions.

Test creation algorithm is closely connected to the chosen method for test delivery. The idea for creating different tests for every student, forced us to apply the model for dynamic test creation. With that idea every student will get different test, with same weight like all other student. These dynamically created tests will have fixed number of questions because this was first time system for automated assessment to be applied at the University. The applied model gives opportunity for students to list the questions one by one, and answer only those which answer they know.

3.4 Marking and Results Reporting

The system for eTesting implements evaluation of the answers at the end of the test, at the moment when the person which knowledge is tested specifies that the entered answers are definite. Because the most questions used in the system are fixed response questions, they are easy to evaluate when final results are displayed. The system displays the final results with an option to see the right answers compared to those entered by the user.

In order to eliminate guessing, we have implemented negative marking. The final score of every test can be displayed on two ways, either using points or percents.

4. Findings

The system is actively in use since 2001 at the University Ss Cyril and Methodius. Until the end of 2007, a series of 789 assessments have been realized with 12861 tests generated. The question database has 14391 questions from 29 courses. The effects of using the system were analyzed in a separate study, both from a teacher and student perspective [7]. In 2005 the system is installed on 3 other faculties from 2 different Universities where it is used also for assessing student knowledge. The creation of 4 different environments opened new question to the research and development team.

Although the system is very comprehensive and covers many aspects in assessment of student knowledge, its current architecture does not allow the system to be compliant with the requirements that any complete conceptual model should in the long run comply to [8]:

- Flexibility: The assessment model can describe assessments that are based on different theories and models:
- Formalization: The assessment model describes assessments and its processes in such a formal way that it is machine-readable and automatic processing is possible. The formalization gives the possibility to extend the model if new developments in assessment arise.
- Reusability: The assessment model supports identification, isolation, decontextualization and exchange of useful objects (e.g. items, assessment units, competencies, assessment plans) and their re-use in other contexts.
- Interoperability and sustainability: The assessment model distinguishes the description standards from the interpretation techniques, thus making the model resistant to technical changes and conversion problems.
- Completeness: The assessment model covers the whole assessment process, including all the typed objects, the relations between the objects and the workflow.

- Explicitly typed objects: The assessment model expresses the semantic meaning of different objects within the context of an assessment.
- Reproducibility: The assessment model describes assessments in such a way that replicated execution is possible.
- Medium neutrality: The educational model for assessment, where possible, supports the use of different media, in different (publication) formats, such as computerized assessments on the web or paper and pencil tests.
- Compatibility: The assessment model matches available standards and specifications.

The results from the analysis of the system towards the above mentioned requirements are presented in Table 1. The score gives a mark indicating the compliance of the system with the requirement in a scale from 1 to 10.

Regarding the flexibility, in theory of learning there is a distinction between three major streams of instructional theories [9]: empiricist (behaviourist), rationalist (cognitivist and constructivist), pragmatist-sociohistoric (situation list). The architecture of the system and its current implementation is compliant only with the behaviourist theory, since the system has predefined learning path through the course material.

The latest developments of the system which tried to incorporate IMS QTI specification, made a significant step towards the support of the requirements of formalization. Still, at its current status, the system is not completely machine-readable.

The complexity of the system and the organization of the courses in tree like structure, separating the learning objects in different nodes of the tree, where each node represent specific learning objective, gives high level of reusability of the assessment content. The learning objects can be reused through the courses and combined in different assessment. The only problem we have faced with, regarding the reusability was the inability to reuse learning object through different installation of the software in different institutions.

Requirement	Score	Requirement	Score
Flexibility	3,33	Explicitly typed objects	2.00
Formalization	2.50	Reproducibility	10.00
Reusability	7.50	Medium neutrality	10.00
Interoperability and sustainability	2.00	Compatibility	5.00
Completeness	10.00		

Table 1 eTesting System compliance with the defined requirements

The current system does provide separation of the content from the interpretation techniques by implementation of question banks, but does not provide complete interoperability between systems installed in different locations. It does not allow searching, using or modifying interoperable questions, neither creation of joint courses where student from different universities can participate. Cross-institutional cooperation by sharing information is a need which arises, because many courses are beginning to be taught collaboratively realizing the concepts of student mobility and lifelong learning.

Analyses on completeness gave good results, since the system itself is complete assessment solution which covers all parts of the assessment process as described in the Chapter 3 above.

Besides the structural organization of the learning units and objects in tree like structure where the learning objects are organizes in meaningful way, the system does not provide any additional semantic meaning of the objects, neither their meaning in the assessment itself. This brings additional problems in the interoperability processes where it is not enough to just transfer assessment content from one system to another but it is highly required that that content has meaning. There is need to extent the description of the learning object and to include additional semantic meaning to them.

Regarding reproducibility and medium neutrality, the system architecture and its three tier architecture gives good possibilities for its use in different platforms, as well as on different devices. Every action in the system is well documented and it is very easy to reproduce the user actions as well as assessments.

The leading specification for the exchange and interoperability of assessments is the Question and Test Interoperability. Our system is still not fully compatible with this specification, but in the past several months tools are being developed which will extend the system and test its compatibility towards this specification.

From the above analyses, several open questions were identified. As can be seen in Table 1, the lowest scores are in interoperability, flexibility, reusability and explicit typing of the objects.

Analysing the current trends in the e-Learning and e-Assessment domain, where lots of emphasis is given to the level of interoperability between systems, exchange of content and data between systems not depending on the platform they work on, by using widely adopted standards, increasing system flexibility and pedagogical diversities supported, I propose remodelling of the existing system with introduction of new Service Oriented Architecture, based on encapsulating existing business functions as loosely coupled, reusable, platform-independent services which collectively realize required business objective.

5. Modern e-Learning system architectures

Although technology has the potential to extend and improve educational and training activities, opposite results can be achieved “not because it (technology) wasn’t effective, but because it … did not adapt to the way people wanted to learn.“ [7]. The potential of the technology can only be fully realized if the activities are built upon a stable and coherent technical infrastructure, and with existence of appropriate widely accepted standards.

The vast majority of the currently used web-based educational systems are powerful integrated systems, like Blackboard [10] or WebCT [11]. Those Institution specific learning systems, unable for cross-institutional sharing of information are the most common problems with which Higher Education Institutions (HEI) are facing with. Those systems are usually vendor specific, lack interoperability and pedagogical flexibility. They do not provide the flexibility a learner needs.

Dagger [12] analyses the evolution of e-learning platforms. Three generation of e-learning platforms are identified.

The first generation consists of mainly monolithic applications which act like black-box solutions. In terms of e-learning evolution, they provided a shift toward modular architectural designs and recognized a need for semantic exchange. In the second generation a separation of content from tools is identified, and the learner information became more distinguished.

The next (third) generation should no longer be monolithic, one-size fits-all solutions, but rather interoperable platforms with a range of e-learning services, letting consumers choose the right combination of services for their requirements. They should offer complete federated exchange among services (information and control), various levels of interoperability (intradomain and interdomain), and service composition (orchestration and choreography).

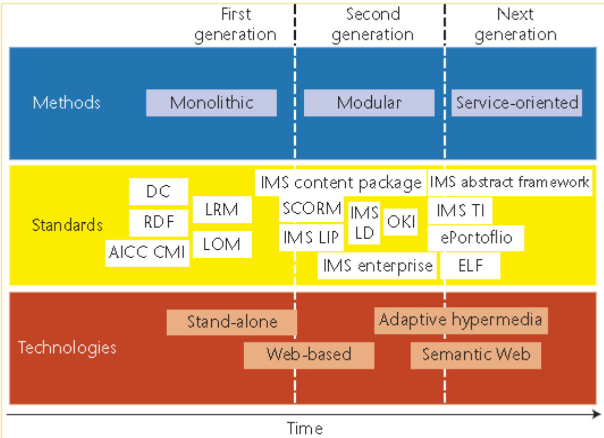


Figure 4 Evolution of e-Learning platforms[12]

Future framework of Learning Management Systems (LMS) will allow the exchange of the learner profile and learning resources with other legacy systems over the Internet.

However, recent standardization efforts in e-learning concentrate on the reuse of learning material, but not on the reuse of application functionalities.

In the past few years, world leading organizations in the e-learning community were focused on creating a joint vision for common technical framework in e-learning area, and in defining international learning technology standards and specifications, in order to allow systems to "support organisational and cross-organisational processes for enabling effective e-learning" [13]. These standards and specifications are supposed to promote interoperability, flexibility and pedagogic diversity in the e-learning process.

As a result of those activities few detailed frameworks were developed. Some of the most successful and comprehensive are:

- JISC e-Learning Technical Framework (ELF) [14]
- IMS Abstract Framework (IAF) [15]
- Open Knowledge Initiative (O.K.I.) [16]
- LeAPP Learning Architecture Project [17]

One common structural issue for which these organizations reached a consensus was the adoption of Service Oriented Architecture (SOA).

The potential of service-oriented software architectures had been recognized previously. Back at a history of distributed communication standards such as DCOM, CORBA or RPC, service-orientation is not a new architectural pattern in itself [18].

Still, the literature shows several differences between COA and SOAs. Component-oriented architectures (COAs), are more finely grained and tightly coupled than SOAs. Changes to individual components typically impact the software those components access, making COAs less flexible and extensible than SOAs.

6. SOA

Service-Oriented Computing is shift from a vision of a web based on the presentation of information to a vision of the web as computational infrastructure, where systems and services can interact in order to fulfil users' requests programmatic view. [19]

The Service-Oriented Computing (SOC) paradigm refers to the set of concepts, principles, and methods that represent computing in Service-Oriented Architecture (SOA) in which software applications are constructed based on independent component services with standard interfaces.

European initiatives such as i2010: European Information Society 2010, supports the implementation of Service Oriented Computing. Also, all of DoD’s major IT initiatives in the past years are based on the SOC paradigm, including the Army’s FCS, the Navy’s FORCEnet, the Air Force’s JBI, and the OSD’s NCES and GIG-ES. [20]

Although there are lots of definitions for SOA we will define it as “An approach for building distributed computing systems based on encapsulating business functions as services that can be easily accessed in a loosely coupled fashion.” [21]. SOA has many advantages, like reusability and flexibility of implementation, higher compatibility with the Grid, “lower overall costs, protection of legacy investment, lower cost of entry, rapid development, potential for business processes to drive technology” [3].

From an institutional point of view it enables collaboration between institutions, faster deployment of new functionality, and support for pedagogic diversity, and avoids lock in to single vendor solutions with the possible attendant costs. From a technical point of view the open interfaces of the components make it relatively simple to connect components in novel and custom ways, encourage interoperability, and facilitate replacing one service with another to provide the same functionality in different ways.

In [22], Willson discusses the pedagogical aspects of SOA e-learning system analyzing 6 pedagogical choices in e-learning, and concludes that "'Brave New World' of web-service driven environments" offers much greater pedagogical diversity than the monolithic systems.

The comparison of abovementioned frameworks shows that they all have layered architecture consisting of a set of services which can be used in e-learning context and collectively realize required business objective.

A Service Orientated Architecture (SOA) is capable of facilitating rapid development of highly customizable systems that can be optimized towards a specific goal or pedagogical requirement. This framework also make it easy to plug in extra components or combine services in novel ways to evaluate their effectiveness.

7. SOA in eAssessment

Although Assessment is present as one of the main services in all mentioned frameworks, JISC [14] as organization developing the E-Learning Framework (ELF), has made significant steps forward in definition of the Assessment domain.

Following its strategy for creation of Reference Models for number of domains and identification of sub services in each domain, identified as one of the 5 prioritized domains in ELF, Assessment is extensively a subject of research in the past few years. Numbers of projects have been funded [23], among which FREMA (Framework Reference Model for Assessment) is the most comprehensive and is concerned with the definition of assessment domain.

The project gave definition of the domain creating "map of resource types that are considered important within the assessment domain, and … concept map of the common processes" [24], identified common usage patterns, developed use cases and defined Web Services in the domain. As a result of the project, these Core Services where identified: Assign, Author item, Author assessment, Validate assessment, Take assessment, Mark assessment, Moderate assessment, Grade Assessment and also 4 Supporting Services: Schedule, Notify, Track, Authorise and Authenticate. [25]

Besides these frameworks which intention is to define the e-learning domain at whole, another project whose main objective is "provide a technical and organizational Infrastructure that can be used by any citizen, team or organization to develop competences" [26] have identified assessment as a main tool for achieving its goal and have intention to develop new assessment model. Analyzing this model we have concluded that the model they have developed consists of these services: Assessment Design, Item Construction, Assessment Construction, Assessment Delivery, Response Evaluation, and Decision Making. During the development of this model no supporting services where identified. [27].

8. Modelling a Common Framework

The lack of standardized and widely adopted model in the e-Assessment area, as well as valuable experience from practical realization and implementation of the existing models or parts of them, results with little experience about the real use of service oriented architecture in the design of e-Learning and e-Assessment systems and the influence it has on the assessment and learning process. Because of that, research on standards and development work is underway in order to see what will be results from the implementation of the proposed models. A number of projects based on web services were funded in this domain in order to develop services or set of services.

Comprehensive overview of assessment projects is given in [28]. Most of them give practical realization of particular service identified by the FREMA, and propose extensions to (or verify) already existing standards. Some projects are more comprehensive, demonstrating the use of multiple services in SOA (ASSIS [29]).

Concentrating on remodelling of the existing eTest system architecture, we have identified Web Services which will collectively realize required business objective of our system and make it compliant with the requirements named in Chapter 4: Item Construction, Test Construction, Test Delivery, Results Collecting, Marking, Decision Making, and Statistical Analysis. Our supporting services are: Schedule, Notify and Announce, Authentication, Track, and User Management.

With the design and practical realization of common widely adopted model, answers to these questions can be given:

Interoperability between different platforms – the concept of service oriented architecture where applications consist of flexible loosely coupled functional components which provide services accessible to the application through standards based interfaces, theoretically should provide better interoperability between existing systems. By implementing tools and modules of the redesigned system we will gather practical results concerning the interoperability between SOA based systems compared with the monolithic systems.

More, we will analyse the possibility to make the assessment systems transparent to the learning environments, meaning that any well designed assessment system can be used within any learning environment without advanced technical knowledge and with be transparency from the user perspective. The design of common model should result with highly interoperable systems which will be able to exchange data not depending on the platform they work on, or on technology they are build with.

Content reuse – centralized question banks, as well as centralized content silos which can be accessible using standardized way, through any e-Assessment system, are challenge to the technology. The redesigned model using SOA should make these question banks accessible, searchable, with easy possibility to retrieve needed items.

Pedagogical diversity – creation of system which supports different dynamical models for test creation and delivery in the assessment and learning process will increase the system adaptability towards the characteristics of the student, at the same time providing higher pedagogical diversity.

Analysing the existing projects objectives and their goals it is noticeable that very few of them in the assessment area had analyzed and discussed the problem of test construction and delivery algorithms and the influence it has on the assessment and learning process.

By updating and fulfilling the existing SOA based models for e-Assessment in the area of test construction and delivery, providing and using different pedagogical models the personalized assessment will be achieved.

According to [30] there are different kinds of test delivery models:

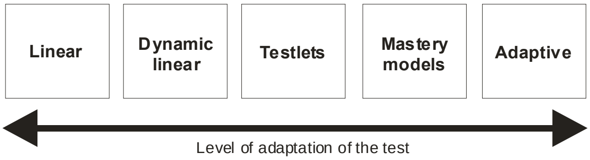


Figure 5 Models for test delivery [30]

In order to support wide pedagogical diversities, any assessment system should be able to provide all models of test delivery. The models can be implemented consequently in the process of learning where student knowledge can be assessed before some content is presented to him. By using different models for test delivery in the learning process we can simulate the world of interactive games, expecting increase of student motivation during the learning process which should result with better results compared to the traditional forms of learning.

9. Conclusions

In this paper we have identified the problems and reasons for redesigning of the existing system for e-Testing. We have shown how our system can be remodelled and upgraded by introducing new structured service oriented architecture.

By identifying and implementing unique functionalities, we expect to update and fulfil the existing SOA frameworks and models, creating a common framework which will provide greater interoperability, exchange of data and content, and greater pedagogical diversity.

In its practical realisation we will use the experiences and results from already developed projects in the assessment area [28], and by researching the possibilities for using different models for test delivery depending on context specifics, we expect to contribute in improvement of the diversity and quality of the learning process.

References

1 Couper, M.P. Web Surveys: the Questionnaire Design Challenge, Proceedings of the 53rd session of the ISI, 2001 at http://134.75.100.178/isi20

2 Davies, W. M. and Davis, H. C. Designing Assessment Tools in a Service Oriented Architecture. In Proceedings of 1st International ELeGI Conference on Advanced Technology for Enhanced Learning BCS Electronic Workshops in Computing (eWiC), 2005, Vico Equense, (Napoli), Italy
3 Blinco K., Mason J., McLean N., Wilson S. Trends and issues in e-learning infrastructure development, A White Paper for alt-i-lab 2004 Prepared on behalf of DEST (Australia) and JISC-CETIS (UK).
4 CAA Centre website at URL http://caacentre.ac.uk/
5 Cristea P, Tuduce R, Automatic Generation of Exercises for Self-testing in Adaptive E-Learning Systems: Exercises on AC Circuits http://wwwis.win.tue.nl/~acristea/AAAEH05/papers/4AIED2K5_W9_PC_RT_1.pdf
6 Sclater, N., Low, B., & Barr, N. Interoperability with CAA: does it work in practice? Paper presented at the Proceedings of the Sixth International Computer Assisted Assessment Conference, Loughborough University, 2002
7 Armenski G, Gusev M, "eTesting Infrastructure", FACTA UNIVERSITATIS (NIS), Series: Electronics and Energetics, Volume 18, Issue No. 2, 181-204p, August 2005. http://factaee.elfak.ni.ac.yu/fu2k52/contents.html
8 Koper, E. J. R. Modelling Units of Study from a Pedagogical Perspective: the pedagogical meta model behind EML, 2001 (http://eml.ou.nl/introduction/docs/pedmetamodel.pdf).
9 Mayes T., de Freitas S. Review of E-Learning Theories, Frameworks, and Models, Joint Information Systems Committee (JISC) e-Learning Models Desk Study, 2004
10 http://www.blackboard.com
11 http://www.webct.com
12 Dagger D., O'Connor A., Lawless S., Walsh E., and Wade V.P. Service-Oriented E-Learning Platforms From Monolithic Systems to Flexible Services, Published by the IEEE Computer Society 2007
13 Wilson S, Blinco K, Rehak D,. Service-Oriented Frameworks: Modelling the infrastructure for the next generation of e-Learning Systems. July 2004. http://www.jisc.ac.uk/uploaded_documents/AltilabServiceOrientedFrameworks.pdf
14 http://www.elframework.org/
15 IMS Global Learning Consortium: Abstract Framework http://www.imsglobal.org/af/index.html
16 http://www.okiproject.org
17 LeAP Project Case Study: Implementing Web Services in an Education Environment http://www.education.tas.gov.au/admin/ict/projects/imsdoecasestudy/LeAPProjectCaseSummary.pdf
18 Leyking K. Service-oriented Knowledge Architectures – Integrating Learning and Business Information Systems
19 Papazoglou M.P., Jan van den Heuvel W. Service Oriented Architectures: Approaches, Technologies and Research Issues, June 2005 http://www.webct.com
20 Tsai W.T., Chen Y., Introduction to Service-Oriented Computing
21 Tsai W. T. Service-Oriented System Engineering: A New Paradigm, Proceedings of the 2005 IEEE International Workshop on Service-Oriented System Engineering (SOSE'05)

22 Wilson S. Can web service technology really help enable ’coherent diversity’ in e-learning? http://www.elearning.ac.uk/features/pedagandws
23 http://cetis.ac.uk:8080/elearningprogram/subjects/assessmentfold/assessment/topic_view
24 Millard, D., Howard, Y., Bailey, C., Davis, H., Gilbert, L., Jeyes, S., Price, J., Sclater, N., Sherratt, R., Tulloch, I., Wills, G., & Young3, R. (2005). Mapping the e-Learning Assessment Domain: Concept Maps for Orientation and Navigation. In Richards, G. (Ed.), Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2005 (pp. 2770-2775).
25 Davis H, Howard Y, Millard D, Wills G, FREMA: e-Learning Framework Reference Model for Assessment http://www.jisc.ac.uk/uploaded_documents/frema_210306.ppt
26 TENCompetence - Building The European Network for Lifelong Competence Development http://www.tencompetence.org/
27 Tattersall C, Hermans H, OUNL’s assessment model, January the 10th 2006 http://dspace.ou.nl/handle/1820/558
28 Davies, W. M., Howard, Y., Millard, D. E., Davis, H. C. and Sclater, N. (2005) Aggregating Assessment Tools in a Service Oriented Architecture. In Proceedings of 9th International CAA Conference, Loughborough
29 ASSIS: Assessment and Simple Sequencing Integration Services http://www.hull.ac.uk/esig/assis.html
30 Patelis, T., An overview of Computer-Based Testing, The college board, RN-09,. http://www.collegeboard.com/research/html/rn09.pdf

Grayscale Image Analysis of Atomic Force Microscopy Cell Images

Ioana Cristina Plajer[1], Detlef Richter [3], Luminita State [2]

[1] *"Transilvania" University, Bulevardul EroilorNr. 30,500007 Brasov, Romania, iplajer@yahoo.com*

[2] *University of Pitesti, Targu din Vale Str Nr.1, 110040 Pitesti, Romania, radus@sunu.mc.ro*

[3] *Wiesbaden University of Applied Sciences, Kurt-Schumacher-Ring 18,D-65197 Wiesbaden, Germany*

This paper describes segmentation and labeling of cerebral endothelial cells in Atomic Force Microscopy (AFM) grayscale images, for estimation of cell volume and surface. Segmentation is done by watershed-based algorithms to separate the cell regions and subsequent thresholding to separate the cells in the segmented regions from the background. After describing the problems caused by the form of the cells and the gray level distribution, the results obtained by the implementation of a classical watershed algorithm and those obtained a modified watershed algorithm using interactively selected seeds are compared. It is shown that the problem of over-segmentation is better solved by the modified algorithm and that the problem of separating the cells from the background cannot be solved in a satisfactory manner by automatic thresholding algorithms. Thresholding by Otsu's procedure and by measurement dependent interactively selected threshold are compared, in order to underline this statement

Keywords

segmentation, watershed, cell separation, seeds, processing queue, threshold.

1. Introduction

Image segmentation is an important task in image analysis and has the objective to partition the image in regions of interest or to separate objects of interest from the background. The regions / objects of interest can subsequently be analyzed and processed in order to gain the necessary information. Cell segmentation is important in research based on microscopy images.

The images segmented and analyzed in this paper are atomic force microscopy (AFM) [1, 2] images of cerebral endothelial cells. The work was part of a larger research program, involving partnership to a group of biophysicists performing cell research the main objective being volume and surface estimation of cells in sequences of images acquired before and after changing the medium of the culture cells over a period of around one hour, in order to estimate their variations. More details about the research aims and methods can be read in [3].

The resolution of the images obtained is typically in the size of 512 × 512 pixels with the gray level resolution of 8-bit. The height of a cell in each point is given by its gray level which is scaled in nm and enables estimation of cell volume. The number of cells in an image is relatively small, approx. up to eight. The kernels of the cells have light gray values, while the gray levels decrease with their distance from the kernel center, merging with the background gray levels. The relative uniform dark gray levels of the background depend on measurement settings. An example of AFM cell image is shown in figure 1.

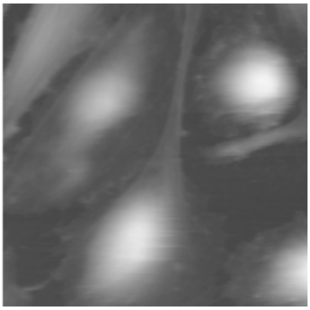


Figure 1 AFM image of totally eight cerebral endothelial cells, which are fully or partially displayed.[4] (Original image 1)

In order to estimate cell surface and volume, first the cells must be identified and labeled in the image. This implies two steps: firstly segmenting the image into regions, each containing a single cell and secondly separating the cells in the segmented regions from their background.

One of the best approaches to segment cell images is the watershed algorithm [5, 6] followed by a thresholding method to separate the cells from the background. Although the steps to obtain the expected results seem to be clear, there are some special problems to be solved.

Classical watershed algorithms are working reasonably well for images in which the objects to be segmented are almost convex. As shown in figure 1 some of the cells have almost queue like extensions. Also some cell extensions belonging to cells outside the image can be seen contained in the image, which make a correct segmentation by the classical watershed algorithm rather difficult. Other difficulties arise by the fact that some of the cells are confluent and by the absence of any defined edges in the image.

The problem of separating the cells from the background by a common thresholding method also proved to be a difficult task. Automatic thresholding methods are mostly based on information extracted from the histogram. For a good separation by thresholding, it is assumed that the histogram is bi- or multimodal, which unfortunately is not the case in these images.

The algorithms and methods proposed in this paper show good solutions for these problems and achieve a segmentation to enable estimations volume and surface of the cells.

2. Methods and Materials

2.1 General Description of Watersheds

The *watershed* notion [5] is based on a topographic interpretation of the gray levels. For each pixel the gray level is considered to be the third. Dark points are considered as valleys and light points as peaks. All slopes surrounding the same valley are called *catchment basins* [5]. The points of the rims, separating the catchment basins are called *watersheds* [5]. The aim of a watershed algorithm is to determine these watersheds and thus to segment the catchment basins. The resulted regions are labeled with distinct labels.

The basic idea of a watershed algorithm is a uniform flooding of the catchment basins with water starting from the lowest valley given by the lowest gray level in the image. In the moment where water tends to spill from one basin into another basin a *dam* is built, which defines a watershed. Starting with gray level 0 an iterative algorithm is used to mark at each step g all the pixels with gray level g.

Considering the image in figure 1, applying the described watershed algorithm would detect watersheds on the light kernels and segment catchment basins around background pixels, which is not the expected result. In order to obtain a correct segmentation, the image has to be firstly inverted and the algorithm has to be applied on the inverted image. An example of inverted image is shown in figure 3, left

2.2 Implementation of a Watershed Algorithm

Based on ideas of [6], the following algorithm was developed to implement the process of flooding.

For sorting the pixels with respect to their gray level, a number 256 S_g, g =0, 1, …, 255 are used. For each pixel $p(x, y)$ of the image, the coordinates are placed in the stack corresponding to the gray level g of $p(x,y)$. After this operation, each stack S_g contains the coordinates of all the pixels of gray level g, as shown in figure 2. Two queues, together with a labeling function are used in the process of labeling the pixels. The *processing queue* contains the pixels of the current gray level, which were labeled and for which the neighbors still have to be computed at the current stage. In the *waiting queue* are placed pixels of gray levels greater than the current gray level, which were marked as belonging to already labeled regions, but for which the neighbors will be computed at a later moment.

The labeling function labels an unlabeled neighbor q of a pixel $p(x,y)$ in the following manner: if q is neighbor to more than one labeled region, then q is labeled with label -1 corresponding to a watershed, else it is labeled with the label of $p(x,y)$. In the beginning all pixels are labeled with 0, which means, that they are not yet part of a segmented region. Each region obtained after segmentation will be labeled with a distinct label greater than 0. Watershed pixels will be labeled with -1.

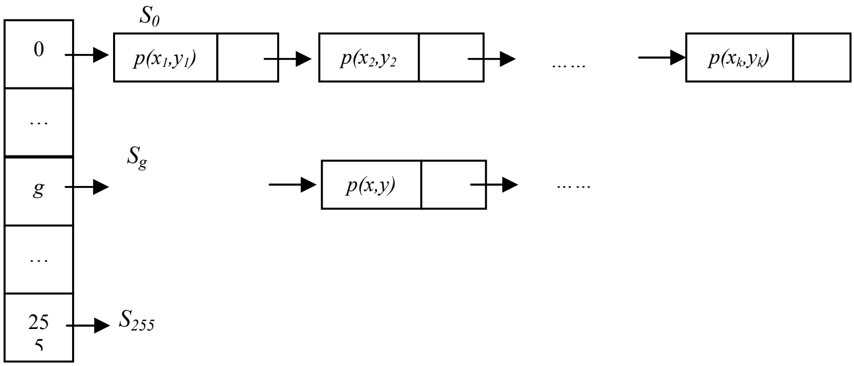


Figure 2 The 256 stacks are dynamically constructed from the pixels of the image and grouped into a table of stacks.

Processing Step 0: In the first step 0, the pixels of gray level g = 0 and their neighbors are labeled as follows. The top pixel p(x, y) in the stack S0 is popped and marked with the label =1, which is in the beginning the current label. Then, using 8-neighbor connectivity, a connected component around p(x, y) of pixels with the same gray level is constructed. This connected component is constructed by a recursive function, which considers each neighbor q of p. and labels it by the labeling function. If q has the same gray level as p, then all its neighbors are labeled recursively. If q has a larger gray level, it is placed in the waiting queue. The marking of its neighbors will thus be done at a later step.

After the first connected component is constructed and labeled the next pixel in the stack S0 is popped. If it was already marked as part of the connected component then it is ignored and the next pixel of the stack is popped. If the pixel is not yet labeled, the current label is incremented and a new connected component constructed and given the incremented label.

The process of popping pixels of S0 and labeling connected components continues until the stack is empty.

At the end of step 0, all the connected components of gray level g = 0 are labeled. The neighboring pixels with larger gray levels are also labeled and placed in the waiting queue.

Processing Step g: In each further step g, g = 1, 2,..., 255, the pixels with gray level g are labeled as follows. First the neighbors of gray level g of the already labeled connected regions are popped from the waiting queue and pushed in the processing queue. While the processing queue is not empty, iteratively a pixel p(x, y) is popped. Each unlabeled neighbor q of p(x, y) is labeled by the labeling function. If q has gray level g, it is placed in the processing queue, so that its neighbors can be labeled at this same step. If q has larger gray level it is placed in the waiting queue.

After all the elements of the processing queue are labeled and so the connected components obtained at step g - 1 are enlarged with pixels of gray level g, new connected components are constructed around the unlabeled pixels of gray level g, extracted from the stack Sg.

To avoid oversegmentation due to irrelevant local minima, the images are first low-pass filtered and then inverted. The watershed algorithm is then applied on the filtered and inverted images and the resulted watersheds are then superimposed on the original image.

2.3 Modified Watershed Algorithm using Seeds

Due to the form of the cells in the images, the algorithm described before produces oversegmentation which needs interactive corrections. To reduce this effect the algorithm was modified defining interactively seeds which will represent the starting points for the labeling. This algorithm does not need low-pass filtering.

The pixels are not sorted at the beginning. The only starting points are the seeds and the number of regions segmented is given by the number of the seeds. The process of labeling works similar to the algorithm described before. It uses also a processing queue and a waiting queue and the same labeling function, but no stack is needed.

Before starting the iterations, the seed pixels are labeled each with a different label. The seeds with the smallest gray level min_s are placed in the processing queue and the other seeds in the waiting queue. In the first step the connected components about the seeds in the proceeding queue are constructed. Then for each gray level g, $min_s \leq g \leq 255$, new connected components around seed pixels of gray level g - if there exist such seeds - are constructed and the already marked connected components are expanded with neighbors with gray levels smaller or equal to g. The main difference to the unmodified algorithm is that when labeling neighbors at each step, not only regions of the current gray level are *flooded*, but also pixels with smaller gray levels. The effect of this labeling is that small valleys around irrelevant local minima, due to noise or gray level repartition within the cells, are merged into one of the cell regions designated by the seed points.

2.4 Thresholding Cells from the Background

The second problem after segmenting the regions corresponding to the cells is to separate the cells from the background. This was achieved by thresholding the segmented image with a threshold obtained by using measurement information. The images acquired during a measurement session, using the same settings, have the same average gray level value in the background. Once the threshold of the first image is set, it can be used for all images.

3. Measurements

Firstly the unmodified watershed algorithm described before was applied on the inverted unfiltered image to observe the degree of oversegmentation.

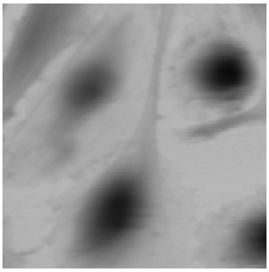
In order to reduce oversegmentation caused by noise and gray level repartition within the cells, the image was smoothened with low-pass filters of different sizes. Firstly binomial filters with sizes of 15 × 15, 21 × 21, 25 × 25 and 27 × 27 and then box filters of the same sizes were used and the results of the segmentation was compared. The filter which gave the best segmentation results was then pointed out.

The modified watershed algorithm was applied on the same image as the unmodified algorithm and the results of the segmentation were compared

The last measurements were made in order to segment the cells from the regions by thresholding. Otsu's threshold algorithm [5] was tested and showed unsatisfactory results. After analyzing the gray scale distribution in the background thresholding with a manually selected threshold was applied to the cell regions.

4. Results and Discussions

Figure 3 right shows the application of the unmodified watershed algorithm on the original inverted image in figure 3 left. A high degree of oversegmentation can be seen in the background area. A lower oversegmentation occurs within the cells.



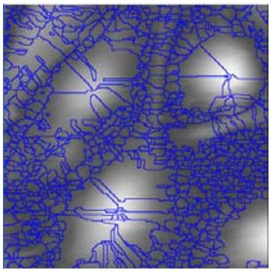


Figure 3 Inverted original Image (left); Oversegmentation resulted after directly applied unmodified watershed algorithm (right)

To reduce oversegmentation the image was first smoothened by low-pass filtering. There were used filter masks of different sizes to obtain optimal segmentation.

In figure 4 are reproduced the results obtained by applying the unmodified algorithm on smoothened images by binomial filters of different sizes. As can be seen from these images, after smoothing with a binomial filter, an important oversegmentation still remain. Enlarging the size of the filter mask does not improve the results.

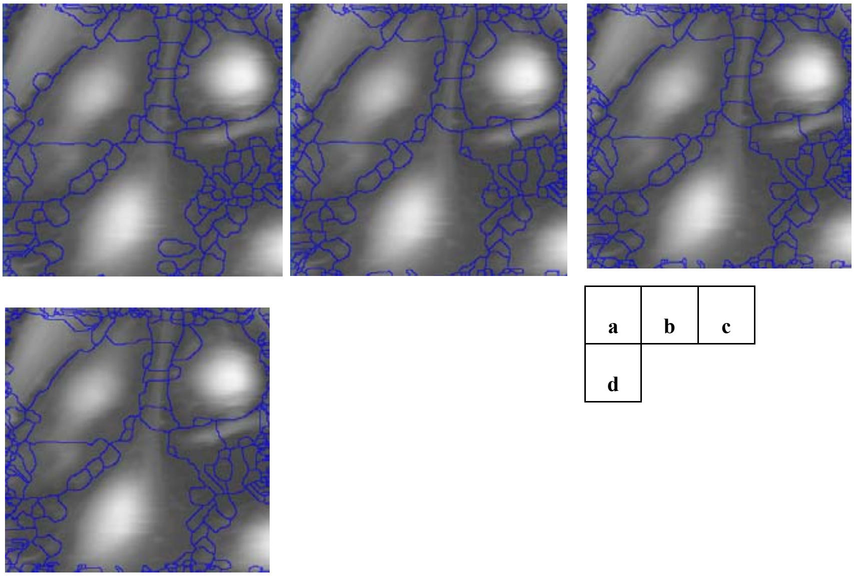


Figure 4 Segmentation after smoothing with binomial filter of size a) 15×15; b) 21×21; c) 25×25; d) 27×27

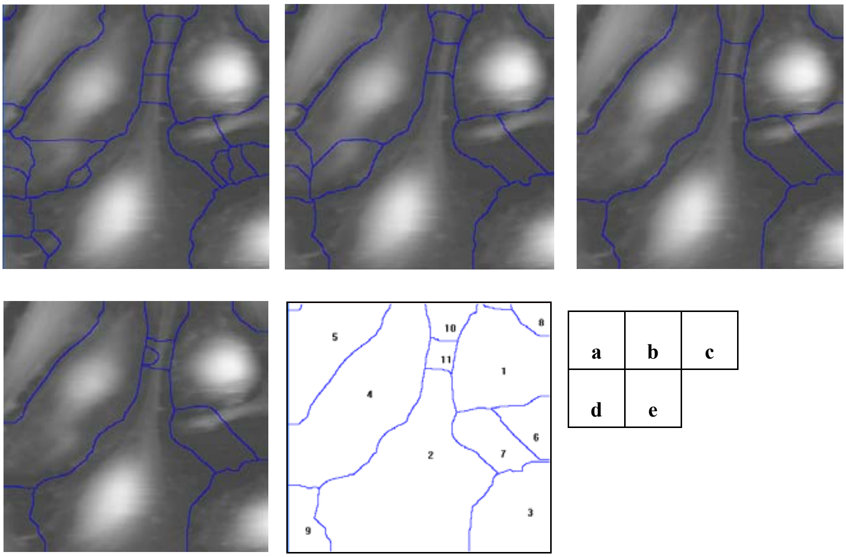


Figure 5 Segmentation after smoothing with binomial filter of size a) 15×15; b) 21×21; c) 25×25; d) 27×27. e) Region map of the regions segmented in c)

Smoothing with box filters reduce the segmentation substantially. Figure 5 shows the results of segmentation after filtering with box filters of different sizes. The best result is obtained with the filter mask of 25 × 25. There still remains a small oversegmentation: regions 4, 5 and 6 from the region map displayed in figure 5 e) belong to the same cell. The same, regions 9 and 10 represent the extension of a cell outside of the image.

For larger mask sizes other segmentation errors occur, like in figure 5 d), where regions 8 and 9 from the region map in figure 5 e) were merged, while they should have been separated, and a new region appeared in region 5 of figure 5 e).

In comparison with the results obtained by the unmodified watershed algorithm, the segmentation with the modified algorithm works better. No oversegmentation can be observed after applying the algorithm on the original image. In figure 6 is shown the result of segmenting the image using interactively selected seeds. The position of the seeds is marked by red circles.

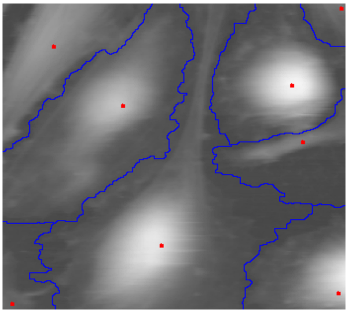


Figure 6 Segmentation with modified algorithm using 8 seed points;

In figure 7 and 8 are shown the results of the modified algorithm on a set of three other cell images.

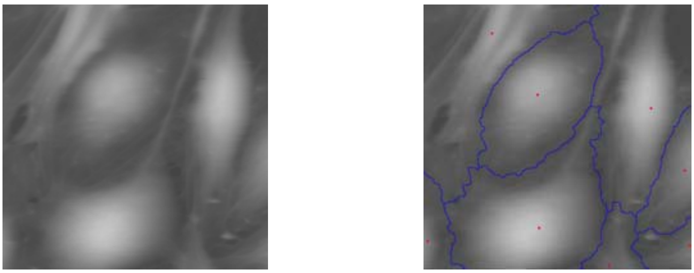


Figure 7 Original image 2 [4] (left). Image segmented with the modified algorithm using 7 seeds (right).

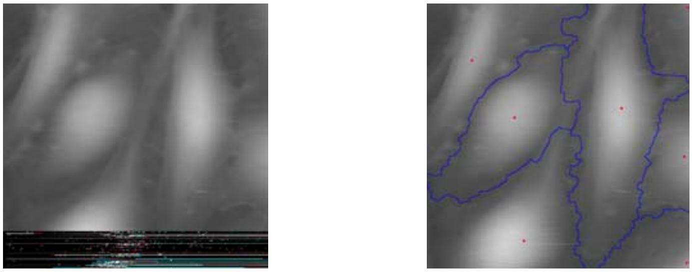


Figure 8 Original image 3 [4] (left). Image segmented with the modified algorithm using 7 seeds (right).

After segmenting the regions, the cells have to be separated from the background, in order to permit volume and surface estimation. Using automatic thresholding algorithms did not give satisfying results, because the histogram of the regions is not bi- or multimodal as shown in figure 9 b). In figure 9 c) and d) can be compared the binarisation results obtained by Otsu's algorithm [5] and by thresholding with threshold depending on measurement settings. It is obvious, that segmentation with Otsu assigns a large part of the cell to the background.

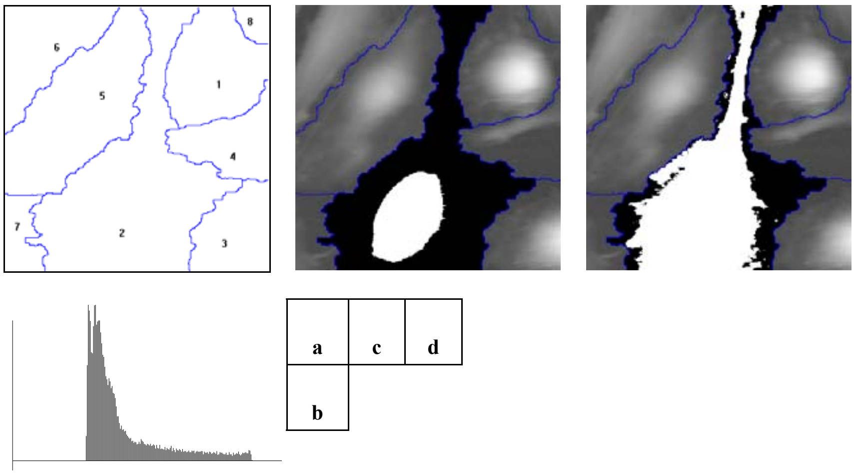


Figure 9 a) region map of segmentation image from figure 6; b) histogram of the region of thresholded cell; c) thresholded cell with Otsu's optimal threshold; d) thresholded cell with threshold set at gray-level 83.

5. Conclusions

Classical watershed algorithms cannot be successfully applied on any kind of cell image. In the case of the AFM cerebral endothelial cells, acquired to estimate volume and surface of the cells, a modified algorithm was elaborated in order to improve segmentation.

The results of the modified algorithm, proved to segment the regions better than the classical algorithm. Moreover there isn't necessary any low-pass filtering, which eliminates the problem of choosing a good filter mask.

The only not yet satisfactory solved problem is the selection of the starting points. Interactively chosen seeds were used.

In some cases, manual corrections have to be done, due to confluence of cells and gray level distribution.

Separation of the cells from the background in the segmented regions by thresholding could not be solved in an automatic way. The binarisation threshold is selected interactively and depends on measurements settings.

References

1 http://www.nanoscience.com/education/AFM.html

2 Dufrêne Y. F. Atomic Force Microscopy, a Powerful Tool in Microbiology. Journal of Bacteriology, October 2002, p. 5205-5213, Vol. 184, No. 19

3 Bálint Z, Krizbai I A, Wilhelm I, Farkas A E, Párducz A, Szegletes Z, Váró G. Changes induced by hyperosmotic mannitol in cerebral endothelial cells: an atomic force microscopy study. European Biophysics Journal, Springer Berlin/Heidelberg, February 2007, Vol. 36, No. 2

4 Courtesy of Prof. Dr. G. Varó, Biological Institute of Szeged

5 Gonzalez R C, Woods R E. Digital Image Processing third edition. Prentice Hall 2007

6 Roerdink J B T M, Meijster A. The Watershed Transform: Definitions, Algorithms and Parallelization Strategies. Fundamenta Informaticae 41, 2001, p. 187-228 IOS Press.

A Fuzzy Reasoning Approach to Automated Detection of Emergent Herd Formations in Computer Aided Simulation

Ognen Paunovski[1], George Eleftherakis[2], Tony Cowling[3]

[1]*SEERC, 17 Mitropoleos Str, Thessaloniki, Greece, ogpaunovski@seerc.org*

[2]*City College, Dept. of Computer Science, 13 Tsimiski Str, 54624 Thessaloniki, Greece, eleftherakis@city.academic.gr*

[3]*The University of Sheffield, Dept. of Computer Science, Western Bank, Sheffield, S10 2TN, UK, a.cowling@dcs.shef.ac.uk*

In recent years multi-agent simulation has been proposed as a method to increase the scientific understanding of emergent phenomena. In this context a vital part of the exploration of emergence is the ability to detect and quantify a manifestation of the phenomenon in the multi-agent models. The formation of herds can be viewed as an emergent phenomenon caused by spatial positioning of animals. As part of a larger study, aimed to explore the causal relations in herd formations, this paper explores the issue of automated detection of herds. The main purpose of the work is to investigate the possibility of developing a mechanism for automated detection of herds in a multi-agent model. The proposed solution is based on fuzzy reasoning approach which incorporates both bottom-up and top-down phases as part of the reasoning process. The evaluation of the proposed reasoning method showed that it can be successfully applied for automated detection of herds in multi-agent simulation.

Keywords

automated herd detection, emergence, fuzzy reasoning, multi-agent model, simulation.

1. Introduction

The basic idea behind emergence was popularized by Anderson in [1], where he elaborated how global features may arise (emerge) as a property of a system and at the same time to be novel to the constitutive components of the system. Emergence can be identified in a variety of systems and processes: from chemical reactions and large scale swarms of social insects to weather patterns and the inner workings of the human brain. Also emergence is associated with social phenomena like culture, patterns in the news, fashion cycles, stock market exchange crashes and so on.

Due to its ubiquity and versatility, the concept of emergence has captured considerable interest by the scientific community in the last couple of decades. Many scientific disciplines have attempted (using different methods) to observe, predict, control and understand emergent phenomena. Nevertheless due to its intrinsic complexity and unpredictability, emergence is difficult to define and quantify in a formal way. Therefore in recent years, agent-oriented modelling and simulation has been used in order to shed light in the process of exploring emergence. The main idea is to use incremental approach which will gradually increase our understanding of a specific manifestation of emergence, by means of observation and analysis of the data generated through simulation. Nevertheless the starting point in this process is to detect and quantify the effects of the emergent phenomenon. This is a vital step because it allows the investigator to establish properties of the emergent effect (manifestation) and observe the impacts that various events in the system have on this effect.

In this context the work presented in this paper is a part of a much larger study [2] which aims to build a causal taxonomy for formation of animal aggregations known as herds. More specifically the issues tackled in this paper are concerned with the exploration of the issue of detection of animal herd formations as emergent behaviour exhibited by animals in nature.

The report is structured as follows. Section 2 offers a brief overview of simulation model developed for the purpose of the study. Section 3 discusses several major issues related to herd detection, while section 4 presents the design of the proposed fuzzy reasoning algorithm and its evaluation. Finally section 5 concludes the paper with a brief summary and planned future work.

2. Simulation Model

The main goal of the herd formation study is to provide insight into how changes in individual's perceptions and responses of its neighbours translate into observable changes in the global herd pattern. The sections that follow examine the theoretical background and the properties of the agent based model as well the simulation environment where the model is executed.

2.1 Theoretical Background for the Model

The Herd Formation model was developed on the basis of the work done by Gueron et al., described in [3]. In the model, the animal is represented by an agent with two main parameters: movement speed and movement direction. In order to simplify the model only three speeds (slow, normal and fast) were taken into account. In addition while having same general direction the animals have the ability to move laterally left or right. The movement speed and direction at time instance *t* for animal *A* are dependent on *A's* speed and direction at time *t-1* and the position of *A's* neighbours at *t-1*.

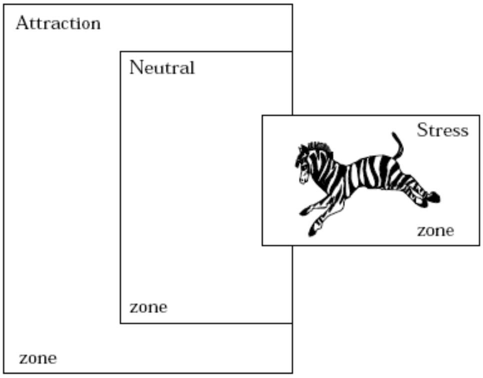


Figure 1 Overview of the neighbour influence zones, from [3]

The neighbour interaction in the model adopts the approach followed in [3] which defines the neighbour interaction in terms of influence zones. An influence zone is a spatial area on a predefined distance relative to the individual in question. This model assumes three zones (stress, neutral and attraction) which are depicted in figure 1. A brief explanation of the zones follows:

- Stress zone or personal zone – the stress zone is of highest importance (compared to the other zones) and it is the primary factor which determines the actions taken by the animal. The need for individual space causes individuals to be repelled by their neighbours when their personal space is invaded [4].
- Neutral zone –is an intermediate zone with no rear dimension. If the neighbours are in the neutral zone the animal does not react, unless all neighbours are on the same side. In this case the agent moves laterally towards the neighbours without changing the speed.
- Attraction zone- the attraction zone influences change in the agent's direction and speed. The presence of neighbours in the attraction zone means that an animal is on the "edge" of the herd. Consequently the instinct is for the agent to move towards its neighbours.

2.2 Model of the Individual Animal and the Simulation Environment

The model of an individual animal was developed using X-machines formal notation and it is presented in figure 2. An X-machine is a general computational machine introduced by Eilenberg [5] and extended by Holcombe [6]. In many ways it looks like a Finite State Machine (FSM), however it is extended with memory. As is visible from the figure, the model does not deal directly with the influence zones, but rather captures the internal states of the animal and the transitions between them. The states represent the movement direction and speed at a particular time instance. While on the other hand the transitions represent a change in the animal's movement speed or direction.

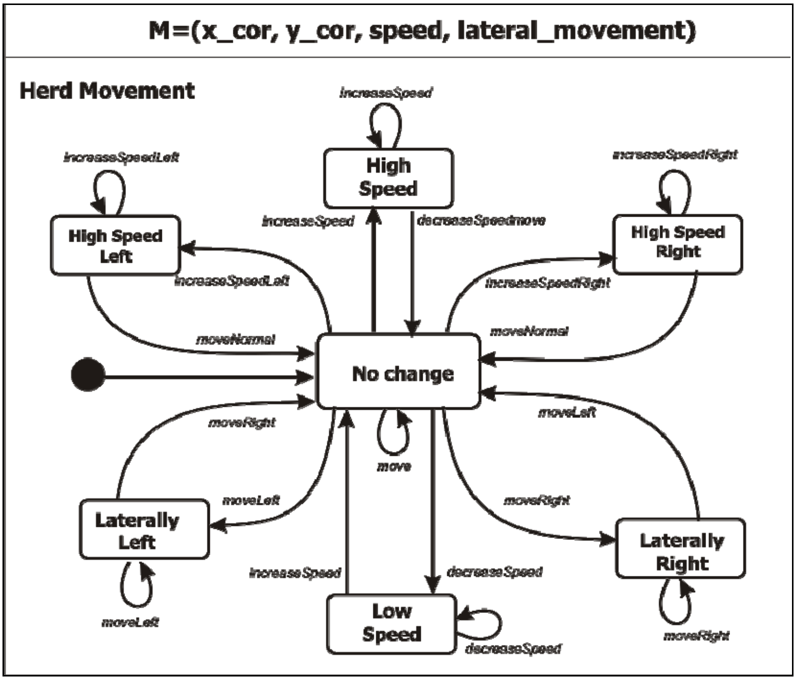


Figure 2 The X-machine model of an animal (agent) in the herd. "M" denotes the memory while the functions in the model are represented as transitions between the states.

Although X-machines are perfect for modelling an individual animal, they lack the expressiveness to deal with the entire multi-agent system. Therefore for the purpose of modelling the multi-agent system as well as execution of the simulation experiments, NetLogo [7] was selected as the most appropriate platform. NetLogo allows modelling and animation of agent like entities in a simulation environment supported by a scripting language, visual animator, simulation controls and data output mechanisms. Figure 3 provides a look at the herd formation model and the NetLogo environment.

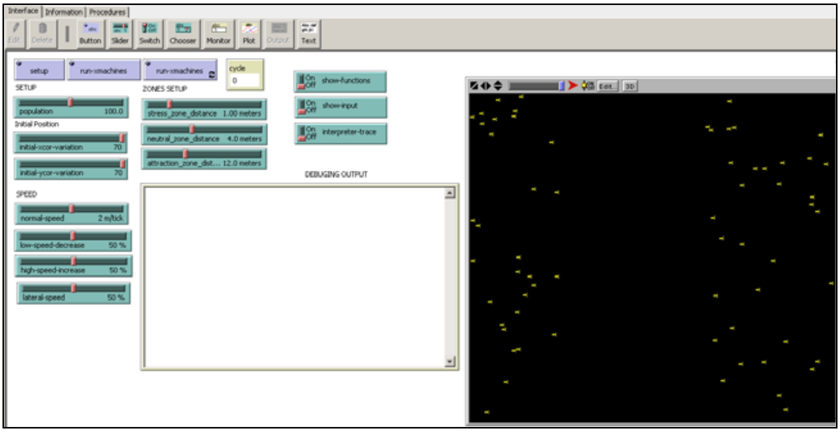


Figure 3 Herd formation screenshot in the NetLogo simulation environment. The simulation setup controls are on the left, while the visual representation of the model is on right

3. Detection of Emergent Herd Formations

In order to achieve the main aim of the study and explore the process that leads to formation of herds, there is a need for a mechanism which would be able to automatically detect groups of animals forming herds. This functionality is vital in order to generate data which will be used in the analysis and delineation of the causes for herd formation. The following sections discuss several issues related to detection of herds as emergent formations caused by spatial repositioning of the animals.

3.1 Herd as Emergent Phenomena

The formation of groups is a common characteristic for many social animals. Shaw [8] examined the grouping in fish, where numerous individuals form a social aggregation known as school. Similar aggregations can be found in birds forming various species of mammals which form herds [9]. Most observations made towards understanding the dynamism of a herd, suggest that both small and large groups rely on local coordination between the individuals to form a herd [3]. Consequently one can argue that a herd, as aggregation of individuals, can be viewed as global pattern which emerges from actions of individuals in their interaction with neighbours. Furthermore there is nothing to suggest in a single animal that when in large numbers these animals would form a herd, thus the formation of a herd is novel behaviour in respect to the individual animals which comprise the herd. In addition the formation of a herd has a functional significance to the system by decreasing the risk of predators and increasing the possibility for mating [10]. Therefore we argue that a herd is an emergent phenomenon due to its functional significance to the system and introduction of novelty in respect to the properties of an individual animal.

3.2 Herd Detection as a Spatial Clustering Problem

The detection of herds in a computer based models can be tackled in different ways. One approach could be to model the group of animals as a separate entity and evaluate the herd from this perspective. It is also possible to adopt a variation on Heylighen's [11] meta-modelling approach, representing the change from loose group to a herd, as transition in the meta-model. However we believe that the most natural and straight forward way to detect a herd is through visual inspection of the spatial distribution of the animals. This approach has a major advantage of being independent from the specific modelling technique and the language in which the model is encoded.

In this manner the detection of herds is essentially a spatial clustering problem, where the goal is to group similar objects into clusters, so that the elements of the cluster have similar properties. A general division of the data-mining spatial clustering algorithms in [12] suggests four main classes: partitioning methods, hierarchical methods, density based methods and grid based methods. Partitioning algorithms like k-means [13] and EM clustering algorithm [14], cluster the object set for k input parameters into k clusters in manner such as to minimize the cluster distribution based on a selected mean point. An interesting approach is followed in EM clustering algorithm where instead of assigning each object to a cluster, a probabilistic membership is computed based on the distance of the object from the mean point. However this method as well as the rest of the partitioning methods, tends to find clusters of spherical shape with similar size, which do not correctly represent the natural distribution of objects [13]. Unlike the partitioning methods, the hierarchical methods, like AGNES and DIANA [15], CHAMELEON [16], are based on hierarchical (tree like) decomposition of object set into smaller sub-sets. While this is a clearly top-down approach, some hierarchical methods can also be applied in bottom-up way (also called "agglomerative") by starting with each object as a separate group and joining them during the clustering process. The algorithms following this clustering approach to not fall in the same trap as partitioning methods, but they tend to make errors in object to cluster membership assignment usually due to over simplistic splitting and merging techniques. Unlike the previous methods the density methods are specifically designed to avoid spherical shape clustering and are best suited in discovering arbitrary shaped clusters. This is achieved by treating a cluster as a dense region of objects. Most famous algorithm in this class is OPTICS [17]. The major problem of density based methods is significantly lowered efficiency especially in the cases when multiple parameters are taken into account. To resolve this problem grid based methods quantize the space into finite number of cells forming a grid structure where the clustering operations are performed. Most famous algorithms are STING [18] and CLIQUE [19]. While these methods perform quite efficiently (in terms of clustering speed and time) they tend to be increasingly error prone with the increase of the number of clustering parameters.

According to Han et al. [12], the selection of an appropriate clustering algorithm is depended on several factors including application goals, trade-off between quality and speed and characteristics of the data. In this context it has to be mentioned that herd formation is a special case due to several reasons. First, the algorithm is supposed to be used in continuous execution rather than single run clustering of data stored in a database, which means that clustering speed is of paramount importance. Second, only a single data attribute (the spatial position) is relevant to the clustering. Also the algorithm should avoid circular shapes and focus on discovery on arbitrary clusters with no mean value. Finally the clustering should be done in such a manner to allow evaluation of the cluster (as a whole) in regards to a "desired state". Given these requirements none of the reviewed spatial data-mining algorithms was found to completely fulfil the requirements. Consequently we moved towards a development of hybrid approach by incorporated ideas from the data-mining algorithms with the requirements imposed by the simulation study. The result was a two way fuzzy reasoning classification algorithm which is described in section 4.

3.3 The Role of the Observer in Detecting Herd Formations

The emergent behaviour relies on stochastic runtime interactions between the elementary components of the system. The process and effects of emergence can only be observed during system's operation (at runtime) and therefore can not be captured with a model of the system. This creates a fundamental difficulty in devising a criteria or metric for identifying and quantifying emergent formation. In addition the dilemma how to judge what is emergence and what is not, leaves the identification of the emergent property or pattern open to interpretations. Therefore one can argue that identifying the effects of emergence is purely subjective process which strongly relies on the nature of the observer, its capabilities, knowledge and judgement.

This reasoning also applies for detection of herds formations. Thus different observers (people) can have different view whether or not the same aggregation of animals forms a herd or not. While on the one hand, herds with strong coherence or extremely loose configurations (which do not form a herd) are immediately distinguishable and clearly identified by most observers. Aggregations on the border between an actual herd and simple collection of animals are difficult to judge. In these cases the same configurations are interpreted differently by different observers. Thus it can be argued that the judgement of the observer directly influences the criteria for detection of herd formations. In addition if we take into consideration the subjective nature of the observer, the question is whether it is possible to clearly define criteria for detection of herds which will be acceptable to all possible observers.

In attempt to avoid going into a philosophical discussions about the nature of an observer, for the purpose of the study, we have developed an idea to incorporate the observers in the reasoning process. The idea is to interpret the results through cut-off threshold boundaries. For example, consider the output variable of the reasoner, called "herd cohesion", which denotes the strength of group. The decision whether this group will be called a herd or not depends whether the value of the output variable is above certain the threshold (boundary) which is defined by the investigator. By modifying the boundary values, the herd reasoner offers the ability to the investigator to incorporate its own personal view (answer to the question what is herd and what is not) into the reasoning process.

4. Fuzzy Reasoning for Detection of Herds

When dealing with a clustering problem, in most of the cases there are clearly defined criteria which allow differentiating between different entities. However in the case of herd formation, the criteria for differentiating between different groupings of animals are far from clearly defined. The main problem is the inability to arrive at a formally quantifiable definition of a herd which will be generally accepted in different contexts by all possible observers. For example one might say a group of animals needs to be "very close" and have "sufficient" members in order to form a herd. However these criteria are a bit unclear and ambiguous. This is the main reason why fuzzy reasoning approach is a perfect candidate for modelling the herd reasoning process. Furthermore fuzzy reasoning allows utilization of inherently inexact concepts in the way humans differentiate herds within an automated computer based process. However the questions still remains whether this process should follow bottom-up or top-down reasoning approach.

Since our study emphasises the fact that a herd is an emergent formation, it is our belief that the best way is to combine the two approaches, as it is suggested in [20]. Therefore in this study, the developed fuzzy reasoning system uses both bottom-up and top-down reasoning in order to identify a herd. A brief overview of the developed reasoning algorithm is presented in figure 4.

1. Determine relevant parameters for single individual (***A***). a. Find the all of ***A***'s neighbours (number of neighbours). b. Find the average distance to ***A***'s neighbour. 2. Use fuzzy reasoning to determine the Herd Belonging Value (HBV) for all animals.	bottom-up phase
3. Identify an individual (***B***) with high herd belonging value a. Identify a group (***B****set*) of ***B***'s strong neighbours (animals with high HBV). b. For the group ***B****set* identified in the previous step, find all neighbours.	transition phase
4. Determine parameters relevant for fuzzy reasoning for ***B****set* . a. Find the number of animals in ***B****set*(herd size). b. Find the average belong value for animals in ***B****set*. c. Find the spatial area covered by the animals in ***B****set*. 5. Use fuzzy reasoning in order to determine the herd cohesion value for ***B****set*.	top-down phase

Figure 4 The developed reasoning algorithm for automated herd detection

Although the operations in the reasoning process are sequentially interconnected, they can be logically divided into three major phases.

- The first phase is the bottom-up reasoning phase, which aims to evaluate the preference of an individual animal to be part of the herd. This reasoning process is depicted by steps 1 and 2 in figure 4. The preference of an animal to be part of the herd is expressed through so called "Herd Belonging Value" (HBV). More details on the practicalities of this phase are discussed in section 5.1.

- The second phase of the reasoning process (depicted by step 3 in figure 4) is the transition from the evaluation of the individual animal, towards reasoning about a group of animals. In this phase the crucial point is the identification of groups of animals which could potentially form a herd which is based on the HBV of a particular animal and its neighbours.

- Once the potential herds are identified, the reasoning process continues with the evaluation of the group coherence. This is in fact the top-down reasoning phase which is depicted by steps 4 and 5 in figure 4. This process is explained in more details in section 5.2.

4.1 Reasoning About an Individual: the Bottom-Up Way

The bottom-up reasoning phase is primarily concerned with identification of the individual preferences for an animal to be part of a herd. Towards this end, two major factors (properties) of an animal are taken into account. These are: average neighbour distance and number of neighbours. Both of them correspond to a fuzzy variable in the reasoning process.

- The *average neighbour distance* represents the average spatial distance to all of the neighbours for a particular animal. Figure 5 (A) represents the fuzzy sets for this variable.
- The *number of neighbours* for particular animal is represented in figure 5 (B).

In the context of the model a neighbour of an animal *A* are all animals which are located either in the stress or neutral zone of animal *A*. The attraction zone is not taken into account since there is an obvious spatial gap between the animals.

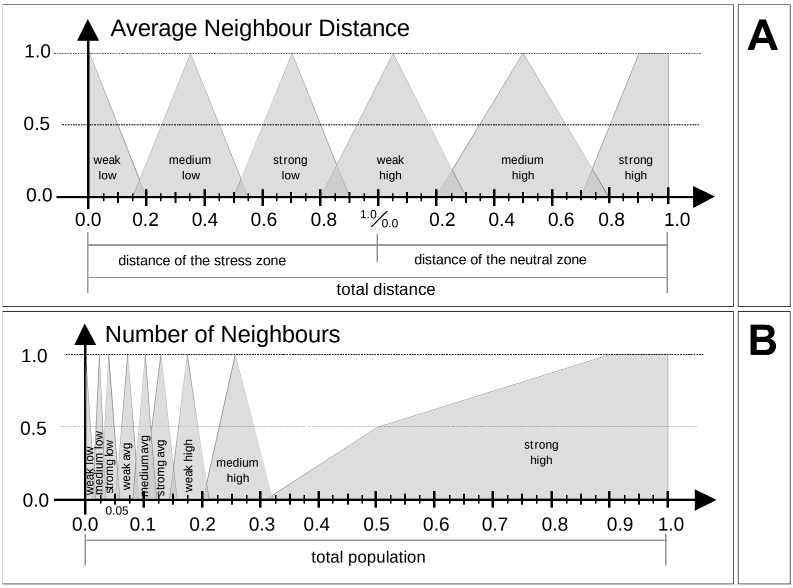


Figure 5 Bottom-Up Input Variables. A (top): Average distance to all neighbours. B (bottom): Number of neighbours for the specific animal.

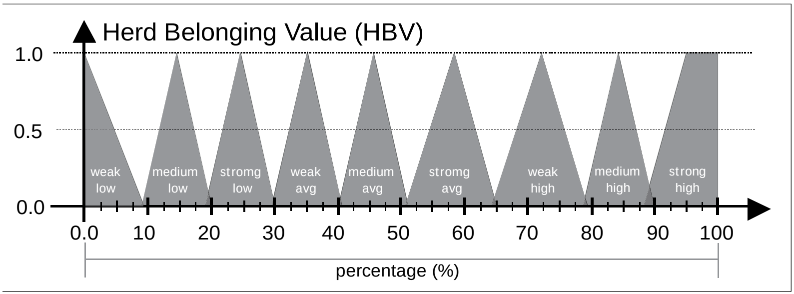


Figure 6 Bottom-Up Output Variable: Herd Belonging Value (HBV).

The two fuzzy variables identified above represent the input parameters in the bottom-up reasoning process. The fuzzy sets for the output variable, the herd belonging value, are presented in figure 6. HBV denotes a preference for particular animal to be a part of a herd.

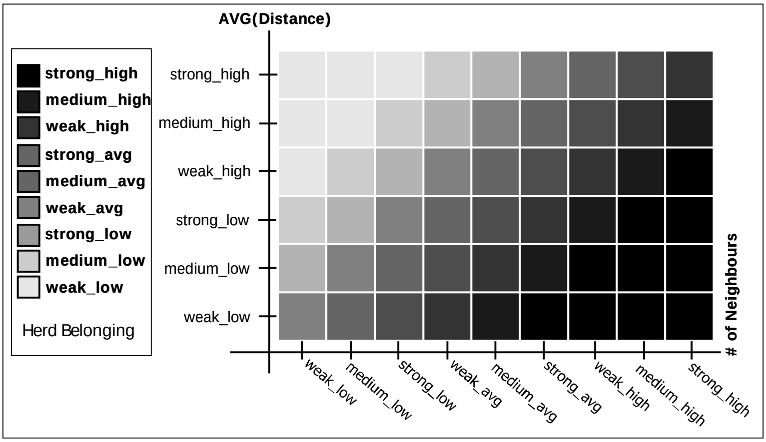


Figure 7 Mapping of the fuzzy sets of the input variables (Average Neighbour Distance and Number of Neighbours) on the fuzzy sets of the output variable (Herd Belonging Value).

Based on the value of the input parameters, the output is determined through a set of reasoning rules. Each rule assigns a single fuzzy set for each of the two input variables and associates it with a corresponding fuzzy set in the output variable. Figure 7 shows the mapping for the input variables and the corresponding fuzzy set in the input variable. As it can be seen from the figure, the value represented by the output fuzzy set has forward correlation with the increase in the number of neighbours (from weak_low to strong_high) and reverse correlation with the increase of average neighbour distance (from stong_high to weak_low). In other words the HBV increases with the increase in the number of neighbours and decreases with the increase of the average neighbour distance. For example if the rule43 (code sample 1) is the only one fired, then the corresponding herd belonging value will be calculated in the "strong_avg" range of 52% to 65% (as it can be seen by mapping the output set from figure 7 to figure 6).

```
FuzzyRule rule43 = new FuzzyRule();

Rule43.addAntecedent(new
FuzzyValue(neighbours_number,"strong_avg"));

Rule43.addAntecedent(new
FuzzyValue(neighbours_distance,"weak_high"));

Rule43.addConclusion(new
FuzzyValue(belongs_value,"strong_avg"));
```

Code sample 1 rule43 - associates "strong_avg" fuzzy set for number of neighbours and "weak_high" set for avg. neighbour distance with the "strong_avg" set in the output HBV.

4.2. Reasoning about a Group: the Top-Down Way

The goal of this reasoning phase is to evaluate whether a group of animals forms a herd or not. In order to achieve this task a vital step is identification of the group members. The identification of the group is primarily depended on the animal's preference to be part of the herd which was evaluated in the bottom-up phase. The HBV is used to determine a set of animals which are called "strong neighbours". Strong neighbours represent a set of neighbouring animals with a high HBV (above the threshold defined by the investigator). The group of strong neighbours forms the skeleton of the possible herd. The process continues with the expansion of the group by adding animals (with low HBV) which are neighbours to the animals in this group. The rationale is that the strong neighbours form the core of the herd while the animals with low HBV form the edge of the herd. In this manner the final result of the process is a group of spatially connected animals.

Once the groups are identified the top-down reasoning process can commence. In addition to the group's average HBV also the group (herd) size and the occupied spatial area are taken into account as input parameters to the reasoning process.

- Average herd belonging value – denotes the average belonging value for all of the animals in the group. Increase in average HBV results in the increase of the cohesion value of the group. Figure 8 (A) depicts the fuzzy sets for this variable.
- Herd size – denotes the number of animals in the group. The appropriate fuzzy sets for this variable are presented in figure 8 (B).
- Herd area – denotes the size of the spatial area that is occupied by the animals in the group. Decrease of the group's area implies increase in the herd cohesion of the group. The fuzzy set for this variable is presented on figure 8 (C).

The output variable (herd cohesion) and the corresponding fuzzy sets are represented in figure 11. The reasoning rules that determine the mapping between the matching fuzzy set in the input variables and the matching fuzzy set in the output variable follow the same principle as the one discussed in section 4. 1.

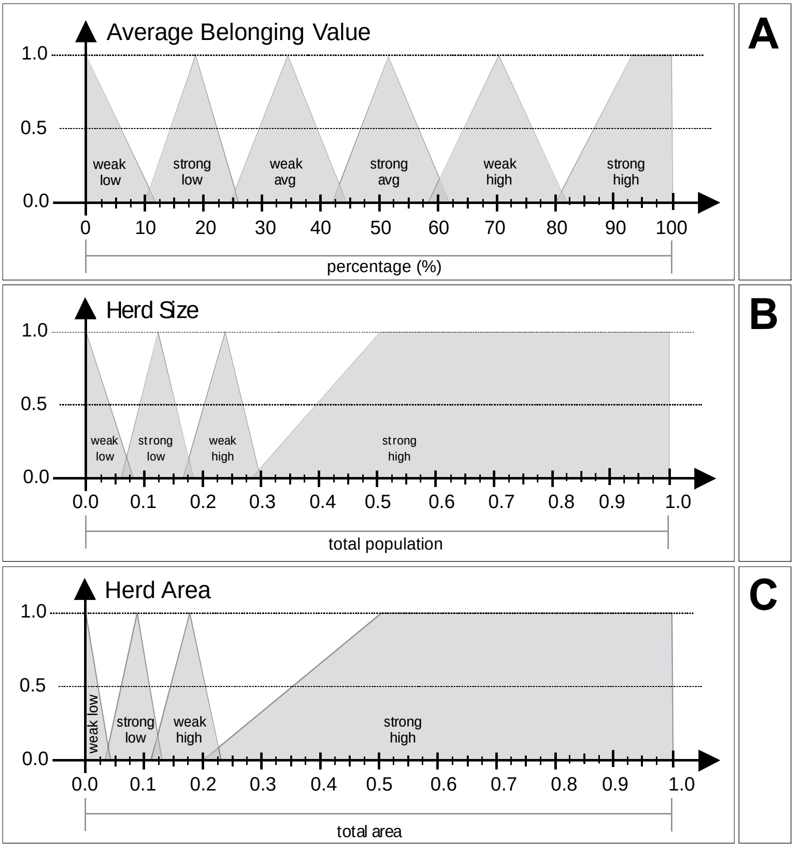


Figure 8 Top-Down Reasoning Input Variables A (top): Average Herd Belonging Value. B (middle): Herd Size. C (bottom): Herd Area.

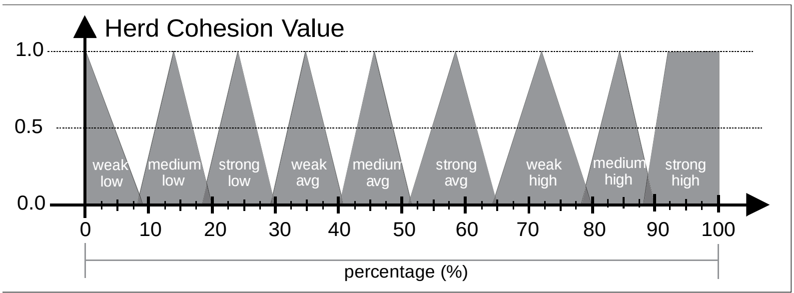


Figure 9 Herd cohesion - output variable in the top-down reasoning process.

4.3 Evaluation and Discussion

The discussed design of the fuzzy reasoner was implemented as a Java-based extension to the Netlogo simulation model (elaborated in section 2). It uses FuzzyJ libraries [21] in order to support the fuzzy reasoning process. The evaluation of the reasoner was done in two steps. First of all the implementation of the reasoner was tested using so called gradation tests. The idea was to use specific sets of data which will check the influence of a single input variable on the output variable for both HBV and herd cohesion variables. An example is presented in figure 10.

TEST SETUP
Simulation Setup Parameters:
Population: 100; Stress Zone: 3m; Neutral zone: 5m;
NEIGHBOURS GRADATION TEST
Test 1: Distance: 5m; Neighbours: 0;
Test 2: Distance: 5m; Neighbours: 2;
Test 3: Distance: 5m; Neighbours: 4;
Test 4: Distance: 5m; Neighbours: 7;
Test 5: Distance: 5m; Neighbours: 11;
Test 6: Distance: 5m; Neighbours: 15;
Test 7: Distance: 5m; Neighbours: 21;
Test 8: Distance: 5m; Neighbours: 28;
Test 9: Distance: 5m; Neighbours: 35;

TEST OUTPUT

```
NEIGHBOURS GRADATION
--------------------------------------------------------
Test 1 Rule 14 matched
Test 1 (dist: 5.0 neigh.num: 0) HBV = 1.50
Test 2 Rule 24 matched
Test 2 (dist: 5.0 neigh.num: 2) HBV = 1.50
Test 3 Rule 34 matched
Test 3 (dist: 5.0 neigh.num: 4) HBV = 14.0
Test 4 Rule 44 matched
Test 4 (dist: 5.0 neigh.num: 7) HBV = 24.0
Test 5 Rule 54, 64 matched
Test 5 (dist: 5.0 neigh.num: 11) HBV = 35.0
Test 6 Rule 64, 74 matched
Tets 6 (dist: 5.0 neigh.num: 15) HBV = 46.0
Test 7 Rule 84 matched
Test 7 (dist: 5.0 neigh.num: 21) HBV = 71.58
Tets 8 Rule 84 matched
Test 8 (dist: 5.0 neigh.num: 28) HBV = 71.83
Test 9 Rule 94 matched
Test 9 (dist: 5.0 neigh.num: 36) HBV = 84.0
```

Figure 10 Gradation test sample with different values for number of neighbours input.

The second phase of the evaluation was significantly influenced by the nature of the simulation study and the ability to validate the results of the reasoning algorithm. The major restriction was the fact that when applying the reasoner on the simulation models a clear “correct” output result can not be precisely determined. Therefore this step of the evaluation process was influenced by the intuition of the investigator how closely the reasoner matched what he considered to be a herd in the given context. Thus the evaluation was done informally through visual observation of the model’s behaviour and the results outputted by the reasoning system. Visual samples of the application of the reasoner to the herd formation model are presented in figure 11.

For most of the simulation scenarios the reasoner performed extraordinary well. It managed to clearly differentiate between loose groups and herd formations. However, several simulation scenarios revealed a conceptual problem in the reasoning algorithm.

The problem occurred due to the constant simulation space and a substantial increase\decrease in the animal population, which resulted in a much higher\lower density of animals. While to some extent the reasoner was designed to compensate for the population flux and performed well for near optimal values, it failed to do so when the population to space ratio was extreme (both high and low). In order to resolve the issue there is a need to express the scales of the fuzzy variables as a function of the population density. Additional problem which become apparent during the evaluation was the identification of groups of animals that might form a herd. In the current process when a group of animals is identified, the evaluation (herd cohesion reasoning) is done for the entire group. This practice eliminates the possibility for a subgroup of this group to be a herd on its own. However changes to the reasoning algorithm to resolve this discrepancy by taking into account all possible sub-groups permutations, showed that the procedure is computationally far too expensive and time consuming while having only minor impact on the simulation data.

Nevertheless, despite the discussed problem the application and evaluation of the proposed algorithm showed that it can successfully detect herd formations while being computationally efficient.

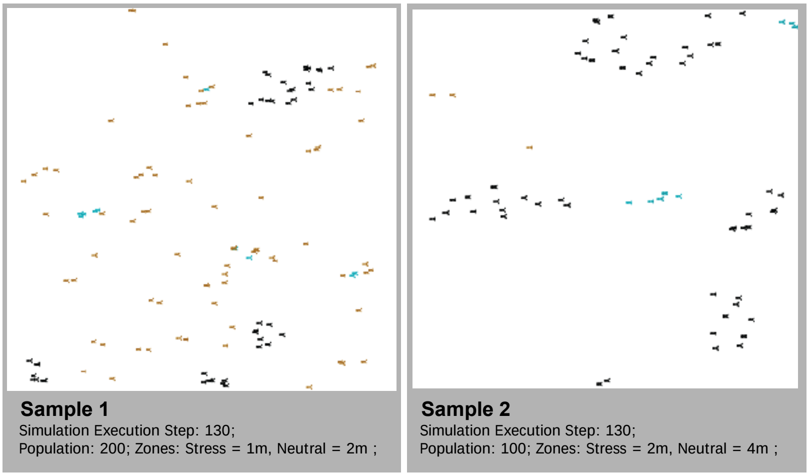


Figure 11 Sample screenshots of the application of the reasoner on the simulation model. Legend: Herd = black, High HBV not in herd = blue, Other animals = red.

5. Conclusions and Future Work

Emergence is one of the most intriguing and at the same time, least understood phenomena of the complex systems. However since it is visible only at runtime, the issue of detecting and quantifying emergent manifestations is a vital part in any attempt to explore and analyze emergence. In this context, the work presented in this report examined issues related to automatic detection of emergent herd formations. In a way the detection of spatial aggregation of animals can be viewed as clustering problem. However while there are numerous studies done in this field, none of them tackles the problem of automatic detection of herd formations. Furthermore the proposed approach is based on two-way reasoning, by utilizing both bottom-up and top-down methods, which is in a way novel method in dealing with a clustering problem.

The starting point for the development of the automated herd detector was the visual inspection of the animal aggregations, which is an approach close to the human way of thinking about herds. This is the main reason why fuzzy logic was used design the reasoner. The implementation of the reasoner was done using FuzzyJ libraries [21] as extension to the herd formation model encoded in NetLogo. The evaluation of the reasoner showed that it can be successfully applied for automated detection of herds in multiagent simulation. Based on the performed tests it performed excellently in most of the test scenarios executed. However in some cases where the spatial/population ratio was extreme it failed to produce the desired results.

The resolution of the identified problem is planned as an initial step of the future work. The main idea is to modify the reasoner in a way that the fuzzy variable scales are expressed through a function of the population density. Once this work is completed, the herd formation study will continue towards experimental investigation of causal relations in emergent herd formation.

References

1 Anderson P W. More is different. Science New Series. 1972; 177(4047): 393-396.

2 Paunovski O, Eleftherakis G, Cowling T. Framework for Empirical Exploration of Emergence within Complex Systems, in proceedings of the 2nd Annual SEERC Doctoral Student Conference (DSC2007), SEERC, Thessaloniki, Greece, 2007.

3 Gueron S, Levin A, Rbenstein D I. The Dynamics of Herds: From Individuals to Aggregations. Journal Theoretical Biology. 1996; 182:85-98.

4 Hediger H. Wild Animals in Captivity. London: Butterworth, 1950.

5 Eilenberg S. Automata, Machines and Languages. New York: Academic Press, 1974.

6 Holcombe M. X-machines as a Basis for Dynamic System Specification. Software Engineering Journal, 1988; 3(2): 69-76.

7 Wilensky U. NetLogo 3.1.3 User Manual. Center for Connected Learning and Computer−Based Modeling, Northwestern University, Evanston, IL, 1999.

8 Shaw E. Schooling fishes. American Scientist. 1978; 66:166–175.

9 Underwood R. Vigilance behaviour in grazing African ungulates. Animal Behaviour. 1982; 79:82-107.

10 Hamilton W D. Geometry for the selfish herd. Journal of Theoretical Biology. 1971; 31: 295–31.

11 Heylighen F. Modelling emergence. World Futures: the Journal of General Evolution, 1991; 31:89-104**.**

12 Han J, Kamber M, Tung A K H. Geographic Data Mining and Knowledge Discovery. In: Spatial Clustering Methods in Data Mining: A Survey. Taylor and Francis. 2001; p. 1–29.

13 MacQueen J. Some Methods for Classification and Analysis of Multivariate Observations. In: Proceedings of Berkly symposium on Mthematical Statistics and Probability. Berkeley, University of California Press. 1967; 1:281-297.

14 Bradley P S, Fayyad U M, Reina C A. Scaling EM (Expectation Maximization) Clustering to Large Databases. Microsoft Research Technical rep. 1990; p. 98-35.

15 Kaufman L, Rousseeuw PJ. Finding Groups in Data: an Introduction to Cluster Analysis. John Wiley & Sons: 1990.

16 Karypis G, Han E H, Kumar V. CHAMELEON: A hierarchical Clustering Algorithm Using Dynamic Modeling. COMPUTER. 1999; 32:68-75.

17 Ankerst M, Breunig M, Kriegel H P, Sander J. OPTICS: Ordering Points to Identify the Clustering Structure. In: Proceedings of international conference on management of data (SIGMOD'99). 1999; p. 49-60.

18 Wang W, Yung J, Munts R. STING: A statistical Information Grid Approach to Spatial Data Mining. In: Proceedings of the International Conference on Very Large Data Bases (VLDB'97), 1997; p. 186-195.

19 Agrawal R, Gherke J, Gunopulos D and Raghavan P. Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications, In: Proceedings of International conference on management of Data (SIGMOD'98). 1998; p. 94-105.

20 Conte R and Castelfranchi C. Simulating multi-agent interdependencies: A two-way approach to the micro-macro link. In: Social science microsimulation. 1995; p. 394-415,

21 National Research Council of Canada's Institute for Information Technology. Java library for Building Fuzzy Systems (FuzzyJ Toolkit). Available online at: http://iit-iti.nrc-cnrc.gc.ca/projects-projets/fuzzyj_e.html

Application of Taguchi Method for the Optimization of Power Consumption in MDF Milling

Stan Cecilia[1]*, Loredana Anne-Marie*[2]

"Transilvania" University of Brasov, Dept. of Machines for Wood Industry, Brasov, 500036 Str. Universitatii, nr.1, c.carp@unitbv.ro[1]*, loredana@unitbv.ro*[2]

This paper presents the findings of an experimental investigation regarding the effects of spindle speed, feed speed and depth of cutting on energy consumption in MDF milling process. In order to perform the experimental study has been used the Taguchi's technique as design of experiment technique and L_{50} orthogonal array design has been used for conducting the experiments. The experimental design reveals that the spindle speed is the significant factor in milling. Obvious that power consumption is minimum (807 W) at the first level of feed, first level of depth of cut and first level of spindle speed. The findings are practical for the industry, both in big companies and in small and medium enterprises. The power consumption decreasing at the same rate of productivity is a wish of each entrepreneur considering the economical aspects involved.

Keywords

MDF, milling, optimization, power consumption, Taguchi method.

1. Introduction

The evolution of medium-density fiberboards in the last decades on the market and the ratio quality/costs aspects led the research presented in this paper. *Medium density fibreboards* are engineered wood-based sheet materials made by bonding together wood fibres with a synthetic resin adhesive, and can be manufactured with either softwood or hardwood species.

The work parameters in machining process such as feed rate, spindle speed and depth of cutting should be selected to optimize the economics of machining operations as assessed by productivity, total manufacturing cost per component or some other suitable criteria.

In investigation of materials machinability, statistical design of experiments is used nowadays quite extensively. Statistical design of experiments refers to the process of planning the experiment so that the appropriate data can be analyzed by statistical methods, resulting in valid and objective conclusions (Montgomery, 1997). DOE methods such as factorial design, response surface methodology and Taguchi methods are now widely used.

2. Taguchi' Method Overview

Taguchi experimental designs, often called orthogonal arrays (OA's), consist of a set of fractional factorial designs which ignore interaction and concentrate on main effect estimation.

Taguchi uses the following convention for naming the orthogonal arrays: L*a*(*b*^*c*) where *a* is the number of experimental runs, *b* is the number of levels of each factor, and *c* is the number of variables. Designs can have factors with several levels, although two and three level designs are the most common. When a design is generated, the levels of each factor are stored in the current database--replacing any data that is already there.[1]

Taguchi robust design methods are set apart from traditional quality control procedures and industrial experimentation in various respects. Of particular importance are the concept of quality loss functions, the use of signal-to-noise (S/N) ratios, and the use of orthogonal arrays.

These basic aspects of robust design methods will be discussed in the following sections. Several books have recently been published on these methods, for example, Peace (1993), Phadke (1989), Ross (1988), and Roy (1990).

Taguchi's parameter design is an important tool for robust design. It offers a simple and systematic approach to optimize design for performance, quality and cost. Two major tools used in robust design are signal to noise ratio, which measures quality with emphasis on variation, and orthogonal array, which accommodates many design factors simultaneously (Park, 1996; Unal and Dean, 1991; Phadke, 1989). Taguchi's design is a fractional factorial matrix that ensures a balanced comparison of levels of any factor. In this design analysis each factor is evaluated independent of all other factors.[2]

3. Experimental Details

3.1 Milling Tests

The cutting tests were performed on a milling machine MNF-10 equipped with a feed system which allows feed speeds in the range 3.7 – 30 m/min. The milling tool used for research has the diameter of 125 mm and provided with synthetic polycrystalline diamond (PCD) plates. The machining parameters for cutting tests are shown in the table1.

Tool characteristics		Milling machine MNF 10 characteristics		Working parameters	
Tool diameter [mm]	125	Spindle speed [rpm]	6621 9717	Spindle speed n [Rpm]	6621; 9717

Tool height [mm]	25	Diameter of tool-holder [mm]	30	Feed speed f [m/min]	4,8,12, 16,20
Clearance angle [°]	15	Power of electric motor [kW]	3,8	Depth of cut dc [mm]	1; 2; 3; 4; 5
Sharpness angle[°]	58	Feed speed [m/min]	3,7-30		
Rake angle [°]	17				

Table 1 Experimental set-up parameters

Determining the power consumption of the machine required certain procedure in order to measure correctly and accurately the overall power draw on the machine. The first step is to locate a place to measure the current flowing into machine from a wire. Then is measured the current without cutting the workpiece both, with the data acquisition system (Velleman) and a multimeter, in order to determine any differences between them and to compute a correction coefficient for the measurements values.

The data acquisition was done in the following conditions:

Time steps: 100 = 1s

Power level: CH1 (channel to register the power consumption for cutting): 6V

The workpieces used in this experimental work, were cutted from MDF boards with 18 mm thickness at the following dimensions: 700 (±0.5) x 300 (± 0,5) mm. The MDF boards has the average humidity of 9 %, density 730 kg/m^3. The experimental tests consist in workpieces milling at different cutting parameters. Each test was performed five times and the means were used in the statistical design. The experimental tests results are presented in table 2 (Pc is power consumption).

Nr.	Pc	dc	f	n	Nr.	Pc	Dc	f	n
1	807	1	4	6621	26	1667	1	4	9717
2	950	2	4	6621	27	1813	2	4	9717
3	1084	3	4	6621	28	1845	3	4	9717
4	1225	4	4	6621	29	2060	4	4	9717
5	1369	5	4	6621	30	2283	5	4	9717
6	861	1	8	6621	31	1555	1	8	9717
7	1079	2	8	6621	32	1922	2	8	9717
8	1281	3	8	6621	33	2282	3	8	9717
9	1519	4	8	6621	34	2642	4	8	9717
10	1720	5	8	6621	35	2988	5	8	9717
11	956	1	12	6621	36	1845	1	12	9717
12	1223	2	12	6621	37	2381	2	12	9717
13	1532	3	12	6621	38	2957	3	12	9717
14	1807	4	12	6621	39	3477	4	12	9717

15	2060	5	12	6621	40	3776	5	12	9717
16	959	1	16	6621	41	1925	1	16	9717
17	1249	2	16	6621	42	2608	2	16	9717
18	1665	3	16	6621	43	3170	3	16	9717
19	2059	4	16	6621	44	3770	4	16	9717
20	2439	5	16	6621	45	4432	5	16	9717
21	1542	1	20	6621	46	2893	1	20	9717
22	1985	2	20	6621	47	3703	2	20	9717
23	2295	3	20	6621	48	4400	3	20	9717
24	2694	4	20	6621	49	4900	4	20	9717
25	3084	5	20	6621	50	5462	5	20	9717

Table 2 Experimental data [3]

3.2 Taguchi Experimental Details – L_{50} Orthogonal Array

In this experimental design we need to minimize the power consumption and for this we chose the case *smaller-the better* and the S/N ratio will be computes as follow:

Eta = -10 * $\log_{10}$ [(1/n) * S(y_i^2)] (1)

Here, *Eta* is the resultant S/N ratio; *n* is the number of observations on the particular product, and *y* is the respective characteristic. The factor *-10* ensures that this ratio measures the inverse of "high power consumption" the more energy is consumed, the greater is the sum of the squared number of power consumption, and the smaller (i.e., more negative) the S/N ratio. Thus, minimizing this ratio will decrease the energy consumption.

An L_{50} orthogonal array design was selected for the present work. This design consists of one factor at 2 levels and up to 11 factors at 5 levels each. There are 50 rows. The Taguchi experimental design detail:

Number of independent factors: 3 :

- dc [mm] (5 levels): 1, 2, 3, 4, 5 [mm];
- f [m/min] (5 levels): 4, 8, 12, 16, 20 [m/min];
- n [rpm] (2 levels): 6621, 9717 [rpm].
- Number of runs (cases, experiments): 50
- Dependent variables: 1; Pc [W]
- Type of problem: Smaller-the-better
- Signal-to-Noise ratio: ETA = -10*log10(SUM(y^2)/N)

Run	dc	f	n	Eta	Run	dc	f	n	Eta
1	1	1	1	-58.1332	26	1	1	2	-64.4398
2	2	1	1	-59.5545	27	2	1	2	-65.1670
3	3	1	1	-60.7022	28	3	1	2	-65.3209
4	4	1	1	-61.7599	29	4	1	2	-66.2782
5	5	1	1	-62.7255	30	5	1	2	-67.1686
6	1	2	1	-58.6980	31	1	2	2	-63.8368
7	2	2	1	-60.6572	32	2	2	2	-65.6760
8	3	2	1	-62.1523	33	3	2	2	-67.1648
9	4	2	1	-63.6300	34	4	2	2	-68.4400
10	5	2	1	-64.7106	35	5	2	2	-69.5082
11	1	3	1	-59.6128	36	1	3	2	-65.3190
12	2	3	1	-61.7500	37	2	3	2	-67.5345
13	3	3	1	-63.7052	38	3	3	2	-69.4164
14	4	3	1	-65.1411	39	4	3	2	-70.8241
15	5	3	1	-66.2757	40	5	3	2	-71.5416
16	1	4	1	-59.6328	41	1	4	2	-65.6886
17	2	4	1	-61.9340	42	2	4	2	-68.3262
18	3	4	1	-64.4293	43	3	4	2	-70.0201
19	4	4	1	-66.2731	44	4	4	2	-71.5273
20	5	4	1	-67.7442	45	5	4	2	-72.9328
21	1	5	1	-63.7617	46	1	5	2	-69.2264
22	2	5	1	-65.9543	47	2	5	2	-71.3701
23	3	5	1	-67.2164	48	3	5	2	-72.8687
24	4	5	1	-68.6080	49	4	5	2	-73.8036
25	5	5	1	-69.7812	50	5	5	2	-74.7475

Table 3 Design summary

	SS	df	MS	F	p
1 dc [mm]	215.6713	4	53.9178	142.3162	0.000000
2 f [m/min]	253.3954	4	63.3489	167.2094	0.000000
3 n [rpm]	356.9996	1	356.9996	942.3011	0.000000
1*2 dc*f	11.9450	16	0.7466	5.583	0.000090
1*3 dc*n	1.1912	4	0.2978	0.7678	0.553238
2*3 f*n	1.0414	4	0.2604	0.6641	0.621014
Residual	15.1544	40	0.3789		

Table 4 Analysis of Variance (Mean= -66.054 Sigma 4.14340)

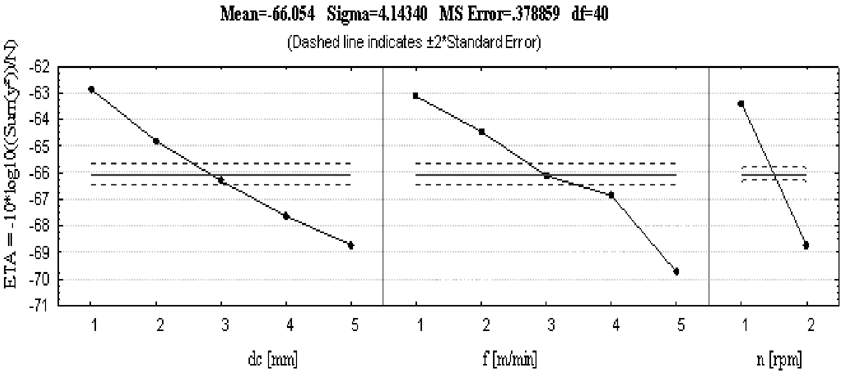


Figure 1 Average Eta by Factor Levels

Effect	Level	Means	Paramet. Estimate	Standard Deviation	Standard Error
Dc	1	-62.8349	3.21890	3.643247	0.339425
	2	-64.7924	1.26143	3.763959	0.345003
	3	-66.2996	-0.24583	3.771472	0.345347
	4	-67.6285	-1.57472	3.733300	0.343595
	5	-68.7136	-2.65978	3.728715	0.343384
F	1	-63.1250	2.92883	3.029002	0.309492
	2	-64.4474	1.60641	3.403331	0.328059
	3	-66.1120	-0.05822	3.848304	0.348847
	4	-66.8508	-0.79704	4.159355	0.362671
	5	-69.7338	-3.67998	3.527992	0.334013
N	1	-63.3817	2.67208	3.214231	0.160355
	2	-68.7259	-2.67208	3.137610	0.158433

Table 5 Average Eta by Factor Levels (Mean = -66.054 Sigma = 4.14340)

Factor	Level	Effect Size	Standard Error
Dc	1	5.4788	0.280010
F	1	4.0842	0.280010
N	1	2.7569	0.280010
Expected S/N		-53.7340	

Table 6 Expected S/N Ration under Optimum Conditions (Mean = -66.054 Sigma = 4.14340)

4. Conclusion

It is obvious that power consumption is minimum (807 W) at the first level of feed, first level of depth of cutting and first level of spindle speed.

The interaction analysis shows that [spindle speed x depth of cutting] and [spindle speed x feed] are the optimal combinations. Thus first level of spindle speed, first level of feed and first level of depth of cut represent the optimal levels of various milling process parameters to yield an optimal value of the power consumption.

The analysis of variance (ANOVA) was performed and is given in Tables 4. The percent contribution of parameters as quantified under column *P* of Tables 4 reveals that the influence of spindle speed and depth of cut in affecting power consumption is significantly larger.

References

1 STATISTICA v7 Electronical Manual

2 Aggarwal, A., et al., Optimizing power consumption for CNC turned parts using response surface methodology and Taguchi's technique—A comparative analysis, J. Mater. Process. Tech. (2007), doi:10.1016/j.jmatprotec.2007.09.041.

3 Carp (Stan) Cecilia (2007) „Dezvoltarea de modele, identificarea metodelor de simulare si optimizare a frezarii MDF-ului" (Models development, indetification of simulation and optimization methods for MDF milling), Research report in frame of Ph.D. thesis.

X-machines Model Checking

Cristina Tudose[1]*, Tudor Balanescu*[2]

The University of Pitesti, Dept. of Computer Science, Str. Târgu Din Vale, No. 1, 110040 Piteşti, România, cristina_ferent@yahoo.com1, tudor_balanescu@yahoo.com2

Model checking is a formal verification technique which determines whether certain properties expressed as temporal logic formulas are satisfied by a system model, usually a finite state machine (FSM). The X-machine model extends the FSM with an attached memory and functions or relations describing transitions between states. An approach to model checking X-machines already exists. It proposes a logic, namely XmCTL, and gives the algorithms for X-machine model checking. This paper presents an algorithm that has as input an X-machine model and an XmCTL formula and as output a Kripke structure (with the same numbers of states like the input X-machine) and a CTL formula.

Keywords
model checking, XmCTL, X-machine.

1. Introduction

The wide use of computers in many safety and business critical domains imposes the need for reliable hardware and software systems. The use of formal methods in the development of systems in safety critical domains can assure that the final product is valid with respect to the user requirements by revealing errors during the development life cycle.

Model checking [1] is a formal verification technique which is based on the exhaustive exploration of a given state space trying to determine whether a given property is satisfied by a system. The model checking problem is as follows: giving a model of the system and a property to verify, model checking return *yes*, if the model satisfies the property and *no* plus a *counterexample*, otherwise. In early 80s two different teams, led by Quielle and Sifakis [2] and by Clarke and Emerson [3], proposed temporal logic model checking algorithms. The property to verify is expressed in a certain temporal logic, usually the *Computation Tree Logic* (CTL). In this context the most usual definition of a model is a Kripke structure K = (Q, R, L) where: Q is a non-empty set of states, R is a binary relation on Q which shows which states are related to other states, L: $Q \rightarrow 2^{Prop}$ is a truth assignment function that shows which propositions are true in each state, where *Prop* is the set of atomic propositions. In CTL each of the temporal operators must be preceded by either A (for all paths) or E (there exists path) path quantifiers. The five basic operators of CTL are:

- X (next) requires that a property holds at the following state

- F (eventually) a property will hold at some state on a path
- G (always) a property holds at every state on a path
- U (until): p U q holds if p holds in a path until q holds
- R (release): p R q holds in the model if the second property holds along a path up to and including the initial state where the first property holds, however without requiring p to hold eventually.

The expressive power of Finite State Machines (FSM) in modelling the control part of systems is demonstrated repeatedly over the last decades. Many model checking techniques [4] have been devised for models expressed as a FSM. These proved to work efficiently in verifying state space with a huge number of states. Finite State Machines are too simple to capture the modelling needs of dynamic systems (e.g. agents) that normally require non-trivial data structures as well as complex control over the states in which these systems can exist. An X-machine [5] is a formal model that satisfies these requirements by introducing a memory structure into finite state machines, as well as computable functions between the states of a model. An approach to model checking X-machines already exists. In [6] a new logic was introduced, namely XmCTL, which extends temporal logic with memory quantifiers, thus facilitating model checking targeted to X-machine models. This paper presents an algorithm that has as input an X-machine model and an XmCTL formula and as output a Kripke structure (with the same numbers of states like the input X-machine) and a CTL formula.

2. X-machine Model Checking

2.1 Stream X-machines

In contrast to FSM, X-machines have functions that receive input symbols and memory values, and produce outputs while modifying the memory values.

A particular class of X-machines is the stream X-machines which is defined as an 8-tuple $M = (\Sigma, \Gamma, Q, M, \Phi, F, q0, m0)$ where:

- Σ, Γ is the input and output alphabet respectively,
- Q is the finite set of states
- M is the (possibly) infinite set called memory,
- Φ is the type of the machine, a finite set of partial functions ϕ that map an input and a memory state to an output and a new memory state, $\phi: \Sigma \times M \rightarrow \Gamma \times M$
- F is the next state partial function that given a state and a function from the type Φ, denotes the next state. *F* is often described as a transition state diagram. $F : Q \times \Phi \rightarrow Q$
- $q0$ and $m0$ are the initial state and memory respectively.

2.2 X-machines Model Checking

An approach to model checking X-machines already exists. In [4] the X-machine is transformed into a Kripke structure (Q, R, L) by exhaustive refinement [7]: Q is the set of all possible states of the X-machine combined with all possible instances of memory in each state; R is the set of transitions between states in Q; L is the truth assignment function, i.e. given a member in Q, it shows which properties are true depending on the values of this memory instance. This process creates two problems: CTL is not expressive enough to describe that a property p holds in some but not all memory instances of all states of the X-machine, and the combinatorial explosion.

In order to overcome those problems, in [6] is proposed a logic (XmCTL) as an extension of CTL, that facilitates intuitive as well as effective model checking of X-machines through the use of operators that quantify memory instances within a single state.

An X-machine state is called an x-state. The set of all memory instances of an x-state q_i is denoted by $M_i=\{M_i^1, M_i^2, ...\}$, where card($M_i$) is the number of memory instances of that x-state and $M_i^j=(M(1)_i^j, M(2)_i^j, ..., M(n)_i^j)$, $1 < j \leq$ card (M_i), where n is the number of memory elements and $M(k)_i^j$ denotes the value of the k-th memory element, $1 \leq k \leq n$, in the j-th memory instance, $1 \leq j \leq$ card (M_i), of the i-th state. In other words, M_i^j is a memory instance corresponding to a state of the exhaustively refined X-machine.

MProp is the set of all predicates composed of instances of memory variables and/or atoms.

Let L: $M \rightarrow 2^{MProp}$ be a function that labels each memory instance M_i^j with all $p \in$ MProp true in that memory instance.

The logic XmCTL contains two memory quantifiers:

- M_x (for all memory instances) requires that a property holds at all possible memory instances of a x-state;
- m_x (there exists memory instance) requires that a property holds at some memory instances of a x-state.

A class of x-state memory formulas (XSM) and a subclass of x-state formulas (XS) are defined inductively, in [8] by the following rules:

XSM1: if $p \in$ MProp then p is an x-state memory formula

XSM2: if a and b are x-state memory formulas, then $\neg$ a, a $\vee$ b, a $\wedge$ b are x-state memory formulas.

XS1. if a is an x-state memory formula, then M_x a and m_x a are x-state formulas.

XS2. if f_1 and f_2 are x-state formulas, then $\neg f_1$, $f_1 \vee f_2$, $f_1 \wedge f_2$ are x-state formulas.

The notation *M*, q $\models$ f, means that f (which is an x-state formula) holds at x-state q in the model *M*. If p is an x-state memory formula, the notation *M*, $M_i^j \models$ p means that p holds in the j-th memory instance of the i-th x-state of the model *M*. Assuming that a, b are x-state memory formulas, f, f_1, f_2 are x-state formulas, and making abstraction of *M,* the relation $\models$ is defined inductively as follows:

XSM1. $M_i^j \models p \Leftrightarrow p \in L(M_i^j)$, where $p \in MProp$

XSM2. $M_i^j \models \neg a \Leftrightarrow \neg (M_i^j \models a)$

$M_i^j \models a \wedge b \Leftrightarrow M_i^j \models a$ and $M_i^j \models b$

$M_i^j \models a \vee b \Leftrightarrow M_i^j \models a$ or $M_i^j \models b$

XS1. $q_i \models M_x\, a \Leftrightarrow$ for all $j \geq 0$, $M_i^j \models a$

$q_i \models m_x\, a \Leftrightarrow$ there exists $j \geq 0$, $M_i^j \models a$

2.3 Model Checking Algorithms

Suppose that we want to check if our X-machine verifies a CTL formula ψ that has m atomic propositions. We need the following definitions:

Definition 1. Label : $M \rightarrow 2^{Subf(\psi)}$ is a function that maps each memory instance M_i^j with subformulas of ψ true in that memory instance.

Definition 2. $Label_x : Q \rightarrow 2^{Subf(\psi)}$ is a function that maps each x-state q_i with subformulas of type m_x a or M_x a where a is a x-state memory formula that holds in that memory instance.

Definition 3. Let q_i be a x-state with $M_i=\{M_i^1, M_i^2, ...\}$. Then we define an equivalent relation ~ on M_i by $M_i^j \sim M_i^k \Leftrightarrow Label (M_i^j) = Label (M_i^k)$, $1 \leq j, k \leq n$ and $j \neq k$.

If $M_i^j \sim M_i^k$ then we say that M_i^j and M_i^k are equivalent.

Definition 4: An x-state q_i is called reduced if for all M_i^j and M_i^k, if M_i^j and M_i^k are equivalent, then $M_i^j = M_i^k$. Given a x-state q, the x-state constructed by merging the memory instances of q that are equivalent will be called the reduced x-state of q and will be denoted by Red (q).

Definition 5: Let q_i be an x-state and $S = \{Mr_i^1, Mr_i^2, ...\}$, card (S) $\leq 2^{card(M)}$. We define $redf_i : M_i \rightarrow S$ to be a reduction function, where $redf_i (M_i^j) = redf_i (M_i^k)$ for all j, k, $j \neq k$ and M_i^j, M_i^k equivalent.

The main idea of the algorithm is summarized in the following steps:

- Label each memory instance of an x-state with memory propositions (atomic propositions), that are subformulas of ψ, valid in that memory instance. A memory proposition is defined using the BNF notation as follows:
  - <memoryProposition>:: = M (<number>) <comparisonOp> <expression> | M(<number>) <comparisonOp> M(<number>), where *<number>* is a natural number*, <comparisonOp> is* any comparison operator and *<expression>* any valid expression that can be compared with a memory element.

- Construct the reduced x-states by emerging all memory instances that are equivalent.
- Label each memory instance of reduced x-states with subformulas of ψ like conjunction, disjunction or negation of memory propositions.
- Label each reduced x-state with all x-states formulas true in that x-state. This formulas are defined below using the BNF notation as follows:
  - <xStateFormula> :: = <MemoryQuantifier> <xStateMemoryFormula>
  - <xStateMemoryFormula> :: = <memoryProposition> | ¬<xStateMemory Formula>|<xStateMemoryFormula><operator> <xStateMemoryFormula>
  - <operator> :: = ∨ | ∧
  - <memoryQuantifier> ::= M_x | m_x
- Transform the X-machine to an equivalent Kripke structure with the same number of states, by considering all x-state formulas as atomic ones.
- Transform XmCTL formula in a CTL formula by replacing each x-state formula with symbols like p,q, ...

3. A Simple Example

We will use the following example taken from [8]:

Consider the case of a machine that accepts coins, 20 and 50 cents of a Euro, until one Euro is entered, as a part of a vending machine. The machine will not return any change.

The X-machine model of the system is shown in Figure 1.

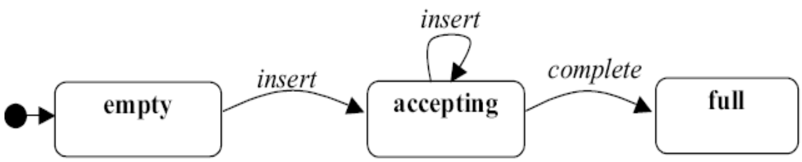


Figure 1 X-machine model of a coin accepter

The model has the following components:

- $\Sigma = \{20,50\}$ is the input alphabet
- $\Gamma = \{20,40,50,60,70,80,90,100\}$ is the output alphabet
- Q = {empty, accepting, full}
- M = {(maxValue,currentValue) / maxValue=100, currentValue ∈ {0,20,40,50,60,70,80,90,100}} is the memory

- $m0 = (100,0)$ is the initial memory
- $q0$ = empty
- The functions $\phi \in \Phi$ are defined below:
  - insert: $\Sigma \times M \rightarrow \Gamma \times M$, insert(c, (maxValue, ValuecurrentValue)) = (c+currentValue, (maxValue, c+currentValue)) if c+currentValue < maxValue, $c \in \Sigma$
  - complete: $\Sigma \times M \rightarrow \Gamma \times M$, complete (c, (maxValue, currentValue)) = (max, (max,max)) if c+currentValue $\geq$ maxValue, $c \in \Sigma$
- F(empty, insert) = accepting, F(accepting, insert) = accepting and F(accepting, complete) = full

The equivalent FSM resulting by exhaustive refinement of this model is shown in Figure 2.

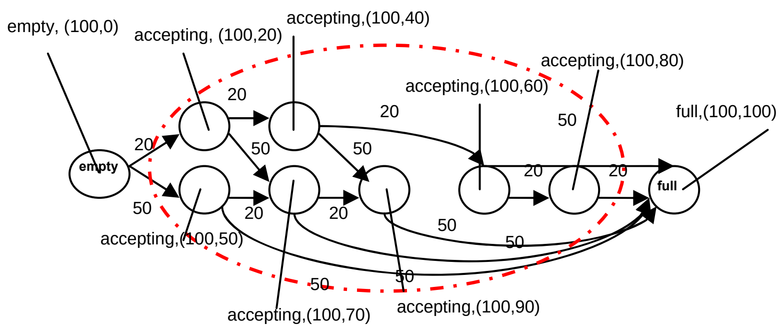


Figure 2 The equivalent FSM to the X-machine model

From this FSM we will obtain the memory instances associated with each x-state of the X-machine. So, for each x-state the set of the memory instances is:

- for *empty* state: $M_1 = \{(100,0)\}$
- for *accepting* state: $M_2 = \{(100,20), (100,40), (100,60), (100,70), (100,80), (100,90)\}$
- for *full* state: $M_3 = \{(100,100)\}$

We want to verify the following XmCTL formula:

EF($m_x(M(2) \leq 60)$) meaning that there is a path and a state in that path, in which there is a memory instance that satisfy $M(2) \leq 60$.

And now we will follow the algorithm described before. For the sake of simplicity, we will consider $M(2) \leq 60$ to be p, $m_x(M(2) \leq 60)$ to be q . Then:

- We will label each memory instances with p if p holds in that memory instance

- L(M1) = {p},
- L(M2) = {p, p, p, ∅, ∅, ∅},
- L(M3) = {∅}

- Now we observe that some memory instances of the second x-state are equivalent, so we could merge all equivalent memory instances in one: L(M2) = {p, ∅}.
- We label each x-state with q.
  - L(M1) = {q}
  - L(M2) = {q],
  - L(M3) = {∅}.
- Now we transform X-machine into a Kripke structure

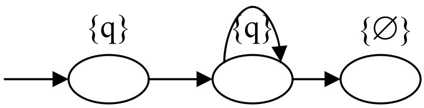


Figure 3 The Kripke structure

- And the XmCTL formula in a CTL formula: EF q

References

1 Clarke EM, Jr., Grumberg O, Peled DA. Model Checking. MIT Press, 2001

2 Quielle JP and Sifakis J. Specification and verification of concurrent systems in CESAR. In Proc. 5th Int. Symp. on Programming, 1981, pp. 337-350.

3 Clarke EM and Emerson EA. Synthesis of synchronization skeletons for branching time temporal logic. In Proc. IBM Workshop on Logics of Programs, Lecture Notes in Computer Science, Vol. 131, Springer-Verlag, 1981, pp. 52-71.

4 Déharbe D. Techniques for Temporal Logic Model Checking. Springer-Verlag, 2006.

5 Eilenberg S. Automata Machines and Languages, Vol. A, Academic Press, 1974.

6 Eleftherakis G, Kefalas P, Sotiriadou A. XmCTL: Extending Temporal Logic to Facilitate Formal Verification of X-machine Models. Analele Universitatii Bucuresti, Matematica-Informatica, Vol. 50, 2001, pp. 79-95.

7 Eleftherakis G, Kefalas P. Towards Model Checking of Finite State Machines Extended with Memory through Refinement. In G. Antoniu, N. Masorakis, and O. Panfilov (Eds.), Advances in Signal Processing and Computer Technologies, World Scientific and Engineering Society Press, July 2001, pp. 321-326.

8 Eleftherakis G. Formal verification of X-machine Models: Towards Formal Development of Computer-Based Systems. Ph.D. Thesis, July 2003.

Computational Modelling of the Development of Human Transactive Memory Systems

Andrea Corbett[1], Mike Holcombe[2], Stephen Wood[3]

[1,2]The University of Sheffield, Dept. of Computer Science, Western Bank, Sheffield, S10 2TN, UK, andrea.smith@sheffield.ac.uk[1], m.holcombe@dcs.shef.ac.uk[2]

[3]The Institute of Work Psychology, University of Sheffield, Sheffield, S10 2TN, s.j.wood@sheffield.ac.uk

Empirical software engineering includes the study of the internal cognitive processes that are utilised by software engineering teams when solving complex problems. The concept of shared cognition has acquired importance in the last 20 years with it being linked to team performance. Transactive memory systems are defined as the beliefs each team member has about the location and quality of their team members' knowledge and expertise. This study is to model the development of transactive memory systems in software engineering teams using an agent-based modelling environment, FLAME. A challenge for this research is to model valid parallels with cognitive processes in human teams. In order to validate the model, data is collected from teams of students carrying out software development projects. This initial model is simple but once proven it can be used as the fundamental architecture for adding further levels of complexity. This paper is intended to describe the early stages of research to develop an agent-based computational model of the cognitive processes that occur in software development teams. It does not endeavour to provide results but to describe the current literature, the planned model concept, and to invite discussion.

Keywords
agent based modelling, empirical software engineering, knowledge worker teams, team cognition, transactive memory systems.

1. Introduction

The cognitive processes that take place in teams are well studied [1] but less work has been carried out relating to the internal cognitive processes that are utilised by software engineering teams when solving complex problems.

Teams are now seen as an essential part of the functioning of organisations and the idea of shared cognition in teams has acquired importance in the last 20 years. Several constructs have been developed to study team cognition, including shared mental models, team situation awareness, transactive memory systems, mutual knowledge and collective mind. [1,2] Each construct has unique aspects but they all incorporate the effect of common knowledge between members.

The nature and quality of shared cognition in teams has been linked with team performance. [3,4,5] Understanding shared cognition in teams impacts team performance in three different ways. Firstly, it explains how members of high performing teams interact with one another and therefore what differentiates high and low performing teams. Secondly, it allows prediction of a team's ability to fulfil a given task and thirdly it can allow the diagnosis of team problems and insight into how to solve them leading to interventions to improve performance. [6]

One concept used to explore team cognition is that of transactive memory systems which model meta-knowledge relating to team members' knowledge and expertise. In this paradigm each team member has beliefs relating to the location and quality of their team members' knowledge and expertise. They gain access to this knowledge by way of communication and interaction with other team members. Lewis [7] defined transactive memory systems within the three dimensions of specialised expertise, credibility and coordination, however, Austin [8] used those of knowledge stock, knowledge specialization, transactive memory consensus and transactive memory accuracy. Both found that transactive memory systems have a positive relationship with group performance.

When more than one person undertakes a joint task, a team is formed and knowledge dependencies are generated that need to be managed by coordination. This is fulfilled by communication between team members. As team members gain expertise over the course of a project they increase their knowledge about each other allowing more efficient coordination and communication. Mcmillan, Entin and Serfaty [5] found that communication efficiency was dependent on knowledge of other team members' activities and this was, in turn, dependent on team preplanning. They also found that more efficient communication among team members resulted in higher team performance. In this study the development of the transactive system started in the preplanning phase which had a direct impact on the efficiency of team communication which led to an impact on team performance.

However, coordination and communication do not stand in isolation. The cognitive functioning of a team is complex and the development of a successful transactive memory system depends on numerous other factors that all interact and ultimately have an effect on team performance.

Previous familiarity with team members and an expectation of continuing to work together has an impact on both face to face and computer mediated team communication. Face to face communication elicits higher information sharing and is more effective in the formation of transactive memory systems. However these differences are reduced with teams that have a history together implying that shared team experience compensates for the perhaps restricted computer mediated communication. It has also been found that the quality of the transactive memory system is directly related to the quality of team decision making and performance. [9,10]

There is a body of work investigating various aspects of teams and their performance. Bell [11] carried out a meta-analysis looking at the relationship between performance and deep level compositional variables of teams, such as personality, values and abilities. She found that in lab settings general mental ability and emotional intelligence were related to team performance but in field settings personality factors and preference for teamwork were strong predictors of team performance. Some previous studies have started to look at the anteceding factors to the development of transactive memory systems and hence team performance. Akgun et al [12] looked at task complexity, team stability, team member familiarity and interpersonal trust. Others include, team structure and resources [5], task interdependence, cooperative goal interdependence and support for innovation [13] and team skills training [14]. This list is not exhaustive but it demonstrates the complexity of the interdependencies that impact on team performance.

Little of this work relates specifically to software engineering teams. The work carried out by software engineering teams requires very specialist skills and is often creative, complex and unstructured requiring teams with high levels of coordination and shared cognition. In fact, expertise coordination has been found to be more important to team performance than input characteristics, presence of expertise and administrative coordination in software development teams.[15] This is an important finding implying that internal cognitive processes have a greater impact on team performance than demographic factors. This is supported in a study by Kang, Yang and Rowley [16] who found that cognitive factors in software development teams had more impact on team effectiveness than the demographic factors of age, tenure and gender.

However, Levesque, Wilson and Wholey [17], studying software development teams found that as role differentiation increased over the course of a project, interaction decreased. They posit that as specialisation increases then communication reduces and this is likely to be an effect in temporary teams who have not had the opportunity to develop strongly shared cognition. This behaviour has been found in other studies looking at temporary task focused teams and may be particularly applicable in the software industry where teams are often are often reformed for each project.[17]

This point shows that it is important to look at the development of these shared cognition concepts over time as they are dynamic and change during the lifecycle of each project and the team itself, potentially impacting on team performance. One method of doing this is to develop a computational model or simulation that can demonstrate changes over time.

This study is to model the development of transactive memory systems developed by software engineering teams, and the factors that may influence their quality, using agent based computational methods. The environment used is FLAME which is an agent based simulation environment, developed at the University of Sheffield. [18] A tailored version of this software is used to model individuals as agents with independent attributes and intentions which communicate with each other through messages.

A software agent is a small software program that has internal memory and decision-making functions, which communicates with other similar agents. As they communicate with each other, agents exhibit emergent behaviour that follows the rules that are encoded within their functions. The rules have to be defined carefully as the choice of rule selection scheme can have a large impact on the way the model works. [19,20]

The behaviour of entities, or agents, interacting with each other in a system will result in some emergent behaviour which may not be expected. Unlike a pure empirical study, an experiment can be performed easily on a model by altering the rules or other parameters of the model yielding a new emergent behaviour. This strategy has proven successful when observing emergent behaviour in ants [21], cells [22] and the human immune system [23].

The intention of the proposed model is not to try and reproduce a facsimile of real life, but to try and produce a dynamic simulation of a set of parameters that interact internally and externally via a group of agents, allowing the researcher to identify dependencies and effects. It is important to recognise that the model is not an end point but a tool that poses questions for further investigation. Also, by modelling a system one is forced to analyse it in detail which in itself poses further questions.

To validate the proposed model, some observable real life data has to be gathered. During the initial stage of this research, in order to assist the development of the first simple model, data is collected using questionnaires administered to four student software engineering teams taking part in a software development project. Data being collected relates to familiarity, transactive memory systems, communication and team performance.

2. Design

2.1 The Agent Based Modelling Environment

The FLAME X-Machine agent modelling environment requires that the following are defined:[18]

- The environment:
- Agents and the memory that each agent holds:
- The different states in which the agents will act:
- The rules:
- The nature of communication:
- Timescales for the model. i.e. the number of iterations or definition of when the modelled project is finished:

The environment in which the agents will operate is known as the project space and it provides them with four locations at which they may operate:

- Meeting room - for team meetings:
- Pair desks - for pair programming:

- Home - for working alone:
- Client – for client meetings:

The project space will also provide the agents with tasks for them to contribute towards the completion of the project.

Each agent will have memory containing the following information:

- Their own knowledge:
- A transactive memory system that can contain the location of other team members' knowledge:
- Their state:
- Their location:
- The progress of the project:
- The rules that they have been given to enable them to process incoming messages:

The typical set of messages that an agent will be able to send will be:

- Request team meeting:
- Request client meeting:
- Request collaboration for a task:
- Accept collaboration for a task:
- Reject collaboration for a task:
- Request knowledge:
- Send requested knowledge:
- Send 'don't know' message:

For some rules or states the succeeding one may be decided by a predefined probability which allows the model to have an element of randomly based behaviour.

A model runs through a number of predefined iterations in order to show the development of emergent behaviour over time. This emergent behaviour can be plotted on a graph or animated.

2.2 The Concept Of Modelling Transactive Memory Systems

For the model, a coding system is employed that uses integers to represent chunks of knowledge. The project requirements are depicted by an arbitrary set of numbers, known as the *start set*. The work carried out by the team is depicted by another set which, as each task is carried out, builds towards matching the start set. This is the completed project and is known as the *end set*. However as not all the original requirements are always completed the end set is a subset of the start set.

At the start of the project each team member, or agent, is allocated a subset of the start set as their set of knowledge. These small sets of knowledge are distributed throughout the team with varying levels of overlap. For example, each agent may have discrete sets, depicting no redundancy, or their sets may overlap depicting shared knowledge. Not all the knowledge in the project is allocated initially and more will be allocated to agents throughout the project indicating their knowledge increasing. The rules for this ongoing allocation may depend on project progress, attendance at meetings, amount of collaboration or other predefined parameters.

Also initially, each agent is given information in their transactive memory system relating to the knowledge each other agent possesses. At the start of the project this may be nothing if they are all unknown to each other but again this is a parameter that can be changed as it may be necessary to model prior knowledge of other agents' knowledge.

Each time an agent or agents work alone, collaborate or attend a meeting the system generates a task that consists of a small group of numbers from the start set that equates to a small set of requirements for knowledge. The agent carrying out the task will check their own set of knowledge for a match with the task and for each that they do not hold they must then check their transactive memory system to establish whether they know the location of the required knowledge. If not then they must find it by requesting the information from other agents by sending knowledge request messages. If any request is successful they can update their transactive memory system to record the location of the knowledge. When they have found all possible knowledge to fulfil the task it is returned to the end set, however this may not be complete as all the knowledge necessary may not be in circulation. As can be seen in Figure 1, all agents are ready to respond to knowledge requests when they are received.

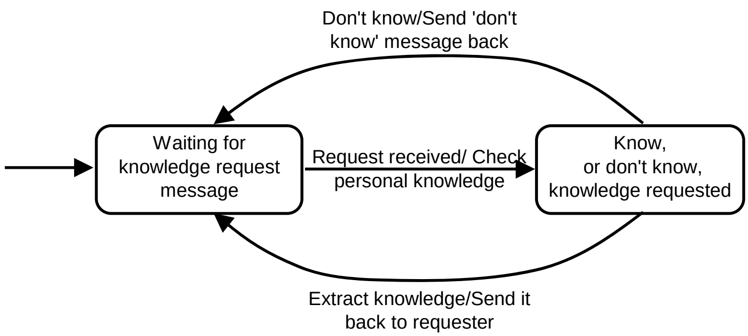


Figure 1 The X-machine for an agent waiting for and processing knowledge request messages

When agents enter the 'task focused working' state they may be working alone or they may be working with one or more other agents in collaboration. Figure 2 shows how the collaborators share knowledge relating to the task and update their transactive memory system. For any knowledge that is missing the agent will send knowledge request messages to all other agents in the team with which they are not collaborating and update their transactive memory system when knowledge is returned. If the knowledge is unknown by the receiving agent they send back a 'don't know' message.

This continues until no more knowledge is available and the task is submitted to the end set.

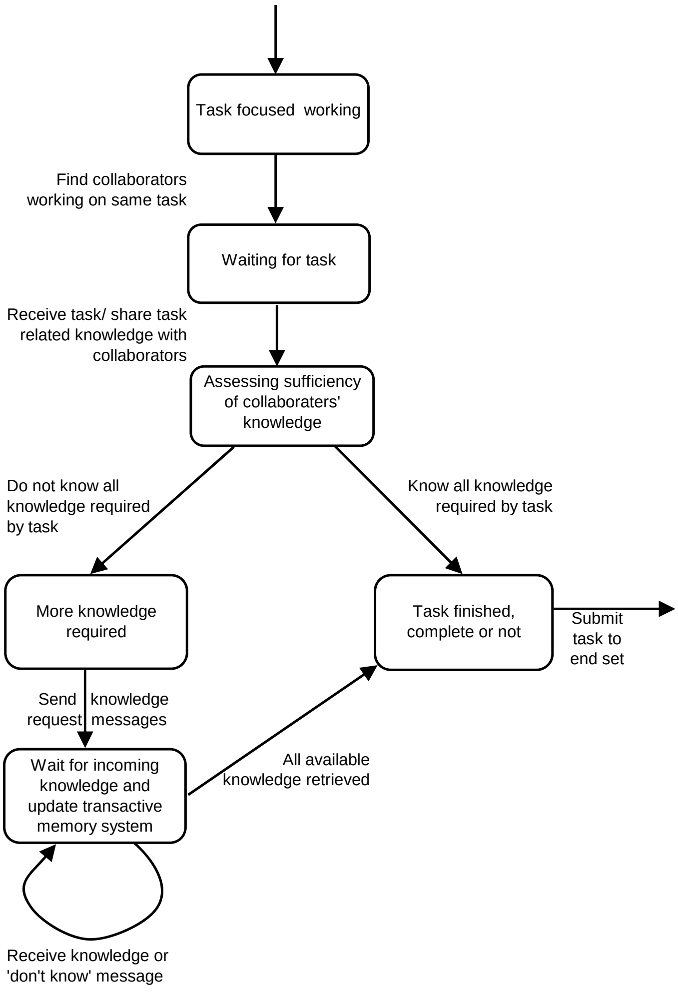


Figure 2 X-machine for processing a task when collaborating or working alone

Agents will be in both task focussed working, and also waiting for, and responding to, knowledge request messages. This mechanism will also be used in meetings where the task numbers, if held, will be passed to other members of the meeting to update their transactive memory systems and no knowledge request messages need to be sent. Agents not present at the meeting will not have the benefit of updating their transactive memory systems.

2.3 Modelling Progress And Performance

One of the challenges for this model is the mechanism for modelling progress and performance. This has to be measurable and have applicability in an actual software engineering project.

In this model project progress is measured by a simple counter incrementing and decrementing according to rules and events. The rules will generate positive, negative or neutral progress depending on factors such as the success of a task or client feedback.

The end of the project needs to be defined and this may happen when an event or threshold is reached. This will be when the following are reached as defined by the individual experiments:

- The end set has reached a predefined level of completeness:
- A specified number of iterations have taken place:
- The progress counter has reached a specified number:
- A specified amount of task work has taken place:

Team performance is measured, depending on the experiment, by using one or a combination of the following parameters, some of which may be dependent on others:

- Completeness of end set:
- Final progress counter:
- Quality and completeness of the transactive memory system:
- Actual number of steps of progress counter:
- Number of knowledge request messages:
- Number of iterations to project completion:
- Time spent on tasks:

2.4 Data Collection to Validate the Model

For this initial phase of the research data is being collected from four teams of student software engineers carrying out a software development project at the University of Sheffield. At the time of writing this paper the data collection is ongoing so no results are presented here.

The data being collected is related to team familiarity, communication, transactive memory systems and team performance. [10,3] At the outset four student teams were identified and each was given the same set of requirements for the software development project taking place over the next 15 weeks. During week one the students were given an introduction and initial briefing. They also met with their client and discussed the client's requirements. Each team had a manager which was a supervisory role fulfilled by a University member of staff.

Familiarity with other team members was measured by each of the students being asked how well they knew other members of their team prior to the start of the project. The students were asked to respond on a four point scale with, 1 = do not know, 2 = acquaintance, 3 = know well and 4 = know very well for each other member of their team [10].

Each student was also asked to estimate the weekly frequencies of both face to face communication and non face to face communication with other members of their team.

Lewis [7] developed and tested a measure to assess transactive memory systems in the field. It is a 15 item scale separated into three categories of specialization, credibility and coordination. This has subsequently been used in other studies [10,13]. This scale was used to collect data on transactive memory systems from the four student teams.

Ancona and Caldwell [3] used measures for team performance which have subsequently been used [8,10,13]. Hence, at the end of the students' project each of the four teams will be assessed based on six items measuring efficiency, quality, technical innovation, timeliness, budgeting and general excellence using a Likert scale.

Table 1 shows the schedule used for collecting the data:

Data	Collected from	Frequency of collection
Initial familiarity with other team members	Student	Once, week 2
Frequency of face to face communication	Student	Weekly
Frequency of non-face to face communication	Student	Weekly
Transactive memory system	Student	Weeks 2,5,6,11
Client evaluation of team performance	Client	Once at end of student project
Manager evaluation of team performance	Manager	Once at end of student project
Team evaluation of team performance	Student	Once at end of student project

Table 1 Student software development project data collection schedule.

Observing the parallels between the data collected and the behaviour of the model will allow the development of methods to validate and develop the model but is unlikely to allow generalisation to other situations without further experimentation. It will give some insight into how to use the data to confirm the model but is unlikely to be sufficient to draw any significant conclusions at this stage.

3. Discussion

3.1 Potential Development of the Model

This initial model is clearly simplistic but once proven it can be used as the fundamental architecture for adding more levels of complexity. For example it can be used to identify some of the precursors that may positively or negatively influence the successful growth of a transactive memory system in a team. The model can also be used to manipulate the team environment and processes to observe the impact on the transactive memory system development. This will ultimately provide insight into the impact of changes of staff, environment and processes on team performance and efficiency.

There are also four key areas in which the model could be improved based on previous research.

Firstly behaviour relating to individual differences of the team members could be modelled. Using theories of personality, types of behaviour of team members and how they interact with others, could be modelled.[24-26] Also, the type and efficiency of communication within a team can be included in the model to predict how it can affect the transactive memory system and team performance.[27,10]

Variables, such as, outset familiarity, management support for innovation, task complexity, team structure, team stability, interpersonal trust, team resources and training environment could be modelled to investigate the impact on transactive memory systems.

The described model simply looks at transactive memory systems in terms of beliefs about each other's knowledge. As accuracy and consensus is important in team transactive memory systems [8] the model could be refined to go one step further and model each team member's beliefs relating to other team members' transactive memory.

This simple model manages tasks by imposing a task on either a single agent or a number of agents collaborating. The introduction of task optimisation would allow agents to select and prioritise tasks based on their transactive memory system. This could be extended further to allow agents to select their collaborators in the same way.

Finally one potential advantage of modelling team behaviour in this way is the possibility of experimenting with extreme boundaries which are perhaps not possible in reality.

3.2 Challenges for this Research

The challenge for this research is to establish the feasibility of modelling cognitive processes using agent based modelling tools. The FLAME framework will be used to model the transactive memory system of software development teams. Whilst FLAME has been used for modelling other environments this is a novel use for it; the study will reveal the suitability of FLAME for modelling cognitive systems in teams.

Furthermore the problem of modelling team progress and performance is one that requires careful design in order to accurately reflect the impact of the parameters being modelled. In addition, empirical testing will take place and it is important to ensure that the data collected is a true measure of the behaviour being modelled.

As mentioned previously, the intention of the proposed model is not to try and reproduce a facsimile of real life, but to develop a tool that can be used to generate questions for further research. To answer these questions further empirical studies will be required which in turn will be used to enhance the model. This technique has lead to significant findings in cell biology. [22]

Acknowledgements

This research is jointly funded by the Engineering and Physical Sciences Research Council (EPSRC) and the Economic and Social Research Council (ESRC). The authors would like to thank Dr Chris Thomson from the University of Sheffield for his helpful criticism.

References

1 Salas E, Fiore SM, editors. Team cognition: understanding the factors that drive process and performance. Washington (DC): American Psychological Association; 2004.

2 Langan-Fox J, Wirth A, Code S, Langfield-Smith K, Wirth A. Analyzing shared and team mental models. Int J Ind Ergonom 2001;28:99-112.

3 Ancona DG, Caldwell DF. Demography and design: predictors of new product team performance. Organ Sci 1992;3(3):321-41.

4 Klimoski R, Mohammed S. Team mental model: construct or metaphor? JOM 1994;20(2):273-83.

5 MacMillan J, Entin EE, Serfaty D. Communication overhead: the hidden cost of team cognition. In: Salas E, Fiore SM, editors. Team cognition: understanding the factors that drive process and performance. Washington (DC): American Psychological Association;2004; p.61-82.

6 Cannon-Bowers JA, Salas E. Reflections on shared cognition. J Organiz Behav 2001;22:195-202.

7 Lewis K. Measuring transactive memory systems in the field: scale development and validation. J Appl Psychol 2003;88:587-604.

8 Austin JR. Transactive memory in organizational groups: the effects of content, consensus, specialization, and accuracy on group performance. J Appl Psychol 2003;88(5):866-78.

9 Alge BJ, Wiethoff C, Klein HJ. When does the medium matter? Knowledge-building experiences and opportunities in decision-making teams. Organ Behav Hum Dec 2003;91:26-37.

10 Lewis K. Knowledge and performance in knowledge-worker teams: a longitudinal study of transactive memory systems. Manag Sci 2004;50(11):1519-33.

11 Bell ST. Deep-level composition variables as predictors of team performance: a meta-analysis. J Appl Psychol 2007;92(3):595-615.

12 Akgun AE, Byrne J, Keskin H, Lynn GS, Imamoglu SZ. Knowledge networks in new product development projects: a transactive memory perspective. Inform Manag 2005;42:1105-20.

13 Zhang ZX, Hempel PS, Han YL, Tjosvold D. Transactive memory system links work team characteristics and performance. J Appl Psychol 2007;92(6):1722-30.

14 Prichard JS, Ashleigh MJ. The effects of team-skills training on transactive memory and performance. Small Gr Res 2007;38(6):696-726.

15 Faraj S, Sproull L. Coordinating expertise in software development teams. Manag Sci 2000;46(12):1554-68.

16 Kang HR, Yang HD, Rowley C. Factors in team effectiveness: cognitive and demographic similarities of software development team members. Hum Relat 2006;59(12):1681-1710.

17 Levesque LL, Wilson JM, Wholey DR. Cognitive divergence and shared mental models in software development project teams. J Organiz Behav 2001;22:135-44.

18 FLAME: flexible large-scale agent modelling environment. [online] 2008 [cited 2008 April 28]; URL: http://www.flame.ac.uk/joomla/

19 Phelan SE. Cognitive capacity as competitive advantage: a simulation test. Simulat Mod Pract Theory 2002;10:455-71.

20 Juran DC, Schruban LW. Using worker personality and demographic information to improve system performance prediction. J Oper Manag 2004;22:355-67.

21 Georghe M, Holcombe M, Kefalas P. Computational models of collective foraging. Biosystems 2001;61(2-3):133-41.

22 Walker DC, Southgate JS, Hill G, Holcombe M, Hose DR, Wood SM, et al. The Epitheliome: modelling the social behaviour of cells. Biosystems 2004;76:89-100.

23 Baldazzi V, Castiglione F, Bernaschi M. An enhanced agent based model of the immune system response. Cell Immunol 2006;244(2):77-9.

24 Matzler K, Renzl B, Muller J, Herting S, Mooradian TA. Personality traits and knowledge sharing. J Econ Psychol 2007;in press.

25 Barrick MR, Stewart GL, Neubert MJ, Mount MK. Relating member ability and personality to work-team processes and team effectiveness. J Appl Psychol 1998;83(3):377-91.

26 Ellis APJ, Hollenbeck JR, Ilgen DR, Porter COLH, West BJ, Moon H. Team learning: collectively connecting the dots. J Appl Psychol 2003;88(5):821-35.

27 Colquitt JA, Hollenbeck JR, Ilgen DR, LePine JA, Sheppard L. Computer-assisted communication and team decision-making performance: the moderating effect of openness to experience. J Appl Psychol 2002;87(2):402-10.

Minimum Flows in Networks using Dynamic Tree Implementation

Georgescu Oana

Transilvania University of Braşov, 50 Iuliu Maniu Str, Braşov, Romania, gescoana@gmail.com

We present an algorithm for finding minimum flow in networks with arcs capacities using the dynamic tree implementation. As a final part, we present an example for this algorithm.

Keywords

dynamic tree, minimum flow problem, network algorithms, network flow.

1. Introduction

The computation of a maximum flow in a graph is an important and well studied problem, both in computer science and operations research. The problem of solving the maximum flow in the special types of networks was treated by the author in a series of the original reports and papers. In order to solve the maximum flow problem, we studied the network flow in planar and bipartite networks [1], in unit capacity and simple capacity networks [2], in serial and acyclic networks [3]. Many efficient algorithms have been developed to solve this problem, (see, e.g., [4]). Sleator and Tarjan [5] developed the dynamic tree data structure and used it to improve the worst-case complexity of Dinic's algorithm from $O(n^2m)$ to $O(nmlog(n))$. Since then, researchers have used this data structure on many occasions to improve the performance of a range of network flow algorithm. Using the dynamic tree data structure, Goldberg and Tarjan [6] improved the complexity of the FIFO preflow-push algorithm from $O(n^3)$ to $O(nmlog(n^2/m))$ and Ahuja, Orlin and Tarjan [7] improved the complexity of the excess scaling algorithm and several of its variants.

The computation of a minimum flow in a directed network has been investigated by Ciurea and Ciupală [8]. In this paper we consider an advanced topic: the use of the dynamic trees data structure to efficiently implement the shortest decreasing path algorithm for the minimum flow problem.

2. Terminology and Preliminaries

As we describe in [9] and [10], we consider a capacitated directed network *G=(N,A,l,c,s,t)* with a nonnegative capacity *c(i,j)* and with a nonnegative lower bound

$l(i,j)$ associated with each arc $(i,j)\in A$. We distinguish two special nodes in the network G: a source node s and a sink node t.

For a given pair of not necessarily disjoint subsets X, Y of the nodes set N of a network G we use the notation: $(X,Y)=\{(i,j) \mid (i,j)\in A,\ i\in X,\ j\in Y\}$ and for a given function f on arcs set A we use the notation $f(X,Y)=\sum_{(X,Y)} f(i,j)$.

A *flow* is a function $f:A\rightarrow \boldsymbol{R}_+$ satisfying the next conditions:

$$f(i,N)-f(N,i)=\begin{cases} v, if\ i=s \\ 0, if\ i\neq s,t \\ -v, if\ i=t \end{cases} \quad \textbf{and } \mathbf{l(i,j)\leq f(i,j)\leq c(i,j)}$$

for some $v \geq 0$ and we refer to v as the *value of the flow f*.

The minimum flow problem consists in determining a flow f for which v is minimized.

A *cut* is a partition of the nodes set N into two subsets S and $T=N-S$; we represent this cut using notation $[S,T]$. We refer to an arc $(i,j)\in A$ with $i\in S$ and $j\in T$ as a *forward arc* of the cut and an arc $(i,j)\in A$ with $i\in T$ and $j\in S$ as a *backward arc* of the cut. Let (S,T) denote the set of forward arcs in the cut and let (T,S) denote the set of backward arcs. We refer to a cut as an *s-t cut* if $s\in S$ and $t\in T$.

For the minimum flow problem, we define the *capacity* $c[S,T]$ of the *s-t cut* $[S,T]$ as the sum of the forward arcs lower bounds minus the sum of the backward arcs capacities. That is: $c[S,T]=l(S,T)-c(T,S)$

We refer to an *s-t cut* which has the maximum capacity among all *s-t* cuts as a *maximum cut*.

An important theorem is the following [8]:

Theorem 1. If there exists a feasible flow in the network, the value of the minimum flow from a source node s to a sink node t in network G equals the capacity of the maximum *s-t cut*.

For the minimum flow problem, the *residual capacity* $r_m(i,j)$ of any arc $(i,j)\in A$, with respect to a given flow f, is given by $r_m(i,j)=c(j,i)-f(j,i)+f(i,j)-l(i,j)$. By convention, if $(i,j)\in A$ and $(j,i)\notin A$ then we add arc (j,i) to the set of arcs A and we set $l(j,i)=0$ and $c(j,i)=0$. The network $G(f)=(N,A_m)$ consisting only of the arcs with positive residual capacity is referred to as the *residual network* (with respect to the flow f).

In the residual network $G(f)=(N,A_m)$, the *distance function* is a function $d_m:N\rightarrow \boldsymbol{N}$. We say that a distance function is *valid* if it satisfies the following conditions: $d_m(s)=0$ and $d_m(j)\leq d_m(i)+1,\ \forall(i,j)\in A_m$.

We refer to $d_m(i)$ as the *distance label of node i* and we refer to $(i,j)\in A_m$ as an *admissible arc* if $d_m(j)=d_m(i)+1$; otherwise it is *inadmissible arc*. We refer to a path from node s to node t consisting entirely of admissible arcs as an *admissible path*. We say that the distance labels are *exact* if for each node i, $d_m(i)$ equals the length of the shortest directed path from node s to node i in the residual network $G(f)$. We refer to a path in G from the source node s to the sink node t as a *decreasing path* if it consists only of arcs with positive residual capacity. Clearly, there is an one to one correspondence between set of decreasing paths in G and the set of directed paths from s to t in $G(f)$.

A *partial admissible path* is a path from some node i to sink node t consisting solely of admissible arcs. In this case the node i is named *current node*.

The minimum flow problem in a general directed s-t network can be solved in two phases:

- establish a feasible flow f, if it exists;
- from a given feasible flow f, establish the minimum flow f_m.

3. The Shortest Path Algorithms for Solving Minimum Flow Problem

There are three approaches for solving the minimum flow problem [8]: (1) using the shortest decreasing path algorithms; (2) using preflow algorithms; (3) using minmax algorithm.

In this paper we refer to an algorithm following the previous approach (1). To obtain the exact distance labels $d_m(i)$ we perform a breadth-first search of the residual network $G(f)$ leaving from source node s. The proposed solution is for the minimum flow problem in residual networks using the implementation with dynamic trees.

The shortest decreasing path algorithm for the minimum flow problem is as following:

```
Program Fm
begin
let be an admissible flow f
determine the residual network G(f)
obtain the exact distance labels dm(i) in G(f)
j:=t
While dm(t)<n do
If exist an admissible arc (i,j) then
begin
advance(i,j)
If j=s then
begin
decrease
j:=t
```

```
end
end
else retreat(j)
end
```

The procedures are as follows:

```
procedure advance(i,j)
begin
succ(i):=j
j:=i
end

procedure retreat(j)
begin
dm(j):=min{dm(i)+1|(i,j)∈Am}
If j≠t then j:=succ(j)
end

procedure decrease
begin
using the successor indices identify an decreasing path Pm
from the sink node t to the source node s
g:=min{rm(i,j)∈P}
decrease g units of flow along path Pm
update the residual network G(f)
end
```

The algorithm maintains a partial admissible path and iteratively performs *advance* or *retreat* operations from the current node *j*.

If the current node *j* has an admissible arc *(i,j)* we perform an *advance* operation and add arc *(i,j)* in front of the partial admissible path; otherwise, we perform a *retreat* operation and backtrack one arc. We repeat these operations until the partial admissible path becomes an admissible path. We repeat this process until the flow is minim.

The algorithm is characterized by the following two theorems [8]:

Theorem 2. The shortest decreasing path algorithm correctly computes a minimum flow.

Theorem 3. The shortest decreasing path algorithm runs in $O(n^2m)$ time.

4. Computing the Minimum Flow using the Dynamic Tree Implementation

Dynamic trees represent a special type of data structure that permits us to implicitly send flow on paths of length n in O(log(n)) steps on average. By doing so we are able to reduce the computational requirement of the shortest decreasing path algorithm for minimum flows from $O(n^2m)$ to O(nmlog(n)).

The dynamic tree data structure maintains a collection D of node-disjoint rooted trees, each arc having an associated value, called *val*. Each rooted tree is a directed in-tree with a unique root. We refer to the nodes of the tree by using the *predecessor-successor* relationship. For instance, each node i (except the root node) has a unique predecessor, which is the next node on the unique path in the tree from that node to the root; we store the predecessor of node i using a predecessor index *pred(i)*. If *j=pred(i)*, we say that node j is the predecessor of node i and node i is the successor of node j. These predecessor indices uniquely define a rooted tree and also allow us to trace out the unique path from any node to the root. Similarly, we define the *ancestors* and the *descendants* of a node. The *descendants* of a node i consist of the node itself, its successors, successors of its successors and so on. We say that a node is an *ancestor* of all of its descendants. Notice that, according to our definitions, each node is its own ancestor and descendant.

We define the following six operations:

find-desc(j): find and return the last descendant of the node j in the tree;

find-value(j): find and return the value of the tree arc incoming node j. If j is a source node, the operation returns the value ∞.

find-min(j): find and return the descendant w of j with the minimum value of *find-value(w)*. In case of a tie, choose the most distant node w from the tree root.

change-value(j,val): add a real number *val* to the value of every arc along the path from *find-desc(j)* to node j and updates the residual network.

link(i,j,val): This operation assumes that i is a tree root and that i and j belong to different trees. The operation combines the trees containing nodes i and j by making node j the parent of node i and giving arc *(i,j)* the value *val*.

cut(j): break the tree containing node j into two trees by deleting the arc joining node j to its descendant and returning the value of the deleted arc. We perform this operation when j is not a tree root.

The following important result lies at the heart of the efficiency of the dynamic tree data structure [4].

Theorem 4. If z is the maximum tree size (i.e., maximum number of nodes in any tree), a sequence of k tree operations, starting with an initial collection of singleton trees, requires a total of $O(k \cdot \log(k+z))$ time.

The dynamic tree implementation stores the values of tree arcs only implicitly. If we want to store these values explicitly, the operation *change-value* on a tree of size z

might require O(z) time (if this tree happens to be a path), which is computationally excessive for most applications. Storing the values implicitly allows us to update the values in only O(log(z)) time.

Dynamic tree implementation of the shortest augmenting path algorithm is as following:

```
program FmAD
begin
let f be a feasible flow in G
determine the residual network G(f)
obtain the exact distance labels dm(i) in G(f)
let D be the collection of all singleton nodes
j:=t
While dm(t)<n do
begin
If j has an admissible arc (i,j)
then tree-m-advance(j)
else tree-m-retreat(j)
if j=s
then tree-m-decrease
end
end.

procedure tree-m-advance(j)
begin
let (i,j) be an admissible arc in Am
link(i,j,rm(i,j))
j:=find-desc(i)
end

procedure tree-m-retreat(j)
begin
dm(j):=min{dm(i)+1|(i,j)∈Am}
For each tree arc (j,k) do cut(k)
j:=find-desc(t)
```

```
end

procedure tree-m-decrease
begin
p:=find-min(t)
g:=find-value(p)
change-value(t,-g)
While find-value(p)=0 do
begin
cut(p)
p:=find-min(t)
end
j:=find-desc(t)
end
```

The first two procedures, *tree-m-advance* and *tree-m-retreat*, are straightforward, but the *tree-m-decrease* procedure requires some explanation. If node *p* is a descendant of node *t* with the minimum value of *find-value(p)* and, among such nodes in the path, it is the most distant to the sink, then *find-value(p)* gives residual capacity of the decreasing path. The operation *change-value(t,-g)* implicitly updates the residual capacities of all the arcs in the decreasing path. This decrease might cause the capacity of more than one arc in the path to become zero. The WHILE loop identifies all such arcs, one by one, and it deletes them from the collection of rooted trees.

We now consider the worst-case complexity of the algorithm. Why is the dynamic tree implementation more efficient than the original implementation of the shortest decreasing path algorithm? The bottleneck operations in the original shortest decreasing path algorithm are the advance and decrease operations, which require $O(n^2m)$ time. Each advance operation in the original algorithm adds one arc; in contrast, the tree implementation adds a collection of arcs using the link operation. Thus, the dynamic tree implementation substantially reduces the number of executions of the link operation. Similarly, while decreasing flow, the tree implementation decreases flow over a collection of arcs by performing the operation *change-value*, thus again substantially reducing the number of required updates.

Theorem 5. The dynamic tree implementation of the shortest decreasing path algorithm solves the minimum flow problem in $O(nm\log(n))$ time.

Although this result establishes the theoretical utility of the dynamic tree data structure for improving the worst-case complexity of the shortest decreasing path algorithm, the practical value of this data structure is doubtful. The dynamic tree implementation reduces the time for performing *advance* and *decrease* operations from $O(n^2m)$ to $O(nm\log(n))$, but simultaneously increases the time of performing *retreat* operations from $O(nm)$ to $O(nm\log(n))$. Empirical experience shows that retreat operation is one of the bottleneck operations in practice. Since the dynamic tree data structure increases the running time of a bottleneck operation, the use of this data structure actually slows down the algorithm in practice. Furthermore, this data structure introduces substantial overhead (i.e. a large-constant factor of work is associated with each tree operation), thus making it of limited practical utility. Therefore, we propose in the next section a practical improvement for the dynamic tree implementation.

4.1 A Practical Improvement for the Dynamic Tree Implementation

The algorithm terminates when $d_m(t) \geq n$. This termination criterion is satisfactory for the worst-case analysis but might not be efficient in practice. Empirical investigations have revealed that the algorithm spends too much time for relabelling nodes and that a major portion of this effort is performed after the algorithm has established a minimum flow. This happens because the algorithm does not know that it has found a minimum flow. We next suggest a technique that is capable of detecting the presence of a maximum cut and so the existence of a minimum flow much before the label of node t satisfies the condition $d_m(t) \geq n$. Incorporating this technique in the algorithm improves its performance substantially in practice.

To implement this approach, we maintain an n-dimensional additional array q, whose indices vary from 0 to $(n-1)$. The value $q(k)$ is the number of nodes whose distance label equals k. The algorithm initializes this array while computing the initial distance labels using a breadth first search. At this point, the positive entries in the array q are consecutive (i.e. the entries $q(0), q(1), \ldots, q(l)$ will be positive up to the index l and the remaining entries will all be zero). Subsequently, whenever the algorithm increases the distance label of a node from the k_1 to k_2, it reduces 1 from $q(k_1)$, adds 1 to $q(k_2)$ and checks whether $q(k_1)=0$. If $q(k_1)$ does equal zero, the algorithm terminates.

In order to see why this termination criteria works, let $S=\{i \in N \mid d_m(i)<k_1\}$ and $T=\{i \in N \mid d_m(i)>k_1\}$. It is easy to verify that $s \in S$ and $t \in T$. Now we consider the *s-t* cut $[S,T]$. The definitions of the sets S and T imply that $d_m(j)>d_m(i)+1$ for all $(i,j) \in [S,T]$. The validity condition $d_m(j) \leq d_m(i)+1$ for every arc (i,j) in the residual network $G(f)$ implies that $r_m(i,j)=0$ for each arc $(i,j) \in [S,T]$. Therefore, $[S,T]$ is a maximum cut and the current flow is a minimum flow.

This practical improvement modifies the algorithm FmAD from previous section in three points: (1) initialize with zero all positions in the *q-array*; (2) transform this vector such as every position $q(k)$ equals the number of nodes whose distance label is k; (3) modify the *tree-m-retreat* procedure as following:

```
procedure tree-m-retreat(j)
begin
k1:=dm(j)
dm(j):=min{dm(i)+1|(i,j)∈A, rm(i,j)>0}
k2:=dm(j)
For each tree arc (j,k) do cut(k)
```

```
q(k1):=q(k1)-1
q(k2):=q(k2)+1
If q(k1)=0
then exit
else j:=find-desc(t)
end
```

5. Example

The network *G=(N,A)* is represented in *Figure 1* and the corresponding residual network $G(f)=(N,A_m)$ is in *Figure 2*. Initially, the dynamic tree contains the collection of singleton six nodes.

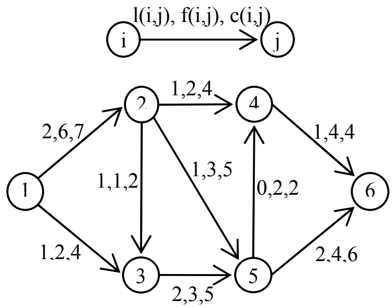


Figure 1 The initial network

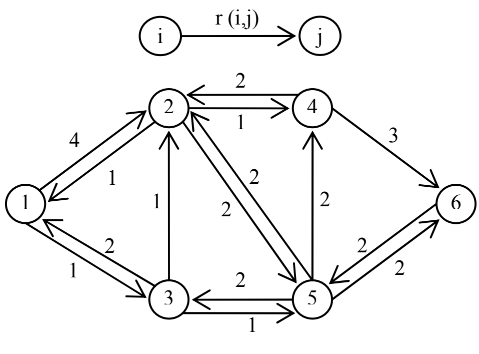


Figure 2 The initial residual network

We determine the exact distance labels for the nodes in the network and obtain $d_m=(0,1,1,2,2,3)$ and $q=(1,2,2,1,0,0)$. We also set the current node as the sink node *t*.

In the first iteration the algorithm selects the sink node *6* and choose *(5,6)* as the admissible arc in the residual network. We apply a *tree-m-advance* procedure; consequently, we add the tree arc *(5,6)* with value *2* by applying *link(5,6,2)* and we set *j:=5* following the operation *j:=find-desc(5)*. In the next iteration we add the tree arc *(2,5)* with value *2* and we set *j:=find-desc(2)=2*. Next, we apply *link(1,2,4)*. Because the current node is the source node, we apply a *tree-m-decrease* procedure on admissible path *(1,2,5,6)*. The tree is represented in *Figure 3*.

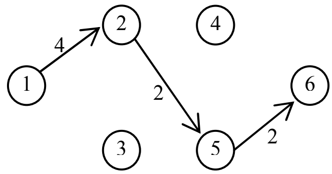


Figure 3 The dynamic tree

We determine value of *p* in the following manner: *p:=find-min(6)=5* with *g=find-value(5)=2*. The operation *change-value(6,-2)* decreases *2* units from every tree arc. In this moment, the arcs *(5,6)* and *(2,5)* have value *0*, so we delete these arcs from the tree (see the *cut(5)* and *cut(2)* operations). The dynamic tree is now represented in *Figure 4*; the update residual network is represented in *Figure 5*.

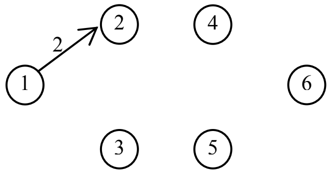


Figure 4 The dynamic tree

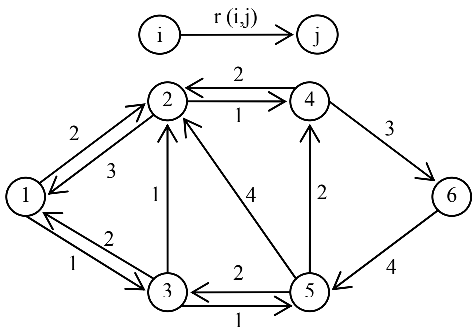


Figure 5 The update residual network

Next we set *j:=find-desc(6)=6*, apply the *link(4,6,3)* and *link(2,4,1)* operations and decrease *1* unit of flow from every arc on path *(1,2,4,6)*. At the end of this iteration the current node is *j=1*, so by start with *tree-m-decrease* procedure, we determine the dynamic tree of *Figure 6* and the residual network of *Figure 7*.

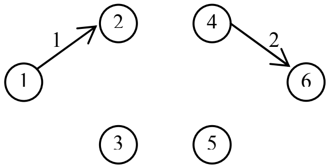


Figure 6 The dynamic tree

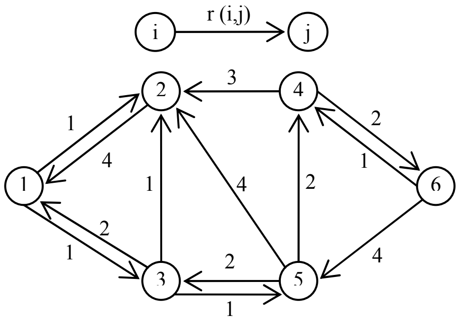


Figure 7 The second update residual network

The current node is now *j:=find-value(6)=4*. In the next iteration we don't have an admissible arc from the node *4*, so we apply the *tree-m-retreat(4)* procedure and determine $k_1=2$, $d_m(4)=min\{d_m(5)+1,\ d_m(6)+1\}=3$, $k_2=3$, $q(2)=2-1=1$, $q(3)=1+1=2$ and $q_1=(1,2,1,2,0,0)$. Now, we delete the arc *(4,6)* from the tree and set j:=6. There are not exist an admissible arc start from node *6*, so we apply again *tree-m-retreat(6)* procedure and obtain $k_1=3$, $d_m(6)=min\{d_m(4)+1\}=4$, $k_2=4$, $q(3)=2-1=1$, $q(4)=0+1=1$ and $q_2=(1,2,1,1,1,0)$.

Now we set *j:=6* and apply the *link(4,6,2)*, *link(5,4,2), link(3,5,1)* and *link(1,3,1)* operations. We obtain the directed path *(1,3,5,4,6)* with *p=3*, *g=1* and apply the *change-value(6,-1)* operation. We delete the arcs *(3,5)* and *(1,3)* from tree. Now, the dynamic tree and the residual network are represented in *Figures 8* and *Figure 9*.

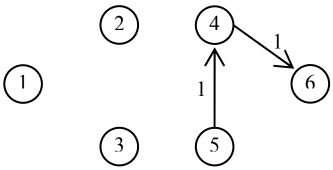


Figure 8 The dynamic tree

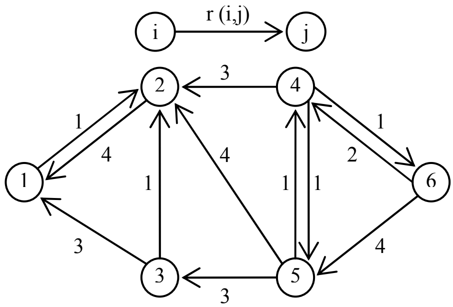


Figure 9 The residual network

The current node is *j:=5*, there not exist an admissible arc start from the *5*, so the *tree-m-retreat(5)* procedure set $k_1=d_m(5)=2$, $d_m(5)=min\{d_m(4)+1, d_m(6)+1\}=4$, $k_2=4$, $q(2)=1-1=0$, $q(4)=1+1=2$ and $q_3=(1,2,0,1,2,0)$ The algorithm is terminated because $q(k_1)=0$. The minimum value of flow is $v=4$. The minimum network flow is represented in *Figure 10*.

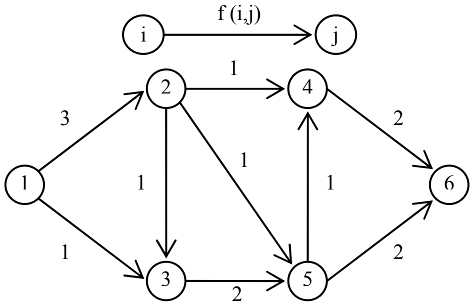


Figure 10 The minimum network flow

6. Conclusions

Our dynamic tree implementation is more efficient than the original implementation of the shortest decreasing path algorithm. An important argument for the value of our result refers the improvement of the running time of the algorithm from $O(n^2m)$ to $O(nm\log(n))$ time.

Usually, this data structure introduces substantial overhead (i.e. a large-constant factor of work is associated with each tree operation), thus making it of limited practical utility. Therefore, we propose a practical improvement for the dynamic tree implementation. Incorporating this technique in the algorithm one may improve its performance substantially in practice. The most representative examples of flow problems in networks refer to: petroleum products in a pipeline network, cars in a road network, messages in a telecommunication network and electricity in an electrical network.

References

1 Georgescu O. Fluxuri în rețele planare şi bipartite. Report as PhD Candidate. Transilvania University of Braşov. 2004.

2 Georgescu O. Fluxuri în rețele cu capacitate unitate şi rețele simple. Report as PhD Candidate. Transilvania University of Braşov. 2004.

3 Georgescu O. Fluxuri în rețele seriale şi aciclce. Report as PhD Candidate. Transilvania University of Braşov. 2005.

4 Ahuja R, Magnanti T, Orlin J. Network Flows: Theory, algorithms and applications. Prentice Hall, Inc., Englewood Cliffs, NY, 1993.

5 Sleator D, Tarjan R. A data structure for dynamic trees. J. Comp. System Sci. 1983; 26: 362-91.

6 Goldberg A, Tarjan R. A new approach to the maximum flow problem. ACMJ 1988; 35: 921-40.

7 Ahuja R, Orlin J, Tarjan R. Improved time bounds for the maximum flow problem. SIAMJ 1989; 18: 939-54.

8 Ciurea E, Ciupală L. Sequential and parallel algorithms for minimum flows. AMCJ 2004; 15(1-2): 53-75.

9 Ciurea E, Georgescu O. Minimum flows in unit capacity networks. Analele Universității Bucureşti 2006; XLV: 11-20.

10 Ciurea E. Algoritmi: Introducere în algoritmica grafurilor. Editura Tehnică Bucureşti 2001

Evaluation Models of Thermal Comfort in Vehicles

Radu Musat[1]*, Elena Helerea*[2]

[1]*Renault Technologie Roumanie, 2/III Pipera-Tunari Str, Voluntari, Romania, r_musat@yahoo.com*

[2]*The University "Transilvania" Brasov, Dept. of Electrical Engineering, 29 Eroilor Str, Brasov, Romania, helerea@unitbv.ro*

At present, efforts are made to estimate the thermal comfort by direct measurement of each parameter microclimate - air temperature inside the vehicle, the humidity of air, mean radiant temperature, air velocity, human level activity and clothing insulation - either through the indices comfort that integrates contribution of each parameter microclimate of the vehicle. The model used for estimate the thermal comfort is proposed by Fanger with the considering of heat balance equation of the human body in vehicle environment. Other models are based on calculations with experimental data and simulations. In this paper are analyzed the conditions extrapolation of the Fanger's model in the vehicle environment. There are analyzed the limitations related to: the thermal steady-state or thermal dynamic state, distinction between local and whole-body thermal comfort, particularities of environmental vehicle that is dynamic and nonuniform.

Keywords

factors and parameters, manikin, modeling, thermal comfort, vehicle.

1. Introduction

In last years, with the trends of reducing costs, weight and safety of vehicles, increased the interest in ensuring an optimal comfort in vehicles. Constructions of vehicles have developed from simplistic to modern cars with last hour technologies, held on the functional and aesthetic criteria, which ensure to the passengers comfort, ergonomics and safety.

Factors that determine the comfort in vehicles are [1]:

- construction factors: the types and dimensions of vehicle interiors, all the installations and equipment of the interior;
- physical factors: microclimate (composition of air, humidity, temperature), lighting, color, noise, aesthetics and hygiene.

Thermal comfort [2] in vehicles represent subjective sensation of heat balance that occurs in the human body when environmental parameters of vehicles - air temperature, air humidity, radiant temperature, air velocity, human level activity and clothing insulation - are in a range of well-defined values.

Thermal comfort is achieved by ensuring temperatures (20 ÷ 22)°C, the resulting temperature air, delimitation areas, humidity and air velocity, consistent with the activity level and clothing insulating of occupants of the vehicle, by avoiding situations in that the occupants come into contact with too cold or too hot surfaces, and avoiding air currents. These requirements must be met along the year, both in winter and in summer. It aims to determine the speed of temperature change in vehicle, in particular, measuring the difference between the temperature recorded at the feet and temperature at the head.

Air velocity inside the vehicle is usually low, with values ranging 0.1 and 0.4 m/s. For values below 0.1m/s, will create a dry air sensation. Because of fluctuations of air velocity, the measurements in a particular item must be carried out over a period of 3 or 5 minutes to obtain a reasonable average value. Appearance of air currents is mostly due tightness.

Mean radiant temperature: the uniform surface temperature of an imaginary black enclosure in which an occupant would exchange the same amount of radiant heat as in the actual non-uniform space (ASHRAE 55). The mean radiant temperature could be measured if walls temperatures and the position for the occupants are known.

Relative humidity: the ratio of the mole fraction of water vapour present in the air to the mole fraction of water vapour present in saturated air at the same temperature and barometric pressure; alternatively, it equals the ratio of the partial pressure of the water vapour in the air to the saturation pressure of water vapour at the same temperature (ASHRAE 55). Relative humidity is measured in only one place inside the vehicle because the pressure of water vapour is uniform in entire vehicle. The comfort sensation is optimal when the relative humidity has the value of 50%.

When all these parameters are measured, can determine their combined effect on the occupants of the vehicle.

Measuring each parameter requires a lot of instrumentation and it is difficult to measure all the parameters in the exact same location and then later calculate the combined influence. Using a transducer that measures the combined effect of all climatic parameters, the equivalent temperature, makes the evaluation much easier.

At present, efforts are made to estimate the thermal comfort by direct measurement of each parameter microclimate, either through comfort indices comprising the combined influence of some of the factors involved, either on the basis of theoretical models [3]-[5].

The thermal environment in a vehicle cabin is very complex and thus difficult to evaluate. These difficulties are due to the influence of convective, radiative and conductive heat exchange created by external thermal loads and the internal heating and ventilation system [6]-[8]. Last, but not least, neither driver nor passengers are able to make up for the asymmetric climate conditions that set high demands for the Instrument and Transducers.

The usual method of evaluating the efficiency of the air conditioning system in vehicles is to apply sensors to measure the air temperature at feet and at head level. The main purpose being to investigate how quickly the system will raise or lower the temperature in a warm or cold vehicle and to study the difference between the temperature at feet and head level. However, using this method only one of the four needed parameters (air temperature, mean radiant temperature and air velocity) that concern the thermal comfort sensation, is measured. This fact is especially unfortunate in vehicles as the mean radiant temperature differs more from the air temperature that in buildings and because the air conditioning system can create high local air velocities.

In the present article is an analysis of the models, Fanger and thermal manikin, used to estimate the thermal comfort in vehicles and settled limitations that occur in estimating the parameters of thermal comfort using different models applicable to vehicles.

2. Fanger's Model in Automotive Environments

The most notable figure to use such a human heat balance equation in the application of thermal comfort analysis was P.O. Fanger. His seminal work, *Thermal Comfort*, published in 1970, outlines the required conditions for thermal comfort and presents an equation to predict what proportion of an average population would find a given environment decidedly uncomfortable [9]. Fanger's model uses the heat balance equation.

Fanger introduces following concept [10]:

- Stress parameters under extreme thermal environment;
- Parameters that include the combined effect of several parameters of the environment;
- PMV (predicted mean vote);
- PPD (predicted percentage dissatisfied).

A vehicle represents a "moderate" thermal environment; as such, Fanger's equation has been the most widely used for automotive research (as Fanger's equation has been validated numerous times for moderate environments), where thermal comfort analyzed by *PMV* (*Predicted Mean Vote*), and thermal discomfort can be analyzed by *PPD* (*Predicted Percentage Dissatisfied*) [9]-[11].

His *predicted mean vote* (PMV) and *predicted percentage dissatisfied* (PPD) became the basis for several standards on thermal comfort, the most significant ones being ISO 7730 (1994), ISO 9920 (1995) [12]-[15] and ASHARE 55 (1992), ASHARE (2001) [16], [17].

Though one could theoretically give this scale to groups of subjects, the ultimate aim of Fanger's model is to use the thermal comfort equation to calculate the PMV for any particular environment. Fanger's model suggests that thermal comfort can be predicted if the values of six parameters are known (air temperature, humidity, air velocity, radiant temperature, occupant clothing level, and occupant activity level).

The PMV is based on the following ASHARE (American Society of Heating, Refrigerating and Air Conditioning Engineers) subjective scale with seven steps [18].

The value of the PMV index has a range from -3 to +3, corresponding to human sensations from cold to hot, respectively, where the null value of the PMV index means neutral. Figure 1 shows a combination of each thermal variable affecting the PMV level.

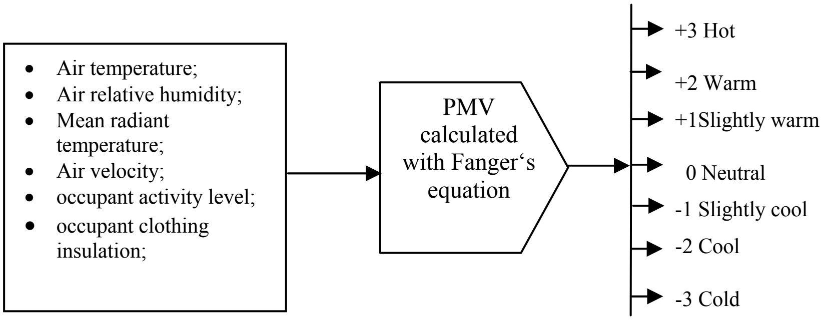


Figure 1 Thermal Sensation Scale

The PMV evaluation method treats the entire body as one object [19-20]. It does not distinguish between different parts of the body. If one side is warm and the other cold, the body would have a zero thermal load, and therefore it obtain a neutral thermal sensation (PMV=0). Because it only calculates the heat transfer for the entire body, it cannot predict local discomfort. Clothing is assumed to cover the entire body uniformly which results in one equal skin temperature across the entire body. Obviously, the local effects of asymmetric conditions, such as an environment with a hot or a cold window, or local air motion around a person's face provided by a fan, are lost in this whole-body model.

There is the distinction between local thermal comfort and whole-body thermal comfort. Whole-body thermal comfort precludes local discomfort; that is, for an individual to be "comfortable," there can be no single part of the body that feels uncomfortable. Local thermal discomfort can be caused by high radiant temperature asymmetry, temperature differences across the body (usually vertical), or contact with hot or cold surfaces, but it is usually considered in terms of *draught*, or local discomfort caused by air movement (ISO 7730). Unfortunately, all of these phenomena are common to a vehicle cabin. For instance, radiant heat from the windows can affect the upper part of the body while air currents toward the floor can cause local discomfort for the feet.

The automobile environment is non-uniform and dynamic and it is likely that most occupants will experience discomfort from more than one source simultaneously.

The equations that led Fanger to develop the concept of PMV and PPD are based on the physiological processes that underlie human heat balance. The interaction between the human body and the environment is described by the heat balance equation between thermal heat developed by metabolism in the human body and the heat transferred through convective, conduction, radiation and evaporation [21]:

$$M - L = Q_{sk} + Q_{res} = \left(C + R + E_{sk}\right) + \left(C_{res} + E_{res}\right) \tag{1}$$

where:

M – metabolic energy production [Wm^{-2}]; L – external work [Wm^{-2}]; Q_{sk} – heat loss through skin [Wm^{-2}]; Q_{res} - heat loss through respiration [Wm^{-2}]; C – convective heat loss through skin [Wm^{-2}]; R – radiative heat loss through skin [Wm^{-2}]; E_{sk} - total evaporation heat loss through skin [Wm^{-2}]; C_{res} – convective heat loss through respiration [Wm^{-2}]; E_{res} – evaporative heat loss through respiration [Wm^{-2}].

The heat production within the human body is related to the heat balance equation:

$$Q_a = M - L - E_{sk} - \left(C_{res} + E_{res}\right) - \left(C + R\right) \tag{2}$$

- if Q_a=0, thermal comfort is properly;
- if Q_a>0, body temperature rises and passenger will have a warm sensation;
- if Q_a < 0, body temperature falls, and passenger will have a cold sensation.

For $Q_a = 0$, the heat balance equation is:

$$\begin{aligned} & M \times (1-\eta) - 0.35 \cdot \left[43 - 0.061 \cdot M \times (1-\eta) - p_a\right] - 0.42 \cdot \left[M \times (1-\eta) - 50\right] - \\ & - 0.0023 \cdot M \cdot (44 - p_a) - 0.0014 \cdot M \cdot (34 - t_a) = \frac{35.7 - 0.032 \cdot M \cdot (1-\eta) - t_h}{0.155 \cdot R_{CL}} \\ & = 3.4 \times 10^{-8} f_{CL} \cdot \left[(t_{CL} + 273)^4 - (t_r + 273)^4\right] - f_{CL} \cdot h_C \cdot (t_{CL} - t_a) \end{aligned} \tag{3}$$

The value of PMV can be determined by following equation:

$$PMV = \left(0.352\, e^{-0.042 \cdot \frac{M}{A_D}} + 0.032\right) \cdot \left[\begin{aligned} & \frac{M}{A_D} \times (1-\eta) - 0.35 \cdot \left[43 - 0.061 \cdot \frac{M}{A_D} \times (1-\eta) - p_a\right] - \\ & - 0.42 \cdot \left[\frac{M}{A_D} \times (1-\eta) - 50\right] - 0.0023 \cdot \frac{M}{A_D} \cdot (44 - p_a) - \\ & - 0.0014 \cdot \frac{M}{A_D} \cdot (34 - t_a) - 3.4 \times 10^{-8} f_{CL} \cdot \\ & \cdot \left[(t_{CL} + 273)^4 - (t_r + 273)^4\right] - f_{CL} \cdot h_C \cdot (t_{CL} - t_a) \end{aligned}\right] \tag{4}$$

$$t_{CL} = 35.7 - 0.032 \cdot \frac{M}{A_D}(1 - \eta) - 0.18 R_{CL} \cdot \left\{ \begin{aligned} & 3.4 \times 10^{-8} \cdot f_{CL}\left[(t_{CL} + 273)^4 - (t_r + 273)^4\right] - \\ & - f_{CL} \cdot h_C \cdot (t_{CL} - t_a) \end{aligned} \right\} \quad (5)$$

$$h_C = \begin{cases} 2.05 \cdot (t_{CL} - t_a)^{0.25} & \text{pentru} \quad 2.38 \cdot (t_{CL} - t_a)^{0.25} > 10.4 \cdot \sqrt{v} \\ 10.4 \cdot \sqrt{v} & \text{pentru} \quad 2.38 \cdot (t_{CL} - t_a)^{0.25} < 10.4 \cdot \sqrt{v} \end{cases} \quad (6)$$

Variables for Fanger equations:

$\frac{M}{A_D}$ - metabolic rate;

M – activity level, $\left[\frac{\text{kcal}}{\text{h} \cdot \text{m}^2}\right]$;

A_D – body area (mean value for adults are between 1.65 and 2 m^2);

η – mechanical efficiency;

R_{CL} – thermal resistance of clothing [clo],

[1 clo] = $\left[0.180 \cdot \frac{\text{C} \cdot \text{m}^2 \cdot \text{h}}{\text{cal}}\right]$;

f_{CL} – clothing area factor (surface area of the clothed body divided by surface area of the nude body);

t_{CL} – mean temperature over the clothed body, [ºC];

t_a – air temperature, [ºC];

t_r – mean radiant temperature, [ºC];

hC – convective heat transfer coefficient, $\left[\frac{\text{kcal}}{\text{m}^2 \cdot \text{h} \cdot {}^\circ\text{C}}\right]$;

v – air velocity, $\left[\frac{\text{m}}{\text{s}}\right]$;

p_a – pressure of water vapour, [Pa].

The PPD *predicted percentage dissatisfied* is a quantitative measure of thermal discomfort of a group of people in a thermal environment. Fanger relates predictable percentage dissatisfaction PPD with the index PMV as follows:

$$PPD = 100 - 95\mathrm{e}^{-\left(0.03353 \cdot PMV^4 + 0.2179 \cdot PMV^2\right)} \tag{7}$$

Theoretically, the index of thermal comfort are the best PMV zero, but standards (for example, ISO-77300) considered the field of thermal comfort corresponds to optimal range of values between -0.5 and +0.5.

When the PMV equals zero (or the environment is rated as "neutral"), the PPD should equal five. This means that for any comfortable environment, approximately five percent of an average population will still rate it as uncomfortable. When the PMV equals –2 or +2, the PPD will equal 80, meaning that 80% of the population will be dissatisfied, and so on.

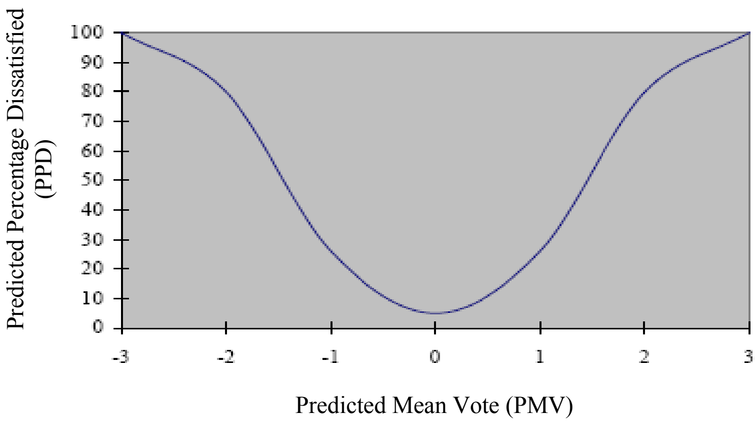


Figure 2 PPD as a function of PMV

The relationship between PMV and PPD is thus a U-shaped curve, as illustrated in Figure 2. It is noted that the optimum value for PPD is 5% and can be obtained only with automatic air-conditioning systems.

Relations (1) - (7) allow the calculation of indices PMV and PPD. Table 1 is the result of calculation presented PMV and PPD indices for all parameters:

- air temperature: $t_a = 24$ ºC;
- mean radiant temperature: $t_r = 24$ ºC;
- pressure of water vapor: $p_a = 10^3$ Pa;
- air velocity: $v = 0{,}15$ m/s.

M	R_{CL}	PMV	PPD
[Wm^{-2}]	[m^2 ºC W^{-1}]	-	[%]
50	0,130	-1,0	27,7
58	0,155	0,0	5,0
66	0,180	0,4	8,8
85	0,130	0,5	10,5
100	0,155	0,9	22,6
115	0,180	1,2	36,4

Table 1 Calculation of PMV and PPD indices

Results of table 1 indicate a significant dependence on the indices *PMV* and *PPD* of thermal comfort in the vehicle characteristics of metabolic passengers.

3. Thermal Manikins

Measurement and assessment of thermal climate using a thermal manikin will make it possible to evaluate the best solution for thermal control. It can also be used to measure clothing and chair insulation.

Thermal manikins are currently available but they are used primarily for measuring the thermal insulation value of clothing. A few thermal manikins have been developed for thermal comfort research but they have very limited capabilities (Wyon 1989, Nilsson 1999, Meinander 1999, Madsen 1999). They utilize the same basic concept in that the heating power required to keep the manikin surface at a constant temperature is measured and used to correlate with thermal comfort. The problem with this approach is that the manikin does not respond to the environment like the human body. A human unconsciously varies the local skin temperature and local rate of heat transfer to control the physiological response, which couples to the sensory input system to affect the perception of thermal comfort. Most current manikins are primarily designed for steady state operation and possess long thermal response times. They also suffer from limited spatial resolution for sensing and response to the environment. The resolution is typically limited to a body segment such as the upper leg or lower leg, but this level is not adequate for highly non-uniform thermal fields, and to develop focused climate control systems to maximize energy efficiency. Most current manikins do not possess a sweating capability and hence only sense dry heat transfer. Evaporative cooling is a critical and often used component of the thermoregulatory system of the body. A thermal manikin should possess this capability in order to accurately simulate the response of the body in all thermal environments. A sweating system is desired that will generate a film of water on the skin surface, affecting the actual rate of evaporation to the environment, or transport into clothing. A thermal manikin that possesses a high degree of sensory spatial resolution, local thermoregulatory responses including sweating, a fast time response, and a feedback loop to continuously react and adjust to a thermal environment like a human has never been developed. An advanced thermal manikin with these capabilities would help industry develop more effective and energy efficient climate control systems for transportation environments, or others where transient and extremely non-uniform thermal environments exist.

A thermal manikin is a useful tool which can be used to realistically and objectively assess the environmental thermal impact on the human body, accurately measure heat exchange with reproducible results, and take cost-effective and comparative measurements.

Many kinds of thermal manikins have been developed in the world and have been used to evaluate the protective qualities, performance and comfort of cold-weather, fire-protection, and biohazard-protection wear, as well as hats, gloves, footwear, and sleeping bags. The development and improvement of thermal manikins as an evaluating system of clothing is significant from the standpoint of industrial health. A thermal manikin needs to have the following properties in order to accurately simulate the human body: (1) correct body shape and size; (2) control of heat emission; (3) control of the distribution of heat across the skin surface; (4) emission of the skin; (5) control of the distribution of perspiration across the skin surface; (6) control of pose and movement; (7) control of core and shell differently to simulate the physiological responses of the human body. So far, no manikin meeting all these criteria has been available. Recently manikins have been developed where perspiration across the skin surface and walking speed can be controlled in some laboratories. However, a two-layer model consisting of a core layer and a shell layer whose temperatures can be independently controlled by a computer has, until now never been developed or tested. Depending on some situation, for example, in the case when wearing cold-weather gear in cold conditions, existing thermal manikins are limited to a uniform temperature distribution across the skin surface even though the human body's extremities experience large drops in skin temperature.

This inevitably leads to an overestimation of heat loss from the extremities in the result obtained using the thermal manikin. The object of this research is to develop the two-layer movable sweating thermal manikins and bring the measurements into closer agreement with what an actual human body would be experiencing.

Measurements of local climate disturbances with a man-sized thermal manikin are well correlated with the thermal sensation experienced by subjects exposed to the same conditions. Criteria for acceptable climatic conditions can be defined in terms of quantities measured with the manikin. The manikin method represent a quick, accurate and reproducible technique for reliable and cost effective assessment of many of the complex details of the climate in a vehicle and their integrated effects on humans.

4. Conclusion

Developed mathematical models offer the possibility of carrying out analysis of the thermal comfort in cars, being a real use in research activity in the automotive field.

Fanger's model combines four physical variables (air temperature, mean air velocity, air humidity and mean radiant temperature) and two individual variables (occupant clothing level, and occupant activity insulation) into an index that predicts the percentage of occupants which has optimal thermal comfort sensation. By measuring all these thermal comfort parameters, can calculate the combined effect on the occupants of the vehicle.

The PMV parameter doesn't estimate properly the thermal comfort sensation. Differences between the actual and the estimated temperature reflect difficulties in obtaining accurate measurements of the thermal insulation of clothing and metabolic rate. Basically, inadequate estimates of these two variables reduce the accuracy of the index PMV. The PMV depends on the context and more specifically in vehicle with air-conditioning system than in natural ventilation and that because of the influence of outside temperature.

Fanger model has limitations related to thermal steady state or dynamic state, distinction between local and whole-body thermal comfort, environmental particularities of the vehicle. Other limitations are related to the local effects of asymmetrical conditions or local air movement around the face of occupants.

A physical model with thermal manikins has limitations related to:

- possess long response time;
- limited resolution on perception;
- limited resolution on response to the environment;
- doesn't possess sweating capacity but only the dry heat transfer.

Although are some limitations, the method with thermal manikins represents a rapid, reliable and efficient model for evaluation the thermal comfort to assessment for disclosure of details related to thermal and climatic comfort of cars and their effect on the human body.

References

1 Brujan, E.,A., Ventilation and Air conditioning.

2 Sârbu, I., Thermal comfort evaluation models, Theoretical considerations, nr.2 (43), 2007, University of Timişoara.

3 S. Atthajariyakul, T. Leephakpreeda, Neural computing thermal comfort index for HVAC systems, School of Manufacturing Systems and Mechanical Engineering, Sirindhorn International Institute of Technology, Thammasat University, Patumthami, Thailand.

4 John P. Rugh, Desikan Bharathan, Predicting human thermal comfort in automobiles, National Renewable Energy Laboratory, NRELl/CP-540-37784, presented at the Vehicle Thermal Management Systems Conference and Exhibition, May 2005, Toronto, Canada.

5 Itsuhei Kohri1, Tohru Mochida, Evaluation Method of Thermal Comfort in a Vehicle with a Dispersed Two-Node Mode Part 1— Development of Dispersed Two-Node Model, Journal of the Human-Environmental System, vol.6, no.1, 2002.

6 McGuffin, R., Burke, R., Huizenga, C., Hui, Z., Vlahinos, A., Fu, G., Human Thermal Comfort Model and Manikin, National Renewable Energy Laboratory, Society of Automotive Engineers, Inc, 2002, Colorado, USA.
7 * * *, Thermal Confort in Automobiles, LumaSense Technologie.
8 ISO 7726 (1998). Ergonomics of the thermal environment—instruments for measuring Physical quantities, Geneva: International Standards Organization.
9 Fanger, P., O., Thermal Comfort: Analysis and Applications in Environmental Engineering, McGraw-Hill Book Company, 1970, USA.
10 Fanger, P., O., Thermal Comfort. McGraw-Hill Book Co., 1973, New York, USA.
11 Joel M. Devonshire, James R. Sayer, The effects of infrared-reflective and antireflective glazing on thermal comfort and visual performance: a literature revie, University of Michigan, Transportation Research Inst., Report no.UMTRI-2002-4 March 2002, U.S.A.
12 ISO 7730. (1994) Moderate thermal environments—determination of the PMV and PPD indices and specification of the conditions for thermal comfort. Geneva.
13 ISO 9920 (1995). Ergonomics of the Thermal Environment: Estimation of the Thermal Insulation and Evaporative Resistance of a Clothing Ensemble. Geneva.
14 Tri Harso Karyono, The applicability of the ISO 7730 (Fanger's comfort model) and the adaptive model of thermal comfort in Jakarta, Indonesia, Agency for the Assessment and Application of Technology, Republic of Indonesia.
15 Parsons,K.,C., Introduction to thermal comfort standards, Loughborough University, UK
16 American Society of Heating Refrigerating and Air Conditioning Engineers, Thermal Environmental Conditions for Human Occupancy (ASHRAE Standard 55),1992, Atlanta.
17 American Society of Heating Refrigerating and Air Conditioning Engineers (ASHRAE), Thermal comfort. In ASHRAE Handbook: Fundamentals, 2001, Atlanta, GA.
18 Charles,K.E. Fanger's Thermal Comfort and Draught Models, Institute for Research in Construction, Report RR-162, October, 2003, Canada, Ottawa.
19 Joel M. Devonshire, James R. Sayer, Radiant heat and thermal comfort in vehicles, The University of Michigan, Transportation Research Institute, Report No. UMTRI-2003-32, November 2003, Michigan, U.S.A.
20 Huizenga, C. Zhang, H., Mattelaer, P., Tiefeng, Yu, Arens, E., Window performance for human thermal confort- final report, The National Fenestration Rating Council, february 2006, Center For the Built Environment, University of California, Berkeley.
21 Parsons, Ken, Human thermal environments – second edition, Taylor & Francis Group, London, ISBN 0-415-23793-9.

Automatic Acquisition of Synonyms Using the Web as a Corpus

Svetlin Nakov

Sofia University "St. Kliment Ohridski", Department of Mathematics and Informatics, 5 James Baucher Bvld., Sofia, Bulgaria, nakov@fmi.uni-sofia.bg

We present an original algorithm for automatic acquisition of synonyms from text. The algorithm measures the semantic similarity between pairs of words by comparing their local contexts extracted from the Web by series of queries against the Google search engine. The results show 11pt average precision of 58.98%.

Keywords

automatic synonym acquisition, semantic similarity measure, Web as a corpus, Web mining.

1. Introduction

Synonyms are important for solving various problems in Natural Language Processing (NLP) such as text summarization, question answering, text generation, search query expansion, etc.

In the present paper, we set the objective to design an algorithm for automatic extraction of pairs of synonyms from a text corpus. The results can be used to create linguistic resources, such as general and domain-specific thesauri and lexicons.

We use the Web as a large corpus which can be efficiently searched. Our approach is based on performing series of queries against a Web search engine and analyzing the returned excerpts of texts (snippets) in order to extract contextual semantic information which we use to measure the semantic similarity between pairs of words and thus to approximate synonymy.

It is considered that the *local context* of a given word (few words before and after the target captured word) contains words that are semantically related to it [Hearst, 1991]. Given a pair of words, we extract their local contexts from the snippets returned by the search engine and we measure the semantic similarity between these words by calculating the similarity between their local contexts. Finally, the measured similarity is used to determine whether the words are likely to be synonyms or not.

The algorithm used for measuring semantic similarity is an adaptation of the algorithm for measuring cross-lingual semantic similarity described in [Nakov et al., 2007a].

In the performed experiments we process Russian texts used for teaching students studying fine arts. We chose Russian and fine arts terminology because of the high volume of such texts available in Internet and the great number of full synonyms in this domain. While the algorithm is general enough and should work for other languages, our present experiments are limited to Russian only.

We start with extracting a list of all terms from the text that are interesting from a linguist's point of view. We can also use a subset of them, e.g. nouns only, or all words in the text.

Using a series of queries against Google, we automatically measure the semantic similarity between each pair of words from the list. Our hypothesis is that synonyms should have higher level of semantic similarity compared to nonsynonyms. The results of our experiments show that this expectation is true in most cases.

In this paper, we show that it is possible (with a minimum human intervention) to extract automatically all pairs of synonyms from a list of terms built from a terminological text. We propose few modifications of the algorithms for measuring semantic similarity using the Web and we study how different parameters affect the quality of the results (precision and recall).

2. Method for Automatic Synonyms Extraction

Our algorithm for automatic extraction of synonyms from a list of words is based on measuring the semantic similarity between pairs of words by querying a Web search engine (e.g., Google) and analyzing the returned results. The semantic similarity is a number between 0 and 1 indicating the degree of similarity.

The words used to find synonyms come as a list. It is possible to process all words in the text or some subset of them. For example, in order to avoid unnecessary computations, we can use grammatical glossary to filter out words belonging to different parts of speech.

The algorithm measures semantic similarity between each pair of words from our list. Our hypothesis is that synonyms, being words with very similar meanings, should have higher semantic similarity than pairs of nonsynonyms. If we order all pairs of words by their measured semantic similarity, we can expect to obtain identical words in the beginning, followed by synonyms, followed by partial synonyms, followed by other words which are less similar by meaning (like hypernyms and hyponyms), and finally all words that are entirely different.

Since our semantic similarity measures to what extent two words have a similar meanings, it is possible to get inaccurate results for some pairs of words and incorrectly to classify them as synonyms. For example *абрис* and *контур* are semantically related because both mean the same concept (*contour* in Russian), but at the same time words like *синий* (*blue* in Russian) and *красный* (*red* in Russian) are also semantically related because both are colors. Therefore, extracting synonyms by measuring semantic similarity only is not possible without human intervention, but our experiments show that this intervention could be minimal.

2.1 Semantic Similarity Measured by Contexts

The algorithm for measuring semantic similarity between two words is based on analysis of the local context in which these words appear and follows the idea that words appearing in similar context should have similar meanings. For example the words *художник* (*artist* in Russian) and *картина* (*panting* in Russian) are semantically related since both appear in sentences about artists, painters, painting, pictures, brushes, tints, art galleries, and other terms from the fine arts.

Some sentences can be quite long, and it is not clear what part of them contains the context of the given word. Most linguists consider only the so called *local context* of the given word in a sentence which consists of few words before and after that word. As an example let us examine the word *painting* in the following sentence:

> You will learn watercolor techniques, oil *painting* techniques, chalks and freehand styles of *painting*, guided by Jane who has over 25 years of experience as a professional portrait artist and painter.

The local context (e.g. three words before and after it) of the word *painting* in the above sentence contains the following words: *watercolor*, *techniques*, *oil*, *techniques*, *chalks*, *and*, *freehand*, *styles*, *guided*, *by*, *Jane*. If we take the basic forms (the lemmata) of these words and remove the repeating words and functional words such as prepositions, conjunction and pronouns, we will end up with the following few words that form the local context of the word *painting* in this sentence: *watercolor*, *technique*, *oil*, *chalk*, *freehand*, *style*, *guide*, *Jane*.

painter

painter	422
painting	262
paint	202
art	167
gallery	94
famous	84
buy	72
big	56
expensive	3
camera	0

painting

painting	461
buy	386
expensive	345
famous	205
gallery	183
big	176
art	188
painter	98
paint	91
camera	2

Table 3 Frequency vectors for the terms "painter" and "painting".

Most of these words are semantically related to *painting*, but some of them are not. If we take the word *painting* and a sufficiently large number of sentences containing that word (e.g., 1,000) and we extract from them all the words from its local context, we could expect that the most frequently appearing words to be semantically related to *painting*. These words should contain terms from fine arts and painting such as *painter*, *paint*, *brush*, *art*, *artist*, *technique*, and *style*. Accidentally found words like *guide* and *Jane* should appear quite rarely if we take a sufficiently large set of arbitrary sentences.

Now let us take two words and extract the frequently appearing words in their local contexts taken from sufficiently large number of sentences. If these two words are semantically similar we could expect their context words and respective numbers of occurrences to be also similar.

We can formalize the above ideas by assigning frequency vectors formed out of the words in the local contexts of the target words and measure the similarity between these vectors. For example for the words *painter* and *painting*, we could have the frequency vectors of the words in their contexts as shown in table 1 (with abridgments).

As dimensions of the vectors we take all words appearing in the contexts of at least one of the words and as coordinates we take their frequencies. For words not appearing in given context, we assume frequency 0. Therefore, we obtain the frequency vectors (with abridgments) shown in Table 2.

word	vector 1 (painter)	vector 2 (painting)
painter	422	98
painting	262	461
paint	202	91
art	167	188
gallery	94	183
famous	84	205
buy	72	386
big	56	176
expensive	3	345
camera	0	2

Table 4 Comparison of the frequency vectors for the terms "painter" and "painting".

We compute the distance between the vectors as the cosine in the n-dimensional Euclidean space. Thus we obtain a number between 0 and 1, which is a numerical measure for the semantic similarity between two words (higher value means more similar words).

2.2 Semantic Similarity Measured by Web Contexts

The World Wide Web (WWW) contains the largest set of text corpora in the world in a number of languages (including Russian) and provides efficient searching capabilities through the Web search engines. This motivates us to use the Web as a source of local context information for measuring semantic similarity between pair of words. We will describe a method for extraction of local context from the Web (web context), similar to the one described in [Nakov et al., 2007a].

For the extraction of the local context for a given word from the Web, we use a query against a Web search engine in which we request 100 results in the target language (in our experiments Russian). Using a sequence of 10 such queries, we can obtain up to 1,000 query results (Google sets explicit limits to never return more than 1,000 results). Each result contains a title and an excerpt (snippet) of text containing the word we searched for. For example, if we search for painting in English, we could get the following list of titles and text snippets, as shown in Table 3.

Painting - Wikipedia, the free encyclopedia
Painting, meant literally, is the practice of applying color to a surface (support) such as, e.g. paper, canvas, wood, glass, lacquer or concrete. ...
Painting - Exterior & Interior House Painters - Faux Finishing ...
Painting information, articles, pictures, **painting** ideas & more. Free price quotes from local exterior & interior **painting** contractors.
About.com **Painting** -- How-To Articles, **Painting** Tips, Projects ...
Whether you're into **painting** with oils, acrylics, watercolors, pastels, or mixed media, here you'll find essential how-to information, tips, ...
...

Table 5 List of titles and text snippets returned by Google for the word "painting".

In the titles and snippets returned by the search engine, we first convert all letters to lowercase and we extract all words.

We then remove all functional words (prepositions, pronouns, conjunctions, particles, interjections, and some adverbs) as well as all words with less than 3 letters. Such words do not bring semantic information about the searched word and should be omitted because they only distort the results.

Then we go through the extracted words sequences and when we find the target word or one of its forms, we take 3 words before and after it (the number 3 here we call *context size*). We consider these words part of the Web context.

We apply lemmatization (replace all words with their basic form), e.g. replace *paintings* with *painting*. For this purpose, we use a rich grammatical dictionary of Russian.

Now we have all words which appear in the local Web context of the target word and their corresponding frequencies (frequency vectors).

We measure the semantic similarity between two words by calculating the cosine between the frequency vectors of these words taken from their Web contexts.

2.3 TF.IDF Weighting

In information retrieval, TF.IDF weighting is a common technique for improving the search quality. The number TF.IDF (term frequency times inverted document frequency) is a statistical measure that shows how important is a certain word for a given document in a set of documents. The importance of the word increases proportionally to the number of its occurrences in the document but decreases proportionally to the total number of documents containing it. It was shown that if words' frequency is weighted according to their importance, the search quality improves [Sparck-Jones, 1972].

To apply TF.IDF weighting in our semantic similarity measure algorithm, we do the following: when we get the first 1000 query results from the Web search engine for a given word w, we directly compute the frequencies $TF[w_i]$ of all words w_i in its context by dividing the occurrences of w_i to the total number of words in the context of w (including duplicates). After that, we compute $IDF[w_i]$ by dividing the total number of documents indexed by Google (we assume they are about eight billions) to the number of occurrences in Google of w_i. Finally, we take $log_2(IDF[w_i])$ and multiply it by $TF[w_i]$ and thus we compute the weighted frequency of the word w_i in the frequency vector of w. The obtained weighted frequency vector we use for more measuring semantic similarity more accurately.

2.4 Semantic Similarity Measured through Reverse Context

When we extract the Web context for a given word, often semantically unrelated words fall in. For example, Internet terminology like *site*, *page*, *blog*, *online*, *forum*, *web*, *network*, *home*, *link*, *menu*, *message*, *download*, etc. are likely to appear in the context of almost any word, despite not being semantically related to it. Removing such words from the context is expected to improve accuracy when measuring semantic similarity with Web contexts [Nakov et al., 2007b].

The *reverse context lookup* technique is based on the idea that if two words are semantically related, the first one should often appear in the context of the second one, and at the same time, the second one should also often appear in the context of the first one.

For example in the context of the word *painting*, words like *painter*, *gallery* and *art* appear often, but so do parasite words like *order*, *news* and *site* as well. If we search the Web for the first three words, we shall convince ourselves that *painting* appears often in their contexts. However, if we search for the last three words, we will find that in their contexts *painting* almost does not appear.

We can formalize this idea as follows. Let $F(x,y)$ be number of appearances of y in the Web context of x. Consider some word w and all the words w_i from its Web context along with their frequencies $F(w,w_i)$. Now let us extract from the Web for each word w_i the number of reverse occurrences $F(w_i,w)$ of the word w in the context of w_i (*reverse context*). Finally, we can obtain a vector of the co-occurrences of the word w with all words from its context. It consists of all words w_i with frequencies:

$$\min(F(w, w_i), F(w_i, w))$$

The obtained frequency vector contains more accurate semantic information than the simple frequency vector because for each word it holds the minimal number of co-occurrences of the word with each word from its context.

When computing the co-occurrence frequency vector we can ignore words that occur in the co-occurrence frequency vector infrequently (e.g., three times or less) because this could have happened by chance. By modifying this parameter (frequency threshold), we can affect the accuracy of the results.

2.5 Synonyms Extraction by Measuring Semantic Similarity

Our method for extraction of synonyms by measuring semantic similarity is based on the hypothesis that synonym pairs should have higher semantic similarity compared to nonsynonyms.

If we are given a set of words and we measure the semantic similarity between each two of them, after sorting the pairs of words in a list in decreasing order by their semantic similarity, we can expect that synonyms are at the beginning of the list, followed by other semantically similar words, followed by words that are unrelated.

3. Experiments and Results

The experiments we preformed focus on studying and analyzing our algorithms for measuring semantic similarity extracted from the Web and their usage for the automatic discovery of synonyms. We performed experiments without and with using the reverse context and TF.IDF weighting and with various thresholds for the minimal frequency of the words in the context.

3.1 Resources Used

For the purposes of our experiments and for the implementation of our algorithms for measuring semantic similarity using the Web, we used the following resources:

- **Online Web search engine Google**[1]. We performed queries for 82,645 Russian words and collected the first 1,000 results for each of them.
- **Grammatical dictionary of the Russian language**, created in the Linguistic Modeling Laboratory, Institute for Parallel Processing, Bulgarian Academy of Sciences [Paskaleva,2007]. The dictionary contains about 1,500,000 wordforms and about 100,000 lemmata. Each dictionary entry contains wordform, corresponding lemma, followed by morphological and grammatical information.
- **List of the functional (stop) words in Russian**. Contains 507 words (prepositions, pronouns, conjunctions, particles, interjections and some adverbs). Created manually.

Our algorithm is general and can be applied to many languages. It does not require resources that are hard to find. The only resource that is not publicly available for any language is the grammatical dictionary. It is good to have it for highly inflectional languages like Russian, but this is less important for languages like English.

3.2 Test Data Set

In the experiments that we preformed, we used a list of 94 words from the Russian fine arts terminology, prepared manually by a linguist based on a set of study texts for students of fine arts. We selected only terms that occur in Google at least 5,000 times in order to have a statistical precision. Terms that occur in too small number of pages on the Web (e.g., just 5 times) cannot be analyzed statistically because the extracted context will be too small and not enough meaningful.

[1] http://www.google.com

Below is an excerpt from our list of 94 words:

> абрис, адгезия, алтарь, амулет, асфальт, вохрение, выжигание, гематит, диамант, жезл, закрепление, ...

There are 50 pairs of synonyms among these words, which we expect to be found by our algorithms.

3.3 Experiments

In all experiments our selected 94 words (terms from fine arts terminology) are processed in pairs and for each of them the semantic similarity is calculated. As a result, we obtained a list of 4,371 word pairs ordered in descending order by their similarity.

We measure the accuracy by *precision* and *recall*, which come from information retrieval. We experimented with few variations of the algorithm for measuring semantic similarity:

- **RAND** – returns a random ordering of all the pairs of words. We use this as a base for comparison with the other algorithms.
- **SIM** – the major algorithm for extraction of semantic similarity from the Web (described in detail in 2.2) with context size of 3 words, without analyzing the reverse context, with lemmatization.
- **SIM+TFIDF** – modification of the SIM algorithm with TD.IDF weighting (described in detail in 2.3).
- **REV2**, **REV3**, **REV4**, **REV5**, **REV6**, **REV7** – modifications of the SIM algorithm using the "reverse context lookup" technique (described in detail in 2.4) with the following frequency thresholds for the context words: 2, 3, 4, 5, 6 and 7.

3.4 Results

There are few well-known metrics for evaluation of information retrieval algorithms:

- **Precision @ n** – specifies what portion of the first *n* results are correct.
- **Recall @ n** – specifies what portion of all correct results are in the set of the first *n*.

The precision and recall are numbers between 0 and 1, and are typically expressed in percentages.

In our case, the algorithms for synonyms extraction using the Web as a corpus return a list of pairs of words and some of them are synonyms while other are not. We compute *precision @ n* by dividing the number of synonyms in the first *n* pairs by the number *n*. We compute *recall @ n* by dividing the total number of synonyms that exist in the data set (50) by the number of synonyms in the first *n* pairs. In practice, to evaluate a given algorithm, we need to know only how many synonyms are found in the first *n* pairs of words in the list. The more the words are, the better the algorithm is.

Table 4 shows an excerpt of the results obtained using the SIM algorithm and their corresponding precision and recall.

n	Word 1	Words 2	Semantic Similarity	Syno-nyms	Precision @ n	Recall @ n
1	выжигание	пирография	0.433805	yes	100.00%	2%
2	тонирование	тонировка	0.382357	yes	100.00%	4%
3	гематит	кровавик	0.325138	yes	100.00%	6%
4	подрамок	подрамник	0.271659	yes	100.00%	8%
5	оливин	перидот	0.252256	yes	100.00%	10%
6	полирование	шлифование	0.220559	no	83.33%	10%
7	полировка	шлифовка	0.216347	no	71.43%	10%
8	амулет	талисман	0.200595	yes	75.00%	12%
9	пластификаторы	мягчители	0.170770	yes	77.78%	14%
10	родонит	орлец	0.168245	yes	80.00%	16%

Table 6 Precision and recall obtained by the SIM algorithm.

The results of all evaluated algorithms are given in Table 5.

Algorithm	1	5	10	20	30	40	50	100	200	Max
RAND	0	0.1	0.1	0.2	0.3	0.4	0.6	1.1	2.3	50
SIM	1	5	8	15	18	23	25	39	48	50
SIM+TFIDF	1	4	8	16	22	27	29	43	48	50
REV2	1	4	8	16	21	27	32	42	43	46
REV3	1	4	8	16	20	28	32	41	42	46
REV4	1	4	8	15	20	28	33	41	42	45
REV5	1	4	8	15	20	28	33	40	41	42
REV6	1	4	8	15	22	28	32	39	40	42
REV7	1	4	8	15	21	27	30	37	39	40

Table 7 Comparison of the algorithms (number of synonyms in the results).

Instead of precision and recall, in table 5, we give the number of synonyms found in the first 1, 5, 10, 20, 30, 40, 50, 100 and 200 results. The best values are given in bold. The last column shows the total number of synonyms found by the corresponding algorithm. This number does not always reach the maximal value of 50 because most of the algorithms return no semantic similarity (value of 0) for large amount of the pairs from the test data set and thus we cannot assign certain positions in the ordered list for them to be able to evaluate the accuracy.

We evaluated the SIM algorithm also using ”11-pt average precision“, a widely used metric in information retrieval which combines precision and recall in a single number

[Salton, 1983]. 11-point average precision is computed by averaging the values in 11 points respectively for recall of 0%, 10%, 20%, … and 100%. The obtained result is 58.98%.

4. Discussion

In table 5, we can see that the proposed algorithms arrange most of the synonyms at the beginning of the produced ordered lists of pairs. The improvement over the random ordering (RAND) is huge, but the algorithms are still not perfect. Below we compare the algorithms in more detail and we discuss what causes the errors.

4.1 Comparison of the Algorithms

Figure 1 shows the precision/recall curves for the algorithms RAND, SIM, SIM+TFIDF and REV4.

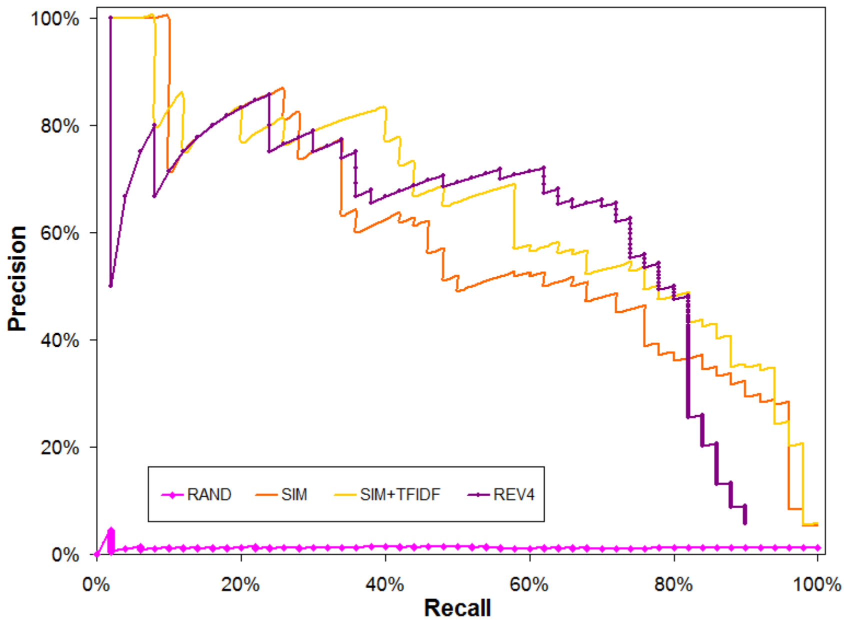


Figure 1 Precision / recall curves for the algorithms RAND, SIM, SIM+TFIDF and REV4.

The major SIM algorithm starts well with 5 correct synonyms and precision @ 5 of 100%. For the top 10 ranked pairs and in the first 20 pairs the precision remains very high: 80%. For the top 40 pairs, the algorithm lists 56% of all the synonyms and its precision is still 70%. For the top 100 pairs, the precision falls to 40%, but the recall is over 80%. For the top 200 pairs, almost all synonyms are listed (recall 96%), but the precision drops to 24%. The SIM algorithm lists almost all synonyms in the first 100 results (which are only 2.11% of all 4,371 pairs in the list).

Applying TF.IDF weighting improves the accuracy of the SIM algorithm and yields better precision and recall. Respectively, the graphic of SIM+TFIDF is located above the graphic of the SIM algorithm most of the time.

Applying the *reverse context lookup* technique improves the SIM algorithm by increasing its precision for the top 40-50 pairs in the list (algorithms REV2 – REV7), but overall decreases the recall. The reverse context lookup technique works better than TF.IDF weighting for the beginning of the list (top 30-50 pairs), but is worse thereafter.

Modifying the threshold of the frequency used in the reverse context lookup has little impact on the accuracy. Lower threshold causes a lower precision at the beginning of the list and an overall higher recall. Higher thresholds improve the precision for the top 40-50 pairs, but lowers the overall recall.

4.2 Known Problems

The accuracy of the obtained results for all algorithms is less than 100%. Below are the most important reasons for that:

- There is an inaccuracy when measuring the semantic similarity by our algorithms because not always words appearing in similar contexts are similar in meaning.
- Using the Web as a corpus limits the extraction of the local contexts to the first 1,000 results only. Since commercial and news sites are typically ranked higher by Google, the top 1,000 results are not a representative sample of all texts on the Web.
- High semantic similarity is typical for synonym pairs, but it is not limited to synonyms only. Our algorithms assume that words appearing in similar contexts are similar, but this does not directly mean that they are synonyms. For example, the colors *blue* and *red* are semantically related (because both mean a kind of color), but are not synonyms. This causes significant errors during synonyms extraction as cen be seen from the obtained results. Fixing this problem would be the most important challenge in our future work.

5. Related Work

Most of the automatic synonym extrcation methods are based on distributional hypothesis, that semantically related words appear in close contexts [Harris, 1954]. This hypothesis provides a key point for many other synonyms retrieval algorithms: *contexts retrieval and comparison*. In its essence, our method is also based on context retrieval and comparison, but we use the Web as a corpus for measuring semantic similarity and in this way we do not depend on other linguistic resources (e.g., large text corpora).

Algorithms, based on the distributional hypothesis, are proposed by [Lin, 1998] and [Curran et al., 2002]. In these papers, the contexts are defined based on predefined grammatical relations that are retrieved from a language corpus. They also take into account the distance between the retrieved contexts.

The main problem of all the above methods is the difficulty to distinguish synonyms from other semantically similar pairs of words such as hyponyms, hypernyms, antonyms, etc. We expect that synonyms, being more strongly related, will have higher degree of similarity than e.g. hyponyms or hypernyms, but this is not always the case. The problem persists in our work as well.

The problem is partially solved by [Plas et al., 2006], who use two parallel corpora aligned at the word level using GIZA++ [Och, 2003], from which the corresponding sentences and all probable translations between word pairs in both languages are retrieved. As a context for a given word in the first language, the set of all its probable translations in the other language are used. Then the semantic similarity between the two words is measured as a distance between their contexts. This approach allows for a more accurate distinction between synonyms and other semantically related words, because antonyms and hypernyms rarely get aligned. The disadvantage of this approach is that it requires a big parallel corpus, which can be unavailable. It will also not work for uncommon words, which are almost not met in the corpus.

[Hagiwara et al., 2007] propose to measure semantic similarity using local contexts extended with indirect retrieval of additional contextual words. In particular, after the local context C for a given word has been retrieved, the words from the local contexts of all words in C are added to the local context of that target word as well. In this way, the semantic information is enriched and thus the accuracy of measuring semantic similarity is improved. The only disadvantage is that this approach of retrieving context from the Web is too expensive because of the high number of search queries needed to retrieve the indirect context words.

The idea of using the Web as a corpus has been used by many scientists solving different problems (see [Kilgarriff et al., 2003] for an overview). Some of them use Web search engines for finding how many times a word or phrase is met and calculating pointwise mutual information [Inkpen, 2007], whereas others directly retrieve context from text snippets returned by Web search engines [Nakov et al., 2007b].

The idea of retrieving information from text snippets returned by a Web search engine is used in [Chen et al., 2006]. The model they introduce is based on the idea that if two words X and Y are semantically bound, then searching for X should cause Y to appear often in the results, and vice versa: searching for Y should cause X to appear often in the results. In this approach, context words are completely ignored (except for X and Y) and their semantics are not used. As it is later discovered [Bollegala et al., 2007], this produces incorrect zero semantic similarity for most of the processed pairs.

[Sahami et al., 2006] use the Web as a corpus to measure the semantic similarity between pairs of short text fragments (search requests), thus gaining automatic requests expansion and offering alternative requests. For this purpose, they retrieve the contexts of the pairs of short texts from the content of the documents returned after searching, and they then compare the most frequent words from these documents. In contrast, we do not compare the content of the documents but only the snippets returned by a Web search engine, which requires much less resources and yields better results since not all words from the document are taken into account but only the ones in the local context.

[Bollegala et al., 2007] combine retrieval of information about the number of occurrences of two words (both together and individually) from a Web search engine, with retrieval of information from text snippets returned by the search engine. They automatically discover lexico-syntactic templates for semantically related and unrelated words using WordNet, and they train a support vector machine (SVM) classifier. The learned templates are used for extracting information from the text fragments returned by the search engine. Finally, the results are combined. The method is more complicated than the one we propose and requires extra resources for training the SVM.

An interesting approach for finding synonyms and lexicalizations from the Web is described in [Sanchez et al., 2005]. They start with a taxonomy of terms relevant to a specific domain built automatically for a given keyword based on series of searches in Google. They then search the Web for the longest multiword terms extracted from the taxonomy after removing the target keyword and assume that synonyms should be found on the same position where the original keyword was. The approach is quite original, but addresses a different problem: find possible synonyms for a given word.

A major advantage of our method is that it does not require large corpora or other resources like WordNet, which are not available for some languages.

References

1 Hearst M. (1991). "Noun Homograph Disambiguation Using Local Context in Large Text Corpora". In *Proceedings of the 7th Annual Conference of the University of Waterloo Centre for the New OED and Text Research*, Oxford, England, pages 1-22.

2 Nakov P., Nakov S., Paskaleva E. (2007a). "Improved Word Alignments Using the Web as a Corpus". In *Proceedings of RANLP'2007*, pages 400-405, Borovetz, Bulgaria.

3 Nakov S., Nakov P., Paskaleva E. (2007b). "Cognate or False Friend? Ask the Web!". In *Proceedings of the Workshop on Acquisition and Management of Multilingual Lexicons, held in conjunction with RANLP'2007*, pages 55-62, Borovetz, Bulgaria.

4 Sparck-Jones K. (1972). "A Statistical Interpretation of Term Specificity and its Application in Retrieval". *Journal of Documentation*, volume 28, pages 11-21.

5 Salton G., McGill M. (1983), *Introduction to Modern Information Retrieval*, McGraw-Hill, New York.

6 Paskaleva E. (2002). "Processing Bulgarian and Russian Resources in Unified Format". In *Proceedings of the 8th International Scientific Symposium MAPRIAL*, Veliko Tarnovo, Bulgaria, pages 185-194.

7 Harris, Z. (1954). "Distributional structure". *Word*, 10, pages 146-162.

8 Lin D. (1998). "Automatic Retrieval and Clustering of Similar Words". In *Proceedings of COLING-ACL'98*, Montreal, Canada, pages 768-774.

9 Curran J., Moens M. (2002). "Improvements in Automatic Thesaurus Extraction". In *Proceedings of the Workshop on Unsupervised Lexical Acquisition, SIGLEX 2002*, Philadelphia, USA, pages 59-67.

10 Plas L., Tiedeman J. (2006). "Finding Synonyms Using Automatic Word Alignment and Measures of Distributional Similarity". In *Proceedings of COLING/ACL 2006*, Sydney, Australia.

11 Och F., Ney H. (2003). "A Systematic Comparison of Various Statistical Alignment Models". *Computational Linguistics*, 29 (1), 2003.
12 Hagiwara M., Ogawa Y., Toyama K. (2007). "Effectiveness of Indirect Dependency for Automatic Synonym Acquisition". In *Proceedings of CoSMo 2007 Workshop, held in conjuction with CONTEXT 2007*, Roskilde, Denmark.
13 Kilgarriff A., Grefenstette G. (2003). "Introduction to the Special Issue on the Web as Corpus", *Computational Linguistics*, 29(3):333–347.
14 Inkpen D. (2007). "Near-synonym Choice in an Intelligent Thesaurus". In *Proceedings of the NAACL-HLT*, New York, USA.
15 Chen H., Lin M., Wei Y. (2006). "Novel Association Measures Using Web Search with Double Checking". In *Proceedings of the COLING/ACL 2006*, Sydney, Australia, pages 1009-1016.
16 Sahami M., Heilman T. (2006). "A Web-based Kernel Function for Measuring the Similarity of Short Text Snippets". In *Proceedings of 15th International World Wide Web Conference*, Edinburgh, Scotland.
17 Bollegala D., Matsuo Y., Ishizuka M. (2007). "Measuring Semantic Similarity between Words Using Web Search Engines", In Proceedings of the 16th International World Wide Web Conference (WWW2007), Banff, Canada, pages 757-766.
18 Sanchez D., Moreno A. (2005), "Automatic Discovery of Synonyms and Lexicalizations from the Web". Artificial Intelligence Research and Development, Volume 131, 2005.

The Implementation of the Nonlinear Wavelet Transform in Image Compression

Indrit Enesi[1], *Betim Cico*[2]

Polytechnic University of Tirana, Telecommunication Department, "Mother Tereza" square, Nr. 4, Tirana, Albania, indrit@mail.com[1]*, bcico@fie.upt.al*[2]

Wavelet transform represents signal in the 3-dimensional time-frequency domain, displaying most distinguished information which is hidden in the time or frequency presentation of the signal. Discrete wavelet series provide sufficient information both for analysis and synthesis of the original signal, with a significant reduction in the computation time. The proposed method introduces nonlinear wavelet transforms into image compression, resulting in efficiently presentation of edges and compact signal energy into lower subbands of the transform. The coding method is based on the algorithm of set partitioning in hierarchical trees (SPIHT), the transformed coefficients are transmitted progressively, controlling either the scale or the bit rate. The experimental results show that the proposed method performs better than JPEG2000, especially at higher compression ratios.

Keywords
lifting schema, nonlinear wavelet, progressive transmission, SPIHT, subbands.

1. Introduction

Wavelet transform represents a three-dimensional display of the signal depending on the time and frequency, providing signal features not feasible during the usual spatial or frequencies transforms [1]. The discreet wavelet series provide sufficient information both for analysis and synthesis of the original signal, highly reducing the computation time. The proposed method introduces nonlinear wavelet transform into image compression resulting in efficiency presentation of the edges and compact the signal's energy into lower subbands of the transform. The coding method is based on the algorithm of set partitioning in the hierarchical tree (SPIHT) [2]. The transformed coefficients are transmitted progressively, controlling either the scale or the bit rate. Experimental results show that the proposed method performs better than the JPEG2000 [3], especially at higher compression ratios.

In this paper it is been treated the wavelet technique, the particularity that it provides in the layout of nonstationary signals as well as a compression and a higher efficient transmission method than in JPEG2000 are been proposed, based on a nonlinear computation algorithm of wavelet transform. The minimization of Heisenberg's uncertainty [4] is rendered through the multi resolutional analysis (MRA) [5], where unlike STFT the spectral components are not selected rateably. MRA provides a high resolution in time for a lower resolution in frequency as well as a lower resolution in time for a higher resolution in frequency. This method is applied perfectly well in images where the spectral components are usually with a higher frequency and short duration as well as spectral components with a low frequency and long duration.

2. Wavelet Transform

2.1 The Continuous Wavelet Transform.

The Continuous Wavelet Transform (CTW), as an alternative method STFT, avoids the resolution problem. Mathematically the continuous wavelet transform is been defined as:

$$CWT_x^{\psi}(\tau, s) = \Psi_x^{\psi}(\tau, s) = \frac{1}{\sqrt{|s|}} \int x(t)\psi^*\left(\frac{t-\tau}{s}\right)dt \qquad \textbf{(1)}$$

The transformed signal ψ is a function of two variables τ and s, that respectively display translation and scale[6]. Mother wavelet $\psi(t)$ is an oscillator, finite in time and a prototype for the generation of other window variables. The term translation displays the position or the window translation by a signal, the scale parameter is defined as the 1/frequency, similar to the scale used in geographical maps.

2.2 The Computation of CWT

The mother wavelet serves as a prototype, the windows we see during the MRA process are an extended or a compressed version of the mother wavelet. The mother wavelet is multiplied and integrated with the signal during all his duration. In order to normalize the energy, the integration result is multiplied to the constant 1/sqrt(s) [6]. Unlike STFT, in which it is used the same resolution all the time and for all the frequencies, the wavelet transform uses a high time resolution and a low spectral resolution in high frequencies and a low time resolution and a high spectral resolution in low frequencies.

2.3 The Rresolution’s Interpretation in Time and Frequency Domain.

All the plane time-frequency points that are located within a rectangular are been represented by a unique value of the function wavelet transform. The rectangles are of different dimensions, but their surfaces are equal. We notice that for low frequencies, the rectangle’s height is low (a high spectral resolution, more accurate frequency values) but with a high longitude (low time resolution, similar time values). When the frequency extends this report differs in the highest frequencies resulting in a good visibility for the moment, lowering the frequency values visibility. Despite the rectangles dimensions, their surfaces in both analyses (STFT and wavelet transform) are equal and are defined by the Heisenberg's principle. The surface can't be lowered endlessly and its lowest limit is $1/4\pi$ [4] .

2.4 The Graphic Interpretation of the Wavelet Transform

The data gathered from the reactions of the patients who suffer from Alzheimer's disease are presented in the framework of a statistical database [7]. In reality the signals are not rectified sinusoids which we can compare with each other, so we need to distinguish between a provided answer by a diseased person and a sane one. For a sane person the answer is provided in the figure 1.

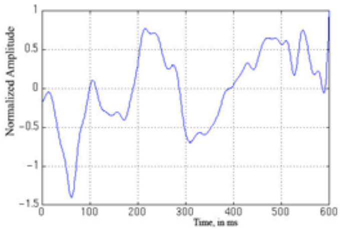


Figure 1 The acquired signal by the sane person

The continuous wavelet transform for this signal rendered by two different viewpoints is presented in figure 2.

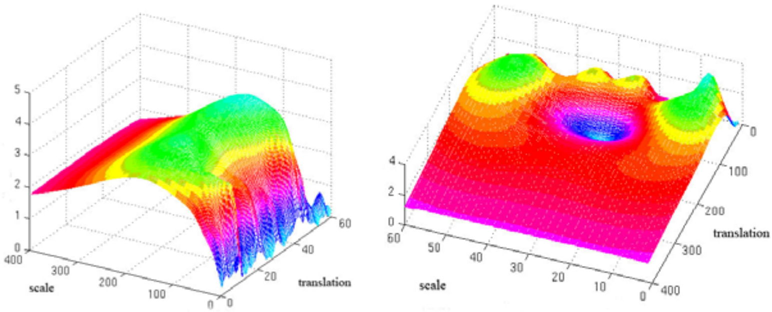


Figure 2 The sane person's signal specter in the scale-translation plane

We consider the reactions from the same database regarding the experience with an Alzheimer diseased person.

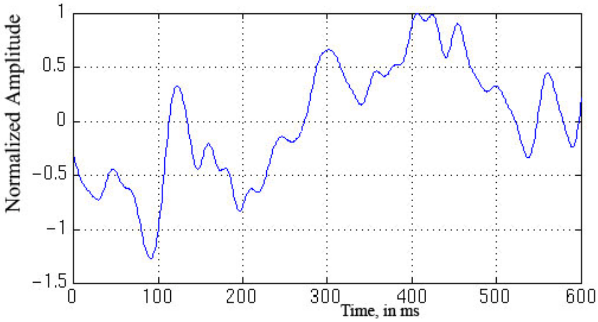


Figure 3 The acquired signal by the Alzheimer disease affected person

The Continuous Wavelet Transform for this signal is presented in figure 4.

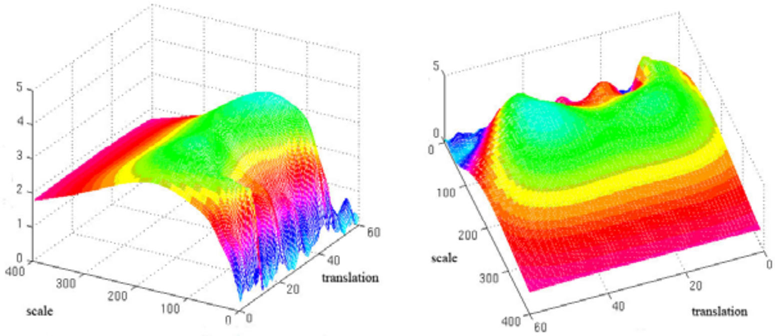


Figure4 The Alzheimer diseased person's signal Specter

Only by a graphic display of the continuous wavelet transform, we can distinguish the characteristics of an Alzheimer diseased person.

3. Nonlinear Wavelet Transform

3.1 The Lifting Schema in the Nonlinear Wavelet Transform

The lifting schema, called otherwise the second wavelet generation, developed by Sweldens [8], unlike the first generation, provides transform adjustment towards the data, enabling the interposition of nonlinearity controlling the wavelet transform characteristics. The typical lifting is acquired by the iteration of three basic operations: Split, Predict and Update as presented in Figure 5.a.

Split: The original data is divided in: the even set and the odd set.

$$x_e[n] = x[2n] \quad \text{and} \quad x_o[n] = x[2n+1] . \qquad (2)$$

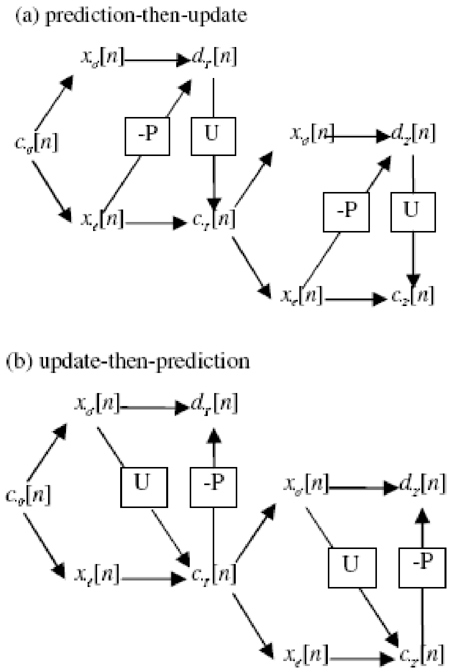


Figure 5 The wavelet transform tree lifted in two levels, a) predict then update, b) update then predict

Predict: We predict $x_o[n]$ by $x_e[n]$ using the prediction operator P. d[n] called wavelet coefficient, is a error in predict. We notice that the prediction filter P leads to a high-pass filter.

$$d[n] = x_0[n] - P(x_e[n]) \qquad (3)$$

Update: $x_e[n]$ and d[n] are combined using the update operator U in order to obtain the scaling coefficients c[n]. The update operator U is carefully defined towards update operator P to display an coarse approximation c[n] towards the original signal x[n].

$$c[n] = x_e[n] + U(d[n]) \qquad (4)$$

3.2 Design Filter of the Nonlinear Transform

When we construct the prediction and update operators using the polynomial data set constraints, the output of update step $c_i[n]$ coming from each iteration step is a coarse approximation (low-pass) [9], [10]. Considered that $c_i[n]$ will be the input to the next iteration step, if the prediction is accomplished by a nonlinear operator, it may be impossible to construct an update operator which can satisfy all the polynomial lifting constraints, because $c_i[n]$ has been affected by P and U before being fed into further iteration. The new structure, first update, then predict resolves this problem as presented in Figure 5.b, since the prediction operator of the new structure is not included in the series and is based only on low-pass coefficients [9]. In regions where the image is locally smooth, predictors of high orders are used, near edges (high frequency) the order and the length of the predictor are reduced. The low-pass filter coefficients are computed using the Haar filter [11]:

1-st order P: $d[n] = x[2n+1] - a_0c[n]$

3-d order P: $d[n] = x[2n+1] - (a_1c[n-1] + a_0c[n] + a_1c[n+1])$

5-th order P: $d[n] = x[2n+1] - (a2c[n-2] + a1c[n-1] + a0c[n] + a1c[n+1] + a2c[n+2])$

7-th order P: $d[n] = x[2n+1] - (a3c[n-3] + a2c[n-2] + a1c[n-1] + a0c[n] + a1c[n+1] + a2c[n+2] + a3c[n+3])$

Where a_i is the coefficient of the lifting scheme of CDF (1,n) wavelet.

4. SPIHT Structure

Coding wavelet technique with the embedded zero tree wavelet [12] (EZW – embedded zerotree wavelet) was first published by Shapiro, optimized by Said and Pearlman was called set partitioning in hierarchical trees (SPIHT) [4]. Both methods EZW[12] and SPIHT[2] depend on the same fundamental idea, if the amplitude of a coefficient is small, the corresponding coefficients in the high frequency domain are also small. So if the coefficient is marked by LH2(x, y) in the subband LH2 does not have a considerable value in comparison to the threshold value in wavelet space, the probability that the corresponding coefficients (LH1(2x,2y), LH1(2x + 1,2y), LH1(2x,2y + 1), LH1(2x + 1,2y + 1)) are also not significant regard to the same threshold is 98.2%.

5. Progressive Image Transmission

Progressive Image Transmission is a class of image transmission techniques where image profile data are transmitted firstly and then the remains data are transmitted progressively to refine the quality of the image. The main goal is to allow the receiver recognize relevant features in an image as quickly as possible at the minimum cost. Our method of progressive image transmission method uses the SPIHT algorithm [2]. Progressive image transmission can be achieved using two methods: first, increase the bit-rate while the scale is kept constant; second, decrease the scale at each stage while the bit-rate is kept constant.

6. The Simulation Results

First, transform the image with nonlinear wavelet, then use one of the progressive transmission image methods to code and progressively transmit those coefficients. We compare the results using the linear 9.7 wavelet that is very popular in image compression and has been used by JPEG2000.

(a) BPP= 0.1, PSNR = 25.42

(b) BPP = 0.2, PSNR = 27.00

(c) BPP = 0.4, PSNR = 28.07

(d) BPP = 0.6, PSNR = 28.82

(e) BPP = 0.8, PSNR = 29.45

(f) BPP = 1, PSNR = 30.03

Figure 6 The images generated in different steps of the progressive transmission in the 4-th scale at different ratios

In the constant scale method, the lifting nonlinear algorithm results are displayed in the table 1 and CDF [13] and JPEG2000 algorithm results using compression 9.7, applied on an image whose dimensions are 256 x 256 for the 4-th scale and with a variable ratio.

Bit – Rate (BPP)	PSNR (dB) Nonlinear	PSNR (dB) CDF 9,7	PSNR (dB) JPEG 2000
0.1	25.4150	25.3450	24.4008
0.2	27.0047	26.8547	25.7876
0.4	28.0696	27.8696	27.3656
0.6	28.8235	28.7535	28.4521
0.8	29.4490	29.3590	29.0139
1	30.0255	29.9155	29.8679

Table 1 The results in different steps of the progressive transmission of the image with scale 4

The data in table 1 are displayed graphically in figure 7 .

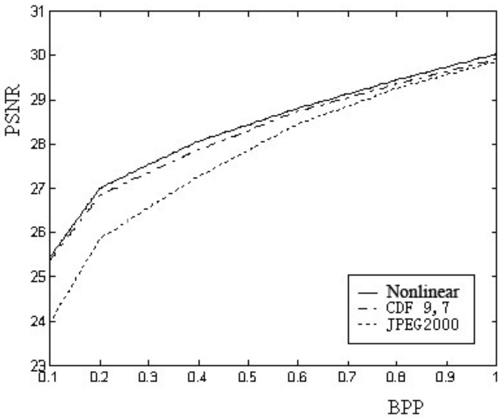


Figure 7 PSNR towards BPP

The simulation results show that the application of the nonlinear wavelet transform in image compression outperforms the results of the linear wavelet transform, especially in higher rates of the compression. The compression performance is computed using the peak of the signal-noise ratio (PSNR)[14] as well as the, subjective qualitative assessment of the image. The PSNR value in this project is defined by the formula (5).

$$PSNR = 20\log_{10}\left(\frac{255}{\sqrt{\frac{1}{MN}\sum\sum\left(x'[i,j]-x[i,j]\right)^2}}\right) \qquad \textbf{(5)}$$

x[.] is the original image whose dimensions are MxN, whereas x’[.] is the newbuilt image.

7. Conclusions

During this approach we were focused on an image compression method using nonlinear wavelet transform. This algorithm efficiently displays the edges and compact the signal's energy into the lower subbands of the transform.

SPIHT can efficiently use and control the resources at all the stages of the transmission. The tests performed with constant scale resulted that nonlinear wavelet transform method is superior over JPEG2000.

The nonlinear wavelet transform provided higher results for the compression image, more advanced monitoring and diagnostic functions.

References

1 Sidney Burrus and Ramesh A. Gopinath. Introduction to Wavelets and Wavelets Transforms. Prentice Hall Inc. pages 300 – 310. 1997.

2 A. Said and W. A. Pearlman, "A new, fast, and efficient image code based on set partitioning in hierarchical trees." IEEE Trans. Circuits Syst. Video Technol., vol. 6, pages 243-250, June 1996.

3 JPEG 2000. [Online] Available: http:// www.jpeg.Org / JPEG2000.htm, December 2007.

4 Thomas Schell , Andreas Uhl. Optimization and assessment of wavelet packet decompositions with evolutionary computation, EURASIP Journal on Applied Signal Processing, v.2003 n.1, pages 806-813, January 2003

5 Abdourrahmane M. Atto, Dominique Pastor, Alexandru Isar. On the statistical decorrelation of the wavelet packet coefficients of a band-limited wide-sense stationary random process, Signal Processing, v.87 n.10, pages 2320-2335, October, 2007.

6 Amara Graps, An Introduction to Wavelets, IEEE Computational Science & Engineering, v.2 n.2, pages.50-61, June 1995.

7 S. G. Mallat. A theory for multiresolution signal decomposition: The wavelet representation. IEEE Transactions on Pattern and Machine Intelligence, faqe 674-693, 2002.

8 W. Sweldens. "The lifting scheme: A custom-design construction of biorthogonal wavelets." Applied Comput. Harmon. Anal.,vol. 3, no. 15, pages. 186-200, 1996.

9 Roger Claypoole, Richard Baraniuk. "Flexible Wavelet Transforms Using Lifting," Proceedings of the 68th SEG Meeting, New Orleans, Louisiana, USA, 1998.

10 Roger Claypoole, Geoffrey Davis, Wim Sweldens, Richard Baraniuk, "Nonlinear wavelet transforms for image coding." Proc. Asilomar conf. Signals, Sytems, Computers, November, 1997

11 Karen Egiazarian , Jaakko Astola, Tree-Structured Haar Transforms, Journal of Mathematical Imaging and Vision, v.16 n.3, faqe.269-279, May 2002

12 J. M. Shapiro, "Embedded image coding using zerotrees of wavelet coefficients," IEEE Trans. Signal Processing, vol. 41, pp. 3445-3462, December 1993.

13 CDF [Online] Available: http://cdf.gsfc.nasa.gov/, January 2008.

14 Roger Claypoole, Geoffrey Davis, Wim Sweldens, Richard Baraniuk, "Nonlinear wavelet transforms for image coding." Proc. Asilomar conf. Signals, Sytems, Computers, November 1997

Appendix

List of Acronyms

CDF	Common Data Format
CTW	Continuous Wavelet Transform
EZW	Embedded Zerotree Wavelet
JPEG	Joint Photographic Expert Group
MRA	Multi Resolutional Analysis
PSNR	Peak of the Signal – Noise Ratio
SPIHT	Set Partitioning in Hierarchical Trees
STFT	Short Time Fourier Transfor

Application of Machine Learning Techniques and Data Mining to Agents

AlketaHyso[1], Betim Cico[2]

[1]*Department of Information Technology, University of Vlora, Vlora, Albania, ahyso@univlora.edu.al*

[2]*Computer Engineering Department, Polytechnic University, Sheshi "Nene Tereza", Tirana, Albania, bcico@fie.upt.al*

In this paper we are focused in the intersection of agent technology and data mining techniques for producing intelligent agents. We can use techniques of machine learning in practical data mining, for finding patterns in data. Useful patterns are transformed into the agent's inference mechanisms that drives the behavior of the agent. Our application shows that using machine learning techniques, we improve the reasoning mechanism of our agent supplying to it a new behavior. So data mining techniques, offer a promising method for the development of intelligent agents.

Keywords

data mining, intelligent agent, machine learning techniques.

1. Introduction

Agent and multi-agent system technologies, methods and theories are currently contributing to many diverse domains. This is not only a very promising technology, it is emerging as a new way of thinking a conceptual paradigm for analysing problems and for designing systems, for dealing with complexity, distribution and interactivity, and perhaps a new perspective on computing and intelligence. Recent years saw a new trend in the combination of the multi-agent system approach and data mining. Technologies of data mining and intelligent agents can complement and benefit from each other yielding more efficient solutions [1].

Integrating agents into data mining systems, or constructing data mining systems from agent perspectives, the flexibility of data mining systems can be greatly improved. Equipping agents with data mining capabilities, the agents are much smarter and more adaptable. In this way, the performance of these agent systems can be improved [2].

Knowledge, hidden in voluminous data repositories, can be extracted by data mining and provide the logic for agents and multiagent systems. This knowledge nuggets constitute the building blocks of agent intelligence.

The dual process of knowledge discovery and intelligence infusion is equivalent to learning, better yet, teaching by experience. Indeed, existing application data are filtered in an effort to distill the best, most successful, empirical rules and heuristics. The process can be applied initially to train ‘dummy’ agent and as more data are gathered, it can be repeated periodically or on demand to further improve agent reasoning [3].

One of the most interesting issues in agent technology has always been the modeling and enhancement of agent behavior. Numerous approaches exists, attempting to optimally reflect both the inner states as well as perceived environment of an agent in order to provide it either with reactivity or proactivity [4].

The data mining techniques can be used for creating intelligent systems that may be able to solve complex problems in adaptive manner . Using data mining techniques based on machine learning makes it possible to develop systems able to learn from and adapt to their environment.

Using various machine learning methods the agents must be able to learn and/or self-learn. The machine learning algorithms allow for an agent to adequately respond to environment changes and improve the behavioral rules or acquire intelligent behavior.

This paper describes the synergy of data mining and agent technology. In section 2 there are treated some concepts on data mining and machine learning. Section 3 presents the methodology to extract pattern models for the agent. In section 4, we give a detailed presentation of the agent training method, the experiment of extracting pattern from a given dataset applying machine learning methods. Finally, section 5 concludes this paper and outlines future work.

2. An Introduction to Data Mining and Machine Learning

The types of tasks Data Mining could accomplish can be roughly divided in two categories: predictive tasks and descriptive tasks.

The first type of tasks try to discover a model that drives the behavior of some variables in a system in order to be able to predict such values in zones not covered by the examples. One of the predictive tasks of Data Mining is the task of finding some form of classifications of the items contained in a set of raw data. When the classification domain is not finite (e.g. when the variable interested by the prediction process is a real number) the operation is regression.

The second type of task tries to find some categorizations of the data producing a shrinked descriptor for wider segments of data. In descriptive Data Mining the task is to discover interesting regularities in the data, to uncover patterns and find interesting subgroups in the bulk of data.

Data Mining techniques could be successfully applied to agent, in order to exploit hidden relations and emergent behavior.

Machine learning, as with automated reasoning, aims to use reasoning to find new, relevant information given some background knowledge. This information may then be used towards completing an intelligent task of more complexity.

Methods used in machine learning are:

- Symbolic Learning
- Neural Learning
- Evolutionary Learning
- Probabilistic Learning
- Other Machine Learning Methods

We can classify the contributions of Data Mining and Machine Learning in agents in two main tasks: Machine Learning can be used to provide the single agent a sort of intelligent behavior. Results are analyzed using Data Mining techniques in order to reveal interesting patterns in data that could help to better model the behavior of the overall systems. The machine learning algorithms allows an agent to learn from its past history in a human similar way, that is to say, by induction, we can choose to create agents with the ability to compute rules and strategies and evolve according to the environment in which they act [5];

3. Methodology

We demonstrate the use of learning techniques and tools to extract models to generate predictions or new behaviors in agents. The stages we have to follow to realize what we forwarded are:

- We apply machine learning techniques on the provided dataset. We interpret machine learning as the acquisition of structural descriptions from examples. We apply the machine learning technique - MultilayerPerceptron, which is a neural network that trains the agent using backpropagation.
- We extract the knowledge model from the dataset. It is a non-linear function approximation.
- In the next step, we incorporate this knowledge model as a reasoning mechanism of the agent.
- We monitor the new behavior of the agent.

4. Application

Our application aims to present the importance of the synergy of data mining methods and agent technology, in building more intelligent agents. In order to exemplify this idea we considered a non-real example. The data we present in dataset are not real and not many. In practice the problems are more intricate, with too much data and noises.

We shall take into consideration an agent which is able to calculate time and to move. After getting the signal from a given sensor, the agent moves through a straight line without obstacles, departing from a position of "waiting", to which corresponds a 0 moment of time. Its moving accelerate with a given acceleration that our agent is not able to calculate. We suppose that in equal distances our agent should leave a print. We should have the agent learn the process of calculating the road it does, only based on its capacity to calculate time.

In order to solve the above mentioned problem, we use a row of data extracted during the movement of the agent. These rows include data about movement: time intervals, the speed of movement in that moment, acceleration and distance.

Acceleration (m/s2)	Time (s)	Distance (m)
2	0	0
	1	1
	2	4
	3	9
	4	16
	5	25
	6	36
	7	49

Table 1 Pattern row

On these data we apply machine learning method – multilayer perceptron neural network, to find a model, which in the following step we will incorporate to our agent. Multilayer feedforward neural networks with sigmoidal activation functions have been termed "universal function approximators". These types of networks can approximate any continuous function to a desired degree of accuracy. This approximation may require an inordinate number of hidden nodes.

Our neural network, has an input layer that sets the values of elements in the hidden layer with 3 nodes with sigmoidal activation function. They set values for the output layer element which is a linear unit. Figure 1, shows the architecture of our neural network.

In MLPs, learning is supervised, with separate training and recall phases. During training the nodes in the hidden layers organise themselves such that different nodes learn to recognise different features of the total input space. During the recall phase of operation the network will respond to inputs that exhibit features similar to those

learned during training. Incomplete or noisy inputs may be completely recovered by the network. So, our agent undergoes a training process.

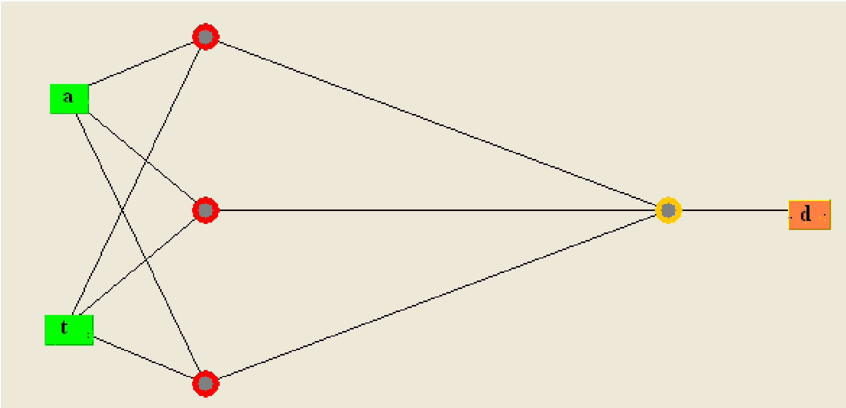


Figure 1 Architecture of neural network.

Among the learning algorithm the one that gives the best solution is the back propagation algorithm (BP) of neural networks which is supported by Multilayer Perceptron method.

BP learns the weights for a multilayer network. It employs gradient descent to attempt to minimise the squared error between the network output values and the target values for these outputs. The learning process has two stages: Forward stage: calculate outputs given input pattern and Backward stage: update weights by calculating delta.

The back-propagation algorithm for updating weights in a multilayer network is presented in [6].

As input parameters we get time and acceleration, as output parameter the completed distance. Number of iterations (for backpropagation), Nit = 100.

We perform the training of the agent according to this algorithm by adjusting such parameters of algorithm as the number of training cycles, learning rate or the given model of the neural network. The final model shows clearly the weights:

weights from input units to output units

4967.103182	-363.815049

output units thresholds

2483.551591

weights from hidden units to output units

-20761.777221	335.640279	-705.166631

weights from input units to hidden units

3.983986	-0.077077	-7.592238
0.226480	0.378405	-1.142265

hidden units thresholds

0.912341 0.335880 -2.606184

Training curve showing the gradual reduction in error as weights are modified over several epochs, for a given set of examples.

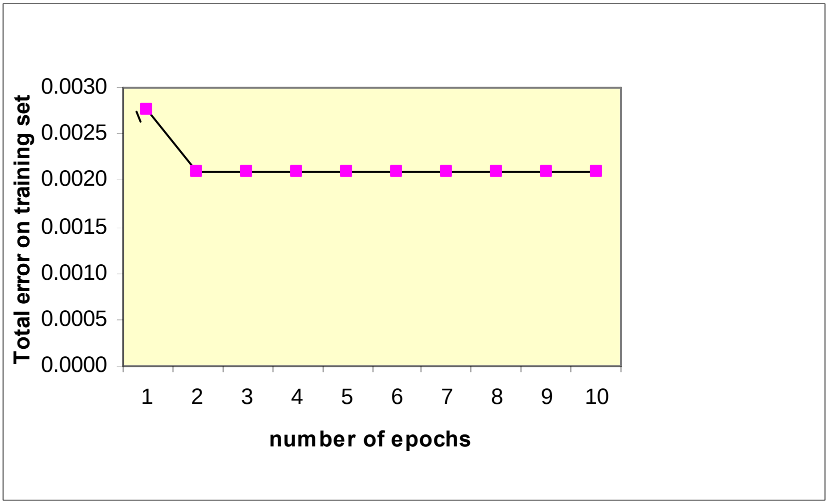


Figure 2 Training curve

The knowledge model that we get is a non-linear function of the type:

$$d = -20761.777221 \cdot \frac{1}{1+\exp^{-(a\cdot 3.983986+\mathrm{t}\cdot(-0.077077)+(-7.592238))}} +$$

$$335.640279 \cdot \frac{1}{1+\exp^{-(a\cdot 0.226480+\mathrm{t}\cdot 0.378405+(-1.142265))}} +$$

$$-705.166631 \cdot \frac{1}{1+\exp^{-(a\cdot 0.912341+\mathrm{t}\cdot 0.335880+(-2.606184))}} +$$

$$a \cdot 4967.103182 + \mathrm{t}\cdot(-363.815049) + 2483.551591$$

In a second phase we build an application which simulates the behavior of the agent where the reasoning mechanism is realized according to the final model obtained from the data mining machine learning methods.

Figure 3, indicate the run distance by the agent and measured distance through reasoning mechanism, which we obtain from the neural network.

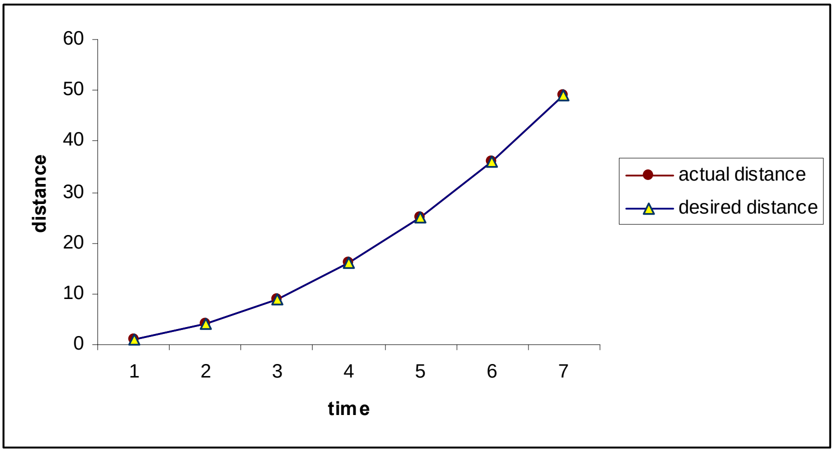


Figure 3 Desired and actual distance measured by the agent.

This agent is able to measure the distance and to predict the time when it arrives a certain distance. By the experiment we see that the agent displays a new behavior: it is able to realize an approximate calculation of distance; by developing an intelligent behavior of the agent.

5. Conclusions and Future Work

This paper aims to present the importance of the synergy of data mining methods and agent technology, in building more intelligent agents.

In this paper we show that Data Mining machine learning tools and techniques offer a promising method for the development of intelligent agents. Using data mining techniques based on machine learning makes it possible to develop agents able to learn from and adapt to their environment. In this way, the performance of these agent systems can be improved.

We demonstrate the use of learning techniques and tools to extract models to generate predictions or new behaviors. By Using a neural network as a reasoning mechanism and by analyzing input and output , our agent will acquire a new int. behavior which it did not possess from the beginning. It will discover its motion low by doing predictions outside the range under analysis.

It is our future intent to apply and assess several different learners to a real problem and compare their performance in order to choose the best among them.

References

1 Symeonidis A, Mitkas P, Agent Intelligence Through Data Mining, Springer, (2005).

2 Zhang C, Zhang Z, Cao L. Agents and Data Mining: Mutual Enhancement by Integration, (Eds) Gorodetsky V, Liu J, Skormin V.A, Autonomous Intelligent Systems: Agents and Data Mining, 50-61, Springer, 2005

3 Mitkas P, Knowledge Discovery for Training Intelligent Agents: Methodology, Tools and Applications, International Workshop AIS-ADM 2005 Russia, June 2005, Proceedings.

4 Symeonidis A, Mitkas P, A Methodology for Predicting Agent Behavior by the Use of Data Mining Techniques, International Workshop AIS-ADM 2005 Russia, June 2005, Proceedings.

5 Remondino M, Correndo G, Data Mining Applied to Agent Based Simulation, Proceedings 19th European Conference on Modelling and Simulation Merkuryev Y, Zobel R, Kerckhoffs E, ECMS, 2005.

6 Russell S, Norvig P, Artificial Intelligence A Modern Approach, Prentice Hall, 1995, 581.

Conceptions of Teaching using Virtual Learning Environments: Findings from A Phenomenographic Inquiry

Petros Lameras[1], Philippa Levy[2], Iraklis Paraskakis[3]

[1,3]South East European Research Centre, Research Centre of the University of Sheffield and CITY liberal studies, 17, Mitropoleos str. 54624, Thessaloniki, Greece, plameras@seerc.org[1], iparaskakis@seerc.org[3]

[2]The University of Sheffield, Dept. of Information Studies, Regent court, 211Portobello str Sheffield S1 4DP, UK, p.levy@shef.ac.uk

This paper reports a sub-set of findings from a phenomenographic investigation into Greek university teachers' conceptions of, and approaches to, teaching and learning using virtual learning environments (VLEs). Analysis of open-ended interviews with a small group of computer science teachers revealed three qualitatively different conceptions of teaching using VLEs, and different pedagogical approaches associated with them. The 'information transfer' conception was associated with information- (or content) - oriented pedagogical approaches to using VLEs. 'Understanding concepts' and 'developing concepts' conceptions were associated with more activity- (or process-) oriented approaches. The data suggest that there may be a relationship between the subject being taught (within the broad discipline of computer science) and different conceptions and approaches to using VLEs. Although based on a very small sample, the pilot study also suggests that pedagogical conceptions and approaches associated with networked learning may not be widely prevalent in computer science teaching in Greece.

Keywords
phenomenography, virtual learning environments, higher education, Greece.

1. Introduction

This paper reports a sub-set of findings from an in-depth, phenomenographic study of Greek university teachers' conceptions of, and approaches to, the use of virtual learning environments (VLEs) such as WebCT, Blackboard or Moodle in teaching. The project is investigating the relationship between computer science teachers' conceptions of, and approaches to, teaching using VLEs and theoretical perspectives of relevance to the e-learning field, in particular those that are associated with the principles and practices of networked learning as a distinctive form of e-learning ([1]; [2]). The paper presents findings from a small-scale pilot study. It concludes by suggesting possible implications for the further development of networked learning practice in the discipline of computer science in Greek higher education.

2. Prior Research and Theoretical Background

The rapid development of ICTs is having significant impact on teaching and learning in universities in Greece as in the wider international context. In the Greek context, international ICT policies mediated by the state are influential in institutional implementations [3]. The use of ICTs in Greek HE encompasses many different kinds of software tools, including VLEs. However, there is a tendency for both development and research initiatives to focus mainly on the technological aspects of the various software applications, rather than on pedagogical issues. Consequently, the use of VLEs (and other e-learning applications) in Greek universities is being taken forward principally by individual academics' enthusiasm to enhance their practice and is based on personal perceptions about pedagogical practice rather than on research-informed development activity.

The theoretical background to this study includes previous phenomenographic research that has explored university teachers' conceptions of, and approaches to, teaching in different discipline areas (e.g. [4, 5]). In a review of thirteen such studies, Kember [6] notes that while there are some differences in terminology, there is also a high level of agreement on findings, in that most findings converge on five conceptions of teaching which can be located on a continuum from a teacher-centered, content-oriented conception to a student-centered and learning-oriented conception of teaching. These are:

- Teaching as imparting information
- Teaching as transmitting structured knowledge
- Teaching as an interaction between the student and the teacher
- Teaching as facilitating understanding on the part of the student
- Teaching as bringing about conceptual change and intellectual development in the student

Previous qualitative research into university teachers' use of ICTs includes investigations of academics' educational beliefs and intentions as related to learning technology, for example using grounded theory. There has been some, but relatively little, phenomenographic work (e.g. [7]). Roberts investigated academics' conceptions of, and approaches to, using the web, firstly using a survey and then a phenomenographic approach. This study identified two distinct orientations focusing on "efficiency" and "effectiveness" respectively, and suggested that networked learning approaches would be more likely to be adopted by teachers whose conceptions and motivations are more oriented towards learning facilitation than resource efficiency.

Research indicates that teachers often hold more than one conception of (and approach to) teaching and that these are influenced by factors such as the cultural and educational environment, technical expertise and staff development opportunities. The ways in which teachers conceive of teaching are likely to be reflected in the way they use ICTs (e.g. as an information bank, or for dialogue in peer learning communities). .

3. Research Approach

A phenomenographic research approach is being adopted for this study. Phenomenography is an interpretive approach that seeks to identify the range of variation in ways of perceiving or experiencing a phenomenon of interest [8], in this case the way in which Greek university teachers understand the educational use of VLEs.

For the pilot study, the sample included 5 computer science academics from 3 universities in Greece. All had been using VLEs in their courses over a period of 1 to 5 years. The interviews (generally 30-40 minutes) concentrated on seeking variations in pedagogical purposes in using VLEs for teaching and learning, and in approaches to their use. The interview questions were open-ended and were developed broadly as follows:

- Please can you describe how you use the VLE in your teaching?
- What are you trying to achieve in using the VLE?
- What do you see as the value of the VLE for teaching and learning?
- How do you help your students to learn with the use of the VLE?

The accepted procedure in a phenomenographic data analysis is to use an iterative process to investigate the relation between meaning (how the phenomenon is experienced) and structure (different ways of experiencing). To identify this relation in this pilot study, a four phase analysis procedure was used based on Marton and Booth's [8] and Bruce's [9] processes of analysis. These phases were:

> *Becoming familiar with the transcripts:* The set of accounts extracted from the interviews were read and re-read several times before any attempt to begin the formal analysis.

Identifying meanings: In addition to the work already done in highlighting key themes and potential variations, the transcripts were loaded into Atlas/TI qualitative analysis software for textual analysis.

Structuring experiences: The outcome of this phase was 'categories of description'. There was a focus on discerning what was in the foreground and background of teachers' awareness in relation to different conceptions. Dimensions of variation emerged that consisted of a number of themes that existed in all categories, but were experienced in qualitatively different ways in each different category.

Establishing the outcome space: This phase involved arranging the categories of description into a logical structure, representing both the meaning and structural aspects.

4. Findings

This section describes teaching and learning using VLEs as conceived by the 5 academics who participated in the pilot study, and the pedagogical approaches associated with different conceptions. Three primary categories were identified along with a number of subcategories. All three categories describe the use of VLEs for teaching and learning as it is seen, experienced or understood by teachers of computer science.

Category 1 Information transfer

- Providing information
- Extending information
- Clarifying information

Category 2 Understanding concepts

- Understanding the topic
- Rethinking the topic

Category 3 Developing concepts

- Developing the topic

Each category is accompanied by a figure depicting the awareness structure where a number of broad areas of focus are presented, some in the foreground and some in the background. In this pilot study, these areas of focus consisted of information, technology, concepts and interacting.

The categories have been illustrated with quotes from the transcripts, referenced as Academic 01 to Academic 05. In relation to each conception, a number of factors, or dimensions of variation as phenomenography terms them, have been identified and described: focus on teaching, focus on learning, role of the teacher, role of the student, focus on use and focus on context.

4.1 Category 1: Information Transfer

The focus is on providing learning content through an information repository, and on directing students to subject information located in various websites. VLEs are also used for providing subject information located in course slides, ebooks, web sites online exercises, online tests etc or administrative information such as course announcements.

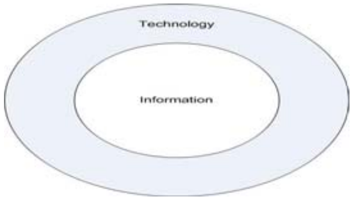


Figure 1 Structure of Awarenes

Transferring information consists of three sub-categories: (1) Providing information (2) Extending information and (3) Clarifying information. Information (content) is the primary focus of awareness, and the technology fulfils the purpose of transmitting information in terms of provision of access to exercises, online tests and lecture notes (see figure 1). This may mean that teachers perceive these resources as ‘information’ in terms of storing them to the VLE for students to access and retrieve them rather than to provide interactive feedback by solving online tests or to engage students in a discussion activity where lecturing notes will be further explained and discussed. Therefore, the focus is on directing students to relevant information without focusing on creating any kind of learning interactions.

4.1.1 Subcategory A: Providing Information

This subcategory describes the online provision of information created to be used off-line first, for supporting face-to-face lectures and other in-class activities. Teachers saw providing this information as a means to ensure coverage of subject matter. It was also highlighted that by using an information repository, students could conveniently access course material beyond the classroom context. Teachers’ conceptions of using VLEs are, therefore, directed towards content-focused features and perceived benefits of the VLE.

> … as an information means with students in a sense of loading up material that is being used during the course like Powerpoint slides, exercises, exercise notifications when the deadline has passed etc, for being aware of the subject matter that needs to be covered (Academic 01)

4.1.2 Subcategory B: Extending Information

Teachers identified the importance of providing extended information, such as giving links to particular websites and including eBooks, in contrast with subcategory A. These learning resources were for on-line use as they were perceived as substantial source of information that could be used to provide access to content not presented during class because of time constraints. Also, the extension of learning content was perceived as significant because the student would not have the opportunity to ask for explanations since the teacher would not be present when the student is accessing the information:

> ...for the material that I am uploading, at least I am trying to be as clear and understandable as possible by including further explanations, further examples and less complex meanings. I think this is the purpose of enhancing material (Academic, 05)

4.1.3 Subcategory C: Clarifying Information

In this subcategory, the VLE was perceived to be useful for providing explanations of small points that were perceived to be important for performing well in assessments. Clarification of information was not accomplished only by the teacher but also by students who could exchange learning content through the use of a bulletin board, to help peers to clarify 'minor' points of information. This process was felt to be important in terms of clarifying transferred information provided both in online resources and in-class for creating a mutual solidarity between students.

> a student may ask his peer to help him with this file or what this exercise says etc he was creating a question and the others were answering this question… mostly the forum was used for resolving simple questions and this was good for creating solidarity between students (Academic, 01)

4.1.4 Dimensions of Variation

Focus on teaching

As described earlier, teaching using VLEs is represented in participants' statements as imparting information.

Focus on learning

Teachers saw learning via VLEs in terms of accessing clearer and richer information.

Role of the teacher

Teachers felt that they had an administrative role in terms of directing students to learning resources and a further role in terms of providing clarification of information.

Role of the student

The role of the student was perceived as using the VLE for receiving subject information which would help the student to grasp and memorise concepts.

Focus on use

Teachers conceived of the VLE as a means of providing access to, or clarification of, content. They approached the use of technology as supportive to face-to-face teaching.

Focus on Context

The contextual focus was on providing information, through the use of the VLE, for passing the course.

4.2 Category 2: Understanding Concepts

The primary focus of awareness is on students' ability to understand both the general concepts of the course (i.e. understanding the holistic issues of a module, e.g. principles underpinning programming) and meanings related to a particular topic (i.e. understanding specific definitions and principles, e.g. of a specific programming language). Information provision is a secondary focus, and technology is in the background. The use of the VLE is experienced as a medium for facilitating material as in Category 1. It was seen as a medium for supporting subject analysis through expressing opinions and ideas, resulting in learning. Engaging students in online discussions was perceived as a way to motivate and increase student participation in this process in comparison with face-to-face discussions. Understanding concepts has two sub-categories: **understanding the topic, and rethinking the topic.**

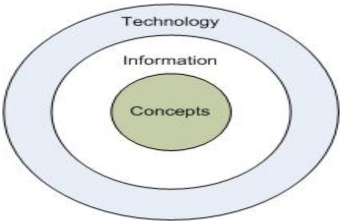


Figure 2 Structure of Awareness

4.2.1 Subcategory A: Understanding the Topic

This sub-category focuses on the ability of students to engage with and understand the structured meanings of the subject. Approaches to using the VLE include providing feedback via emails, triggering students' interest by explaining concepts using various tools of the VLE (e.g. whiteboards, Web links,) and monitoring progress through online tests and quizzes

> …my effort is to keep the interest of my students alive around the meaning of my subjects […] if it is needed to show it in detail in slides or demonstrating it through the VLE while they are working with a particular software (Academic 03)

> Sometimes I prepare to [the VLE] some tests for formative evaluation that the students want in order to have a hint on where they stand (Academic, 04)

4.2.2 Subcategory B: Rethinking the Topic

In this sub-category the VLE is used to help students reflect on the topic by applying theoretical concepts in online exercises, and by developing awareness of how other students conceive the same issues. For some, online discussions were perceived as useful in conjunction with face-to-face, for students to further explore previously expressed ideas and opinions (off-line) and feed in to further offline discussions.

> Each student can see particular meanings from other students and therefore, a more complete learning experience is created through historical references (Academic, 03)

4.2.3 Dimensions of Variation

Focus on teaching

Teaching using VLEs is represented as transmitting structured knowledge, for example in terms of explaining concepts through giving feedback.

Focus on learning

Teachers saw learning via VLEs in terms of providing a medium for reflection on concepts and access to feedback.

Role of the teacher

Teachers felt they had a role in helping the student to understand concepts and in organising the online learning environment to ensure it included opportunities for access to content, online questions, provision of feedback and online exercises.

Role of the student

The role of the student was perceived as using the VLE to individually construct meaning for understanding the topic.

Focus on use

Using the VLE for provision of information was not a primary concern. Teachers conceived of the VLE as a means of engaging students in interactions of various kinds that would provide feedback. The VLE was seen as a means of increasing participation in discussions in comparison with face-to-face interactions.

Focus on context

The contextual focus was on using the VLE to promote greater, more holistic, understanding of the topic.

4.3 Category 3: Developing Concepts

The primary focus of awareness is on students developing their own ideas and concepts by sharing and adding new ideas and opinions, and at the same time contributing to the development of the topic. Teachers' use of the VLE is primarily to enable students to interact with the teacher and peers. As in category 2, activities such as online discussions are used but the focus is on sharing and criticising ideas for developing concepts. This category has one sub-category: **(1) Developing the topic**

> I think that this is the only way to develop their thoughts, if you share because even if you are very clever if nobody knows what you know then you are alone, I mean students have to share… if you read what other people say maybe you get some ideas and say oh yes this is true and expand your ideas. (Academic, 02)

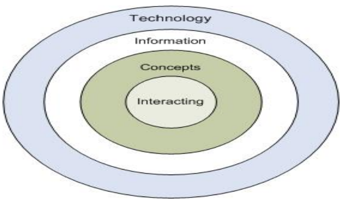


Figure 3 Structure of Awareness

4.3.1 Subcategory A: Developing the Topic

In this subcategory, teachers were using the VLE to help students to contribute and share opinions and ideas for developing the concepts of the topic. The VLE was used with the purpose of promoting ‘sharing’ interactions and reflection through discussions. To achieve that, the aim was to motivate students to participate in online discussions by posting interesting subjects.

> …with a kick off point the discussion can direct you wherever you like (Academic, 04)
>
> I think if it is really interesting for them, they will take part so that is why I have to see what is interesting for them I mean in the frame of the subject (Academic, 02)

4.3.2 Dimensions of Variation

Focus on Teaching

Teaching using VLEs is represented as facilitating further development of students’ conceptions they already hold.

Focus on learning

Learning using VLEs is focused on students developing concepts and the topic by sharing opinions and ideas.

Role of the teacher

The teacher’s role in the VLE is as participant in the environment as well as organiser, motivator and provider of information.

Role of the student

The role of students in using the VLE is as self-motivated explorer and developer of concepts which might also feed in to development of the module.

Focus on use

The VLE was conceived as a means of providing an environment for creating interactions for reflection and criticism. Using the VLE for and discussing about issues in more socially-focused subjects felt it was important for developing their concepts and the topic

Focus on Context

The contextual focus was on using the VLE to promote critical engagement with the topic and development of professional expertise for the workplace.

5. Discussion and conclusions

In the first category, conceptions of, and approaches to, the use of VLEs were based on information transfer and access - ‘imparting information’ in terms of Kember’s [6] typology. That is, teachers were directing students to relevant learning resources either firstly offline through the uploading of learning content or by recommending relevant Web sites where students could acquire information through online use. This content-oriented approach reflects an 'imparting information' conception of teaching as the VLE is used for presenting, memorising, and reviewing information that will enable quicker and better understanding.

In the second category, the conceptual focus is on using the VLE to enable students to understand a module's concepts. Teachers used the VLE for online exercises in applying theory to practice and for the provision of feedback in order to communicate structured information, or promote interaction between teacher and student, in Kember's [6]) terms. Online interaction is focused on provision of access to previous ideas and opinions.

In the third category, the conceptual focus is more on using the VLE to support personal meaning-making through social negotiation, perhaps more closely reflecting the principles of networked learning. The VLE is used for enabling learning relationships from the student perspective. The perception is that reflection and criticism could be developed by expressing and sharing views online. Presenting interesting subjects that would motivate students to share their opinions was regarded as an important pedagogical strategy. All of this may be consistent with a 'facilitating understanding' conception of teaching.

Taking into consideration previous research on conceptions of teaching using the Web (e.g.[7]) there is broad consistency between the categories of description. In Robert's study, the "subject information" conception was identified as content-focused while the "self-paced learning" and "dialogue" conceptions were identified as more focused on learning processes. Similarly in this study, the "information transferring" conception was focused on provision of, and engagement with, subject information, while the "understanding concepts" and "developing concepts" conceptions were more oriented towards activity-based pedagogy. In the pilot study the 'transferring information' conception was more prevalent in the context of the teaching of more computational courses and the 'developing concepts' conception in the context of the teaching of more open-ended, socially-focused subjects. It may prove useful to further explore these associations in the full study, as it is possible that subject area may impact on the way in which the use of VLEs in teaching and learning is conceptualised and instantiated within the discipline of computer science. More generally, although based on a very small sample, the pilot study suggests that pedagogical conceptions and approaches associated with networked learning may not be widely prevalent in computer science teaching in Greece.

References

1 Goodyear, P., *Psychological foundations for networked learning*, in *Networked Learning: Perspectives and Issues*, C. Steeples, Jones, C, Editor. 2003, Springer: London.

2 Levy, P., *Living theory: a pedagogical framework for process support in networked learning.* ALT-J, Research in learning technology, 2006. 14(3): p. 225-240.

3 Kokosalakis, N., *Lifelong learning in Greek Universities: policies, practices and prospects*. European Journal of Education, 2001. 36(3): p. 329-339.
4 Prosser, M., Trigwell, K, Taylor, P, *A phenomenographic study of academics' conceptions of science learning and teaching*. Learning and Instruction, 1994. 4: p. 217-231.
5 Prosser, M., Trigwell, K, *Understading learning and teaching: the experience in higher education*. 1999: Buckingham, Society for Research into Higher Education & Open University Press.
6 Kember, D., *A reconceptualisation of the research into university academics' conceptions of teaching*. Learning and Instruction, 1997. 7(3): p. 255-275.
7 Roberts, G., *Teaching using the Web*. Instructional science, 2003. 31: p. 127-150.
8 Marton, F., Booth, S, *Learning and Awareness*. 1997, NJ: Lawrence Erlbaum Associates.
9 Bruce, C., *The seven faces of information literacy*. 1997, Australia: Auslib Press Pty Ltd.

Understanding Students' Cultures from South East Europe Studying in Greek Higher Education

George Katakalos[1], Jose Miguel Nunes[2], Panayiotis Ketikidis[3]

[1]SEERC, 17 Mitropoleos Str, Thessaloniki, Greece, gkatakalos@seerc.org

[2]The University of Sheffield, Dept. of Information Studies, Western Bank, Sheffield, S10 2TN, UK, j.m.nunes@sheffield.ac.uk

[3]City College, Dept. of Computer Science, 13 Tsimiski Str, 54624 Thessaloniki, Greece, ketikidis@city.academic.gr

Higher Education within the European Union is changing both in terms of pedagogical models, International reach of the universities and an increasing utilization of information and communication technology for teaching and learning purposes. In fact, with the introduction of the World Wide Web in formal educational settings, an increasingly diverse cohort of students is being asked to interact with each other in multicultural environments for which they may be ill-prepared. This raises the rapidly increasing need for the consideration of cultural differences in communication, collaboration and education. But, dramatic geo-political changes in the South East European (SEE) region have constituted previous research outputs concerning the identification and comparison of national cultures inadequate for SEE region. This paper presents and discusses the differences of applying Hofstede's questionnaire on students coming from SEE countries and studying in Greece.

Keywords
eLearning, South East Europe, culture, student interaction.

1. Introduction

Higher Education (HE) in European Union (EU) was initially established by the Maastricht Treaty which provided an initial framework facilitating the development and exchange of information and experiences between Higher Education institutions across the Union. In 1998, HE ministers from a subset of the EU had started what today is known by the Bologna process in response to the Sorbonne Declaration. This process aimed at facilitating and enhancing student and teacher mobility and transparent recognition of courses and modules across the different educational systems through a common degree-level system. The noble aim was to develop a European area of HE supported by an underlying European cultural dimension, namely through the creation of integrated programmes of study.

With the introduction of the World Wide Web (WWW) in the formal education settings physical student mobility is not necessarily implied. In fact, WWW with the rapid development of Information and Communication Technologies (ICT) in HE has caused changes in the way teaching and learning is viewed. More specifically, ICT started to remove the traditional time and geographical space barriers of formal education that limit access to the different educational systems [19]. Without these traditional barriers, distance learning, eLearning and remote access to learning materials are being increasingly accommodated in the formal education and are even causing the education to be viewed as an export factor between countries [5]

However, this vision of student mobility will imply an increasing diversity in the student cohort. Thus, students will increasingly be asked to interact with other students in multicultural environments for which they may be ill-prepared. Nevertheless, globalisation is seen as the main force that causes a rapidly increasing need for the consideration of cultural differences in communication, collaboration and education [19, 14, 15, 11].

Previous researches like [8] identified cultural characteristics at a national country level making possible the comparison between cultures seem now to be inadequate. Specifically, concerning the South East European (SEE) region dramatic changes have taken place making the democratisation difficult to establish [1]. These dramatic geo-political changes are attributed as a post-war effect and a post-communism effect [12]. Therefore, new studies should be made in order to produce up-to-date results reflecting the changes taking place in SEE region. This paper aims at presenting and discussing the findings of a replication of Hofstede's study in students coming from SEE countries and studying in Greek HE indicating the differences with Hofstede's original data that seem to reflect the changes undergone in the SEE region.

2. Literature Review

Defining culture is not an easy task. Culture does not remain committed to our birth culture and can be affected by many ways [7]. Due to the numerous ways with which culture can be affected a variety of definitions exist. For example, in Western languages, culture commonly means "civilisation or refinement of the mind and in particular the results of the refinement including education, art, literature" [4, 7]. This is a narrow definition of the culture that is restricted on a particular viewpoint, the Western view. The most pervasive view defines culture to be a demonstration of ways in which an identifiable group adapts to its changing environment [19, 7]. In other words culture can refer to individual culture, and/or organisational culture and/or national culture, and/or academic culture. Therefore, people may belong to more than one cultural groups and thus, possessing a subset of each culture's identifiable characteristics.

One aspect that most definitions of culture seem to have in common is the view that it is learnt and not innate, and that it is a response to peoples' environments.

Anthropologists agree on three characteristics of culture, which “it is not innate, but learned; the various facets of culture are interrelated - you touch a culture in one place and everything else is affected; it is shared and in effect defines the boundaries of different groups”[4]. He further argues that it is those aspects of culture that are least recognised that “have the greatest influence on behaviour.” [4].

From the previous discussion, it can be deduced that culture is continuously changing and does not remain the same throughout our lives. Thus, the process of defining culture is a problematic task. For the purpose of this paper culture is defined at a national/country level referring to behaviours displayed by the majority of individuals inside a country.

In order to specify cultural charactaristics Hofstede’s research was used. Hofstede studied a large body of survey data of IBM employees in 50 different countries. Hofstede’s study revealed that employees faced common problems but with solutions differing from country to country in the following areas [8]:

- Social inequality, including the relationship with authority;
- The relationship between the individual and the group;
- Concepts of masculinity and femininity;
- Ways of dealing with uncertainty.

These four areas identified in Hofstede’s study were already identified by [10] and represented dimensions of culture. A dimension according to Hofstede is “an aspect of a culture that can be measured relatively to other cultures” [7]. Therefore, in his research in 1980, Hofstede named the above four areas as: power distance (from small to large), collectivism versus individualism, femininity versus masculinity and uncertainty avoidance. A fifth dimension has also been identified, long term orientation vs. short term orientation. This had not been discovered before because of a “Western” way of thinking by the researchers, but has been revealed by a following study on the IBM data of people’s values around the world using a questionnaire composed by “Eastern” minds [8].

It could be argued at this stage that Hofstede’s research was used in IBM to measure employees’ problems and not students’ problems in Higher Education (HE). As explained in the introduction the influence of globalisation in education reflected in the European Union’s treaty’s and directives, has caused changes in the ways we perceive formal educational learning by embracing terms like distance learning, eLearning and remote access to learning materials leading to the development of a European area of HE. Thus, students are working in multicultural teams. Moreover, HE urges learning to be authentic and embedded in realistic and relevant real-world contexts [16]. In respect to this statement, a student can be viewed as an “employee” in training mode since the aim of HE is to train students for real-world situations. Therefore, a replication of Hofstede’s questionnaire for HE can provide us with better understanding of the students’ cultural dimensions.

3. Research Question and Results

As stated in the introduction, this paper presents and discusses the findings of the application of Hofstede's questionnaire on students from South East Europe studying in CITY Liberal studies, a private university located in Greece. Many researchers and authors are still referring to Hofstede's research dated back in 1960 as the only good example [13, 11]. The questionnaires discussed on this paper were used as a tool for better understanding cultures. The research question answered in this paper is: "how students' culture coming from different SEE countries may hinder the students' interaction" For the purpose of this paper, the authors have selected students studying in Greece and coming from SEE countries as their sample of applying Hofstede's questionnaire. More specifically, the students were studying for a first degree of Computer Science and Business Administration and Economics departments of CITY Liberal Studies. The results of our study showed that new measurements of these values are needed in order to accurately reflect major changes taking place in the SEE region.

In total forty questionnaires have been administered to students. The questionnaire administered is Values Survey Modules (VSM) '94 which is an altered version of the original questionnaire used by Hofstede in his initial research [6]. Until VSM 94, the questionnaire passed through several revisions as follows:

- VSM 81 experimental version was issued;
- VSM 82 which included forty seven questions plus six demographic;
- VSM 94 which included 26 questions plus six demographic. It is in the VSM 94 that a fifth dimension was added.

The students participating in this study were coming from Greece, FYROM, Serbia, Albania and Bulgaria and were undergraduate students of Computer Science and Business Administration and Economics Departments. The data resulted from questionnaires was analysed by using the formulas provided by Hofstede. More specifically, Hofstede's research has provided four dimensions of culture:

- *Power Distance Index (PDI)* represents "the extent to which the less powerful members of institutions and organisations within a society expect and accept that power is distributed unequally" [7]. Countries scoring low are more likely to de-emphasise differences between power and wealth.
- *Individualism (IDV)* typifies societies in which "the ties between individuals are loose: a person is expected to take care of himself or herself and his or her immediate family" [7]. A low Individualism ranking typifies societies that are more collectivist in nature with close ties between individuals.
- *Masculinity (MAS):* refers to the distribution of the roles between the genders [7]
- *Uncertainty Avoidance (UAI)* is defined as "the extent to which the members of institutions and organizations within a society feel threatened by uncertain, unknown, ambiguous or unstructured situations" [7]. A low score indicates that a country has weak Uncertainty Avoidance.

The results for each dimension are presented in Figure 1. It should be noted that long term orientation (LTO) dimension was not included as part of the analysis because the overall students are only studying for three years, an insufficient time to measure accurately LTO.

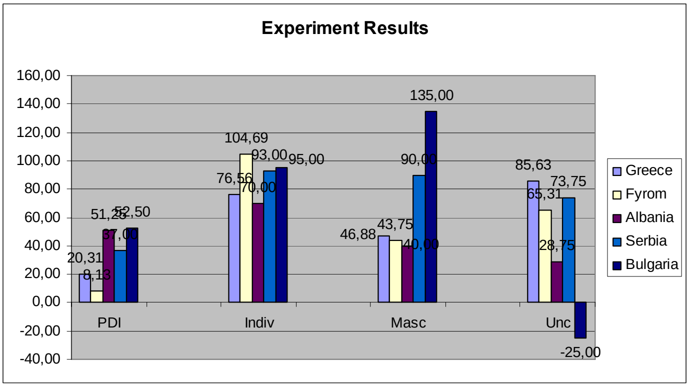


Figure 1 Experiment Results

The results of the analysis as represented in Diagram 1 showed that there is a variation from 8.13 for FYROM to 52.50 for Bulgaria concerning the Power Distance Index (PDI). This difference in scores indicates that student' cultures coming from FYROM and Bulgaria could be considered as antithetic. Concerning the individualism (IDV) dimension the participating countries' scores are almost the same ranking from 70 for Albania and 104.69 for FYROM. These scores indicate that all the countries are individualistic in nature looking more for the individual and less for the collective groups. In the masculinity (MAS) dimension is observed the highest difference. The minimum score is 40 for Albania and 135 for Bulgaria. This indicates the second group of antithetic cultures comprised by the countries scoring low (Greece, FYROM and Albania) and those scoring high (Serbia and Bulgaria). Finally, in the uncertainty avoidance dimension (UAI) differences in scores can be clearly observed. The minimum score was -25 for Bulgaria indicating that Bulgarians are more likely to follow uncertain situation and procedures with the tendency to be more prone to risk, and the maximum was 85.63 for Greece indicating a country that do not prefer uncertain situations and avoid risk. Concerning the score of Bulgaria (-25) indicates that the country is weaker on uncertainty avoidance, since the theoretical ranges according to this dimension are -90 to 120 [9].

4. Findings: Differences with Hofstede's Results

As stated in section 2, Hofstede's results may not be accurate enough to reflect the changes happening in SEE. In the following subsections each dimension will be compared with Hofstede's original data, highlighting the changes reflected in our results. It should be noted that Hofstede calculated the dimensions for Yugoslavia which was consisted by FYROM, Serbia, Montenegro, Bosnia and Herzegovina, Croatia and Slovenia.

4.1 Power Distance (PDI)

The results presented in Figure 2 indicate a change in PDI scores. This change can be attributed to the post-war and post communism effect [12]. More specifically as argued by Barlett after the fall of the communist regime in the SEE region, democratisation process has begun with the help of EU [1]. But democratization in the region is difficult to be achieved as there are open issues between the newely created countries like the case of Serbia and Montenegro resulting in economic embargos between the involved countries [1]. This situation is clearly reflected in the results of our study since the scored indicate a small decrease. The only exception is FYROM indicating a huge difference from Hofstede's results that cannot only be explained by the separation of Former Yugoslavia. This difference clearly reflects the situation as described by Barlett that Croatia and FYROM are rapidly implementing and accommodating democratisation and economic reforms [1].

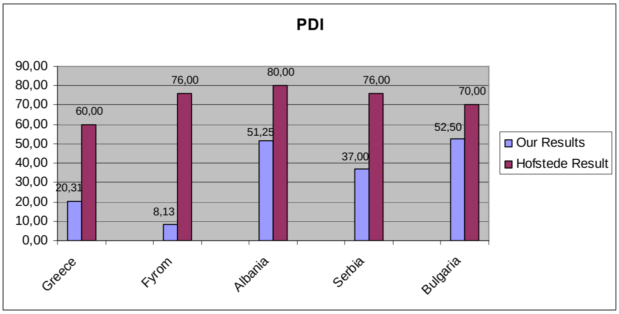


Figure 2 PDI Comparison

4.2 Individualism versus Collectivism (IDV)

Hofstede in his work stated that PDI and IDV are likely to be negatively correlated [7]. Our findings indicate this negative correlation between the two dimensions. It should be noted that all the participating countries in this study scored all high on this index. As argued also in the section 4.2, this high scoring can be explained by the fall of the communism that caused all the involved countries in the regions to rapidly adapt to the democratic regime [12, 1]. According to Limset this caused the creation of a strong state authority [12]. Therefore, it could be expected that these countries would have scored high on the individualism as a post communism effect.

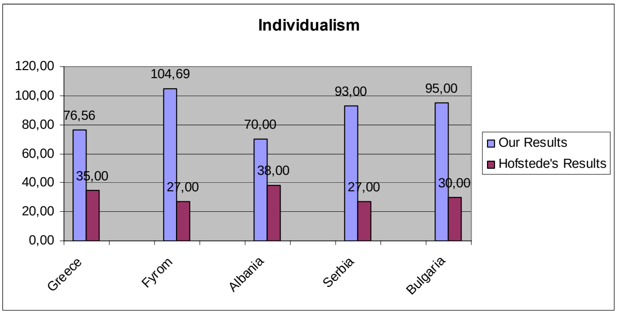


Figure 3 Individualism Comparison

4.3 Masculinity versus Femininity (MAS)

As stated by Hofstede in a masculine society gender roles are clearly distinct: men are supposed to be assertive tough and focused on material success and women to be modest, tender and concerned with the quality of life [7]. A low score on this dimension indicates a country feminine in nature.

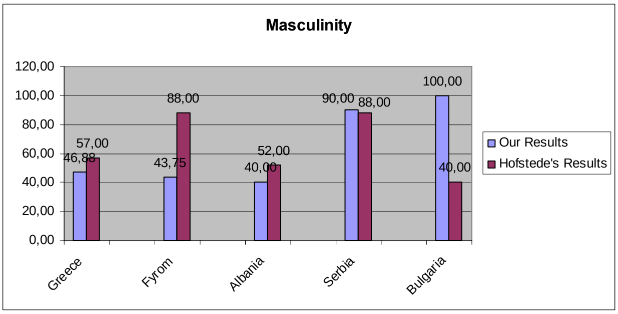


Figure 4 Masculinity Comparison

As presented in Figure 4 the scores for Greece, Albania and Serbia are almost the same with Hofstede's original values. The great difference is observed in the scores for Bulgaria and FYROM. Bulgaria scored high on this dimension because only male Bulgarian students from Computer Science participated in our study. More male students enroll on Computer Science course indicating that this phenomenon is common to all Computer Science departments in Europe and the researchers could not locate Bulgarian students in the other department [17, 3]. Thus, Computer Science department in our university is no exception. The difference of scores concerning FYROM reflects the argumentation made in section 4.1 stating that Crotia and FYROM are democratising in rapid way [1]. Thus, in democratic countries equality between gender roles is fostered.

4.4 Uncertainty Avoidance (UAI)

As indicated in Figure 5, the only country score similar to Hofstede's is for Greece. The differences in scores for the other countries can reflect the instability in SEE region. Albania is so far below the EU level concerning the political and economical infrastructure clearly affecting the individuals uncertainty levels [2]. Concerning Bulgaria, the EU's accession negotiations has made considerable impact on the Bulgaria's political agendas by harmonising policies to the EU standard [18]. According to Spendzharova this continuous EU accession has created a feeling of uncertainty for Bulgarians. FYROM and Serbia scores can be attributed to the post-war effect [18].

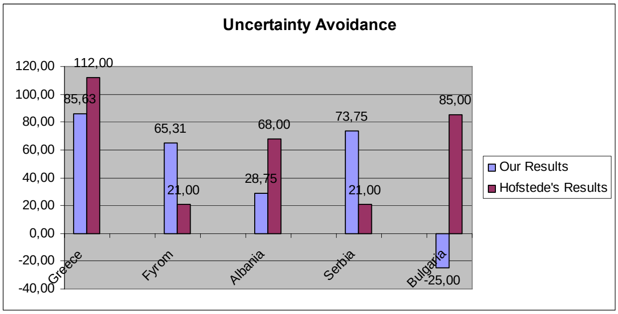


Figure 5 Uncertainty Avoidance Comparison

5. Conclusions

Higher Education is influenced by globalisation. EU has set the aim to develop a European area of HE supported by an underlying European cultural dimension, namely through the creation of integrated programmes of study. WWW combined with ICT started to remove the traditional time and geographical space barriers of formal education that limit access to the different educational systems. This has caused a consideration of cultural differences in communication, collaboration and education. Previous researches like Hofstede in 1980 identifying cultural characteristics at a national country level making possible the comparison between cultures seem now to be inadequate specifically for the South East European (SEE) region.

A replication of Hofstede's questionnaire was made in order to identify potential changes in the original values for the countries dimensions. The results of the analysis have shown that values of Hofstede's dimensions have dramatically changed for Greece, FYROM, Albania, Bulgaria and Serbia. The new values identified in this research seem to comply with the current unstable situation taking place in SEE region. The differences between our results and Hofstede's can be summarized in two main areas: the fall of the communism regime and post-war effect of former Yugoslavia. Unfortunately the results of our studies cannot be generalised so as to cover the national/country culture. For the generalisation to be made possible a national scale research should be performed.

To summarise, the results of this study showed that differences were exhibited in three of the four cultural dimensions: Power Distance (PDI), Masculinity versus Femininity (MAS) and Uncertainty Avoidance (UAI). Based on this study's results the following countries can be characterised as potentially antithetic in respect to the cultural dimensions: for PDI: FYROM with Bulgaria, for MAS: Albania with Bulgaria, for UAI: Greece and Albania with Bulgaria. Concerning our research the conclusions that were drawn in respect to the dimensions were:

- **PDI**: Potential communication problems may probably take place between FYROM and Bulgaria. On the one hand FYROM students may follow a more student-centred approach to teaching and learning assuming more responsibilities and leading roles. On the other hand, Bulgaria students may prefer a more teacher-centred approach to teaching and learning by requiring the presence of teacher leading and guiding the discussion;
- **IDV**: No communication problems are expected to arise since the countries' scores are close and high;
- **MAS**: Potential communication problems may arise between Greece, FYROM and Albania with Serbia and Bulgaria. On the one hand, students coming from Greece, FYROM and Albania will probably follow general simple answers but not sacrifice quality for quantity. On the other hand, Serbians and Bulgarians are more likely to display an ego-centric behaviour in their posts, get the leading role of the conversation and prefer a teacher as a leader in our case a moderator of the discussion;
- **UAI**: Potential problems of communication may probably arise in the discussion between Albanians and Greeks. On the one hand, Albanians may probably prefer open-ended learning situations and discussions; they don't expect to know everything and results are attributed to own abilities On the other hand Greeks are more likely to prefer structured learning situations, may be concerned if an answer is right or wrong, and expect the teacher to have all the answers.

The values calculated by Hofstede's questionnaires provided the basis for understanding cultures. More specifically, the findings presented in this paper can provide an initial guide for devising a case study with potential not "culturally conflicting" questions that may hinder the interaction between students from different countries.

References

1 W. Bartlett, Samardzija, V.,, *The Reconstruction of South East Europe, the Stability Pact and the Role of the EU:An Overview*, MOCT-MOST: Economic Policy in Transitional Economies, 10 (2000), pp. 245-263.

2 European Commission, *The Western Balkans In Transition*, European Economy Enlargement Paper No.23 December 2004, Brussels: DG for Economic and Financial Affairs (2004).

3 T. Greening, *Gender stereotyping in a computer science course*, *The proceedings of the thirtieth SIGCSE technical symposium on Computer science education* ACM, New Orleans, Louisiana, United States, 1999.

4 E. Hall, T., *Beyond Culture*, Anchor Books, New York, 1977.
5 L. Harasim, Hiltz, S.R., Teles, L., and Turoff, M., *Learning Networks: A Field Guide to Teaching and Learning Online*, Cambridge, MA: MIT Press, 1995.
6 G. Hofstede, *VSM 94 - Values Survey Modules 1994*, *Available at: http://feweb.uvt.nl/center/hofstede/VSM.html [Last Accessed: 15/12/2007]*, 1994.
7 G. Hofstede, Hofstede, J. G.,, *Cultures and Organisations: Software of the Mind*, 2nd ed. London, McGraw Hill, 2005.
8 Hofstede G., *Culture's Consequences: International Differences in Work-related Values.*, Beverly Hills, California: Sage Publications, 1980.
9 Hofstede G., *Cultures and organizations: Software of the mind.*, London: McGrawHill, 1997.
10 A. Inkeles, Levinson, D.,.J., *National character: the study of modal personality and sociocultural systems.*, in Lindsey and Aronson, ed., *The handbook of social Psychology*, Reading MA: Addison-Wesley, 1969.
11 J.-E. Joo, *Cultural issues of the Internet in classrooms.*, British Journal of Educational Technology, 30 (1999), pp. 245-250.
12 S. Limset, M.,, *The Indispensability of Political Parties*, Journal of Democracy (2000), pp. 48-55.
13 Z. Mansfeldova, Sparschuh, V., Wenninger, A., *Patterns of Europeanization in Central and Eastern Europe*, Asia Europe Journal, 3 (2005), pp. 577-579.
14 C. McLoughlin, *Culturally responsive technology use: developing an on-line community of learners.*, British Journal of Educational Technology, 30 (1999), pp. 231-243.
15 C. McLoughlin, and Oliver, R.,, *Instructional design for cultural difference: A case study of indigenous online learning in a tertiary context*, *In: Winn J, ed., ASCILITE'99: Responding to Diversity Proceedings of the 16th annual conference of the Australian Online Socialisation Society for Computers in Learning in Tertiary Education*, Brisbane: Queensland University of Technology, 1999, pp. 229-238.
16 M. Mcpherson, Nunes, B., M.,, *Developing Innovation in Online Learning: An Action Research Framework* RoutledgeFalmer, 2004.
17 B. Schinzel, *Cultural Differences of Female Enrollment in Tertiary Education in Computer Science*, *Proceedings of the IFIP 17th World Computer Congress - TC9 Stream / 6th International Conference on Human Choice and Computers: Issues of Choice and Quality of Life in the Information Society*, Kluwer, B.V., 2002.
18 A. Spendzharova, B., *Bringing Europe In? The Impact of EU Conditionality on Bulgarian and Romanian Politics*, South East European Politics, 4 (2003), pp. 141-156.
19 M. Wild, and Henderson, L.,, *Contextualising Learning in the World Wide Web: accounting for the impact of culture*, Education and Information Technologies, 2 (1997), pp. 179-192.

A standalone SCORM Content Repository framework for Healthcare Education

Stathis Th. Konstantinidis, Panagiotis D. Bamidis

Aristotle University of Thessaloniki, Medical School, Laboratory of Medical Informatics, Thessaloniki, Greece, cs@med.auth.gr, bamidis@med.auth.gr

Information and communication technologies are integrated incrementally with all facets of educational procedures. International standards such as SCORM were introduced by educational technologists to enhance reusability and interoperability of courses. In addition, the IEEE LOM/Healthcare LOM was introduced to enhance the proper description of healthcare courses. However, their exploitation remains limited.. To account for this problem, educational content repositories were engineered based on specific architectures like CORDRA. Only a few proposals however exist dealing with communication architectures for repositories with clear decisions on copyright and availability of the learning objects content in the field of Medical Education. Herein, we propose a standalone SCORM Content Repository framework for Healthcare Education comprised of three basic modules: the standalone repository, the communication API and the meta-search engine. The latter is deviced in a way to fulfil the special needs of this specific education area, in parallel to closing certain gaps left open in the original proposal of the CORDRA architecture. Such gaps concern the determination of context-specific metadata, the indexing of learning objects in search, the level of presentation to users and the provenance of content packages.

Keywords

SCORM, Healthcare LOM, content repository, e-learning, CORDRA, medical education

1. Introduction

Information and communication technologies are integrated incrementally with all facets of educational procedures. Tutors are using e-learning environments as web references and web content management systems. New trends in educational technology necessitate full online courses, especially in the field of medical sciences. In terms of Continuing Medical Education the online distance learning is a demand.

Nowadays, any contemporary material that is going to be used for online courses or seminars should be fulfilling requirements such as Interoperability, Accessibility, Reusability, Durability, Maintainability and Adaptability, so that the material remains credible, up-to-date and tracks changes and developments of medical techniques and standards through time. Central to this activity is the employment of the emerging e-learning standard, namely, SCORM, Shareable Content Object Reference Model [5]. SCORM has been designed as a set of eXtensible Markup Language (XML) based specifications that can define, manage, access and deliver modular educational objects so that they are easily shared among different e-learning management systems. Until now SCORM was used usually at pilot courses [1]-[3].

Healthcare educational course material compliant with SCORM Content Packages may be additionally described by means of IEEE Learning Objects Metadata (LOM) or Healthcare Learning Object Metadata (LOM).

Consequently, courses that are compatible and conformant with SCORM can be described and may be exchanged by teachers. For this to successfully occur there is a need for web-based repositories, such as the one proposed by the Advanced Distributed Learning Initiative of the US government, namely, CORDRA [4]. There exist also other similar efforts [5]-[7] where the use of IEEE LOM for describing the content of the learning objects obey to educational requirements put forward by other scientific areas.

Limited numbers of research attempts focus, however, on communication architectures between the repositories or the emerging copyright issues or the decisions of the learning objects healthcare content availability. To fill this gap, we propose a framework for SCORM content repositories using Healthcare LOM. In so doing, we utilise the architecture for communication between separate systems proposed in [8] for standalone separate repositories, together with a meta-search engine enabling the metadata search within these repositories by means of appropriate APIs.

The remainder of this paper is structured as follows. In section 2 we provide a brief account on related work. In section 3 an overall architecture of the standalone SCORM Content Repository is described, while in the following section user definitions and security issues are given. A possible scenario of use is introduced in section 5, followed by a discussion of key issues.

2. Setting the scene

New trends in educational technology necessitate the use of international educational Standards, such as SCORM, IEEE LOM and Healthcare LOM. SCORM was developed from a U.S. Government's initiative in Advanced Distributed Learning (ADL) [9]. ADL created SCORM in order to enable the production of learning material compliant to requirements of interoperability, Accessibility, Reusability, Durability, Maintainability and Adaptability, so that the learning material remains credible, up-to-date and tracks changes and developments of medical techniques and standards through time. Furthermore, Healthcare Learning Object Metadata (LOM) [10] was conceptualised as an extension of IEEE LOM [11], including all the appropriate information for medical education as proposed by the MedBiquitous Consortium. These metadata are written in XML following the "http://ltsc.ieee.org/xsd/LOM healthcarelom.xsd" XML Schema Definition and is considered a proper way to describe SCORM content.

Despite the existence of those standards, SCORM has not been widely implemented and used in the medical domain [12], [13]. It is used mainly in pilot courses [1]-[3] where the appropriate needs were defined and minor results were obtained.

The creation of high quality online educational material by higher education institutions and medical associations is often hindered by the extra development cost, time required and the lack of know-how on international education standards. Educational content repositories came into focus, in order to solve this problem.

The USA government with the "Advanced Distributed Learning Initiative", a non-profit "Corporation for National Research Initiatives – CNRI", and academia "Learning Systems Architecture Lab – LSAL" developed an open, standards-based model for designing and implementing software systems for the purposes of discovery, sharing and reuse of learning content through the establishment of interoperable federations of learning content repositories, called Content Object Repository Discovery Registration/Resolution Architecture (CORDRA) [4].

Despite these efforts, some stand alone repositories where proposed [5]-[7] taking under consideration the use of IEEE LOM or similar metadata in order to describe the content of the learning object and in some cases taking into account the specific varying requirements of necessities of separate scientific fields.

To this extend, some content repositories permitting queries by keywords are: GEODE - Global Education Online Depository and Exchange [14], MLX - Maricopa Learning Exchange [15] and CAREO - Campus Alberta Repository of Educational Objects [16]. There also exist other repositories containing complementary education material about a specific scientific field such as the National Learning Network: Materials [17], PBS Teachers [18], MERLOT [18] and others. In addition, repositories that contain multimedia learning material for use by K-12 educators such as Learn-Alberta [19], and LearningLanguages.net [20], specify certain object technologies..

Regarding the field of medical science there are also a few existing relative repositories such as HEAL - Health Education Assets Library [21] providing resources such as images, cases and quiz questions, but these resources are not conformant and compliant with SCORM or the Healthcare LOM standard.

3. Overall architecture of the standalone SCORM Content Repository

In this paper a framework comprising of the three basic modules is proposed. These are the standalone repository, the communication API and the meta-search engine (cf. Figure 1).

3.1 Standalone SCORM content repository

The standalone repository consists of a database system that holds location and metadata of the SCORM content object or smaller learning objects.

A few years ago, a similar standalone repository was proposed as a long term vision for Grid-enabled healthcare within the e-DiaMoND project. In that project, a prototype for a Grid-enabled national (UK) database of mammograms [8] was attempted, and an abstract view of the e-DiaMoND system was proposed, considering communication between hospitals, and taking into account sensitive information security requirements.

Extending that abstract view of the e-DiaMoND system, every separate unit (Hospital, or other) can be considered as a standalone repository of SCORM content objects. In consistency with the description in [8], the owner of the content repository decides which data are available outside the system and, in terms of this framework, available to the meta-search engine. The communication with the metasearch engine is made through the "External Communication Module", which is responsible to translate the messages from communication API types, to internal messages from the repository and vice versa.

The architecture of the standalone content repository could be varied in order for threads to be used instead of a database, as long as this is compatible with the communication API.

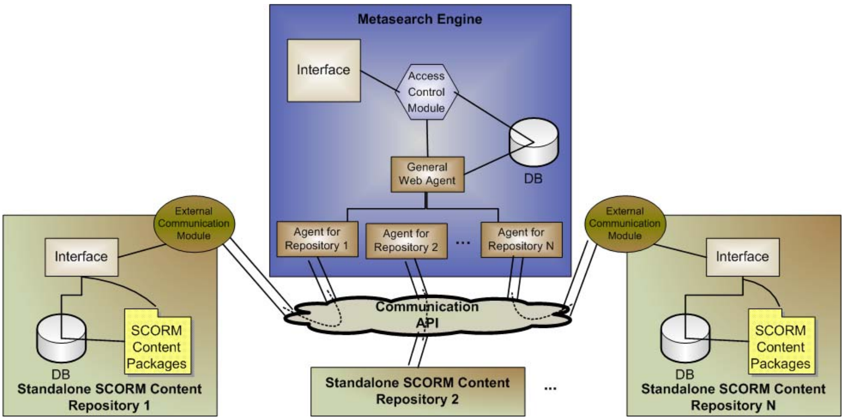


Figure 1 The overall architecture of the standalone SCORM Content Repository

3.2 Communication API

The communication API is responsible for asking and retrieving information from the content between standalone repositories and the meta-search engine. Furthermore, appropriate functions of the API about privileges on a specific SCORM content package or learning object and policies about using and delivering them are dealth with the copyright arranger according details provided by the owner of the repository.

The communication API should incorporate some basic functions concerning the opening of the connection between the metasearch engine and the standalone repository, the bidirectional exchange of data, the question about the existence of a SCORM content object, the verification of its existence, the access privilege (copyright) on the SCORM content object and its provision to the metasearch engine, as well as, the closing of the communication.

It is noted that these functions are fundamental. Other functions may exploit the fundamental functions ensuring the communication safety between the standalone repository and the metasearch engine. A summary of the fundamental functions is provided in Table 1.

Function	USE
Init()	Initialization of communication
ObjectExist()	Object Request
ObjectVerify()	Object Verification
AskUserPrivilege()	Request of user access privileges
UserPrivilege()	Sending user access privileges
AskObjectPrivilege()	Request of object access privileges
ObjectPrivilege()	Sending object access privileges
GiveObject()	Object Provision
Terminate()	Termination of communication

Table 1 Basic functions of communication API.

3.3 Metasearch engine

The meta-search engine is comprised of three parts. The first part regulates the personalized search engine and the user access control model. The second one consists of the general web agent that is responsible for communicating with the web agents for each separate SCORM content repository and supervising the retrieval of appropriate data from each one through the communication API. The last part is the meta-search engine database where all the necessary data about users, agents and general meta search functions are stored.

The Role-Based Access Control (RBAC) model is utilised as the access control model of choice herein, as it suits well to collaboration environments [22]–[27].

A proper way to supervise the communication in a metasearch engine is by the use of web agents. Upon a need for data retrieval, the general web agent commands each separate web agent responsible for a standalone SCORM content repository to fetch the necessary data but always matching the end user priveleges. All the agents provide their information to the general web agent which is responsible for providing the data to the interface and interacting with the end user.

The last but very important part for the overall function of the system is the meta-search engine database. In there, permanent and temporary data, about all the above are stored, while some crucial parts of it are encrypted.

4. User definition and security issues

4.1 User definition

One basic step of system design is to define possible users in order to ascertain their needs. First of all the key role of administrator is separated in two: “admin” of the standalone SCORM Content Repository and “admin” of the metasearch engine. The “admin” of each separate standalone SCORM Content Repository is responsible for the proper function of it by using the communication API functions as end functions. The “admin” of the metasearch engine is responsible for administering user accounts and distributing the web agents.

“Tutor” is the second user role. “Tutor” has two main actions: (i) searching for a SCORM Content Object and (ii) uploading material into the system. The material could be uploaded in two ways. The most common way is through the interface of the metasearch engine. The second way prerequisites that a standalone SCORM content repository would have an independent (of the system) interface. The “Tutor” will login into the standalone SCORM content repository and upload educational material. In both cases, copyright policy and access is determined by the “Tutor” who uploads the material.

4.2 Security issues

The first step in providing access control to the system is authentication. It should involve validation of the identity of a user via username and password login. Session variables will ensure that no user would jump to another page viewing sensitive data or accomplishing a non proper search. The second step is to determine the user permissions. Thus, authorization of the user determines the user’s role which is used to provide access to certain modules and processes of the system. As mentioned above, for fulfilling security requirements the RBAC may be used as the access control strategy. In addition, confidentiality of data is achieved through data encryption by means of SSL/TLS protocols.

5. Scenario of use

Figure 2 illustrates the scenario of use. We assume that a teacher wishes to search for a SCORM content course, about Electronic Health Records. The teacher provides the metasearch engine with keywords about the content of the course: “Hospital Information Systems”, “Integrated Health Care”, and about the copyright policy “Use within a University Postgraduate course”. The teacher can also provide some additional information in order to launch an advanced search. The topics of the advanced search are defined indirectly by Healthcare LOM tags. The latter may be: the desirable duration of the course, types of the course material (e.g. slides, narrative text etc), as well as, educational objectives set for the course (e.g. "Make students aware of its importance and possible use within the health care system").

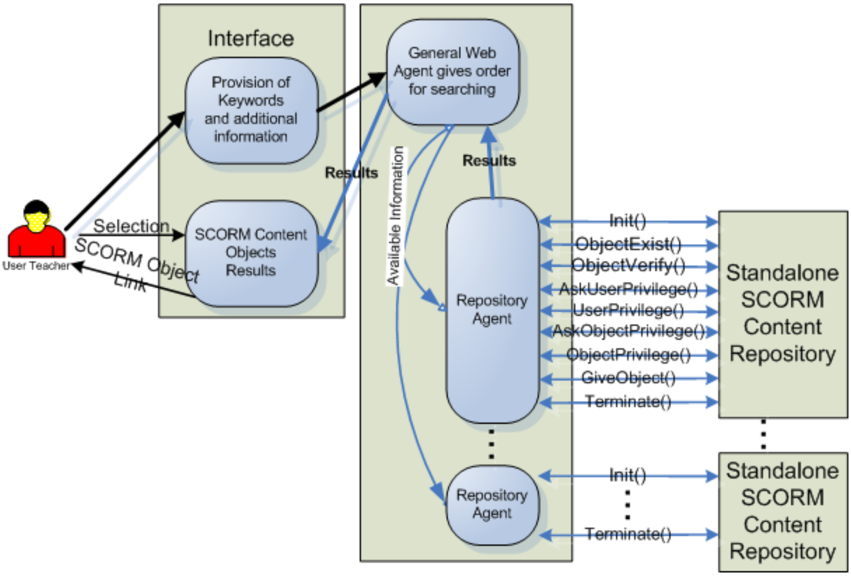


Figure 2 Scenario of retrieving SCORM Content Object

The general web agent takes this information and forwards it to each separate agent one of each standalone Content Repository. Each agent, by use of the communication API (Init(), ObjectExist()), transfers the research to the standalone Content Repository. Each repository matches the keywords with the Healthcare LOM information stored for each SCORM content object. The search and retrieval ways may be varied and adjusted to the needs and the abilities of each separate repository. The standalone content repository provides, through the use of the communication API ObjectVerify(), the Healthcare LOM XML file to the agent, which in turn forwards it to the general web agent. The general web Agent provides the end user with all the Healthcare LOM XML files (Figure 3) through appropriate interface transforms. At the same time, communication API functions about privileges are used for an interaction between the metasearch engine and a standalone SCORM content repository. Once the communication is established additional information about extra privileges is released between them. The Healthcare LOM XML file provides information about the copyright policy of the learning objects. This XML file also informs the user about the privileges that he/she has on each separate SCORM content object. After the transitions of the data the communication is terminated by the use of Terminate() function.

Figure 3 A sample extract from the Healthcare LOM XML file

The user “Tutor” selects the SCORM content object. The metasearch engine identifies the standalone SCORM content repository that contains the specific object through the metasearch engine database. The general Web Agent queries the agent of this specific standalone repository to provide a downloadable link of the specific object. The agent establishes the communication (Init()) within the metasearch engine and this specific standalone SCORM content repository. Finally the link provided to user “Tutor” (GiveObject()) and the session is terminated (Terminate()).

5. Discussion and conclusion

5.1 Research limitation/implication

The limitation of the framework could be summarized in universal open issues, such as the quality of the content or the metadata that can not be easily assured and evaluated without human interference.

A common architecture, in terms of using web agents for retrieving data from remote databases was followed by this framework. This approach is generally characterized robust and reliable in literature.

5.2 Practical implication

The practical implications can be assumed within the first step, that is, the development of a portal providing SCORM content packages in various fields of healthcare. In the latter, the use of standards and information may change frequently while the need for updated knowledge by doctors is increasing.

5.3 Originality/value

Despite the fact that some standalone repositories where proposed [5]-[7] using the IEEE LOM none of them focuses in the specific field of Healthcare and in the use of Healthcare LOM and SCORM that fulfil the special needs of this scientific area.

The CORDRA architecture which is a similar to the proposed framework leaves some open issues. The first of these regards the determination of context-specific metadata, indexing and uses in search, as well as, the level of presentation in users. The second open issue is concerned with the provenance of content packages, such as the determination of the ids that should be used and the policy about rights, as well as, the versioning of content packages and their metadata. Last but not least, the way users search the main CORDRA repository is enigmatic if not problematic. On the contrary, our framework covers the majority of the above issues. It also administrates the policy and the copyright issues in a rather dynamic form depending on copyright holders. The Helathcare LOM XML file accounts copyright management. In comparison with other repositories the metasearch engine is provided through the Internet. It is accessible in a common way to the end user, sustaining security.

Last but not least content object and metadata versioning is a resolved issue within our framework. Each standalone repository has the commitment and, at the same time, the flexibility to version its own content object and metadata.

References

1 S. T. Konstantinidis, M.M. Nikolaidou, P.D. Bamidis, C. Pappas, “Distance continuing medical education in cross border areas of Greece and Bulgaria – The CrossBorderHealth project example”, in Proceedings of the 4th International Conference in Open and Distance Learning ICODL 2007, Part B: pp.73-79

2 S. S. Marolov, S. T. Konstantinidis, P. D. Bamidis, M.M. Nikolaidou, C. Pappas, “Cross-border collaboration between Greece and Bulgaria: The use of ICT in improving social medicine education”, in Proceedings of the 3rd Balkan Conference in Informatics BCI 2007, 2: 281-290.

3 D. Hoa, A. Micheau, and G. Gahide, "Creating an Interactive Web-based e-Learning Course: A Practical Introduction for Radiologists," Radiographics, vol. 26, pp. e25-, 2006.

4 CORDRA - Content Object Repository Discovery and Registration/Resolution Architecture, Advanced Distributed Learning Initiative, (ADL) the Corporation for National Research Initiatives, (CNRI) and the Learning Systems Architecture Lab (LSAL). http://cordra.net/, visited at 20/4/2008

5 T. Kleinberger, L. Schrepfer, A. Holzinger, and P. Müller, "A Multimedia Repository for Online Educational Content," presented at ED-MEDIA 2001, Tampere, Finland, 2001.

6 D. R. Rehak, P. Dodds, and L. Lannom, "A Model and Infrastructure for Federated Learning Content Repositories," presented at The International World Wide Web

Conference - WWW' 05 Workshop on Interoperability of Web-Based Educational Systems, Chiba, Japan, 2005.
7 T. K. Shih, C.-C. Chang, and H. W. Lin, "Reusability on Learning Object Repository," in Advances in Web Based Learning – ICWL 2006, 2006, pp. 203-214.
8 D. J. Power, E. A. Politou, M. A. Slaymaker, and A. C. Simpson, "Towards secure Grid-enabled healthcare: Research Articles," Softw. Pract. Exper., vol. 35, pp. 857-871, 2005.
9 "SCORM® 2004 3rd Edition Overview Version 1.0 , © 2006," Advanced Distributed Learning(ADL) 2006.
10 V. Smothers, "Healthcare Learning Object Metadata Specifications and Description Document, v0.9," MedBiquitous Consortium, http://www.medbiq.org/working_groups/learning_objects/HealthcareLOMSpecification.pdf, visited 1/5/2007
11 "Draft Standard for Learning Object Metadata" Learning Technology Standards Committee of the IEEE, 2002. http://ltsc.ieee.org/wg12/files/LOM_1484_12_1_v1_Final_Draft.pdf, visited 11/10/2007
12 I. Hege, K. Radon, M. Dugas, E. Scharrer, and D. Nowak, "Web-based training in occupational medicine," International Archives of Occupational and Environmental Health, vol. 76, pp. 50-54, 2003.
13 A. G. Picciano, "Beyond Student Perceptions: Issues Of Interaction, Presence, And Performance In An Online Course," JALN, vol. 6, pp. 21-40, 2002.
14 GEODE - Global Education Online Depository and Exchange, University of Wisconsin - Milwaukee's Institute of World Affairs., http://www.uw-igs.org/, visited at 20/9/2006
15 MLX - Maricopa Learning Exchange, Maricopa Community Colleges in Arizona, http://www.mcli.dist.maricopa.edu/mlx, visited at 20/4/2008
16 CAREO - Campus Alberta Repository of Educational Objects, Universities of Alberta and Calgary, http://www.ucalgary.ca/commons/careo/, visited at 20/9/2007
17 PBS Teachers, Public Broadcasting Service (PBS), http://www.pbs.org/teachers/, visited at 15/3/2008
18 MERLOT - Multimedia educational Resource for Learning and online teaching, http://www.merlot.org/merlot/index.htm, visited at 20/3/2008
19 Learn-Alberta, Government of Alberta, http://www.learnalberta.ca/, visited at 22/3/2008
20 LearningLanguages.net, Internet Scout Project and the University of Wisconsin College of Letters and Sciences, http://www.learninglanguages.net, visited at 25/3/2008
21 HEAL - Health Education Assets Library, National Sciences Digital Library, http://www.healcentral.org/index.jsp, visited at 23/3/2008
22 William Tolone, Gail-Joon Ahn, Tanusree Pai, Seng-Phil Hong, Access Control in Collaborative Systems, ACM Computing Surveys (CSUR), March 2005, Volume 37 Issue 1. Publisher: ACM Press.
23 SANDHU, R. AND SAMARATI, P. 1994. Access control: Principles and practice. IEEE Communications 32, 9, 40–48.
24 Thomas, R. K., and Sandhu, R. S., Task-based Authorization Controls (TBAC): A Family of Models forActive and Enterprise-oriented Authorization Management,

Proceedings of the IFIP WG11.3 Workshop on Database Security, Lake Tahoe, California, August 11-13, 1997

25 Ferraiolo, D. F., Barkley, J. F., and Kuhn, D. R. 1999. A role-based access control model and reference implementation within a corporate intranet. ACM Trans. Inf. Syst. Secur. 2, 1 (Feb. 1999), 34-64. DOI= http://doi.acm.org/10.1145/300830.300834

26 Mary Thompson, William Johnston, Srilekha Mudumbai, Gary Hoo, Keith Jackson, Abdelilah Essiari, Certificate-based Access Control for Widely Distributed Resources, Pp. 215–228, Proceedings of the 8th USENIX Security Symposium, August 23-26, 1999, Washington, D.C., USA.

27 EDWARDS, W. K. 1996. Policies and roles in collaborative applications. In ACM Conference on Computer-Supported Cooperative Work. Cambridge, MA.

Universities go Mobile – Case Study Experiment in Using Mobile Devices

Majlinda Fetaji[1], Bekim Fetaji[2]

South East European University, Tetovo, FYR Macedonia, {m.fetaji[1], b.fetaji[2]}@seeu.edu.mk

The objective of this research was to investigate the possibilities of using mobile applications and mobile devices in University environment. The research integrates several existing technology adoption models and then for the needs of the research conducted and develop a case study experiment. The objective was to raise the communication level and accessibility, as well as the dissemination of knowledge and learning. The significance of the research is based in the fact that today, almost every student has a mobile device at all times while not everyone has a computer and internet connection at all times. In order to realize this, a case study experiment was realized. It involved investigation into the factors that influence mobile applications.

Keywords
Mobile application, wireless devices, human-computer interface, usability

1. Introduction

What are the possibilities of improving and increasing accessibility, communication and learning in University environment through mobile usage and wireless applications? Will this have a substantial impact on these factors above mentioned? Will the students find this useful? The research has attempted to answer to these questions primarily above all.

While the opportunities that mobile and wirless devices present us are new, however the challenges are quite old, smaller screen sizes, limited processing power, reduced input capabilities. These challenges mean that adapting existing e-learning services and content to m-learning is not a trivial task.

Because wireless devices are compact and relatively easy to handle, the Information technology focus is quickly shifting from PCs to PDAs, cellular phones and pagers.

The transition to wireless, mobile Internet devices will fundamentally alter the landscape and architecture of communication networks. Especially since wireless devices are becoming more and more customizable.

2. Research Design and Methodology

The research design consisted of three phases. The first phase is explorative and consisted of a literature study of the context of the mobile phone user as well as the factors that influence wireless devices especially mobile phones, mobile computing, and usability issues in mobile wireless devices. The findings from the literature study are our response to the first research sub-question.

In the second phase these findings have been integrated to propose a model of factors that influence mobile wireless devices. This model is a response to the second research sub-question. Then based on our model we have developed and conducted an experiment to see what software development Life Cycle model is most appropriate in the context of mobile environment. Afterwards, the focus have been set on several testing procedures in order to address the identified usability issues and investigate what does and does not work in particular situation and develop guidelines and recommendations for those cases.

- The research is conducted firstly based on fundamental research and grounded theory research and then afterwards action research. For this reason a case study experiment was conducted. Then exploratory research and constructive research to build the solution as well as quantitative research to study the relationships.

- The data collection is realized through surveys focus groups and feasibility study of the realized needs analyses, and then through using the developed software life cycle model.

3. Background Research

Wireless devices and especially mobile phones have been researched from a variety of perspectives, for example, information systems [6] and from Human-Computer Interaction (HCI) [2,3,4]. What is, however, lacking is a model that integrates the factors that influence wireless devices especially mobile phones. Technology adoption, in general, has been widely studied and several models of technology adoption have been proposed and tested in [9].

Therefore, the research is focusing in investigating mobile software development models as a strategy to match mobile phone design to user's technological needs and expectations focusing in usability and HCI.

ased on the literature study the research integrates several existing technology adoption models and then for the needs of the research conducted and develop a case study experiment as mobile application.

The research seeks to present usability guidelines by grounding the user interface on usability theoretical framework, possible constraints, and unique properties of mobile computing.

Three categories of usability have been set as the focus:

1) user analysis,

2) interaction and

3) interface design.

Usability guidelines were suggested in aiming for designing highly efficacious, user friendly and usable mobile interface to support dynamicity of mobile and handheld devices.

According to [8] usability means the measure of the quality the users' experience when interacting interface. Moreover, usability is not a surface gloss which applied at the last minutes or before the releases of the system or product; but it is deeply affected by every stage of the analysis, design, and development [8].

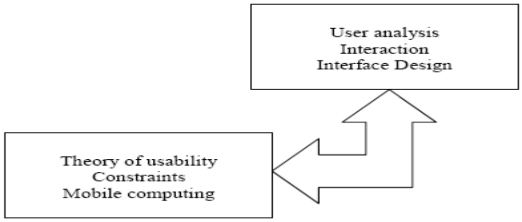


Figure 1 Mapping of usability categories to mobile software development

Usable systems are easy to learn (learnability), efficient to use (efficiency), easy to remember (memorability), not error-prone (errors), and satisfactory in use (subjective satisfaction) [8]. The ultimate goal of usability is meeting the needs of to users' satisfaction. Moreover, advantages of usability encompass increased productivity, enhanced quality of work, improved user satisfaction, and reductions in support and training costs [3]. The reduction in costs has attracted many project managers and interface designers to employ the theory of usability when designing the interfaces as reported in [4].

4. Modeling the Factors that influence mobile applications

Improving the knowledge transfer level by improving software processes is a major concern. M-learning is e-learning that uses mobile devices to access learning content and wireless networks. m-learning uses the same technologies as e-learning, where e-learning uses a computer instead. As usual, in a university environment the computer networks are not wireless, that many times depend on the technical equipment. That might cause problems in connectivity. There are no problems of the type in wireless connections that mobile devices use. Although the development of m-learning environments does not differ from development of e-learning environments in the used techniques and technologies, it differs in the user interface design of the application. M-learning differentiates from e-learning, as being ubiquity, convenience, location awareness and personalization [7]. The main differences are categorized as:

- the mobile phone itself

- network accessibility
- the end-user
- context of usage
- usability

Concerns with the conception, development and verification of a software system have been based on the previous investigation points. Identifying, defining, realizing and verifying the required characteristics of the resultant software. These software characteristics include: functionality, reliability, maintainability, availability, testability, ease-of-use, portability, and other attributes. We will address these characteristics by preparing design and technical specifications that, if implemented properly, will result in software that can be verified to meet these requirements.

5. Research Instrument Development

Major challenge for mobile researchers is to assess the m-learning effectiveness. In order to do that used is the methodology that we previously defined in e-learning systems, called ELUAT (E-learning Usability Attributes Testing) [2] which combines an inspection technique with user-testing based on 4 usability attributes we have set.

The usability attributes we have set as the most important factors that influence mobile apps are:

1)Time to learn, 2) Performance speed; 3) Rate of errors; 4) Subjective satisfaction.

This methodology is necessary for presenting the m-learning in an efficient aspect. The theoretical bases are the pedagogical conceptions defined from [3] as following: 1) Learning according to the constructivist perspective, 2) usability of the m-learning environment and 3) research about user opinions.

The measuring instrument is based on the use of predefined evaluation tasks (PET) defined from our previous study [2], which precisely describe the activities to be performed during inspection in the form of predefined tasks, measuring previously assessed usability attributes. Using this technique evaluated usability attributes using evaluation tasks for a particular scenario. Evaluation tasks in this technique are determined throw designing several user scenarios and choosing the scenarios that include the most of the provided options of the software solution.

6. The Experiment

Improving user satisfaction level by improving software processes is a major concern. In the development process guidelines from (Pressman, 2005) regarding the software life-cycle process have been closely followed. We have chosen the spiral software life-cycle as our model for the software development of the solution.

The actuality of mobile learning is based in the fact that almost every user has a mobile device at all times while not everyone has a computer and internet connection at all times. Providing information on real time when needed increases user accessibility and satisfaction of the offered services in real time. Thus based on the student feedback we believe that, developing a mobile accessible learning environment increases the accessibility of the electronic learning content (e-content), user convenience and immediate feedback of the request not depending on the computer equipment, network connections and bandwidth. In order to achieve that, we developed the MobileView application intended for the students of Computer sciences to view the core and elective subjects they can choose, view short lesson description, view the announcements, exam dates and exam results. The development of the MobileView application is based on students needs, preferences and context.

The software solution was developed in Visual Studio.NET 2003 using ASP (Active Server Pages).

The users are using their mobile devices specifically their mobile browsers using GPRS in order to access the content. The students log on to the web server using their ID number for identification, after which they have three options: To view their exam results; 2) to see the new announcements and news; 3) to see the teaching content for a particular subject.

These options have been derived based on previous student needs analyses that were based on a web survey and a focus group from the Communication Sciences and Computer Technologies Faculty at South East European University. The students have defined these functionalities and options as important for them. The mobile view was compared at all times with the computer (browser) view in order to asses its usability and accessibility.

After loging in the students have been given three options given below.

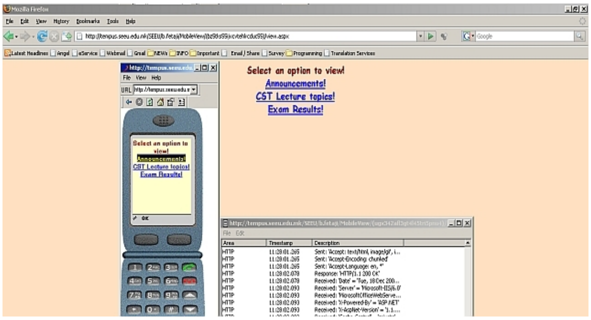


Figure 2 Mobile application : Options

The announcment section was provided as combo box and while being fine in computer (browser view) it was quite not practical and assesed as very bad choice from the testers. This option has to be addressed in other way and the solution that we have been agreed and prefered from student side is the List form.

Also the student were given the opportunity to select and check their exam results without the need to come to the University using their mobiles. The past exams optionality is limited in exams that occured in the time period of past 2 years.

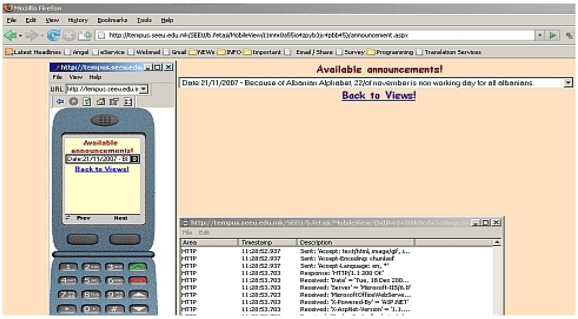


Figure 3 Mobile application interface: Announcements option

The option to view the exams was defined by the subject name and date of the exam as well as the type. Three types of exams have been defined : midterm, final, make-up exam. Below is given a visual representation of the views in mobile view and in browser for comparison reasons.

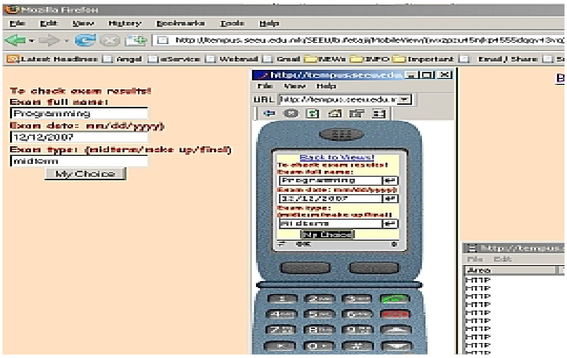


Figure 4 Mobile application interface: View Exam Results by subject name and date of exam

Also the students requested the option to access a short description of their subject topics organised in one semester consisted of 15 weeks.

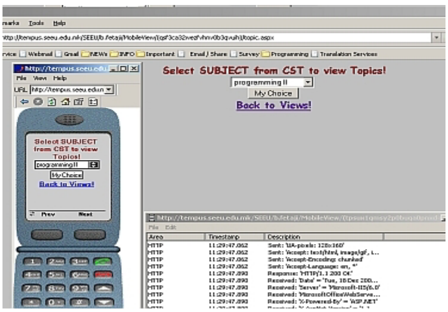


Figure 5 Mobile application interface: View Lecture topic

The week topics consisted of short discription of that week lecture in the form of an abstract with goals and objectives and clearly defined outcomes. From the student side it has been evaluated as very informative and just enough information needed since they do not expect to learn the entire lecture using this option. They prefer this option only to be aware of the content of that particular lesson topic. The students still prefer the printed and hard copies of lectures in the learning process and according to them this will hardly change in favor of electronic views.

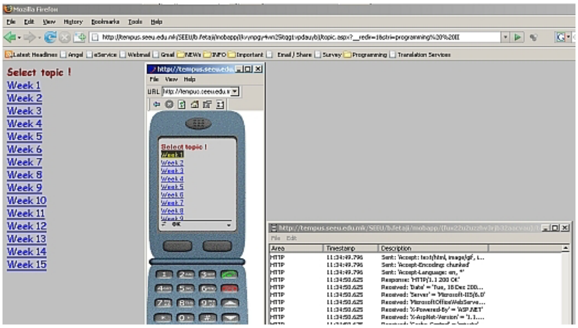


Figure 6 Mobile application interface: Lecture topics organised in semester: 15 weeks per semester

7. Results

The usability testing process was divided into three phases and we have followed the guidelines as defined from [5, 10]: planning, acquisition and execution with evaluation. After the usability test collected the data from the 10 participants were 5 of them were experts while the other 5 novices. The mobile devices used the tester have varied (Nokia, Sony Ericson, HP ipaq). However, they where all the new generation with year of production 2005 to 2007 with GPRS browsing capabilities. In order to handle those data the triangulation technique from [1] was used. There we look at all data at the same time to see how the different data supports each other.

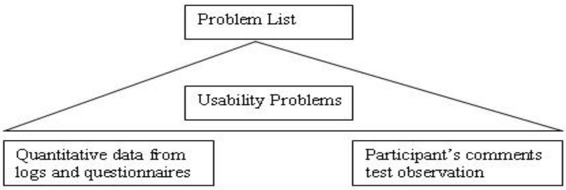


Figure 7 Triangulation technique, [1]

Also tabulated were the data for the performance measurements using the next usability attributes: time to learn, speed of performance, rate of errors, Subjective satisfaction, and Frustration for the both classes of users Experts and novices. Here is the tabulated data sheet for time to learn, and speed of performance as well as the general usability requirements measures.

Usability Attribute	Value to be measured	Current Level average	Worst acceptable	Planned target level	Best possible
Time to learn to use	Time to learn to use the app (seconds)	8	20	10	5
Speed of performance	Time to complete all tasks (seconds)	17	40	20	10
Rate of errors	Number of errors	2.37	4	1	0
Subjective satisfaction	Satisfaction degree of users	4.03	3	4	5
* legend: very high 5 high 4 average 3 low 2 very low 1					

Table 1 Results - Usability Measurements for Class Expert

Usability Attribute	Value to be measured	Current Level average	Worst acceptable	Planned target level	Best possible
Time to learn to use	Time to learn to use the app (seconds)	14	30	15	5
Speed of performance	Time to complete all tasks (seconds)	37	60	30	10
Rate of errors	Number of errors	4.53	5	3	0
Subjective satisfaction	Satisfaction degree of users	3.06	1	3	5
* legend: very high 5 high 4 average 3 low 2 very low 1					

Table 2 Results - Usability Measurements for Class Novices

8. Conclusion

Analysis led to a series of recommendations for changes to methods and procedures currently employed in developing mobile software solutions that are to be used in learning and offer learning services.

Nowadays mobile phones are rapidly becoming increasingly powerful (both from hardware and software point of view) however their screens remain comparatively small. Often also the navigation is hard. Equipped with a small phone-style keyboard or a touch-screen the users lose more time in searching where on the page the information they need is than also in reading it and manipulating to view it.

The student feedback on the developed mobile application is that it offers quite important options and increases accessibility on real time and when needed however usability of the presented information and e-content remains still a challenge and an issue to be addressed.

8.1 Positive Outcomes

It was found that mobile access of the e-content and provided services at real time is far more effective vehicle in dissemination of knowledge and knowledge transfer primarily based on the user motivation and recent popularity the mobile and wireless devices are gaining in general.

Students generally agree that this approach with mobile devices is more intrinsically motivating, it encourages collaboration, is more challenging; focused on higher-order skills and reflective learning.

Generally the mobile software is very much appreciated and well welcomed specially from the more experienced expert group of users. Based on the problems identified by the students the appearance of the interface is changed and is made more usable according to the usability test and users comments and that is believed to increase the learning outcome specifically knowledge and understanding.

Those involved in executing and managing the process gained a substantial increase in understanding the overall learning process using mobile devices and attitudinal changes occurred and the role each group played in its successful completion lead to increased goal congruence.

8.2 Negative Outcomes

It was found that graphical representations in mobile applications are not preferable at all and should be avoided as much as possible. Use of computers and mobile devices and applications to teach theoretical Computer Science is dangerous as students may think they don’t have to learn how to solve problems, but just how to use computer packages to solve them. What effected substantially their motivation in using the system because of the amount of work additionally added to maintain courses? We are in a opinion that this could prove also very successful for different institutions/departments since it does not require any particular effort to set it up, almost no need for maintenance and training.

We have organized and analyzed the problems appeared from the testing in two dimensions:

1) Scope (how widespread is the problem) and 2) Severity (how critical is the problem)

Global problems by scope: The mobile screens had different resolutions and screen dimensions and generally not all provided the same view of the application especially those with smaller screens. For example some of the used mobile devices for testing purposes had the following characteristics and not all of them provided the same view.

Nokia 8800: 208 x 208 pixels, 31 x 31 mm

Sony Ericsson P990: 240 x 320 pixels, 41 x 56 mm

Sony Ericsson P900: 208 x 320 pixels, 40 x 62 mm

Nokia e61i: 320 x 240 pixels, 2.8 inches

Motorola w380: 128 x 160 pixels

Recommendation is to accept and develop for resolution which is in the middle 208x208 since the resolutions constantly are being upgraded. However there is still no standard resolution that all of the mobile manufacturers would support which is a consideration to be raised from the developer community.

Global problems by severity: Several objects like combo box option that works well in computers has not been proven as accessible and usable from the mobile user's perspective. Having several screens should be also avoided because of the time it takes for the content to load and the expenses that users have in such cases. The content should be organized in as few as possible screens and should be provided as hyperlinks in order to interconnect and be more usable and in this manner to increase the navigation.

As conclusion in general mobile applications and mobile devices offer very important increase in accessibility and communication level at real time when needed. However, usability still represents a major concern and primarily depends in the need to define standardization in screen resolution and screen dimensions among the mobile manufacturers that at the moment is left without addressing in order for the developer community to be able to offer more usable interfaces.

References

1 Dumas, J. S., & Redish J. C. (1999) "A practical guide to Usability Testing" revised edition, Pearson Education Limited, pp.55-62

2 Fetaji, M., Loskovska, S., Fetaji, B. & Ebibi, M. (2007). Assessing and developing a graphical user interfaces for Virtual Environments. In C. Montgomerie & J. Seale (Eds.), Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications 2007 (pp. 843-848). Chesapeake, VA: AACE.

3 Hyyppä, K., Tamminen, S., Hautala, I., Repokari, L. (2000): The effect of Mental Model Guiding User's Action in Mobile Phone Answering Situations. In Electronical proceedings from The 1st Nordic Conference on Computer Human Interaction. 2000. Stockholm, Sweden: NordiCHI 2000.

4 Iacucci, G., (2001) Bridging Observation and Design in Concept Development for Mobile Services. In Third International Symposium on Human Computer Interaction with Mobile Devices, IHM-HCI 2001. Lille, France

5 Jones, M.,Marsden, G. (2005): Mobile Interaction Design. John Wiley and Sons, Hoboken, NJ,

6 Kiljander, H. (2004): Evolution and Usability of Mobile Phone Interaction Styles. Thesis, Department of Computer Science and Engineering, Helsinki University of Technology

7 Kukulska-Hulme, A., & Traxler, J. (2005). Mobile teaching and learning. In A. Kukulska-Hume & J. Traxler (Eds.), Mobile Learning: A Handbook for Educators and Trainers. London: Routledge.

8 Nielsen, J. (2000). Designing Web Usability: The Practice of Simplicity. New Riders Publishing, Indianapolis, ISBN 1-56205-810-X

9 Palen, L., Salzman, M., Youngs, E. (2000) : Going wireless: behavior & practice of new mobile phone users. In Proceedings from the 2000 ACM Conference on Computer Supported Cooperative Work. 2000. Philadelphia, PA: ACM Press.

10 Pressman, R. (2005) “Software Engineering: A practitioners approach 6 Ed” McGraw-HILL, inc, pp.81-117

Affective computing: state-of-the-art, challenges and application areas

Andrej Luneski[1], Panagiotis D. Bamidis[2], Roger K. Moore[3]

[1]*SEERC, 17 Mitropoleos Str, Thessaloniki, Greece,* anluneski@seerc.org

[2]*Lab of Medical Informatics, Medical School, Aristotle University of Thessaloniki, Greece,* bamidis@med.auth.gr

[3]*The University of Sheffield, Dept. of Computer Science, Western Bank, Sheffield, S10 2TN, UK,* r.k.moore@dcs.shef.ac.uk

Affective computing is defined as "computing that relates to, arises from and deliberately influences emotions". It is an interdisciplinary field of research that incorporates emotion recognition and synthesis, affective HCI, emotion theories and many other reearch domains. Since its introducation, it increased the awareness within the scientific community of the significance of emotions as well as the benefits of injecting affect into computer applications and computer-mediated interaction. This paper presents the state-of-the-art in affective computing field, the most significant research directions application domains. Finaly, possibl e future direcations and challenges are discussed.

Keywords

Affective computing, emotions, emotional intelligence, learning, medical informatics, collaborative networks, appraisal theory.

1. Introduction

The interaction between humans and computers has been a subject of improvement since the introduction of the first computers. The primary goal leading the process of enhancement of human-computer interaction (HCI) has been to turn it into a more natural, user-acceptable and efficient form of communication. Perhaps one of the most appropriate approaches is getting it closer to the human-human interaction [1]. Within the context of 'copying' the interaction between humans, the computer has been equipped with human abilities, where intelligence is the most significant of them. The HCI has experienced the introduction of facial and speech recognition, natural language recognition, as well as, software intelligent agents than can learn and reason on their actions [2]. Nowadays we are witnesses of computers that can recognize speech, images (gestures and expressions), or convey such abilities through one or more communication channels. What has been considered a science fiction a few decades ago is now a reality.

In the quest for more efficient and human-like interaction, emotions have been constantly avoided in the past. They were considered as an obstacle for rational and intelligent expression [3]. However, during recent years many proofs have been given on the importance of emotions in intelligent human-like behaviour. The arguments in the favour of emotions gave birth to the area of "emotional intelligence", defined as "the capacity to understand emotional information and to reason with emotions" [3]. The introduction of emotional intelligence boosted the research into the theories and concepts of human emotions, affective interaction and methods for injecting affect into HCI or computer-mediated human interaction. The final result is a broader interdisciplinary field of affective computing, formally defined as "computing that relates to, arises from and deliberately influences emotions" [4].

This paper provides an overview of the affective computing area, describing the key aspects and views on human emotions. By presenting the state-of-the-art in affective computing, we discuss its current level of maturity. Moreover, the challenges of the research in the affective computing field are shown along with the application of affective technologies in various domains. Finally, an effort is made to point out the possible future directions and impact of affective computing …..

2. Affective Computing

Introduced only a decade ago, affective computing is relatively young scientific area. Nevertheless, it has already proven to be very successful in the fusion of the various domains linked with emotions. As an interdisciplinary area it is comprised of the following research fields:

- Theories for defining human emotions;
- Techniques and methods for understanding affective reactions;
- Affective HCI and computer-mediated human interaction (affect recognition and synthesis);
- Emotional awareness and intelligence by the computer;
- "Copying" the natural affective human-human communication;
- Affective computing applications;
- Technologies that facilitate all the above potentials;

Figure 1 illustrates the affective computing area and the comprising domains, as visioned by Picard and the Affective Computing Group at MIT.

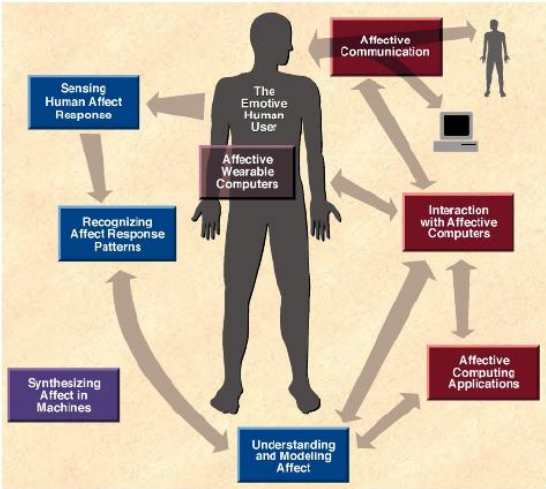


Figure 2 Affective computing sub-domains and research areas

2.1 Theories and views on emotion

The complex nature of emotions makes the definition and explanation of emotional human process very difficult. There exist many views on emotions and many valid theoretical explanations. The research directions within affective computing mainly depend on which of the existing theories will be accepted as the basis of affect definition. Of the many views we can focus on the most significant that the majority of the research community has accepted.

One of the earliest and perhaps best known to the scientific community is the discrete theory of emotion suggested by Ekman [5]. As the name itself suggests, the theory distinguishes several basic emotions (happiness, sadness, anger, fear, surprise and disgust), and each of them is characterized by a unique and representative physiological response pattern as a reaction to external stimuli. Additionally, each of the basic emotions is known to be connected to distinct facial expression.

Another view on emotions is the dimensional theory [6,7]. According to this theory, emotions are mapped into X-dimensional space, The two primary dimensions are affective valence (ranging from pleasant to unpleasant) and one of arousal (ranging from calm to excited). A third, less strongly-related dimension is variously called 'dominance' or 'control'.

The most interesting and most recent view on emotions is so called componential model based on the appraisal theory [8]. According to this theory, the emotions are perceived as an outcome of the organism’s appraisal on the external event or stimuli and the significance of that event for the organism (for its well being). The main motivation behind this characteristic view results from the main inconsistencies or shortcomings of the existing theories. First of all there is the temporal and individual difference in emotional response to the same event [9], where these theories fail to explain the different emotional reactions by different individuals on the same event or even different reactions by the same individual in the same situation in different time periods. Moreover, the exact trigger inside the organism of the emotional response is yet unexplained i.e. what is the core function or part in the organism that triggers the emotional response. Finally, yet unexplained is the appropriateness of emotional responses to the situations in which they occur [9]. There are even more problems that cannot be or haven't been explained by the theories before the proposition of the appraisal view on emotions. The strongest argument for the success of the appraisal theories is that with this approach emotion is placed in wider context of affective states such as mood and personality, and not only just as a reaction to external events. As such appraisal theories can be a very promising theatrical foundation for new innovative applications on emotional speech.

2.2 The significance of human emotions and the need for affective interaction

Emotions directly influence human physical and especially mental states and this in turn has an impact on the performance in the outer environment. According to the 'broaden-and-build model', positive emotions play a key role in increasing an individual's creativity [10]. Moreover, emotions can control the motivation for achieving goals [11] as well as improving a person's problem solving ability [12]. A person in a happy state is most likely to accomplish his/her pre-set goals, and will perform better than if he/she was in a negative emotional state such as anger or depression. Empathy is considered a strong mechanism for reducing frustration and generally improving the emotional state and mood in a natural human-like communication [13]. In an affective interaction, people are capable of expressing their empathy towards the computer (in HCI) or towards other people, thereby increasing the closeness and the levels of trust among them.

The naturalness of the affective communication in computer-mediated environments brings about increased interactiveness as it sets the base for the sense of "social presence". Social presence is defined as "the salience of the other in a mediated communication and the consequent salience of their interpersonal interactions" [14]. The expression of emotion, feelings, and mood in a mediated interaction has been considered as a defining characteristic of social presence [15].

The impact emotions have on human health has been reported for centuries, even since the ancient Greek philosophers and scientists, such as Socrates and Hippocrates, who considered emotion as a determinant of human health and diseases [16]. With the increase of lifestyle dynamicity, more focus has been given to the stress, depression, frustration and their negative influence on the human health. Namely, stress or depression weaken the human immune system [17]. It has been stated that depressive people catch a cold more frequently than people in normal emotional state [18]. Stress causes faster and harder heart pumping, which can lead to the amount of blood reaching heart through the coronary arteries not to be enough to support the faster work of the heart, or a condition known as myocardial ischemia [19].

The significance of emotions in the various aspects of social surroundings proves the need for affective computing technologies. With the advancement of information and telecommunication technologies (ICT) and their constantly increasing role in our live, affective technologies can serve as a mechanism for maintain the naturalness, effectiveness and pleasantness in the HCI or computer-mediated human communication.

2.3 Affective interaction

In a natural interaction affect is mainly expressed through speech, facial expressions and body gestures and these communication channels are the most obvious way of interpreting emotions. Additionally, human physiology is affected by emotional reactions and by employing specific sensors (for measuring heart bit rate, respiration, skin conductance etc.) it can serve as additional affective channel. Each of the channels has its own specifics and therefore acts as a separate sub-domain of the affective computing area.

2.3.1 Vocal Emotion Communication

Emotional information is carried in the human speech semantics as well as in the prosody. Semantics (what has been said) is a more obvious expression of emotion, while prosody holds more detailed emotional information. The term prosody combines nonsemantic cues in spoken language, such as: fundamental frequency (pitch), loudness, rhythm, formant structure of speech sounds, intonation etc. In the process of encoding and decoding in vocal communication the focus is on the six basic emotions (anger, joy sadness, fear, disgust and surprise) [20, 21, 22]. The average success rate in the recognition process is between 70% and 80%, which is higher that the human recognition rate of around 60%, [23, 24].

The quality of synthetic speech has rapidly increased and it could be argued that its "ineligibility is approaching that of human speech" [25]. The naturalness of the speech is perhaps the main requirement for human-like 'appearance' of the computer through human-like character. From the many methods of affective speech synthesis, concatenative unit selection method is considered as the most efficient one, producing most natural speech [25]. For intelligent affective speech predication dynamic alteration is necessary. Humans alter their speech "as a function of the communicative context" [26]. The speakers tune up their performance according to communicative and situational demands and therefore Moore suggests that for successful speech synthesis, the system has to be aware of the output it produces (from the feedback loop) and adjust it according to the needs of the situation.

2.3.2 Affective facial expressions and gestures

The "universal" human emotions such as happiness, anger, sadness, surprise, fear and disgust are characterised by distinct facial expressions, which help in identifying the expressed emotions [27]. Visual emotion expression by the computer is realised using human-like character (avatars). Avatars have been used as a visual representation for more natural human-computer interaction (see examples in figure 2). They are seen as "the physical representation of the self in virtual reality" [28]. Studies have shown that users interact with avatars as with other human beings [29]. Avatars have been mostly used in collaborative virtual environments [30], in HCI to express empathy and therefore improve the current emotional state of the user [31] etc.

(anger) (happiness) (sadness)

Figure 3 Affect expression of an avatar

2.3.3 Affective psychophysiology

Emotional reactions are accompanied with physiological disturbances in the human body - alteration of brain states, increase of the heart-beat rate, hands sweating, (face) muscle movements, increase of the respiration rate etc. Monitoring these types of emotional expressions requires employment of specific sensors and recordings of their output such as Electroencephalogram (EEG), Electrocardiogram (ECG), Skin Conductnce or Electrodermal Activity (EDA), Electromyogram (EMG) etc. The usage of physiological signals in identifying a user's emotional state has became popular in recent years due to the advanced development and availability of unobtrusive sensors that can provide constant and reliable monitoring of a user's internal emotional reaction [4]. Sensors have been integrated into clothing and jewellery; in shoes, in earrings, in a sports bra and numerous others [32]. A multi-channel framework for experimenting with physiological sensing of human emotion has been proposed in [33].

3. Affective computing applications

As we have seen, emotions have enormous impact in various domains of human everyday activity. In computer-mediated interactions and HCI computers employ sensors for full awareness and monitoring of user's affective state. The actual benefit of emotional-awareness can be seen by answering the question:

What can computers do and how can they be of help, once they have detected the correct emotional state of the user?

The application of affective computing technologies in various domains is a process that started only recently and yet, much success has been accomplished thus far.

Collaborative networks (CN) are one of the emerging areas that have attracted great attention by the research community. CN can be seen in various manifestations such as virtual communities, virtual enterprises, organisations etc [34]. Being a geographically dispersed network of people with similar goals, it can benefit from injecting affect in the computer-mediated interaction, bringing closeness, naturalness empathy, trust and pleasantness in the communication among its members [35].

Affective computing has been included in the research within the area of Ambient Intelligence (AmI), also referred to as Pervasive or Ubiquitous Computing. AmI combines unobtrusive and possibly invisible computing and advanced networking technology and specific interfaces that are aware of human presence, that can meet certain needs and can engage into intelligent interaction (spoken, gesture based etc.) [36]. AmI is considered as one of the most important emerging technologies that will be interrelated with affective computing in the future [4]. The significance of AmI, together with AC, has been reported in the final report by the IST Advisory Group (ISTAG) regarding "Scenarios for Ambient Intelligence in 2010" [37].

Tele-home health care (THHC) is one of the areas that have received great attention by the affective computing community. Internet-based communication technologies have enabled patient monitoring without the need of physical presence by the caregiver. Current systems facilitate collection of vital sign data remotely, verification of compliance with medicine regimes, assessment of mental or emotional status and more. Communication between the caretaker and care recipient through emotional channels in such environments has shown to be of vital importance to the patient [38]. Existing systems employ multi-modal interface including avatars (able to express emotions), to remind the patient for a medication, show empathy to the user when certain negative emotion is detected etc [39].

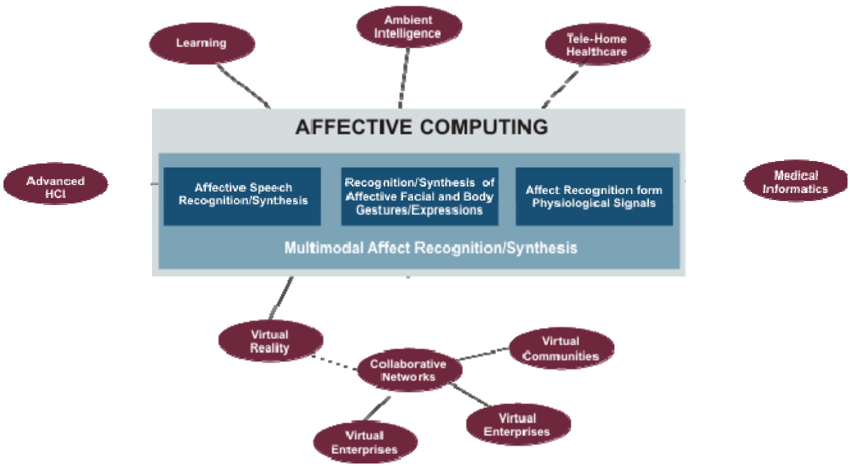


Figure 4 Affective computing application domains

Recently, there are research attempts for employing affective computing technologies in the learning and education fields. The idea arises from the influence of emotions and mood on the learning performance and decision-making process. Picard and the Affective Group are working on Affective Learning Companion[1] project that uses software-based interactive application that will recognize the affective and cognitive state of the learner and respond in an appropriate manner (e.g., can adjust the pace, difficulty, complexity). The existent affective technologies such as virtual agents that can visually express emotions, affective speech synthesis and recognition etc, enable the successful application of AC in e-learning.

Figure 3 depicts only few of the research domains that have began realising the potential benefits of affective computing.

4. Discussion

Affective computing is an interdisciplinary area and it has many research directions as well as various application areas. As we have seen, certain domains, such as medical informatics, collaborative networks, virtual reality, learning etc. have already stared realising the potential of affect-aware interactions and gaining its benefits. Affect-recognition and expression are areas where most of the early research has been performed in and, even though some of the basic emotions can distinctly be recognised/expressed, much progress still needs to follow. Moreover, emotion theories and views are under constant revision and new approaches are frequently emerging. The advancement tempo of affective computing is highly dependent on the ICT technologies as a facilitator for affective communication through wearable recognition sensors and advanced emotion expression methods.

Appraisal theory of emotions is one of the most promising research fields and since it is the most recent view of emotions, the amount of research is scarce. Appraisal theory is a solid base ground for research on the intelligent emotion production and expression through all the communication channels, with regards to the current situation. We are preparing an innovative experiment at the medical lad of the Aristotle University that will examine how the appraisal of the situation and the feedback from the other person in the communication influences emotional expression in speech. The results of the experiment can be beneficial for further research and applications in both computer-mediated human-human interaction and HCI, in environments such as collaborative networks.

Furthermore, we believe that multi-modal emotion interaction using all the communication channels can be the most appropriate way of affective interaction with the computer. With multi-modal interaction, the individual shortcomings of emotion recognition or expression in each of the channels can be overcome by combining them. Using avatars as a tool for visual emotion expression, we can have natural affective interaction in many domains such as virtual environments, computer-mediuated education, health applications etc. Apart from the avatar, affective synthetic speech is a

[1] http://affect.media.mit.edu/projectpages/lc/

challenging research field that requires further development and can provide great benefits in the process towards human-like affective interaction.

Apart from the various domains that have already been related to affective computing, there are many more that remain unexplored. With the continuous advancement of affective computing technologies new research directions are opened and new potential applications are unwrapped.

References

1 B. Reeves, C. Nass, The Media Equation: How People Treat Computers, Cambridge Univ. Press, 1996.

2 C. Ghauoi, Human computer interaction encyclopaedia, Idea Group Inc., London, UK, 2006.

3 D. Goleman, *Emotional Intelligen*ce, Bantham Books, New York, NY, 1995.

4 R.W. Picard, "Affective Computing", MIT Press, Cambridge, MA, 1997.

5 P. Ekman, "Basic emotions", In T. Dalgleish and M. Power (Eds.). *Handbook of Cognition and Emotion,* John Wiley & Sons, Ltd., Sussex, U.K., 1999.

6 W. Wundt, *Principles of Physiological Psychology,*1874, (English translation in 1904).

7 A. Mehrabian & J. A. Russell, *An approach to environmental psychology,* M.I.T. Press, Cambridge, MA, 1974.

6 B. L. Fredrickson, *The value of positive emotions*. American Scientist. 2003: 91. pp 330-335.

7 Cooper B, Brna P, Martins A. "Effective Affective in Intelligent Systems – Building on Evidence of Empathy in Teaching and Learning". in Ana Paiva (Ed.) Affective Interactions: Towards a New Generation of Computer Interfaces. Springer Verlag, 2000. pp 21-34.

8 Scherer et at., Appraisal Processes in Emotions: Theory, Methods and Research, Oxford University Press, 2001.

9 I. J. Roseman and C. A.. Smith, "Appraisal Theory: Overview, Assumptions, Varieties, Controversies" in *Appraisal Processes in Emotions: Theory, Methods and Research*, Oxford University Press, 2001, pp. 3-20.

10 B. L. Fredrickson, "The value of positive emotions", *American Scientist,* vol 91, pp 330-335, 2003.

11 B. Cooper, P. Brna and A. Martins, "Effective Affective in Intelligent Systems – Building on Evidence of Empathy in Teaching and Learning", in Ana Paiva (Ed.) *Affective Interactions: Towards a New Generation of Computer Interfaces,* Springer Verlag, 2000, pp 21-34.

12 A. Damasio, *Descarte's Error: Emotion, Reason and the Human Brain,* Avon, New York, 1994.

13 K. Hone, "Empathic Agents to Reduce User Frustration: The Effects of Varying Agent Characteristics", *Interacting with Computers*, vol. 18, issue 2, pp 227-245, 2006.

14 J. Short, E. Williams, B. Christie, *The social psychology of telecommunications*, Wiley, 1976. Toronto, ON.

15 R. Garrison, T. Anderson and W. Archer, “Critical inquiry in a text-based environment: Computer conferencing in higher education”, Unpublished manuscript, 2000.
16 J.T. Cacioppo, “Introduction: Emotion and Health”, in *Handbook of Affective Stress*, 1st ed., editors: Richard J. Davidson, Klaus R. Scherer, Hill H. Goldsmith, Oxford University Press, 2003.
17 R. Glaser, D. K. Pearl, J. K. Kiecolt-Glaser W. B. & Malarkey, “Plasma cortisol levels and reactivation of latent Epstein-Barr virus in response to examination stress”, *Psychoneuroendocrinology*, vol. 19, pp. 765-772, 1994.
18 S. Cohen , W. J. Doyle, D. P. Skoner, B. S. Rabin & J. M. Gwaltney, “Social Ties and Susceptibility to the Common Cold”, *JAMA* 227, vol. 24, pp. 1940-1944, 1997.
19 M. M. Burg, “Stress, behavior, and heart disease”, in: *Heart Book*, B. L. Zaret, M Moser, L. S. Cohen eds., Yale University School of Medicine, 1992.
20 R. Banse, K. R. Scherer,“Acoustic profiles in vocal emotion expression”, *J. Pers. Soc. Psychol*, vol 70, issue 3, pp 614-636, 1996.
21 R. Cowie, E. Douglas-Cowie, “Automatic statistical analysis of the signal and prosodic signs of emotion in speech”, *In: Proc. ICSLP_96*, Philadelphia, 1996, pp. 1989-1992.
22 J. A. Bachorowski, “Vocal expression and perception of emotion”, *Current direct. Psychol. Sci.*, vol 8, issue 2, pp 53-57, 1999.
23 F. Yu et al., “Emotion Detection From Speech To Enrich Multimedia Content”, *In the 2nd IEEE Pacific-Rim Conf. on Multimedia,* October 24-26, 2001, Beijing, China.
24 F. Dellaert F, T. S. Polzin and A. Waibel, “Recognizing emotion in speech”. *Int. Conf. on Spoken Language Processing (ICSLP)*, Pittsburgh, 1996.
25 M. Schroder M, “Emotional Speech Synthesis - A Review”, In *Proc. Eurospeech 2001,* Aalborg, pp 561-564, 2001.
26 R. K. Moore, “Spoken Language Processing”, *In Speech Communication, Special Issue on Bridging the Gap Between Human and Automatic Speech Processing,* 2007.
27 P. Ekman, “Universals and cultural differences in facial expressions of emotion”, *In J. Cole (Ed.), Nebraska Symposium on Motivation 1971,* vol 19, University of Nebraska, Lincoln, NE, pp 207-283, 1972.
28 E. Castronova E. "Theory of the Avatar". CESifo Working Paper Series No. 863, Fevruaty 2003.
29 J. Bailenson, A. Beall, J. Blascovich, M. Raimundo, M. Weishbush, “Intelligent agents who wear your face: User’s reactions to the virtual self”. Technical Report. Center for the Virtual Environment and Behaviors, Department of Psychology, University of California, Santa Barbara, 2000.
30 M. Fabri, D. J. Moore, “The use of emotionally expressive avatars in Collaborative Virtual Environments”, in Proceedings of the AISB Symposium on Virtual Social Agents, University of Hertfordshire, UK, pp. 88-95, April 2005
31 K. Hone, “Empathic Agents to Reduce User Frustration: The Effects of Varying Agent Characteristics”, Interacting with Computers 18, 2006, p. 227-245
32 R. W. Picard,, “Affective Medicine: Technology with Emotional Intelligence”. In Future of Health Technology, R. G. Bushko ed. OIS Press, 2001.
33 P. D. Bamidis, A. Luneski, A. Vivas, C. Papadelis, N. Maglaveras and C. Pappas, “Multi-channel physiological sensing of human emotion: insights into Emotion-

Aware Computing using Affective Protocols, Avatars and Emotion Specifications". In Proceedings of Medinfo 2007. Brisbane, Australia. pp 1068-72.
34 L. Camarinha-Matos, H. Afsarmanesh, "Collaborative Networks: A New Scientific Disipline", *Journal of Intelligent Manufacturing*, vol 16, issue 4-5, pp 439-452, 2005.
35 A. Luneski, P. D. Bamidis and R. K. Moore, "Affective Computing and Collaborative Networks: Towards Emotion Aware Interaction", in Proceedings of PRO-VE 2008, Poznan, Poland, to appear.
36 G. Riva, F. Davide, F. Vatalaro & M. Alcañiz, *Ambient Intelligence: The evolution of technology, communication and cognition towards the future of the human-computer interaction*, IOS Press, Amsterdam, 2004.
37 K. Ducatel, M. Bogdanowicz, F. Scapolo, J. Leijten & J-C. Burgelman, "Scenarios For Ambient Intelligence in 2010", *ISTAG Final Report*, February 2001.
38 C. Lisetti et al. "Developing multimodal intelligent affective interfaces for tele-home health care", *International Journal of Human-Computer Studies,* vol. 59, pp. 245-255, 2003.
39 H. Prendinger & M. Ishizuka, "What Affective Computing and Life-like Character Technology Can Do for Tele-Home Health Care", *Workshop on HCI and Homecare: Connecting Families and Clinicans*, *in conj. with CHI-04*, Vienna, Austria, 2004.

Reliable Web Service Publication and Discovery through Model-Based Testing and Verification

Ervin Ramollari[1], *Dimitris Dranidis*[2], *Anthony J. H. Simons*[3]

[1]*SEERC, 17 Mitropoleos Str, Thessaloniki, Greece, sname@seerc.org*

[2]*City College, Dept. of Computer Science, 13 Tsimiski Str, 54624 Thessaloniki, Greece, dranidis@city.academic.gr*

[3]*The University of Sheffield, Dept. of Computer Science, Western Bank, Sheffield, S10 2TN, UK,a.simons@sheffield.ac.uk*

Currently, the issues of trust and dependability on third-party Web services are becoming key factors to the adoption of service oriented computing in industrial environments. As a result, robust service testing and verification techniques are highly important in order for consumers to build confidence on third-party Web services. In this paper we propose modelling the behaviour of a Web service using stream X-machines, in order to derive a complete test set and to perform model-based testing. We apply these techniques in a novel publication and discovery approach involving all three main actors in a SOA environment, i.e. the service provider, the service broker, and the service consumer. The provider augments the service interface description (WSDL) with a stream X-machine (SXM) model reflecting the Web service behaviour. This model is both utilised by the broker during publication to derive a test set and verify Web service behavioural conformance, and by the consumer during discovery to perform service selection based on model validation.

Keywords
model-based testing, SOA, stream X-machines, validation, Web services.

1. Introduction

Service Oriented Computing is a new computing paradigm that utilizes services as the key abstraction to support the development of rapid, low-cost and easy composition of distributed applications even in heterogeneous environments [1]. Services are loosely coupled, reusable, and implementation-independent software modules with well-defined interfaces. They can be *described*, *published*, *discovered*, and dynamically *assembled* for developing massively distributed, interoperable, evolvable systems.

Services are made available by service providers within or outside the boundaries of an enterprise, and invoked by service consumers.

Currently, the prevailing alternative to implement a Service-Oriented Architecture (SOA) is the Web services framework, which is founded on widely accepted standards, such as WSDL for the service interface description, SOAP for the communication protocol, and UDDI for service discovery. Interaction between the three main parties that are involved, that is, service consumers, service providers, and service brokers, occurs as follows: service consumers discover Web services in a UDDI service registry maintained by service brokers. They retrieve WSDL descriptions of Web services offered by service providers, who previously published those WSDL descriptions in the UDDI registry. After the WSDL has been retrieved, the service consumer binds to the service providers by invoking the service through SOAP.

Given the increasing number of Web services that are being offered by third-party providers, the issues of *trust* and *dependability* on Web services are becoming increasingly important and considered as key factors to the adoption of service oriented computing in industry. As a result, *robust service testing, verification, and validation techniques are crucial* in order for consumers and integrators to build confidence on third-party Web services. In other words, service consumers need to ensure that advertised Web services are what they need, and that their implementations have been verified. However, one major obstacle in achieving these goals is that the WSDL standard lacks support for capturing the semantics relating to functional and non-functional aspects of a service. Therefore, it is not possible to guarantee that a discovered service advertisement matches a service request in all respects, and this may lead to inappropriate bindings.

One of the important aspects that the WSDL specification lacks is Web service *behaviour*, i.e. the definition of preconditions and effects of each Web service operation as well as the implied sequencing of these operations. Behavioural specifications are especially useful in cases of Web services assuming an interaction protocol (stateful Web services) and Web services operating on persistent data. In this paper we propose an approach which tries to fill this gap by introducing formal specifications of Web service behaviour to its interface description. For this purpose we choose stream X-machines (SXMs) [2], which we consider an intuitive and powerful formalism. Not only do SXMs allow for unambiguous specification of Web service behaviour, but they are also highly useful to perform model-based testing of a Web service implementation under test (IUT). Research on stream X-machine offers a test generation method, which under certain assumptions is proven to find all faults in the implementation [3, 4]. Our approach makes use of these benefits requiring the cooperation of all three main stakeholders in a SOA environment (SOA triangle), i.e. the service provider, the service broker, and the service consumer. The provider's role is to create a SXM model reflecting the behaviour of the provided Web service implementation, and add it to WSDL during the publication process. Based on this model, the broker is able to derive the necessary test cases, which are run in order to verify behavioural equivalence between the advertised model and the implementation. Only services with successful test results are accepted (i.e. checked-in) in the registry. On the other hand, during the discovery process, the consumer is provided with a number of service candidates fulfilling the request. Through the provided SXM models, the consumer can validate the behaviour of candidate services against consumer needs, a process that aids in the selection of the most appropriate service. Therefore, our approach ensures that clients bind with Web services providing a suitable behaviour, and a verified implementation.

The rest of this paper is structured as follows. Section 2 presents a summary of related work in model-based testing of Web services and in related publication/discovery approaches. Section 3 provides a description of the used stream X-machine formalism, the Web service modelling process, and the associated complete functional testing method, illustrated with a simplified shopping cart case study. Section 4 provides an overview of the approach for reliable Web service publication and discovery based on stream X-machines, described from the perspectives of the service provider, the service broker, and the service consumer. In the end, Section 5 concludes the paper by summarizing the main points of the presented work, and suggesting directions for future work necessary to realise the described approach.

2. Related Work

A number of approaches have been proposed for applying model-based testing to verify Web services. The authors in [5] propose an algorithm, which translates Web service descriptions annotated in WSDL-S into an equivalent Extended Finite State Machine representation, which extends simple Finite State Machines with the addition of a multi-dimensional structure (memory) and the modification of the state transition function so that it maps a (input, initial state, initial memory) tuple to a (output, new state, new memory) tuple. The WSDL-S document is enhanced with references to OWL-S concepts, as well as to SWRL rules, which explicate the behaviour of individual Web service operations in the form of inputs, outputs, preconditions, and effects (IOPE). The resulting EFSM model, consisting of a *single* state, is then exploited, using *any* appropriate test case generation algorithm, to derive an effective test set for verifying a Web service behavioural conformance.

Keum et al [6] present another model-based testing approach using Extended Finite State Machines, which extend finite state machines with memory, and with computing blocks and predicate conditions for state transitions. A procedure is described for semi-automatically deriving the EFSM model from a WSDL specification and additional user input. The model covers behavioural aspects of *stateful* Web services, and the resulting test cases represent sequences of invocations of Web service operations. The authors provide experimental results showing that their method has the potential to find more faults compared to other methods, but notably, with a resulting test case set that is much larger and takes more time to execute.

Some other works have proposed the application of model-based verification of Web services in the context of more complete approaches. Bertolino et al [7] describe an envisaged registry-based Web service verification framework. The provider augments the WSDL document with behavioural descriptions in a UML 2.0 Protocol State Machine (PSM) diagram, which is then translated to a Symbolic Transition System (STS). On the other hand, the broker utilises the attached STS model to automatically generate the test cases and run them on the provided Web service for behavioural conformance verification. Upon successful test results the Web service is published in the UDDI registry as a certified service. For this reason, the authors call their approach an "Audition" framework, where the Web service undergoes a monitored trial before being put "to stage".

Heckel and Mariani [8] use graph transformation rules to model the behaviour of Web service operations and apply them in a reliable Web service publication and discovery approach. Both the behaviour of the advertised service by the service provider *and* the requested service by the service consumer need to be modelled in terms of graph transformation rules. A test case derivation method is employed to verify that the actual service implementation conforms to the provided model. This verification is performed by the service broker before services are accepted in the registry, resulting in what the authors refer to as high-quality service discovery agencies. In addition, during discovery, the service broker enables matchmaking of request and advertisement models that are expressed as graph transformation rules, in order to return service candidates satisfying the consumer's behavioural constraints.

In the approach we propose we find some advantages relative to the aforementioned approaches. The employed X-machine complete functional testing method can be proven to reveal all implementation errors, under certain design-for test conditions that the model and the implementation have to satisfy [3, 4]. Additionally, in contrast to the approach proposed by Heckel and Mariani, this approach does not require models of both the service request and service advertisement during service discovery, since service selection is performed through behavioural validation at the consumer site. Indeed, it is impractical to assume that the consumer knows in advance the detailed behaviour of the requested Web service and can create a formal model of that behaviour.

3. Modelling Web Service Behaviour with Stream X-Machines

3.1 Stream X-machines

Stream X-machines (SXMs) [2] is a computational model capable of modelling both the data and the control of a system. SXMs are special instances of X-machines introduced by Eilenberg [9]. They employ a diagrammatic approach of modelling the control by extending the expressive power of finite state machines. In contrast to finite state machines, SXMs are capable of modelling non-trivial data structures by employing a memory, which is attached to the state machine. Additionally, transitions between states are not labelled with simple input symbols but with processing functions. Processing functions represent internal system transitions triggered by input symbols under specific memory conditions, and produce output symbols while modifying the memory. The benefit of the addition of a memory structure is that state explosion is avoided and the number of states is reduced to those states which are considered critical for the correct modelling of the system's abstract control structure. A divide-and-conquer approach to design allows the model to hide some of the complexity in the transition functions, which are later exposed as simpler SXMs at the next level.

A stream X-machine is defined as an 8-tuple, $(\Sigma, \Gamma, Q, M, \Phi, F, q_0, m_0)$ where:

- Σ and Γ is the input and output finite alphabet respectively;
- Q is the finite set of states;
- M is the (possibly) infinite set called memory;
- Φ, which is called the type of the machine SXM, is a finite set of partial functions (processing functions) φ that map an input and a memory state to an output and a new memory state, $\varphi: \Sigma \times M \rightarrow \Gamma \times M$;
- F is the next state partial function that given a state and a function from the type Φ, provides the next state, $F: Q \times \Phi \rightarrow Q$ (F is often described as a transition state diagram);
- q_0 and m_0 are the initial state and memory respectively.

The sequence of transitions (path) triggered by the stream of input symbols is called a *computation*. The computation halts when all input symbols are consumed. The result of a computation is the sequence of outputs symbols produced by this path.

Apart from being formal as well as proven to possess the computational power of Turing machines [3], SXMs have the significant advantage of offering a testing method [3, 4] that ensures conformance of an implementation to a specification. This method generates test sets for a system specified as a SXM whose application ensures that the system behaviour is identical to that of the specification provided that the system is made of fault-free components and some explicit design-for-test requirements are met.

In order to allow for specifications of stream X-machines, the XMDL (X-Machine Definition Language) language was introduced in [10] and fully developed in Kefalas [11]. XMDL serves as an interlingua for the development of tools supporting Stream X-machines [12]. An extension of XMDL to support an object-based notation was suggested in [13]. The object-based extension, called XMDL-O, enables an easier and more readable specification of Stream X-machines and is employed in this paper for the specification of the example Web service.

3.2 Shopping cart example

We illustrate our modelling and test set generation method with a simplified version of a Web service that is intended to provide the backend functionality of a shopping cart to consumers, also described in a previous paper [14]. Similar Web services are already being used and made available over the Web, such as the Amazon Shopping Cart Web service (http://developer.amazonwebservices.com). The `ShoppingCart` Web service provides the following operations:

- The `login` operation allows authentication for using the service. It is invoked with the input message `LoginRequest` consisting of the username and the password of the user. The request message is represented as `LoginRequest(user, pwd)`. The operation sends back the response message `LoginResponse(result)`, where result is a boolean value; true indicates successful authentication.

- The `addToCart` operation adds an item to the shopping cart. It is invoked with the input message `AddToCartRequest` consisting of the identifier of the item to be added. The request message is represented as `AddToCartRequest(itemId)`. The operation sends back the response message `AddToCartResponse(itemId)`. It is assumed that all item identifiers are valid and correspond to products that may be purchased.

- The `clearCart` operation removes all items from the shopping cart. It is invoked with the simple request message `ClearCartRequest` represented as `ClearCartRequest()` and it sends back the response message `ClearCartResponse()`.

- The operation `checkout` completes the shopping process. It is invoked with the simple request message `CheckoutRequest` represented as `CheckoutRequest()` and it sends back the response message `CheckoutResponse()`.

The ShoppingCart service is an example of a *stateful Web service*. This implies that the availability of Web service operations depends not only on the input, but also on the internal state of the service, which in turn results from previous operation invocations. For instance, the client is not allowed to perform any operation before authenticating, and checking out only makes sense with a non-empty cart. Viewed from another angle, stateful Web services are a form of conversational Web services, which assume an interaction protocol defining rules for a suitable sequencing of operation invocations. Understanding and verifying this interaction protocol is crucial for the interoperability between a consumer system and a provided Web service. In addition, the ShoppingCart service operates on *persistent data* (in contrast to a *transformational* Web service, which accepts some user input and returns a result). The presence of persistent data implies that the result of invoking a Web service depends on the state of the persistent data (such as the existence of a user account in an accounts database), in addition to the user input (such as login data). Stream X-machines are suitable to formally model both the behaviour of single operations (pre-conditions and effects on persistent data), as well as the expected sequencing of operations for successful interaction.

Some parallels can be drawn between a stateful Web service and a stream X-machine, given that both accept inputs and produce outputs, while performing specific actions and moving between internal states. SXM inputs correspond to request messages, outputs correspond to response messages, and processing functions correspond to operation invocations in distinct contexts. In addition, the service provider has to define the memory structure, not only as a substitute for internal state, but also to supply genuine test data that can become part of the generated test sequences. Figure 5 is the diagrammatical representation of the stream X-machine model of the ShoppingCart service. It has to be noted that the transitions on the diagram do not correspond to operations or messages of the Web service but to processing functions as defined later on. Furthermore, some transitions that represent exceptional behaviour are not shown in the diagram for the sake of clarity. For instance, attempting to invoke the operation `addItem` while the service is found at state waiting, will exercise the self-transition `faultyAddItem`. Similar transitions exist for the rest of the operations and the states.

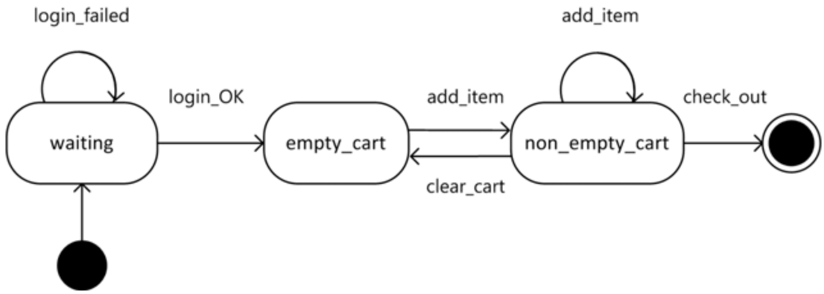


Figure 5 Stream X-machine model of the shopping cart Web service example

The memory in the ShoppingCart service example is used to store information about valid user accounts and the contents of the shopping cart. The following XMDL-O code shows the definition of accounts as a set of Account objects and the cart as a set of item identifiers (strings). For the purpose of testing the system we assume that there are two valid user accounts.

```
#class Account {
     username: string,
     password: string,
}.

#objects:
     account1: Account,
     account2: Account,
     accounts: set_of Account,
     cart: set_of string.

#init_values:
     account1.username <- "usr1",
     account1.password <- "pwd1",
     account2.username <- "usr2",
     account2.password <- "pwd2",
     accounts <- {account1, account2},
     cart <- emptySet.
```

State transitions in SXMs are labelled with processing functions. A processing function is triggered by an input event, when a specified guard condition holds, produces some output, and potentially updates (modifies) the memory. The updating of the memory consists of a sequence of assignments as specified in the update part of the processing function definition. The following XMDL-O code shows the definition of processing functions. When modelling Web services, the inputs and the outputs of the processing functions correspond intuitively to request and response messages of Web services respectively.

```
#fun loginOK( LoginRequest(?usr, ?pwd) ) =
     if ?account \= null and ?pwd = ?account.password
     then ( LoginResponse(true) )
     where
```

```
          ?account <- head (select(username = ?usr, accounts)).

#fun loginFailed( LoginRequest(?usr, ?pwd) ) =
     if ?account = null or ?pwd \= ?account.password
     then ( LoginResponse(false) )
     where
          ?account <- head (select(username = ?usr, accounts)).

#fun addItem( AddToCartRequest(itemId) ) =
     then ( AddToCartResponse() )
     update
          cart <- itemId addsetelement cart.

#fun clear( ClearCartRequest() ) =
     then ( ClearCartResponse() )
     update
          cart <- emptySet.

#fun checkOut( CheckOutRequest() ) =
     if cart \= emptySet
     then ( CheckOutResponse() ).
```

3.3 Test set derivation

A main strength of modelling systems with SXMs is the existence of a test generation method which under certain assumptions [3, 4], is proven to find all faults in the implementation. Examples of faults that can be detected in the implementation include erroneous transition labels, erroneous next-states, missing states, extra states, etc [15]. The testing method is a generalization of the W-method [16]. It works on the assumption that the system specification and the implementation can be both represented as stream X-machines with the same type (i.e. both specification and implementation have the same processing functions) and satisfies the following design for test conditions: completeness with respect to memory (all processing functions can be exercised from any memory value using appropriate inputs) and output distinguishability (any two different processing functions will produce different outputs if applied on the same memory/input pair).

When the above requirements are met, the Stream X-machine testing method may be employed to produce a complete test set of input sequences which can be used for the verification of the implementation under test. In fact it is proved that only if the specification and the implementation are behaviourally equivalent, the test set produces identical results when applied to both of them. Otherwise it is guaranteed that it will reveal the faults in the implementation.

The first step to constructing the test set of input sequences is based on the application of the W-method to the associated finite state automaton of the SXM, by considering processing functions as simple inputs. The test set X for the associated automaton consists of sequences of processing functions and it is given by the formula:

$$X = S(\Phi^{k+1} \cup \Phi^{k} \cup \ldots \cup \Phi \cup \{\square\})W$$

Where W is a characterization set, S a state cover of the associated finite state automaton, and k is the estimated difference of states between the implementation and the specification. A characterization set is a set of sequences of processing functions for which any two distinct states of the machine are distinguishable and a state cover is a set of sequences of processing functions such that all states are reachable from the initial state. The W and S sets in the ShoppingCart Web service example are:

```
W = {<hloginOK>, <addItem>, <checkout>}
S= {<□>, <loginOK>, <loginOK, addItem>, <loginOK, addItem,
checkout>}
```

The derived test set X, e.g. for $k = 0$, is the following (note that it is not completely presented):

```
X = {<loginOK>, <addItem>, <checkOut>, <loginOK, loginOK>,
<loginFailed, loginOK>, <addItem, loginOK>, <clearCart, loginOK>,
<checkOut, loginOK>, <loginOK, addItem>, <loginOK, checkOut>,
<loginOK, loginOK, loginOK>, <loginOK, loginFailed, loginOK>,
<loginOK, addItem, loginOK>, <loginOK, clearCart, loginOK>,
<loginOK, checkOut, loginOK>, <loginOK, loginOK, addItem>,
<loginOK, loginFailed, addItem>, <loginOK, addItem, addItem>,
<loginOK, clearCart, addItem>, <loginOK, checkOut, addItem> …}
```

The above test-set X consists of sequences of operations. These sequences have to be converted to sequences of inputs. This is achieved by the fundamental test function as described in [3]. For instance, the sequence of operations `<loginOK, addItem, addItem>` is converted to the following sequence of inputs:

```
<loginRequest("usr1, "pwd1"),
addToCartRequest("912"),
addToCartRequest("875")>
```

To complete the process of test set generation and enable a testing engine to execute the test cases, these abstract test cases have to be mapped to executable test cases that the testing engine can understand.

4. Reliable Web Service Discovery and Publication Approach

The method described in the previous section for modelling and model-based testing of Web services using stream X-machines has applicability in a range of scenarios involving various stakeholders. In this section we describe the use of stream X-machine formal models in a reliable Web service publication and discovery approach [17, 18]. The approach is founded on the idea that augmenting Web service interface descriptions (WSDL) with formal behavioural specifications is beneficial in registry-based testing of provided Web services during publication, and in service selection by the consumer during discovery.

Figure 6 provides an overview of the proposed approach, which requires the cooperation of the three main stakeholders in a SOA environment: the service provider, the service broker, and the service consumer. The following subsections describe the steps involved in the approach from the perspectives of each of these three stakeholders.

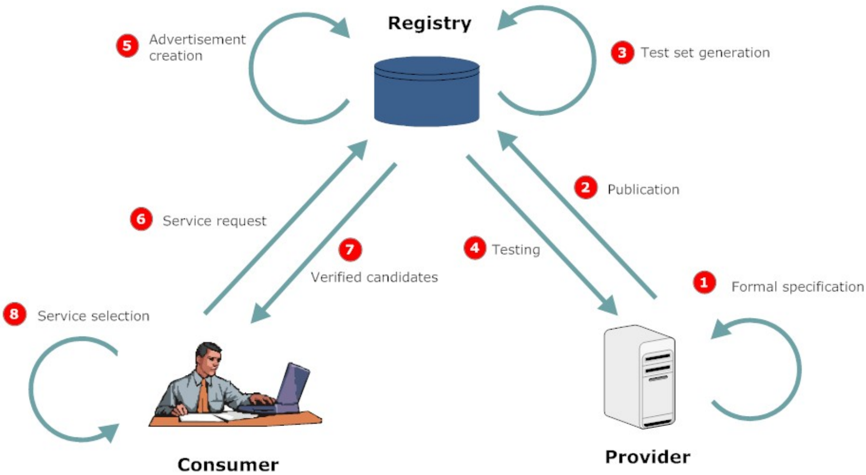


Figure 7 Overview of the publication and discovery approach

4.1 Provider's perspective

The service provider goes through data-level and behavioural-level analysis to derive a formal model reflecting the behaviour of the Web service that is to be published, using the stream X-machine (SXM) formalism [2]. The SXM model, expressed in a markup language such as XMDL [11], is then linked to the WSDL document of the Web service. Practically, this may be achieved by adding an SAWSDL annotation [19] that references the URI of the SXM markup document. The next step by the provider is the publication of the Web service to a service registry maintained by a broker. The publication query, which references the semantically annotated WSDL document at the provider site, initiates the publication procedure at the broker site.

4.2 Broker's perspective

A key role of the service broker in this approach is to verify the behaviour of the provided Web service implementation through model-based testing, and upon successful test results, to accept it in the service registry. This step is necessary to ensure that the implementation of the provided Web service really conforms to the advertised behavioural specifications. It is possible that this might not be the case, either because of insufficient testing at the provider site, or because of malicious intent. With the attached SXM specification, the broker is able to derive the test sequences for verification automatically. The theory of complete functional testing from X-machines offers a method for deriving a complete, finite set of test cases, which is proven to find all faults in the implementation under test [3].

The input sequences and the expected output sequences produced by the testing algorithm are at the same level of abstraction as the stream X-machine model, so they need to be mapped to concrete data types (XSD), which can be understood by the Web service. This is possible if the provider uses a mechanism to link the abstract types in the XMDL model with the XSD types in the WSDL document. The test cases are then written in a form of executable tests, which are interpreted and run by a testing engine that communicates with the Web service under test through SOAP messages. If the test results are successful, i.e. the expected and produced outputs match, then the Web service implementation has been shown to be free of faults with respect to the behavioural specifications. In such a case, an advertisement of the Web service is created and added to the service registry, otherwise the Web service is rejected as faulty. The benefit of performing the verification procedure at the broker site, as opposed to performing it at the consumer site upon discovery, is that it needs to be done only once. Since only successfully tested Web services are accepted by the broker, consumers are ensured that the Web services they discover have been verified with respect to their specifications.

4.3 Consumer's perspective

As a first step during discovery, the service consumer formulates a service request and submits it to the service registry. In response, the service broker returns a set of annotated service descriptions that match the service request. Notably, our approach is not bound to any particular matchmaking mechanism, so that any existing mechanism may be employed to perform syntactic or semantic matchmaking between the service request and the service advertisements. The service consumer can take advantage of the SXM behavioural model provided with each service candidate, in order to perform service selection. This is a validation process where the consumer ensures that a service model satisfies his or her requirements. An important validation technique is model animation, during which the user feeds the model with sample inputs and observes the current state, transitions, processing functions, memory values, and last but not least, the outputs. For example, X-System is a prolog-based tool supporting the animation of stream X-machine models [10]. In addition, model checking may be employed on the SXM model to check for desirable or undesirable properties, which are specified in a temporal logic formula. Research on X-machines offers a model-checking logic, called XmCTL, which extends Computation Tree Logic (CTL) with memory quantifiers in order to facilitate model-checking of X-machine models [20]. Alternatively, if the consumer has a SXM model of the required service, it can be validated by state and transition refinement against the published SXM of the provided service [21].

5. Conclusions

The approach described in this paper is supported in fragments by a number of existing tools, which have been developed during previous research. However, numerous gaps exist in the required supporting infrastructure, and future research will address the consolidation of techniques and tools into a single framework with industrial applicability. The main focus will be on the broker infrastructure, which requires more substantial work to support automated Web service testing and, possibly, behavioural matchmaking. We will base our work on a semantically-enhanced, UDDI-based service registry supporting the SAWSDL specification, as part of the EU-funded STREP project FUSION [22]. In order to support model-based testing of Web services, we are planning to integrate the semantic registry with tools for test case generation from XMDL specifications, and with capabilities for runtime testing of a Web service implementation. Additionally, in order to support behavioural matchmaking, we are planning to define an abstract query language for the service consumer and extend the current matchmaking algorithm of the semantic service registry to match the behavioural query with the advertised stream X-machine models.

References

1 Papazoglou, M. P., Traverso, P., Dustdar, S., Leymann, F. (2006). Roadmap of Service Oriented Computing. http://infolab.uvt.nl/pub/papazogloump-2006-96.pdf. March 2006.

2 G. Laycock. The Theory and Practice of Specification-Based Software Testing. PhD thesis, Dept of Computer Science, Sheffield University, UK, 1993.

3 Holcombe, M. and Ipate, F. (1998). Correct Systems: Building Business Process Solutions. Springer-Verlag, Berlin.

4 Ipate, F. and Holcombe, M. (1997). An integration testing method that is proven to find all faults. International Journal of Computer Mathematics, 63, pp 159-178.

5 Sinha, A. and Paradkar, A. (2006). Model-based functional conformance testing of Web services operating on persistent data. In TAV-WEB'06, pages 17{22, Portland, Maine, USA, 2006. ACM.

6 Keum, C., Kang, S., and Ko, I.Y. (2006). Generating test cases for web services using extended finite state machine. In TestCom 2006, pp 103-117. Springer, 2006.

7 Bertolino, A., Frantzen, I., Polini, A., and Tretmans, J. (2006). Audition of web services for testing conformance to open speci_ed protocols. Architecting Systems with Trustworthy Components, LNCS 3938, 2006.

8 Heckel, R. and Mariani, L. (2005). Automatic conformance testing of web services. In FASE 2005, pp 34-48. Springer, 2005.

9 Eilenberg, S. (1974). Automata, languages and machines. Academic Press, New York.

10 Kapeti, E. and Kefalas, P. (1999). A design language and tool for X-machine specification. In Proceedings of the 7th Panhellenic Conference on Information Techology, Greek Computer Society, Ioannina, 1999.

11 Kefalas, P. (2000). X-machine description language: User manual, version 1.6. Technical Report WP-CS07-00, CITY College, 2000.

12 Kefalas, P., Eleftherakis, G., and Sotiriadou, A. (2003). Developing tools for formal methods. In Proceedings of the 9th Panhellenic Conference in Informatics, pages 625-639, November 2003.

13 Dranidis, D., Eleftherakis, G., and Kefalas, P. (2005). Object-based language for generalized state machines. Annals of Mathematics, Computing and Teleinformatics (AMCT), 1(3):8-17, 2005.

14 Dranidis, D., Kourtesis, D. and Ramollari, E. (2007). Formal Verification of Web Service Behavioural Conformance through Testing. In Proceedings. of the 3rd South-East European Workshop on Formal Methods, Thessaloniki, Greece, November 2007.

15 F. Ipate and R. Lefticaru. State-based testing is functional testing. Testing: Academic and Industrial Conference Practice and Research Techniques - MUTATION, 2007. TAICPART-MUTATION 2007, pages 55-66, 10-14 Sept. 2007.

16 T. S. Chow. Testing software design modelled by finite state machines. IEEE Transactions on Software Engineering, 4:178-187, 1978.

17 E. Ramollari, D. Kourtesis, D. Dranidis, and A.J.H. Simons. Towards Reliable Web Service Discovery through Behavioural Verification and Validation. In Proceedings of the 3rd European Young Researchers Workshop on Service Oriented Computing (YR-SOC 2008), London, UK, June 2008 (forthcoming).

18 D. Kourtesis, E. Ramollari, D. Dranidis, and I. Paraskakis. Discovery and Selection of Certified Web Services through Registry-Based Testing and Verification. In Proceedings of the 9th IFIP Working Conference on Virtual Enterprises (PRO-VE'08), Poznan, Poland, September 2008 (forthcoming).

19 J. Farrell and H. Lausen. Semantic Annotations for WSDL and XML Schema. W3C Candidate Recommendation, January 2007. Available at: http://www.w3.org/TR/sawsdl/.

20 G. Eleftherakis, P. Kefalas, and A. Sotiriadou. Xmctl: Extending temporal logic to facilitate formal verification of x-machines. Matematica-Informatica, 50:79-95, 2002.

21 A. J. H. Simons. A theory of regression testing for behaviourally compatible object types. Software Testing, Verification, and Reliability, 16(3):133-156, August 2006.

22 Kourtesis, D. and Paraskakis, I. (2008). Web service discovery in the FUSION semantic registry. In Proceedings of the 11th International Conference on Business Information Systems (BIS 2008), May 2008.

RISK, WELL BEING AND COGNITION

Quality of Life Factors among University Students Abstaining, Experimenting and Habitually using Drugs.

Mateusz Hędzelek[1], Paulina Wnukiewicz, Marcin Wnuk, Krystian Siłakowski, Paweł Makarczyk, Janusz Czapiński[2]

Warsaw University Faculty of Psychology, Department of Social Psychology, Ul. Stawki 5/7 00-183 Warsaw, Poland,{mhedzelek[1], czapinski[2]}@psych.uw.edu.pl

The research concerns a comparison of a conglomerate of quality of life, sense of life and sense of coherence factors between university students abstaining, experimenting and habitually using drugs. Following factors have been measured: desire to live, passion for life, feeling of happiness, global contentment of life concerning the past, present and future, particular satisfaction rates of contentment of different life areas. The aim of the study was to explain the psychological conditions of quality of life among university students in different stages of drug use. The results were to show how drugs effect a complex range of quality of life factors among students using them. The fundamental hypothesis was that with the development of drug use (abstaining - experimenting – using habitually) the quality of life measures will decrease. In order to evaluate this phenomenon one hundred-thirty-three students were researched. The comparison has been done between three groups of participants standaryzed in socio-demographic data. An analysis of factors decreasing and increasing their quality of life has been done. The research indicated that there were no differences in quality of life and sense of coherence measures between these three groups. Statistically important differences have not been noted. There were statistically significant correlations within the quality of life factors in separate groups and in the researched population as a whole.

Keywords
quality of life, drugs, university students.

1. Introduction

Quality of life, broadly analyzed by such fields of science as: philosophy, sociology, psychology, medicine, economy, political science is still an open to discussion topic with intriguing issues and recent research discoveries. Based on those listed above paradigms of approach there are formulated more or less specific definitions of this complex phenomenon. The most current trend is to reduce its aspects to one particular variable based on what scientific field is used to describe and evaluate the topic. The outcome of reducing the complexity is profitable - more stable data for quantitative and qualitiatve analysis is obtained.

The pioneer in terms of broad population quality of life research of its citizens, are the United States of America. Their project carried out in the sixties was aimed to find the most handicapped social groups and work with enhancing their status. Even though already at that time there were numerous groups of people suffering from the illness of alcoholism and drug addiction there has been surprisingly very little research done on those groups. Surprising is the fact that these illnesses were so clearly represented by significant amounts of people in the US population and that the destructive effect of addictions deteriorates the lives of those ill themselves but also has negative impact on their families and social surroundings [1].

Stages of drug use include: abstinence, experimenting, habitual use, abuse and addiction. The first three phases are not considered as problematic although there is apprehension given to the fact that they might lead to the development of addiction. Drug addiction is treated as a chronic and deadly disease, which has declining effect on the quality of life of the addict and his family. The fundamental requisite to stop its development is abstinence. Not using drugs is obligatory for starting treatment [2],[3].

2. Drug Addiction as a Disease

2.1 Substance Dependence Development

In order to abuse and further on become addicted to drugs – these substances have to be administrated to the human blood stream. Routes of administration are various. Starting with swallowing, through smoking, snorting, inhaling and finishing with the intravenous injections. The annual prevalence of drug use within the global population was estimated at 200 million – 4,8% of the 15-64 of age population. The basic three phases of the development of any addiction (substantial related or non substance related are the same: initial use associated with the experimenting stage and afterwards habitual using. Some people stop at this level of drug use and keep their habit stably non negative in terms of different life areas. Some of the habitual users stop taking the substance and become abstinent. Some never try illicit substances and thus always are in the stage of abstinence, but according to the United Nations Office on Drugs and Crime 2007 report a group of 27 million people between the age of 15-64 become problem users. The term problem users applies to two psychiatric categories: abusers and addicts.

Extent of drug use (annual prevalence*) estimates 2005/6 (or latest year available)

	Cannabis	Amphetamine-type stimulants		Cocaine	Opiates	of which heroin
		Amphetamines	Ecstasy			
(million people)	**158.8**	**24.9**	**8.6**	**14.3**	**15.6**	**11.1**
in % of global population age 15-64	3.8%	0.6%	0.2%	0.3%	0.4%	0.3%

* Annual prevalence is a measure of the number/percentage of people who have consumed an illicit drug at least once in the 12 month-period preceding the assessment.

Figure 1 Illegal drug use at the global level (2005/2006)

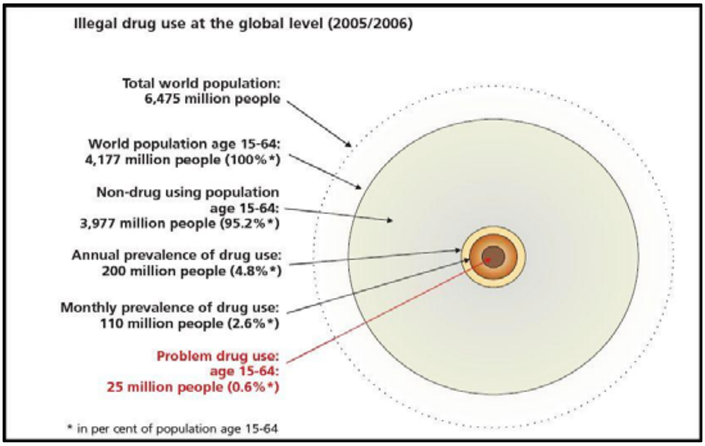


Figure 2 Extent of drug use (annual prevalence) estimates 2005/2006

2.2 Criteria for Substance Abuse

The essential feature of substance abuse according to the fourth edition of the American Psychology Association Diagnostic and Statistical Manual of Mental Disorders (DSM-IV) is a maladaptive pattern of substance use manifested by recurrent and significant adverse consequences related to the repeated use of substances [5]. There may be repeated failure to fulfill major role obligations, repeated use in situations in which it is physically hazardous, multiple legal problems, and recurrent social and interpersonal problems (Criterion A). These problems must occur recurrently during the same 12-month period. Unlike the criteria for substance dependence the criteria for substance abuse do not include tolerance, withdrawal, or a pattern of compulsive use and instead include only the harmful consequences of repeated use. A diagnosis for substance abuse is preempted by the diagnosis by the diagnosis of substance dependence if the individual's pattern of substance use have ever met the criteria for dependence for that class of substances (Criterion B). Although a diagnosis if substance abuse is more likely in individuals who have only recently started taking the substance, some individuals continue to have substance-related adverse social consequences over a long period of time without developing evidence of substance dependence. The category of substance abuse does not apply to caffeine and nicotine. The concrete criteria for substance abuse are:

A. A maladaptive pattern of substance use leading to clinically significant impairment or distress, as manifested by one (or more) of the following, occurring within a 12-month period:

(1) recurrent substance use resulting in a failure to fulfill major role obligations at work, school, or home (e.g., repeated absences or poor work performance related to substance use; substance-related absences, suspensions or expulsions from school; neglect of children or household)

(2) recurrent substance use situations in which it is physically hazardous (e.g., driving an automobile or operating a machine when impaired by substance use)

(3) recurrent substance-related legal problems (e.g., arrests for substance-related disorderly conduct)

(4) continued substance use despite having persistent or recurrent social or interpersonal problems caused or exaverbated by the effects of the substance (e.g., arguments with spouse about consequences of intoxification, physical fights)

B. The symptoms have never met the criteria for substance dependence for this class of substance.

2.3 Criteria for Substance Dependence

The fourth edition of the American Psychology Association Diagnostic and Statistical Manual of Mental Disorders (DSM-IV) describes substance dependence as a cluster of cognitive, behavioral, and physiological symptoms indicating that the individual continues use of the substance despite significant substance-related problems [5]. There is a pattern of repeated self-administration that usually results in tolerance, withdrawal, and compulsive drug-taking behavior. A diagnosis of substance dependence can be applied to every class of substances except caffeine. The symptoms of dependence are similar across the various categories of substances, but for certain classes some symptoms are less salient, and in a few instances not all symptoms apply (e.g., withdrawal symptoms are not specified for hallucinogen dependence). Although not specifically listed as a criterion item, „craving" (a strong subjective drive to use the substance) is likely to be experienced by most (if not all) individuals with substance dependence. Dependence is defined as a cluster of three or more of the symptoms listed below occurring at any time in the same 12-month period.

A maladaptive pattern of substance use, leading to clinically significant impairment of distress, as manifested by three (or more) of the following, ocuring at any time in the same 12 month period:

(1) tolerance, as defined by either of the following

(a) a need for markedly increased amounts of the substance to achieve intoxication or desired effect

(b) markedly diminished effect with continued use of the same amount of the substance

(2) withdrawal, as manifested by either of the following:

(a) the characteristic withdrawal syndrome for the substance

(b) the same (or closely related) substance is taken to relieve or avoid withdrawal symptoms

(3) the substance is often taken in larger amounts or over a longer period than was intended

(4) there is a persistent desire or unsuccessful efforts to cut down or control substance use

(5) a great deal of time is spent in activities necessary to obtain the substance (e.g., visiting multiple doctors or driving long distances), use the substance (e.g., chain smoking), or recover from its effects

(6) important social, occupational, or recreational activities are given up or reduced because of substance use

(7) the substance use is continued despite knowledge of having a persistent or recurrent physical or psychological problem that is likely to have been caused or exacerbated by the substance (e.g., current cocaine use despite recognition of cocaine-induced depression, or continued drinking despite recognition that an ulcer was made worse by alcohol consumption)

3. Quality of Life among Drug Addicts

Unfortunately the increase in interest of the quality of life phenomenon does not lead to growth of research intensivity among drug addicts and especially those in recovery in the Narcotics Anonymous program.

The Polish Bureau of Drug Addiction Prevention in their official publications underline the decrease of quality of life among people addicted to drugs as a result of:

- danger of contamination with HIV and hepatitis
- functioning in a criminal setting and dangers deriving from that (e.g., involvement in criminal activities, violence, moral deterioration)
- the constant need to obtain the substance
- the negative effect of the substance itself (e.g., overdose, infections, psychosis)
- hazardous sex
- functioning under the influence of drugs and decrease of control over behavior

In alcoholism there is a stable relation indicating that with the growth of alcohol dependence the quality of life decreases and the problems related to drinking increase [6]. There is also a significant difference in quality of life between people dependent and abusing alcohol – the quality of life of dependents is lower than those of abusers [7]. Research carried with the use of the Life Situation Survey (LSS) and Nothingham Health Profile (NHP) indicated that 80% of alcoholics had lower results than healthy people in both surveys. The Rotterdam Symptom Checklist (RSCL) was used to assess the mental and physical quality of life factors of addicts and people with cancer – the outcome indicated that it took three months of abstinence to equal the quality of life status between those two groups.

As a result of addiction development quality of life among drug users significantly decreases reaching the critical point of total mental and physical impairment. Addiction being a chronical and progressing disease as part of its unpleasant features in the final run leads to death so the quality of life decreasing influence is a rather obvious conclusion. Observing teenagers with few years of amphetamine use curriculum the horrifying effects of drugs on their minds and bodies are an urgent call for more increased activities leading those addicted to treatment and those endangered to adequate and effective prevention programs.

4. Research Methodology

The aim of the study was to explain the psychological conditions of quality of life among students in the initial three phases of addiction development. An analysis of factors decreasing and increasing their quality of life has been done. The whole project was carried out in an ex post facto one stage modality. There were three groups researched. One of the groups consisted of students experimenting with drugs (this was the experimental group 1). Another group consisted of students using habitually drugs. The last group consisted of students abstaining from all illegal drugs (this was the control group).

4.1 Participants

There were 133 researched participants divided by the frequency of drug use criteria into three groups (abstaining N=43, experimenting N=60, using habitually N=30). The criteria of pairing were socio-demographic variables – age and sex. The researched students were recruited from different universities in Poland. The questionnaire was distributed by psychology students from the Warsaw University among their friends that were suspected of experimenting and habitually using illegal drugs. The positioning in the particular groups was not based on the subjective declaration of being in one or another phase, but by evaluating the frequency of drug use stated in the beginning demographic questions. The criteria for abstinence was not using drugs during the past year and longer. The criteria for experimenting was using up to six times drugs during the past year and the criteria for habitually using was over six times used drugs during the past year.

The data gathered from 133 people was evaluated. The youngest student was 17 years old and the oldest was 50 years old. The average age for the researched students was 22,27 years. The biggest amount of students were from following faculties (arranged in quantitative order): psychology 33,3%, business 7,4%, law 6,9%, computer sciences and political science equally 4,4%.The tobacco smoking rate was 44,8% were smokers, 55,2% were nonsmokers. Within the 150 participants researched the drugs most commonly used while experimenting were: marihuana 53.3%, hallucinogens 17,2%, amphetamines 16,1%, cocaine 11,2%. Non of the surveyed students declared that they are experimenting with heroin. Within the 133 participants researched the drugs most commonly used while habitually using were: marihuana 34,8%, amphetamine 8,1%, cocaine 6,1%. Only two participants declared they are using heroin 1,5%. In the topic of legalizing marihuana 51,9% were anti legalization, 31,1% were pro legalization and 14,8% answered that they do not know are they against or pro.

4.2 Variables

The research measured and evaluated four specific groups of variables:

4.2.1 Quality of Life

The sense of life scale consists one test position in which the researched participant evaluates on a scale from 0 to 9 his desire to live (0 – expresses no desire to live at all – the answer „I don't want to live" and 9 – expresses the strongest desire to live – the answer „I have a very strong desire to live"). The average score for the Polish population is M=7,41; SD=2,38 [15].

The passion of life scale also consists one question. The researched participant chooses from one of three alternative answers (0-boring, 1-average, 3- passionate). The average score for this scale M=1,25; SD=0,56 [15].

In the scale of happiness in recent days the researched participant chooses one of the four answers (0-not too happy, 1-rather happy, 2-happy, 3-very happy). The average score for this scale is M=1,72; SD=0,68 [15].

The Beck Scale of Hopelessness is the most frequently used scale to assess quality of life. It consists 24 sentences with which the researched participant expresses agreement or disagreement (yes-no). The score is a sum of points obtained in particular answers. A number of 0-20 points can be obtained. The higher the score, the higher the sense of hopelessness. The average score for this scale is M=6,53; SD=4,59 [15].

The Cantrill Ladder is one of the most popular methods used to measure sense of quality of life. The researched participant assesses from 0-10 his overall life satisfaction. In this project the ladder was used in three variants: - retrospective measure (evaluation of how life seemed five years ago); - present measure (evaluation of current situation); - prospective measure (evaluation of the situation five years from now).

4.2.2 Sense of Coherence

The Sense of Coherence (SOC 29) questionnaire consists of twenty nine questions regarding these three areas: understanding life, sense of life and life steering. The researched participant choses his answer on the scale from 1-7. In this tool 7 represents the maximum power of a character involved with these three listed above areas and 1 represents the minimum power of that character. By adding the total amount of points from different questions we obtain an overall score. The bigger the score the higher the sense of personal coherence. The psychometric parameters of this questionnaire are proved to be desirable. The alfa-Cronbach validity factor for this tool is between 0,82 and 0,95 [17]. The conducted research on the Polish population proved that the alfa-Cronbach validity variable was between 0,85 and 0,88 and was also assessed to be high.

4.2.3 Socio-demographic Variables

The socio-demographic variables consist 3 factors: age; sex; education level.

4.3 Statistical Analysis

4.3.1 Difference Analysis

In order to explore the difference hypothesis the average scores in groups were counted for all variables listed above. The obtained scores were compared with the control group using the t-Student test for quantitative data and the U Mann-Whitney test for ordinal data or One Step Analysis of Variance (ANOVA) [19].

Quality of life	Group	N	Average	Standard abrevation	F	Significance
Desire of living	students abstaining	42	8,119	1,400	4,060	0,019
	students using habitually	30	7,066	2,211		
	students experimenting	60	7,916	1,393		
Life passion	students abstaining	43	1,511	0,505	0,548	0,580
	students using habitually	30	1,366	0,556		
	students experimenting	61	1,442	0,646		

Happiness during the last days	students abstaining	43	1,744	0,658	0,898	0,410
	students using habitually	30	1,566	0,858		
	students experimenting	61	1,524	0,941		
Beck-sense of hopelessness	students abstaining	43	4,302	2,385	4,393	0,014
	students using habitually	30	6,500	4,876		
	students experimenting	60	4,400	3,320		
Cantrill – retrospective	students abstaining	43	6,325	2,327	2,704	0,071
	students using habitually	30	6,033	2,341		
	students experimenting	61	7,082	2,115		
Cantrill – present	students abstaining	43	7,674	1,267	1,208	0,302
	students using habitually	30	7,133	1,547		
	students experimenting	61	7,524	1,597		
Cantrill – prospective	students abstaining	43	8,767	0,947	3,524	0,032
	students using habitually	30	8,066	2,083		
	students experimenting	61	8,852	1,180		

Table 1 Anova table - statistics for groups and significance of quality of life differences between students abstaining, experimenting and habitually using drugs.

There were no significant differences in quality of life measures between the three researched groups. The level of: desire of living, life passion, happiness during the last days, Beck sense of hopelessness, Cantrill ladder retrospective, present, prospective has not been proven to be significantly statistically different among students abstaining, experimenting and using habitually drugs. The results clearly indicate that within these three stages of drug use there are no noted subjective signs of quality of life enhancement or deterioration. This proves that either the researched group was precisely chosen in terms of placement in the specific phase of addiction and thus did not yet suffer any quality of life distortions, or the mechanisms of denial inherent for the disease of addiction have already started influencing the users perception of self and life.

		Substraction of squares	df	Average square	F	Significance
SOC Understanding	Between groups (All together)	427,106	2	213,553	2,527	,084
	Within groups	10899,705	129	84,494		
	Overall	11326,811	131			
SOC Sense	Between groups (All together)	647,294	2	323,647	6,643	,002
	Within groups	6285,221	129	48,723		
	Overall	6932,515	131			
SOC Steering	Between groups (All together)	105,739	2	52,869	,741	,479
	Within groups	9205,140	129	71,358		
	Overall	9310,879	131			
SOC Overall	Between groups (All together)	2779,687	2	1389,843	3,043	,051
	Within groups	58912,942	129	456,689		
	Overall	61692,629	131			

Table 2 Anova table - statistics for groups and significance of sense of coherence differences between students abstaining, experimenting and habitually using drugs.

There were no significant differences in sense of coherence measures between the three researched groups. The level of: SOC Understanding variable SOC Sense variable, SOC Steering variable and SOC Overall has not been proven to be significantly statistically different among students abstaining, experimenting and using habitually drugs. The results clearly indicate that within these three stages of drug use there are no noted subjective signs of sense of coherence enhancement or deterioration. As above this proves that either the researched group was precisely chosen in terms of placement in the specific phase of addiction and thus did not yet suffer any quality of life distortions, or the mechanisms of denial inherent for the disease of addiction have already started influencing the users perception of self and life.

4.3.2 Correlation Analysis

In order to explore the correlations between measured variables the Pearson correlation factor was used for quantitive data and the Spearman correlation factor was used for ordinal data [19].

	Sense of Coherence
Desire to live	,559(**)
Passion for life	,124
Happiness during the last days	,278(**)
Beck sense of hopelesness	-,652(**)
Cantril retrospective	,286(**)
Cantrill present	,501(**)
Cantrill prospective	,475(**)

* Corelation is significant at level 0,05

** Corelation is significant at the level 0,01

Table 4 Corelations between Quality of life measures and overall Sense of Coherence.

The results indicate that sense of coherence correlates strongly (significance level 0,01) with following quality of life factors: desire to live, happiness during the last days, negatively with Beck sense of hopelessness, Cantrill retrospective, Cantrill present and Cantrill prospective. The only quality of life measure that does not correlate significantly with sense of coherence is passion for life. The results indicate that almost all quality of life factors have a positive connection with the sense of coherence - this among many other conclusions proves that the research tools were properly gathered and used.

5. Conclusions

The research indicated that the three groups of students in early phases of addiction did not differ in terms of quality of life. I believe that it is a very interesting result clearly stating two important facts. The first one is: you can experiment with drugs and take them habitually and not have negative effects in terms of quality of life. Various research indicates that using drugs of any kind and in any extent is something horrible for ones life and the psychological consequences appear immediately. Our research shows that this might not be the truth. There is an over diagnosis in Poland stigmatizing all people having any contact with drugs as addicts. They are forced to start treatment and pressured to admit that they are addicts. For many it is a completely mistaken intervention which leads them to opposition to therapists and therapy and for some to using drugs extensively even though before the diagnosis they did not have a problem with substance abuse. The second important fact deriving from this research project is that for some the immanent for the disease of addictions mechanisms of denial are already working in these early phases of dependence (denial mechanisms in this approach are defined as unconciouss strategies of the ego to defend itself from the awareness that the person is doing things that pose a threat to his or hers psychological, social and physical well-being). This phenomenon explains how some dug users are not decreasing their drug use despite the negative consequences it inflicts on them. There are various theories of reasons and risk groups trying to explain why some continue their substance abuse habit and others stop but this research may give very concrete evidence why some develop an extensive drug habit.

Another non difference factor were the sense of coherence measures. Sense of coherence is defined in psychology as ones understanding of life, finding sense in it and being able to steer it. It is often used and interpreted within patients in therapy as a factor determining if the patients will continue treatment. In this research it can be perceived as a factor that can partly determine continuing drug use.

The correlation results indicated that almost all quality of life measures have a bond with the sense of coherence. The fact that the correlations were very strong (0,001 level) proves that the tools used in this research project were properly assigned and that the responders answered the questionnaire in a thorough and valid manner.

6. Discussion

Psychoactive substances have been present in the human society for thousands of years. Some where used in religious activities, others just as part of recreation. Some were made legal and others were abolished by government institutions. Some were discovered not intentionally and others were produced as an effect of a long scientific process. Some are treated as medication and some are treated as poison. No one in our present times can deny their effect on the human population. As most of things in life they also bring profits and at the same time they can have a negative effect. The border between this influence matter for some people is not precise. The fact that some cross this border and then return to the healthy area can be a popularizing intervention promoting these kinds of behavior. The society reacts to that with criticism and fights strongly against these effects, but we scientists have to be based in empirical data even if it can bring hazard. In this research we have shown that experimenting and habitually using illegal drugs does not have to deteriorate once quality of life. As is written in the conclusions part we are not certain if the results were not distorted by the denial mechanisms. This proves that there should be a more objective rather than subjective research carried out. With these two approaches we could prove more precisely how ones life is effected by the early phases of drug use that may lead to active addiction. For now after analyzing data gathered among university students we did not find any evidence that their subjective quality of life has been decreased by the fact of experimenting and even habitually using drugs.

References

1 Wnuk, M. Quality of life among members of Alcoholics Anonymous. Unpublished MA thesis. University of Adam Mickiewicz Poznań 2005.

2 Mellibruda, J. People with alcohol problems. Warsaw: Center for Development of Social Service 1996.

3 Niewiadomska, I. The role of values in being abstinent by Alcoholics Anonymous. W: Cz. Cekiera (red.), Addictio prevention a path to human freedom (p.165-177). Lublin: Scientific Association of KUL 2001.

4 Branicki, W. The miracle of change. On the 30th anniversary of Alcoholics Anonymous Fellowship in Poland. Poznań: Medzia Rodzina 2004.

5 American Psychiatric Association. Diagnostic and statistical manual of mental disorders (4^{th} ed., rev.) Washington DC 1994.

6 Polak, K. Quality of life evaluation of soldiers with alcohol dependence diagnose. Polish Psychiatry 2001, 6, (p.1057-1070).

7 Foster, J.H., Powell, J.E., Marshall, E.J., Peters, T.J. (Quality of life in alcohol dependent subjects – review. Quality of life research 1999, 8, (p.255-261).

8 Narcotics Anonymous. Narcotics Anonymous – Basic Text. World Service Office. CA 1983.

9 Narcotics Anonymous. It Works: How and Why: The Twelve Steps and Twelve Traditions of Narcotics Anonymous. World Service Office. CA 1993.

10 Brown, S. Treatment of alcoholics: developmental model of recovery. Warsaw: National Medical Publishing Companies 1990.

11 Cierpiałkowska, L. Alcoholism. Marriages in recovery. Pozań: Sientific Publishing UAM 1997.

12 Cekiera, Cz. Psycho-prevention of addictions, therapy and resocialization of addicts: methods, programs, models, treatment centers, rehabs, fellowships. Lublin: Scientific Association KUL 2001.

13 Woronowicz, B.T. Without secrets about addictions and their treatment. Warsaw: Institute of Psychiatry and Neurology 2001.

14 Wojdyłło, E. The beginning of the path: psychological lectures at an addictions psychiatric ward. Warsaw: Institute of Psychiatry and Neurology 1990.

15 Derbis, R. (red.) Quality of development versus quality of life. Częstochowa 2000.

16 Cekiera, Cz Toxicomania. Drug addiction. Medication addiction. Alcoholism. Nicotinism. Warsaw: National Scientific Publishing 1985.

17 Antonovsky 1995

18 Brzeziński, J. Methodology of psychological research. Warszaw: National Scientific Publishing 1996.

Factors Related to Compliance with Antismoking Regulation among Albanian Health Professionals

Erika Melonashi[1], *Angelos Rodafinos*[2], *Richard Eiser*[3]

[1]*SEERC, 17 Mitropoleos Str, Thessaloniki, Greece, emelonashi@seerc.org*

[2]*City College, Dept. of Computer Science, 13 Tsimiski Str, 54624 Thessaloniki, Greece, rodafinos@city.academic.gr*

[3]*The University of Sheffield, Dept. of Computer Science, Western Bank, Sheffield, S10 2TN, UK, j.r.eiser@sheffield.ac.uk*

The present study assessed differences in attitudes towards the policy, health risk beliefs and health professional role beliefs in smoking and nonsmoking Albanian health professionals. Additionally the study examined the relationship of compliance with antismoking regulation to attitudes towards the policy, health risk beliefs, health professional role beliefs and perceived compliance norms. Participants were 203 staff members of two Albanian hospitals. The measure used was a questionnaire constructed for the specific purpose of the study. Results supported three out of the seven hypotheses as significant differences were found between smokers and nonsmokers in attitudes towards the policy, health risk beliefs, and health professional role beliefs. However, no significant correlations were found between compliance and attitudes towards the policy, health risk beliefs, health professional role beliefs and perceived compliance norms. Additionally, the frequency and extensiveness of comments from participants prompted a qualitative analysis, which provided important insights. It was concluded that focusing on individual factors like attitudes and beliefs is probably not the best way to understand compliance in the Albanian context. According to qualitative data, external factors like formal deterrents are crucial in affecting compliance rates. These findings have practical implications for policy makers.

Keywords
Albania, antismoking regulation, compliance, healthcare.

1. Introduction

Albania is a developing country in South Eastern Europe, which has undergone major socio-political and economic changes in the last 20 years. This society in transition has provided a fertile ground for tobacco industries, which have invested considerably in the country. In a time where unemployment rates were inconceivably high (early 90s), Philip Morris was paying three times the average salary for cigarette sales and was even distributing cigarettes for free [1]. Since then, the situation has not changed a lot; tobacco production, smuggling and very high rates of consumption especially by young people are all phenomena characterizing the tobacco epidemic in the country [2].

In this context, several legal attempts towards tobacco control have been formulated. Antismoking regulations have been introduced in the Albanian legislation since 2001, when the country was involved in the Third Action Plan 1997-2001 for a Tobacco Free Europe, an initiative of the WHO. Subsequently, the Framework Convention on Tobacco Control was signed in June, 2004 and ratified in April, 2006 [3]. More specifically, as regards workplace smoking, Albania is legally bound by the International Labor Organization convention on occupational safety and health, which has identified smoking as the cause of disability, disease and death [4]. Therefore, presently in Albania there is a smoking ban in healthcare, education and government facilities, all indoor workplaces and offices, theatres and cinemas.

Nonetheless, studies have suggested that the implementation of antismoking regulation faces great difficulties [5] and the impact of noncompliance might become particularly important in settings like healthcare facilities. The healthcare system is the main structure involved in smoking prevention and cessation. It has been argued that doctors and nurses are the best advocates in the battle against smoking, since they are considered as the most authoritative health-promoting source [6-8]. Recognizing the crucial role of the health professionals in the battle against smoking it has claimed that antismoking campaigns cannot be successful if the agents that are supposed to represent health and well-being are actually smokers [9]. There is evidence that the smoking status of the physician is an important factor determining his/her suitability to act as a non-smoking advocate. Research findings across countries indicated that nonsmokers are more likely to practice cessation counseling with their patients as compared to smokers [10,11]. Furthermore, health professionals who smoke provide an unhealthy role model to patients, as demonstrated by studies documenting increased rates of smoking among patients observing personnel smoking [12].

Therefore, the nonsmoking health professional has been an important target in the battle against smoking. Although studies from Western countries suggest that smoking prevalence among healthcare members has constantly decreased [13] research evidence from Eastern Europe or the Balkans is still scarce [14] These countries deserve particular attention, since they have now become ‘the epicenter’ of the tobacco problem [15,16]. Furthermore, a consideration of Western examples right after the introduction of the tobacco regulations, suggests that implementation faces great problems, especially during the initial stages. For instance, a study of Italian hospitals in 1998 suggested that 87% of smokers actually smoked in the hospital environment, despite the antismoking regulation. Furthermore, almost 90% of the staff admitted that they were exposed to passive smoking during their working hours [5]. Evidence from China is even more striking, as 37% of doctors who smoke admitted that they had even smoked in front of their patients [17].

These high rates of noncompliance have also been characterized by a negative attitude towards the policy. A review published by the Organization for Economic Co-operation and Development [18] on regulatory governmental policies has identified attitude towards the policy as one of the main factors influencing voluntary compliance in the absence of external incentives. Comparative studies have shown that although there is some heterogeneity in the attitudes of smokers towards the policy [19, 20] they are generally more negative as compared to nonsmokers [21, 22] The present study aimed to readdress this issue in the Albanian context by examining differences in attitude towards the policy among smokers and nonsmokers; however, it additionally determined whether attitude bears any relationship to compliance with antismoking regulation among smokers.

Programs designed for promoting compliance with the policies also suggest important insights on other determinants of compliance. These programs have shown effectiveness in reducing workplace smoking and increasing cessation rates and have focused on two main aspects, including knowledge of health effects and the health promoter role of professionals [23]. The influence of knowledge on health behavior is still a widely disputed issue since substantial evidence has suggested a dichotomy between knowledge and behavior even among health professionals, i.e. knowledge of the direct health consequences of smoking shows no relationship to smoking behavior [24]. However, there is evidence that smoking behavior could in turn be predicted through a consideration of health beliefs about consequences of passive smoking, i.e. the tendency to minimize the health risk of passive smoke is associated to greater likelihood of smoking [25].

The emphasis on the health risks of passive smoking is especially important and should not be taken for granted even among health professionals. In fact, a study conducted in Chinese hospitals concluded that only 50% of the doctors acknowledged the negative health effects of passive smoking [17]. Additionally, health-risk beliefs are affected by the physicians' smoking habits as smoking physicians are less likely than nonsmokers to admit the causative role of passive smoking on several diseases [26]. Hence, there is evidence that smokers and nonsmokers differ in health risk beliefs associated to passive smoking. The present study readdressed this question by comparing smokers and non-smokers on health risk beliefs. However, it additionally determined whether health risk beliefs actually bear any relationship to compliance with antismoking policy among smokers.

The second component of training programs (i.e. the health promoter role) has received considerable attention since the practical applications extend beyond smoking, into health behavior in general (exercising, healthy eating etc). In fact, contemporary medicine views the health professional first and foremost as a promoter of healthy life [27, 28]. However, as mentioned above, there is evidence that the health promoter role is influenced by the smoking status of health professionals [10]. The present study readdressed this question by investigating differences in health promoter role beliefs in smoking and non-smoking health professionals. Additionally, the study determined whether health-promoter role beliefs associated to compliance with the antismoking policy in smokers.

The scope of the study extended further in considering reported compliance norms (defined as perceptions of other people complying with regulations). This decision was especially motivated by the observed massive noncompliance in the particular cultural context. It has been suggested that massive noncompliance undermines the authority of the law, and in these circumstances noncompliance might acquires the status of a social norm guiding behavior [18]. The present study tested this proposition by investigating the relationship between perceived compliance norms and self-reported compliance.

To conclude, the aim of the present study was to investigate differences between smoking and non-smoking health professionals in attitudes towards the ban, health risk beliefs and health professional role beliefs. More specifically it was expected that compared to smokers, non-smokers would be more positive towards the antismoking policy, more likely to agree with the health consequences of passive smoking and also more likely to agree with the health promoter role of professionals (Hypotheses 1 to 3). Furthermore, the study would investigate the relationships between compliance with the policy and attitudes towards the ban, health risk beliefs, health professional role beliefs and compliance norms in smokers. More specifically, it was expected that higher self-compliance would be associated to more positive attitude, more agreement with health risk beliefs and health professional role beliefs, and higher observed compliance (Hypotheses 4 to 7). Finally, the study also comprised an exploratory aspect, including the examination of reasons for noncompliance and additional free comments of participants on different issues. This part of the study enabled insight into the way that smokers and non smokers think about the issue and provided supplementary understanding to research questions.

2. Method

2.1 Participants

Questionnaires were administered to 250 health professionals employed in two hospitals in a town of South East Albania. The return rate of the questionnaires was 81.2%, which amounted to a final sample of 203 participants. More specifically, the sample consisted of 41 doctors, 26 males and 15 females (*Mage* = 42.5 years, *SD* = 11.97), 126 nurses, 46 males and 80 females (*Mage* = 33.1 years, *SD* = 11.7), 12 administrative staff, 3 males and 9 females (*Mage* = 44.5 years, *SD* = 10.1) and 24 other staff members (e.g. social workers, psychologists, cooks, ancillary staff etc.) (4 males, 20 females, *Mage* = 36.9 years, *SD* = 13.9).

2.2 Materials

A questionnaire was constructed for the specific purpose of the study. The measure consisted of 32 items, comprising five subscales, which measured reported self-compliance, perceived compliance norm, attitude towards nonsmoking regulation, health professional role beliefs and health risk beliefs. The last two subscales were adopted from existing scales while the rest of the questionnaire was constructed for the specific purpose of the study. Additionally, the questionnaire incorporated individual items. Examples include the first four items assessing demographic data (age, gender), profession, and smoking status of the participant. Smoking status was assessed through the item “Which of the following best describes your smoking behavior?” with response options including 1. I have never smoked /2. I have quit smoking / 3. I currently smoke occasionally / 4. I currently smoke everyday. The type of existing policy was also examined through an individual question “What sort of smoking policy is in place at the hospital?” with response options including 1. No smoking policy in place / 2. Smoking rooms available / 3. No smoking allowed at all on premises. Reasons for noncompliance were also measured through individual items. Sample items included: “Smokers consider the restrictions as unfair to them” with response options ranging from 1 = Strongly Agree to 5 = Strongly Disagree.

Self-reported compliance was assessed through seven items addressing only current smokers (regular and occasional). Items recorded measured smoking behavior in different areas of the hospital (e.g., hall, office etc.). Sample items included: “How often do you smoke in each of the following areas: office/corridor etc.” and response options ranged from 1 = Almost Always to 5 = Never, with higher scores showing higher compliance. The subscale showed good internal consistency (standardized alpha = .73)

Items 7 through 11 composed the second subscale measuring perceived compliance norm. Sample items included “How often do the hospital staff members ignore the smoking restrictions?” with response options ranging from 1 = Almost always to 5 = Never, and higher scores showing higher compliance. Reliability analysis showed good internal consistency (standardized alpha = .75).

The third subscale assessed the attitude of participants towards the antismoking policy and included six pairs of semantic differentials in a scale from one to eight, with higher scores showing more positive attitude. Sample items included “Introducing a smoking-free policy in the hospital is: worthless (1)/useful (8).” However, after the internal consistency analysis one of the items (Difficult-Easy) was dropped because it showed poor correlations with the overall scores. After this procedure the standardized alpha reached the value of .72.

The fourth subscale was composed of seven items and measured health professional role beliefs. This subscale was adopted from the Global Health Professionals’ Survey [29]. Sample items included “Heath professionals should routinely advise their smoking patients to quit smoking” with response options ranging from 1 = Strongly Agree to 5 = Strongly Disagree. Reliability analysis showed unacceptable internal consistency for this subscale (alpha value = .35). Therefore, the option of selecting single items for analysis was adopted; Item 23 “Health professionals should set a good example by not smoking” was considered as the most appropriate choice. The decision was motivated by the specific scope of the study, i.e., assessing beliefs regarding the health profession in the context of smoking. Analyzing the content of the items, it was concluded that those mentioning specific behaviors like asking/advising about smoking/quitting were not representative of the broader scope of the investigation (e.g., “Health professionals should routinely advise patients to avoid smoking around children”). On the other hand, item 22 was too broad, in the sense of not being specific to the smoking behavior (“Health professionals serve as role models for their patients and the public”). Hence, Item 23 was the most appropriate in the sense of being neither too specific, nor too broad but addressing exactly the issue of interest.

The final subscale assessed health risk beliefs associated to passive smoking and was adopted from the ITC survey [30] with a few additional items. Sample items include “Exposing a pregnant woman to other people’s smoke is dangerous for the baby” with response options ranging from 1 = Strongly Agree to 5 = Strongly Disagree. The subscale showed good internal consistency (standardized alpha = .81).

The questionnaire was translated in Albanian and back-translated into English to ensure correctness of translation by certified translators proficient in both languages. The guidelines proposed by Hambleton and De Jong [31] on improving the accuracy of translation were carefully considered. However, no major discrepancies were encountered among the two versions of the questionnaire. The small discrepancies were addressed through minor alterations in wording; these changes although irrelevant to the core meaning, increased the clarity of expression. Additionally, prior to its administration the questionnaire was completed by three health professionals. No problems with comprehensibility were evident during this procedure.

2.3 Design

The design of the present study comprised four separate parts. The first part involved between groups comparisons (smokers vs. nonsmokers), and included three between-subject variables: attitudes towards the policy, health risk beliefs and health promoter role beliefs. The second part was a correlation design assessing the relationships between self-reported compliance with antismoking policy (criterion variable) and four predictor variables including attitude towards the policy, health related beliefs, perceived norms of compliance and attitude towards the health promoter role in smokers. The third part involved descriptive statistics in summarizing the beliefs of smokers and nonsmokers about reasons of noncompliance with antismoking regulations. The last part was qualitative and involved content analysis of additional comments from participants.

2.4 Procedure

Preliminary contacts were established with the directors of the hospitals in order to obtain official permission to conduct the study in those settings. The directors and other staff members were very cooperative and greatly facilitated the whole procedure. The researcher was not physically present during the study; instead, the two social workers and the psychologist of the hospital were engaged in the data collection process. The researcher provided them with all the necessary information (both written and through phone) for an ethical data collection process. Furthermore the personal e-mail address and phone number of the researcher were provided to all participants for any queries they could have. Additionally, participants were provided with information about the topic of study through the information sheet prepared by the researcher. They were also assured about their right to withdraw at any time and that the results would be anonymous and confidential. The individuals who agreed to participate were given the consent forms to sign and subsequently the questionnaires were distributed. Questionnaires were either completed in the hospital settings or were taken home and returned at a latter date.

3. Results

Before the analysis the data were checked for normality of distribution, skewness and kurtosis. The skewness and kurtosis values for attitudes, health risk beliefs, health professional role beliefs, perceived compliance norms and self-compliance were within the acceptable range (values in the range from 0 to 1) suggesting that the distribution of data did not violate the assumption of normality.

Nonetheless, considering that Levene's tests for homogeneity of variance were significant, nonparametric tests were considered a better option to t-tests. Therefore, Man-Whitney test was used to compare smokers and nonsmokers attitudes, health risk beliefs and health promoter role beliefs. Results showed a significant difference in attitude towards the policy between smokers and nonsmokers (U = 2472, N1 = 159, N2 = 42, p < .01, 2-tailed significance) with nonsmokers having a significantly more positive attitude than smokers (See Mean Ranks, Table 1). Man-Whitney test for differences in health risk beliefs also showed a significant difference between smokers and nonsmokers (U = 2674, N1 = 159, N2 = 43, p < .05, 2-tailed significance), with nonsmokers reporting significantly more agreement with health related consequences of smoking. Finally, Man-Whitney test for differences in beliefs on the health professional role showed a significant difference between smokers and nonsmokers (U = 2730.5, N1 = 159, N2 = 43, p < .05, 2-tailed significance).

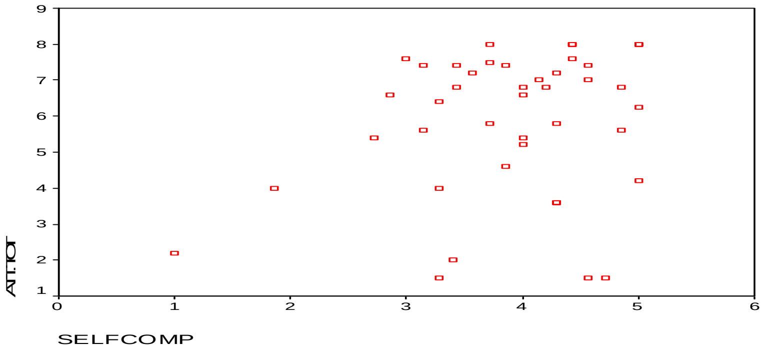
Variables	Nonsmokers Mean ranks	Smokers Mean Ranks	U-value	p-value
1. Attitude	106.45	80.36	2472.0	.009
2. HealthBelief	96.82	118.80	2674.5	.027
3. Health Prof.	97.17	117.50	2730.5	.012

Table 1 Between-group comparisons.

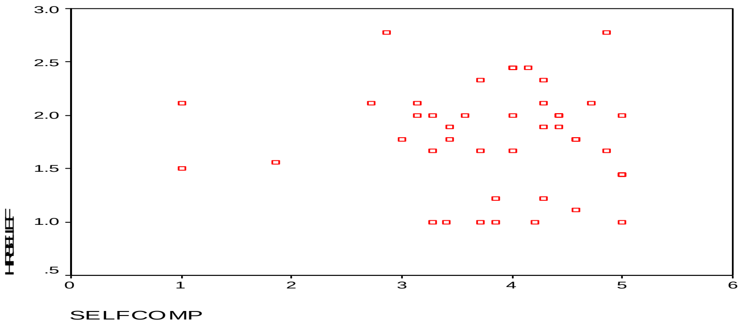
The second part of the analysis involved testing of four hypotheses through Pearson's correlations. There were no significant correlations between self-compliance and any of the following: attitudes towards the policy, health risk beliefs, health professional role beliefs and perceived compliance norms (see Scatterplots). Table 2 shows means, standard deviations and intercorrelations among the study variables.

Variables	M	SD	1	2	3	4	5
1.Attitude	5.84	1.98	1.00				
2. HBelief	1.79	.49	.08	1.00			
3. HealthPr	1.63	.85	.11	.17	1.00		
4.Oth.Comp	2.20	.80	-.01	.03	.04	1.00	
5.Self.Comp	3.82	.95	.24	-.04	-.07	.10	1.00

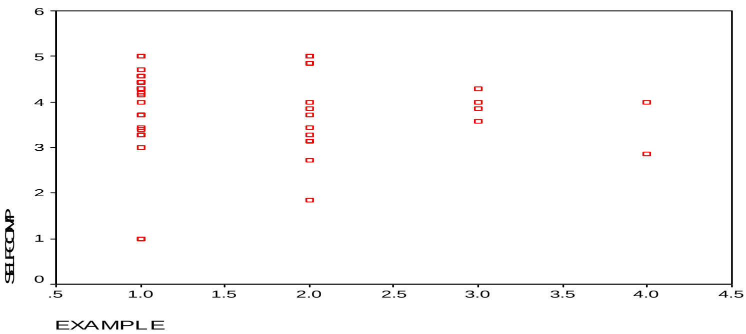
Table 2 Means, Standard Deviations And Intercorrelations for Smokers.



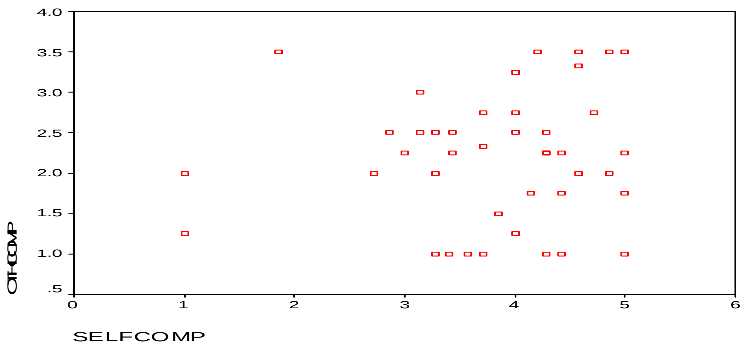
Scatterplot 1 Self-compliance vs. Attitude



Scatterplot 2 Self-compliance vs. Health-related beliefs



Scatterplot 3 Self-compliance vs. Health Professional Role



Scatterplot 4 Self-compliance vs. Compliance Norm

The third part of the study was exploratory; descriptive statistics were used to investigate on the type of the existing antismoking policy. The existing policy was reported as being 'total ban' by 49.8% of the participants, 'places for smoking' by 40.4% of the participants and 'no policy' by 9.9%. Additionally, frequencies (percentages) of agreement or disagreement with reasons for noncompliance were also examined. Descriptive analysis suggested that the majority of smokers agreed with the following reasons for noncompliance: perceived unfairness of regulation, its perceived violation by other people, addiction to smoking and smoking as a coping mechanism. Conversely, the majority of smokers disagreed with statements referring to lack of knowledge, lack of disapproval from other people and lack of consequences to violation of the regulation as reasons for noncompliance. Table 3 gives a detailed account for each of the items involved.

Reasons	Agree Smokers/Nonsmokers		Don't know Smokers/Nonsmokers		Disagree Smokers/Nonsmokers	
Unfair	46.4%	49.7%	26.8%	35.9%	26.9%	14.4%
Harm	26.2%	49.3%	21.4%	14.5%	52.4%	36.2%
Disapprove	38.4%	38.5%	12.8%	17.0%	48.7%	44.4%
Others	90.2%	79.5%	7.3%	14.6%	2.4%	6.0%
Coping	56.1%	64.4%	9.8%	13.8%	34.1%	21.7%
Addiction	70.8%	77.0%	17.1%	12.5%	12.2%	10.5%
Noconsequences	38.9%	37.8%	16.7%	13.6%	44.4%	48.6%

Table 3 Reasons for Smoking.

4. Discussion

The present study assessed differences in attitudes towards the policy, health risk beliefs and health professional role beliefs in smoking and nonsmoking Albanian health professionals (Hypotheses1 to 3). Additionally, the study examined the relationship of compliance with antismoking regulation to attitudes towards the policy, health risk beliefs, health professional role beliefs and perceived compliance norms (Hypotheses 4 to 7). The first hypothesis was supported by the results as there were significant differences between smokers and nonsmokers in attitude towards the policy. Therefore, nonsmokers were more positive towards the antismoking policy as compared to smokers, a finding that is in line with previous research [21,22]. An insight into reasons of the smokers' more negative attitude comes from the descriptive data as the majority of smokers cite "unfairness of the policy" as one of the reasons for noncompliance. Therefore, negative attitude is explicable in terms of perceived restrictiveness of the policy and interference with personal freedom. Indeed there is evidence that perceptions of restrictiveness might lead to normalization of noncompliance [18]. Nonetheless, the group of smokers is not homogenous in terms of attitude and it was expected that those with a more positive attitude would be more likely to comply. Results showed that this was not actually the case.

In fact, the fourth hypothesis was rejected as attitude towards the policy showed no relationship with reported compliance. Although past research has suggested that attitude is not a very good predictor of behavior [32], the present findings (i.e. no relationship at all) are rather extreme. It could be suggested that although a general measure of attitude towards the policy is sensitive enough to distinguish between smokers and nonsmokers it is less useful when it comes to predicting behavior. In this context, it might be important for future research aiming to predict compliance behavior to consider more specific variables like 'attitude towards compliance with the antismoking regulation'. In fact, research evidence suggests that specificity of attitude definition leads to better predictions of behavior [33, 34].

However, the present findings suggest that general attitude towards the policy and compliance are independent of each other. In fact, an examination of the Scatterplot illustrates the absence of any relationship and even instances of smokers with the same attitude but different levels of compliance are evident. These findings could be explicable in terms of perceived behavioral control (see Theory of Planned Behavior) [32]. Therefore, smokers with the same attitude might behave differently due to the amount of control they perceive on their behavior. Physical and psychological dependence from smoking might be important considerations in this context, taking into account that in the present study 71% of smokers cite "addiction" as reason for violating the regulation and another 56% cite "coping with stress." Hence, the consideration of self-efficacy in complying with the antismoking regulation might be an important suggestion for future studies.

The second hypothesis was instead supported by the results as there were significant differences between smokers and nonsmokers in health-related beliefs. Smokers tended to show less agreement with health-related consequences as compared to nonsmokers, findings that are in line with previous research [26]. These results are explicable in terms of cognitive dissonance; hence, smokers underestimate the health effects of passive smoking because they are engaged in an activity, which harms other people [35]. An important question to address at this point is whether health-related beliefs are somehow related to compliance behavior (Hypothesis 5). The findings of the study suggested that this is not the case.

In fact, the fifth hypothesis was rejected as there was no relationship between health-related beliefs and reported compliance. Therefore, smokers violate the regulation independently of how strongly they believe that it harms other people. An examination of the Scatterplot2 shows instances of people with the same beliefs, who nonetheless display very different rates of compliance. Descriptive data provide further support to the above finding, as more than half of the smokers disagree that knowledge of health effects influences compliance. These findings though might be related to the specific sample of the study (i.e. they are health professionals who supposedly have knowledge).

The third hypothesis was instead supported by the results as non-smokers tended to show significantly more agreement than smokers with the following statement "Health professionals should set a good example by not smoking." These findings could be explained through the phenomenon of 'role distance' whereby people are aware of their roles only as a specific part of their identity, which differs from the whole [36]. Therefore, the personal and professional aspects are well differentiated from each other to the extent that the requirements of profession do not dictate behavior. Cognitive dissonance could also explain the findings as smokers might tend to resolve the contradiction smoker vs. health professional, by changing their beliefs about the promoter role of the health professional.

Even though smokers and nonsmokers differ in their beliefs about the health promoter role, this variable shows no relationship to compliance with antismoking regulations (Hypothesis 6 was rejected). The examination of the Scatterplot3 suggests instances of smokers with the same level of agreement to the health professional role but remarkably different levels of compliance. This evidence suggests that although smokers and nonsmokers might be differentiated in terms of health professional beliefs, these beliefs in turn do not prove to be useful in understanding compliance behavior.

Finally, the seventh hypothesis was also rejected as there was no relationship between compliance and perceived compliance norm. Scatterplot4 suggests instances of smokers with the same level of perceived compliance norm but different levels of compliance. However, descriptive data suggest that 90% of smokers agree that violation of the rules by other people is a reason for noncompliance. These two apparently contradicting findings might be combined in suggesting a different level of analysis for perceived compliance norm. Instead of investigating compliance norm in general, it could be considered specifically within the particular social group (e.g., smoking doctors reporting on doctors). Indeed, since smoking behavior has a social character, it is sensitive to the identification with a particular social group and behavioral modeling [37]. Additionally, there is evidence that perceptions and attitudes towards smoking differ between healthcare professionals (doctors, nurses) and therefore it is important that future studies distinguish between these categories and also consider the strength of identification with the particular social group [38-40].

4.1 Discussing Qualitative Data

The frequency and extensiveness of comments from participants prompted the need for qualitative analysis. However, for the purpose of the present study only a brief summary of the most important issues will be provided. A recurring theme mentioned by both smokers and nonsmokers categories was "the need for stricter regulation," including specific measures taken for people who violate the law and authorized structures checking on implementation. It might seem rather controversial that smokers might ask for stricter measures but those who do, additionally say that they are trying to quit. Therefore, the quest for stricter regulation is in the sense of a helping tool for quitting, in a context where other means (e.g., quitting programs and counseling) are inexistent. Hence, formal sanctions are important not only for nonsmokers, but for smokers as well. Although studies suggest that informal sanctions (e.g., social norms against smoking) are better deterrents than formal ones, under conditions where the application of informal sanctions is unlikely, formal sanctions are necessary [41]. Studies have shown that formal enforcement of the policy is efficacious in decreasing the prevalence of smoking [42-44].

The suggestions for strictness of the law are contrasted with quests for increasing its reasonability through providing "special areas for smoking." According to this group of smokers, a total ban can never work, but the existence of special areas can at least protect nonsmokers. This pessimistic view is also realistic in the sense that quitting programs are inexistent and smokers do not receive any kind of help. In this context it has been suggested that the responsibility is upon the Ministry of Health and medical societies in helping to promote a smoke-free healthcare system, e.g., by providing smoking cessation programs for healthcare professionals.

Another suggestion for increasing compliance refers to the ambiguity of the regulation; Descriptive data clarify the reason for this comment, since only half of the population has reported the existence of a total ban. This issue is important to consider in the context of the study's limitations. In fact "compliance" might be defined differently depending on whether people are reporting "total ban" or "places for smoking." Apparently participants are aware of this inconsistency and suggest the need for informative/educational campaigns to raise public awareness of the law.

Willingness to become antismoking advocates was another recurring theme especially among nurses. This is a very important finding since nurses have great potential for fulfilling an effective health education functioning and results compare favorably to outcomes from programs involving general practitioners [45, 46]. Future studies might be designed in order to assess the baseline (i.e. what is happening now) and also the readiness for being involved in quitting programs [47, 48].

To conclude, apart from providing important directions for future studies, these findings have obvious practical implications for policy makers. Since the most efficient way of increasing effectiveness of public policies is that of consulting with target populations [18] the feedback provided by this study acquires major importance.

5. Conclusions

This was the first study aiming to assess the implementation of antismoking regulation in Albania. Several limitations must be acknowledged including generalization of findings (only two hospitals were considered), social desirability effects (self-reports) and measures used (the questionnaire is not standardized in Albania). Despite these limitations the present findings are useful in suggesting directions for future research. By combining findings from both quantitative and qualitative data several conclusions can be made. First of all it could be suggested that focusing on individual factors like attitudes and beliefs (e.g., health consequences, health professional role) is probably not the best way to understand compliance with antismoking regulation in Albanian healthcare. In fact, according to qualitative data, external factors like formal deterrents are crucial in improving compliance, ensuring a safe and healthy environment for nonsmokers and helping smokers quit. Therefore, the appeal goes to policy makers and controlling structures to enforce the implementation of the law and make 'compliance to regulation' the norm rather than the exception. Furthermore, the stricter implementation of the law should be supplemented by other important moves like general antismoking campaigns and especially quitting programs. Recognizing the crucial role of health professionals in smoking prevention and cessation future studies could be conducted to assess their readiness to be involved as antismoking advocates.

References

1 Bouchecher P. Rendez Vous with Sarah Bogdani, youth advocate for a tobacco free Albania. http://www.tobacco.org/News/rendezvous/bogdani.html April, 27, 2008.

2 World Health Organization. The European report on tobacco control policy: Review of the implementation of the third action plan for a tobacco-free Europe 1997-2001. Copenhagen: WHO Regional Office for Europe; 2002.

3 World Health Organization. Updated status of the WHO Framework Convention on Tobacco Control http://www.who.int/tobacco/framework/countrylist/en/index.html April, 2008.

4 World Health Organization. WHO Framework Convention on Tobacco Control. Switzerland: WHO Publications; 2006.

5 Zanetti F, Gambi A, Bergamaschi A, Gentilini F, De Luca G, Monti C. et al. Smoking habits, exposure to passive smoking and attitudes to a non-smoking policy among hospital staff. Public Health 1998; 112: 57-62.

6 Awad M A, Loughlin J O. Physician delivery of smoking prevention counseling to young patients in the United Arab Emirates. Patient Educ Couns 2007; 67:151-156.

7 Klien J D, Levine L J, Allan M J. Delivery of smoking prevention and cessation to adolescents. Arch. Ped. Adolesc. Med 2001; 155: 597-609.

8 Vokes N I, Bailey J M, Rhodes K. "Should I give you my smoking lecture now or later?" Characterizing emergency physician smoking discussions and cessation counseling. Ann Emerg Med 2006; 48: 406-414.

9 World Health Organization. WHO urges health professionals to engage in tobacco control. http://www.who.int/mediacentre/news/releases/2005/pr22/en/index.html April, 2008.
10 Gunes G, Karaoglu L, Genc M F, Pehlivan E, Egri M. University hospital physicians' attitudes and practices for smoking cessation counseling in Malatya, Turkey. Patient Educ Couns 2005; 56: 147-153.
11 McKenna H, Slater P, McCance T, Bunting B, Spiers A, McElwee G. Qualified nurses' smoking prevalence: their reasons for smoking and desire to quit. J Adv Nurs 2001; 35: 769-775.
12 Dawley H H, Carrol S F, Morrison J E. The discouragement of smoking in a hospital setting: The importance of modeled behavior. Int J Adict 1981: 16: 905-910.
13 Davis R M. When doctors smoke. Tob Control 1993; 2: 187-88.
14 Nelson D E, Giovani G A, Emont S L Trends in cigarette smoking among US physicians and nurses. J Am Med Assoc 1994: 271: 1273-1275.
15 Lotrean L M, Ionut C, De Vries H. Tobacco use among Romanian youth. Salud Publica Mex 2006; 48: 107-112.
16 Thun M J, Silva V L. Introduction and overview of global tobacco surveillance. In Shafey O, Dolwick S, Guindon G E, editors. Tobacco control country profiles. New York: The American Cancer Society; 2003. p. 7-12.
17 Jiang Y, Ong M K, Tong E K, Yang Y, Nan Y, Gan Q. et al. Chinese physicians and their smoking knowledge, attitudes and practices. Am J Prev Med 2007; 33: 15-22.
18 Organization for Economic Co-operation and Development. Reducing the risk of policy failure: Challenges for regulatory compliance. UK: OECD Publishings; 2000.
19 Poland B D, Cohen J E, Ashley M J, Adlaf E, Ferrence R, Pedersen L et al. Heterogeneity among smokers and non-smokers in attitudes and behavior regarding smoking and smoking restrictions. Tob Control 2000; 9: 364-371.
20 Samuels N. Smoking among doctors in Israel and their attitudes regarding anti-smoking legislation. Public Health 1997; 111: 285-288.
21 Brenner H, Born J, Novak P. Smoking behavior and attitude towards smoking regulations and passive smoking in the workplace. Am J Prev Med 1997; 26: 138-143.
22 Hodgetts G, Broers T, Godwin M. Smoking behavior, knowledge and attitudes among family medicine physicians and nurses in Bosnia and Herzegovina. Fam Pract 2004; 5: 27-39.
23 Battle E, Boixet M, Agudo A, Almirall J, Salvador T. Tobacco prevention in hospitals: long-term follow-up of smoking control programme. Br J Addict 1991; 86: 709-717.
24 Siques P, Brito J, Munoz C, Pasten P, Zavala P, Vergara J. Prevalence and characteristics of smoking in primary healthcare workers in Iquique, Chile. Public Health 2006; 120: 618-623.
25 Tewolde S, Ferguson B S, Benson J. Risky behavior in youth: An analysis of factors influencing smoking decisions in Canada. Subst Use Misuse 2006; 41: 467-487.
26 Parna K, Rahu K, Rahu M. Smoking habits and attitudes towards smoking among Estonian physicians. Public Health 2005; 119: 390-399.

27 Coups E J, Gaba A, Orleans T. Physician screening for multiple behavioral health risk factors. Am J Prev Med 2004; 27: 34-41.
28 Nawaz H, Adams M L, Katz D L. Physician-patient interactions regarding diet, exercise, and smoking. Prev Med 2000; 31: 652-657.
29 World Health Organization. The Global Health Professionals Survey. WHO Publishings; 2002.
30 International Tobacco Control Survey. http://www.bcg.com/publications/files/ITC_Survey_rpt__Dec03.pdf April, 2008.
31 Hambleton R K, De Jong H J Advances in translating and adapting educational and psychological tests. Lang Test 2003; 20: 127-134.
32 Ajzen I. The theory of planned behavior. Organ Behav Hum Decis Process 1991; 50: 179-211.
33 Ajzen I, Fishbein M. The influence of attitudes on behavior. In Albarracin D, Johnson B T, Zanna, P, editors. Handbook of attitude and attitude change: Basic principles. USA: Lawrence Erlbaum Associates; 2005. p. 173-221.
34 De Vries H, Backbier E, Kok G, Dijkstra M. The impact of social influences in the context of attitude, self-efficacy, intention, and previous behavior as predictors of smoking onset. J Appl Soc Psychol 1995; 25: 237–257.
35 Sacker A. Smoking habits of nurses and midwives. J Adv Nurs 1990; 15: 1341-1346.
36 Eiser J R. Social Psychology: Attitudes, cognition and social behavior. UK:Cambridge University Press; 1996.
37 Myers D G. Social psychology (9th Ed.). US: McGraw-Hill; 2008.
38 Braun B L, Fowles J B, Solberg L I, Kind E A, Lando H, Pine D. Smoking-related attitudes and clinical practices of medical personnel in Minnesota. Am J Prev Med 2004; 27: 316-322.
39 Stillman F A, Hantula D A, Swank R. Creating a smoke-free hospital: attitudes and smoking behaviors of nurses and physicians. Am J Health Promot 1994; 9: 108-114.
40 Terry D J, Hogg M A. Group norms and the attitude-behavior relationship: A role for group identification. Pers Soc Psychol Bull 1996; 22: 776-793.
41 Paternoster R, Simpson S. Sanction threats and appeals to morality: Testing a rational choice model of corporate crime. Law Soc Rev 1996; 30: 549-583.
42 Griesbach D, Inchley J, Currie C. More than words? The status and impact of smoking policies in Scottish schools. Health Promot Int 2002; 17: 31-41.
43 Moore L, Roberts C, Tudor-Smith C. School smoking policies and smoking prevalence among adolescents: multilevel analysis of cross-sectional data from Wales. Tob Control 2001; 10: 117-123.
44 Wakefield M A, Chaloupka F J, Kaufman N J, Orleans C T, Barker D C, Ruel E E. Effect of restrictions on smoking at home, at school and in public places on teenage smoking: Cross-sectional study. Br Med J 2000; 321: 333-337.
45 Clark J M, Haverty S, Kendall S. Helping people to stop smoking: A study of the nurse's role. J Adv Nurs 1990; 16: 357-363.
46 Katz D A, Brown R, Muehlenbruch R N, Fiore M C, Baker T B Implementing guidelines for smoking cessation: Comparing the efforts of nurses and medical assistants. Am J Prev Med, 2004; 27: 411-416.
47 Ferketich A K, Khan Y, Wewers M E. Are physicians asking about tobacco use and assisting with cessation? Results from the 2001-2004 national ambulatory medical care survey (NAMCS). Prev Med 2006; 43: 472-476.

48 Park E R, DePue J D, Goldstein M G, Niaura R, Harlow L, Willey C, et al. Assessing the Transtheoretical Model of Change constructs for physicians counseling smokers. Ann Behav Med 2003; 25: 120-126.

Sound assessment of arteriovenous fistula in hemodialysis patients

Pavlos Malindretos[1,2], Christos Frantzidis[1], Panagiotis Bamidis[1], Pavlos Nikolaidis[2]

[1] *: Laboratory of Medical Informatics, Aristotle University of Thessaloniki*

[2]*: Renal Unit, 1st Medical Department, AHEPA University Hospital*

Purpose: Arteriovenous fistula (AVF) represents a permanent vascular access in end stage renal disease patients (ESRD). The purpose of our study was to distinguish functioning from nonfunctioning fistulas based on the produced bruit from blood circulation, with the assist of a computerized method.

Design/methodology/approach: An electronic stethoscope connected to a Lap top computer was used. Forty hemodialysis patients participated in the study. Twenty fistulas were functioning, while the rest were not. The bruit was recorded, sample rate was 48,000. Fast Fourier Transformation (FFT) was used, Frequencies obtained were divided in 40 frequency intervals, 250Hz apart. These 20 components served as inputs to a feed-forward neural network. The network architecture consists of an input layer, a hidden layer and a single output neuron which provided a unique result regarding the functioning status of each fistula

Findings: Mean amplitudes of the frequency bands obtained with FFT algorithm were significantly higher in the functioning ($p<0.00001$). The neural network classified successfully all the fistulas used for evaluation ($p<0.0001$).

Originality/value: This is the first attempt to discriminate functioning from non-functioning fistula with the use of an intelligent system based on neural networks.

Keywords
Fistula, Fast Fourier Transformation, FFT, hemodialysis, Neural Networks

1. Purpose

Chronic kidney disease (CKD) represents a major public health problem worldwide. The concern is even greater since, CKD is more frequent in the elderly. The median prevalence in persons aged 30 years or older is estimated to be 7,2%, while the prevalence of CKD in persons older than 64 years reaches 35,8% (5 times more frequent).[1] The mean incidence of patients suffering from end stage renal disease (ESRD) in Europe rose from 79 to 117 new patients per million population per year the last decade,[2] unfortunately even greater proportion is seen in some European countries.[3] Worldwide data regarding CKD are comparable with those met in European Union. By the end of the year 2001, 1,5 million people were receiving hemodialysis treatment due to ESRD, showing a 7% increase compared with year-end 2000, and those receiving peritoneal dialysis showed a 5% increase. Additionally, the majority of patients suffering from ESRD are being treated with hemodialysis (69%), followed by organ transplantation (23%) and peritoneal dialysis(8%).[4] The observed trend seems to be rather constant, likewise, the increase ranged between 6 and 7% compared with year-end 2003 regarding the total numbers of ESRD patients receiving hemodialysis and organ transplant.[5]

Patients undergoing hemodialysis are in need of a vascular access able to provide adequate blood supply presenting at the same time an acceptable number of adverse ivents. Up to present day arteriovenous fistula (AVF) represents the permanent vascular access of choice in ESRD. AVF is formed surgically, by performing an anastomosis between an artery to an adjacent native vein. Most frequently preferred locations are the wrist, the forearm, the elbow and the upper arm area.[6;7] An arteriovenous fistula is considered adequate for hemodialysis treatment (mature) when it has a measured flow of 600ml/min or greater, diameter greater than 0.6 cm and is approximately 0.6 cm from skin surface.[8] It is estimated that up to 30-60% of the performed fistula will eventually fail. [9] Even though AVF represents the preferred form of permanent hemodialysis access, it is characterized by both early and late failure problems, leading in a necessity for monitoring and surveillance of fistula functioning.[10-13] Failure of an AVF to develop adequately or failure within 3 months from dialysis initiation is considered early failure.[14] The associated histological lesion is represented by neointimal hyperplasia in both early and late failure, while early failure is also characterized by adverse vascular remodeling represented by vasoconstriction or inability to dilate adequately.[15] The traditional methods of AVF function assessment are represented by clinical evaluation,[16] Doppler ultrasound,[13;17] Magnetic resonance flow measurement,[18] Computed Tomography,[19;20] Ultrasound dilution techniques,[21-25] Venography[26] and dialysis adequacy methods (URR and Kt/V).[8;25] Obviously, the easiest method of estimating fistula functioning is clinical examination, which is based on assessing a thrill and listening to the associated bruit at the site of arteriovenous anastomosis. Even though this represents the easiest, at the same time it is a subjective assessment, dependent on experience bias and even though its cost is minimal it is used less frequently.[27] Clinical examination is based on assessing the extent of the fistula, its diameter and depth, the associated pulse and thrill and listening to bruit at the site of arteriovenous anastomosis.

The purpose of the present study was to distinguish functioning from nonfunctioning fistulas based on the produced bruit from blood circulation, with the assist of a computerized method.

2. Design/methodology/approach

For the estimation of bruit generated from blood circulation within the AVF, an electronic stethoscope was used. Its sensitivity was 20-20,000Hz. The stethoscope was connected to a portable Lap top computer (Intel 2.00GHz CPU, 1GHz RAM). Forty hemodialysis patients participated in the study. All measurements were made with patients in resting supine position. Twenty fistulas were functioning and the patients were receiving standard hemodialysis treatment through them, while the remaining 20 were not functioning and the patients were already receiving hemodialysis through another access (other fistula or a central venous catheter). After the determination of the surface where the perceived bruit was more intense, the bruit was recorded, without using any sound filtering. For those fistulas where no bruit was heard the recording was performed at the nearest, to the nonfunctioning anastomosis, palpable pulse. The sample rate was initially decided to be 48,000. Sound files created were saved in wav type lossless format. In order to assess possible differences between functioning and nonfunctioning fistulas, Fast Fourier Transformation (FFT) was used. Fourier Transformation was preferred, since it can be particularly useful in revealing periodicity in input data, mapping its frequency behavior as well as the strength of its components. Periodicity in fistula bruit is reasonably suspected since the bruit is generated from blood flow originated from the heart and thus it follows the pulse frequency of the patient. Frequencies obtained with FFT were divided in 40 frequency intervals, 250Hz apart. This way separate sound frequencies generated by the heart and the anastomosis bruit could be easier revealed. Fast Fourier Transformation has been also used by other investigators in analyzing and differentiating breath sounds.[28] Statistical analysis followed Fourier Transformation. Descriptive statistics were used for patient demographic and clinical data. The x-square method was used to assess any differences in comorbid conditions between groups, while Man Whitney U test was used to estimate possible differences in sound frequencies between the two groups.

After the statistical analysis we attempted to classify the two groups of patients according to the amplitude of the mean frequency bands obtained as described above. We chose the frequency bands between 1 KHz and 6 KHz. These 20 components served as inputs to a feed-forward neural network. The network architecture consists of an input layer, a hidden layer and a single output neuron which provided a unique result regarding the functioning status of each fistula (close to zero for the non functioning and close to one for the functioning ones). The transfer function for each group was log-sigmoid, whereas as training function was chosen the Batch Gradient Descent (traingd). A medium learning rate (0.01) and a moderate training goal (0.001) was chosen in order to ensure the good network performance and avoid network over-fitting. The 75% of each group was used for training the network, while the remaining data served as test for the network performance evaluation. Neural networks[29;30] as well as other methods[31-33] have been used earlier in various sound discrimination attempts. All mathematical analysis was performed using the Matlab environment and the Neural Network toolbox.

3. Findings

There were no differences regarding clinical data and demographic characteristics between groups. Mean amplitudes of the 40 frequency bands obtained with FFT algorithm were significantly higher in the functioning fistulas compared to the nonfunctioning group ($p<0.00001$), Figure 1.

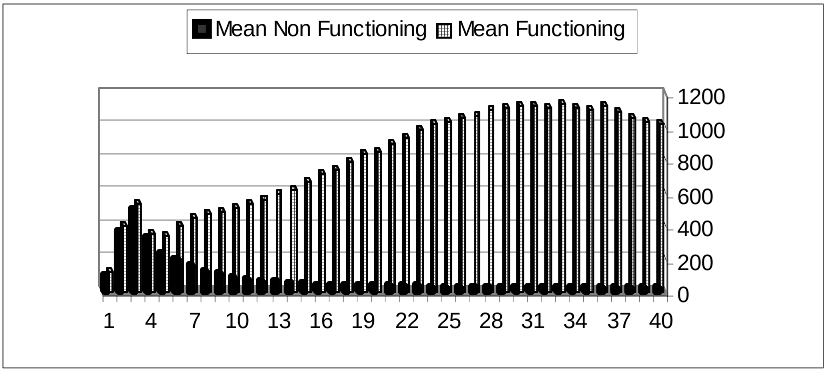


Figure 1 Comparison of the frequency bands, obtained from Fast Fourier Transformation, between non-functioning and functioning fistulas.

Additionally Fourier Transformation revealed that differences in frequencies were more pronounced in the higher frequency group. Taking into consideration that sound generated from the heart is normally below 500Hz, makes the later differences reasonable. The neural network classified successfully all the fistulas used for evaluation. The obtained results were significantly different for the two groups ($p<0.0001$), Figure 2. The output neuron values are depicted in table1.

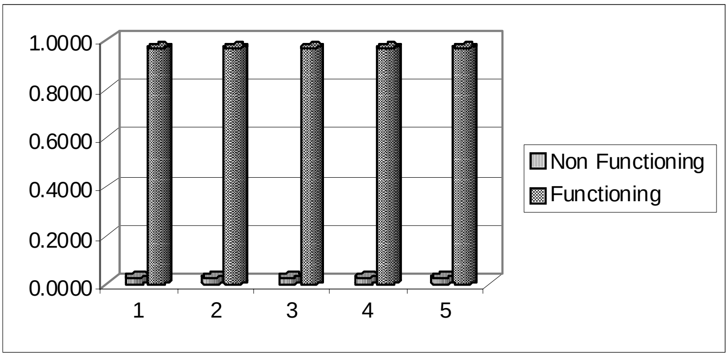


Figure 2 Comparison of results from Neural Network between non-functioning and functioning fistulas.

Functioning_1	0,965254308322810
Functioning_2	0,965254308336262
Non Functioning_1	0,028153596426492
Non Functioning_2	0,028153542929936
Functioning_3	0,965254308336161
Non Functioning_3	0,028153542058562
Functioning_4	0,965254308392872
Non Functioning_4	0,028153542012214
Non Functioning_5	0,028153541992550
Functioning_5	0,965254308336145

Table 1 Neural Network Classification

4. Research limitations/implications

Measurements were performed in a small patient sample. All sound recordings were made by the same investigator, possible bias when different persons perform measurements cannot be excluded. Nevertheless, differences were so intense that even biased results from various doctors are highly unlikely to alter significantly the observed results.

5. Practical Implications

Even though clinical estimation of a fistula's condition is part of every day practice in nephrology, computer assisted sound estimation will probably increase diagnostic accuracy, without adding too much to the expenses. Other methods of assessment probably are more accurate but sound assessment is much easier, as well as much cheaper.

6. Originality/value

To our knowledge this is the first attempt to use computer assisted sound assessment for the discrimination between functioning or non-functioning of arteriovenous fistula with the use of an intelligent system based on neural networks.

References

1 Zhang QL, Rothenbacher D: Prevalence of chronic kidney disease in population-based studies: systematic review. BMC Public Health 2008;8:117.

2 Stengel B, Billon S, Van Dijk PC, Jager KJ, Dekker FW, Simpson K, Briggs JD: Trends in the incidence of renal replacement therapy for end-stage renal disease in Europe, 1990-1999. Nephrol Dial Transplant 2003;18:1824-1833.

3 Couchoud C, Stengel B, Landais P, Aldigier JC, de CF, Dabot C, Maheut H, Joyeux V, Kessler M, Labeeuw M, Isnard H, Jacquelinet C: The renal epidemiology and information network (REIN): a new registry for end-stage renal disease in France. Nephrol Dial Transplant 2006;21:411-418.

4 Moeller S, Gioberge S, Brown G: ESRD patients in 2001: global overview of patients, treatment modalities and development trends. Nephrol Dial Transplant 2002;17:2071-2076.

5 Grassmann A, Gioberge S, Moeller S, Brown G: ESRD patients in 2004: global overview of patient numbers, treatment modalities and associated trends. Nephrol Dial Transplant 2005;20:2587-2593.

6 Brescia MJ, Cimino JE, Appel K, Hurwich BJ: Chronic hemodialysis using venipuncture and a surgically created arteriovenous fistula. N Engl J Med 1966;275:1089-1092.

7 Shenoy S: Innovative surgical approaches to maximize arteriovenous fistula creation. Semin Vasc Surg 2007;20:141-147.

8 National Kidney Foundation: DOQI clinical practice guidelines for vascular access: update 2006. Am J Kidney Dis 2006;S177-S247.

9 Barone GW, Wright CF, Krause MW, Brosnahan GM, Portilla D, Banerjee S, McCowan TC, Culp WC, Yousaf M: Hemodialysis access success: beyond the operating room. Am J Surg 2007;194:668-671.

10 Besarab A, Asif A, Roy-Chaudhury P, Spergel LM, Ravani P: The native arteriovenous fistula in 2007. Surveillance and monitoring. J Nephrol 2007;20:656-667.

11 Patel ST, Hughes J, Mills JL, Sr.: Failure of arteriovenous fistula maturation: an unintended consequence of exceeding dialysis outcome quality Initiative guidelines for hemodialysis access. J Vasc Surg 2003;38:439-445.

12 Dixon BS, Novak L, Fangman J: Hemodialysis vascular access survival: upper-arm native arteriovenous fistula. Am J Kidney Dis 2002;39:92-101.

13 Pietura R, Janczarek M, Zaluska W, Szymanska A, Janicka L, Skublewska-Bednarek A, Szczerbo-Trojanowska M: Colour Doppler ultrasound assessment of well-functioning mature arteriovenous fistulas for haemodialysis access. Eur J Radiol 2005;55:113-119.

14 Beathard GA, Arnold P, Jackson J, Litchfield T: Aggressive treatment of early fistula failure. Kidney Int 2003;64:1487-1494.

15 Roy-Chaudhury P, Spergel LM, Besarab A, Asif A, Ravani P: Biology of arteriovenous fistula failure. J Nephrol 2007;20:150-163.

16 Beathard GA: An algorithm for the physical examination of early fistula failure. Semin Dial 2005;18:331-335.

17 Besarab A, Sullivan KL, Ross RP, Moritz MJ: Utility of intra-access pressure monitoring in detecting and correcting venous outlet stenoses prior to thrombosis. Kidney Int 1995;47:1364-1373.

18 Doelman C, Duijm LE, Liem YS, Froger CL, Tielbeek AV, Donkers-van Rossum AB, Cuypers PW, Douwes-Draaijer P, Buth J, van den Bosch HC: Stenosis detection in failing hemodialysis access fistulas and grafts: comparison of color Doppler ultrasonography, contrast-enhanced magnetic resonance angiography, and digital subtraction angiography. J Vasc Surg 2005;42:739-746.

19 Karadeli E, Tarhan NC, Ulu EM, Tutar NU, Basaran O, Coskun M, Niron EA: Evaluation of failing hemodialysis fistulas with multidetector CT angiography: Comparison of different 3D planes. Eur J Radiol 2007.

20 Neyman EG, Johnson PT, Fishman EK: Hemodialysis fistula occlusion: demonstration with 64-slice CT angiography. J Comput Assist Tomogr 2006;30:157-159.

21 Tonelli M, Klarenbach S, Jindal K, Harries S, Zuidema S, Caldwell S, Pannu N: Access flow in arteriovenous accesses by optodilutional and ultrasound dilution methods. Am J Kidney Dis 2005;46:933-937.

22 McCarley P, Wingard RL, Shyr Y, Pettus W, Hakim RM, Ikizler TA: Vascular access blood flow monitoring reduces access morbidity and costs. Kidney Int 2001;60:1164-1172.

23 Tessitore N, Mansueto G, Bedogna V, Lipari G, Poli A, Gammaro L, Baggio E, Morana G, Loschiavo C, Laudon A, Oldrizzi L, Maschio G: A prospective controlled trial on effect of percutaneous transluminal angioplasty on functioning arteriovenous fistulae survival. J Am Soc Nephrol 2003;14:1623-1627.

24 Tonelli M, Hirsch D, Clark TW, Wile C, Mossop P, Marryatt J, Jindal K: Access flow monitoring of patients with native vessel arteriovenous fistulae and previous angioplasty. J Am Soc Nephrol 2002;13:2969-2973.

25 Roca-Tey R, Samon R, Ibrik O, Martines-Cercos R, Viladoms J: Functional vascular access evaluation after elective intervention for stenosis. J Vasc Access 2006;7:29-34.

26 Brown PW: Preoperative radiological assessment for vascular access. Eur J Vasc Endovasc Surg 2006;31:64-69.

27 D'Cunha PT, Besarab A: Vascular access for hemodialysis: 2004 and beyond. Curr Opin Nephrol Hypertens 2004;13:623-629.

28 Elphick HE, Ritson S, Rodgers H, Everard ML: When a "wheeze" is not a wheeze: acoustic analysis of breath sounds in infants. Eur Respir J 2000;16:593-597.

29 Crovato CD, Schuck A: The use of wavelet packet transform and artificial neural networks in analysis and classification of dysphonic voices. IEEE Trans Biomed Eng 2007;54:1898-1900.

30 DeGroff CG, Bhatikar S, Hertzberg J, Shandas R, Valdes-Cruz L, Mahajan RL: Artificial neural network-based method of screening heart murmurs in children. Circulation 2001;103:2711-2716.

31 Hadjileontiadis LJ: Discrimination analysis of discontinuous breath sounds using higher-order crossings. Med Biol Eng Comput 2003;41:445-455.

32 Hadjileontiadis LJ: Wavelet-based enhancement of lung and bowel sounds using fractal dimension thresholding--Part I: methodology. IEEE Trans Biomed Eng 2005;52:1143-1148.

33 Hadjileontiadis LJ: Wavelet-based enhancement of lung and bowel sounds using fractal dimension thresholding--Part II: application results. IEEE Trans Biomed Eng 2005;52:1050-1064.

Cooperation among Students in the Dormitory: How the Experience of Reciprocal Altruism Defines the Ones You Share a Room with

Peter Hlohinec[1], *Martin Kanovsky*[2]

Commenius University, Faculty of Social and Economic Sciences, Odbojarov 10/A, P. O. Box 129, 820 05 Bratislava 25, Slovakia, {hlohinec[1]*, kanovsky*[2]*}@fses.uniba.sk*

The possibility of evolution of cooperation is considered to be a puzzle. Various evolutionary approaches were proposed to solve this puzzle assuming kin selection, direct reciprocity or indirect reciprocity as crucial elements in cooperation. Recently, prominent models of human cooperation suggested reputation as mechanism for sustaining cooperation. In this paper I apply two of reputation based models to explain cooperative behaviour among students on dormitories. I use data obtained from ethnographic fieldwork performed in Bratislava, Slovakia. Results describe process through which certain students make decision about potential partners. Thus, I provide an explanation of behaviour which leads into cooperation on the basis of observing reputation of potential partners. Two main crucial problems are being discussed in respect to theory. The first one is how previous experience influences the decision making for potential partners. The second problem is under what conditions an individual selects partners for cooperation from a broader group.

Keywords
evolution of cooperation, indirect reciprocity, reputation based choice.

1. Introduction

The existence of cooperation in human societies is one of the essential issues of social sciences. Many historical and current achievements of human kind are by all means the outcome of ability of humans to generate cooperative relationships for the benefit of both, individual as well as groups. Cooperation also exists on the level of other species but none of them attains the complexity characteristic for man. From the perspective of evolutionary theory, any behavior demonstrating traits of cooperation and altruism is problematic. The target of natural selection should be that an individual behaves with respect to his own fitness. Evolution is characterized by competition and not by cooperation. Therefore, the puzzle of human cooperation arises. Various different approaches have been elaborated to answer this puzzle. Among them, approach of kin-selection [1, 2], direct reciprocity [3] or indirect reciprocity [4] attempt to provide the answer to individuals' expensive cost.

Recently, prominent models of evolution of cooperation were proposed assuming indirect reciprocity in its center [5-8]. On the basis of indirect reciprocity there must exist a mechanisms which prevents the exploitation of the system. One of the mechanisms is the reputation. How an individual is perceived in the society and how he perceives others can be a way by which cooperation is established and sustained.

In this paper I apply models of establishing cooperative networks [7] based on observing reputation [8] for explicating cooperation among students in dormitories. I draw attention on the process through which an individual student decides with whom he will share a common room. Two main crucial problems are being discussed in respect to theory. The first one is how previous experience influences the decision making for potential partners [7]. The second problem is under what conditions an individual selects partners for cooperation from a broader group [8]. A long-term fieldwork made it possible to attain the necessary data. My aim is to interlink the existing evolutionary theory on cooperation with observed and recorded relations in the real world setting.

The paper is divided into four parts. The part following the introduction provides a brief overview of the problem of cooperation. The fieldwork process and data described in the second part which is followed by a section presenting the overall argument of this paper. Lastly, the conclusion summarizes my line of reasoning.

2.

Favoritism of others rather than of the individual's own reproductive abilities is nothing surprising. The essence of the problem is the existence of a sacrifice, in the form of giving up own reproductive abilities in favor of other organism. Indeed, this same challenge confronted also Darwin in the *Origin of species* [9]. This behavior is in contrast to the basics of the entire evolutionary theory, which is constructed on heritability and reproduction.

Evolutionary theory presumes an intensive and stark competition among organisms and only those individuals should survive, whose pursuance is selfish [10]. The existence of cooperation based on altruism is hardly compatible with this presumption. Therefore, does it mean, natural selection prefers individuals who directly enhance by their behavior others' fitness? Traditionally, it is referred to this problem as the puzzle of human cooperation.

At this point, the usage of the expression - puzzle is legitimate as the history of evolutionary approach indicates. In other words, the history of evolutionary approach towards cooperation is the history of the attempt to reform the puzzle into a problem subsequently to provide its solution. The most widely spread technique how to reformulate the confusing puzzle of cooperation into a problem is to assume that cooperative behavior of an individual is not the product of evolutionary attained altruism. Indeed, the argument behind this thinking is based on the image of the selfish individual, which acts only in respect to his own interest. Thus, subsequent explanation seeks those reasons for pursuance, which would explain altruism not as a selfless act, but as an act producing benefits.

William Hamilton in his work *Genetic evolution of social behavior* [1] proposes an evolutionary explanation of altruism. According to him, altruism is the product of behavior preferring genetic relatives. This concept is called the *theory of kinship* and in its center is the postulate of Hamilton's rule. The idea behind the concept is that natural selection will prefer cooperation if the donor and beneficiary of altruistic acts are genetically related. This relatedness (r) is higher than the rate of costs of the donor(c) and benefits of the beneficiary (b):

$$r > \frac{c}{b} \quad \textbf{[1]}$$

Deriving from this rule, the higher is the proportion of genes shared by two organisms the more frequent is the degree of altruistic acts higher in rate and aggression lower in proportion. Later on, Dawkins [2] with his thesis on the selfish gene and Hamilton [1] managed to explain that the level on which evolutionary selection takes place is the level of genetic. According to Hamilton, not the individual fitness of each organism (which is the object of natural selection) has to be considered, but the inclusive fitness of genes.

In fact, the theory of kinship has to face to problems because it does not explain altruism in case when interaction occurs among individual organisms without shared genes. A question arises, how to explain cooperation among individuals, whose relatedness is low or zero [11]. On that account, a theory providing explanation on only cooperation among related organisms is not sufficient in order to comprehend the complexity of human cooperation.

The deficiency of the theory of kinship led Robert Trivers to claim that saving one's own drowning child is not at all an example of altruism, but simply a contribution to the survival of one's own genes invested in the child [3]. Trivers in this famous article *Evolution of reciprocal altruism* proposed the model of reciprocal altruism for the explanation of cooperation. This model works on the basis of presumed interactions between two individuals, when it arrives to mutual altruistic behavior. Richard Dawkins calls this relation a relation of symbiosis [2].

The exchange of costs of altruism directly benefits individuals. The same donor receives them back from the same beneficiary. The possibilities of this type of relation can be very well formalized by the prisoner's dilemma. Consequently, the prisoner's dilemma has become a standard method in evolutionary biology, economics and international relations while examining the possibility of the evolution of cooperation between two integrated agents [12].

The game of prisoner's dilemma has the following scenario: Each of the two players has at his disposal the choice of two strategies. Thus, there are four possible results of the dilemma. If they both cooperate the benefits are reciprocal. If none of them cooperates, then both face the costs. These benefits and costs are relatively low in comparison to the two other possibilities. If only one player decides to cooperate then only he carries the costs and the other player who stayed out of cooperation receives the net benefit, and this applies also vice versa. Costs and benefits in this case are quite considerable. The ways of solving the game are also labeled as reward (R), if both cooperate. Punishment (P) is applied in case if nobody cooperates. Temptation not to cooperate (T) is labeled, when one player does not cooperate and his opponent acts the opposite, and naivety of the sucker (S) is in the opposite case. The player can just assume which strategy his opponent is going to choose. From the above described, it derives that the most advantageous strategy for the player is to select the not to cooperate choice. In other words, if his opponent cooperates the player receives the net benefit or a comparatively small punishment if the opponent does not cooperate. Applying this strategy the player avoids high costs which would be in case of using the cooperative strategy while his opponent would stay out of cooperation.

However, from the long run the most advantageous strategy for both players is when both of them cooperate. For the emergence of cooperation, the repeated prisoner's dilemma is used when the same players are playing the same game. They have at their disposal the information of previous decisions of the opponents and based on this information they decide on their strategies. Moreover, the repeated prisoner's dilemma is the tool for examining strategies of individuals, which are in contact over a longer period of time.

In the aftermath, when Axelrod and Hamilton [13] formulated Trivers's assumption mathematically on the basis of prisoner's dilemma, a series of questions have emerged concerning the hypothesis of reciprocal altruism. One of the questions is also which strategy is the most suitable in the repeated interaction of two individuals, expressed by the prisoner's dilemma. It is a strategy, in the sense of attempting to find equilibrium in the game, when individuals will not change their pursuance. In the case of evolution of cooperation, it means to find an evolutionary stable strategy:

An ESS *(Evolutionary Stable Strategy)* is a strategy such that, if all members of a population adopt it, then no mutant strategy could invade the population under the influence of natural selection (p: 12 1[14]).

Without undertaking an analysis of various proposals for evolutionary stable strategies, I attempt to summarize an objection against these efforts. In the core there is the objection stating that reciprocal altruism focuses only on the repeated interactions of two identical individuals and does not refer to the complexity of human cooperation expressed for example by reputation or communication [5].

However, the most serious argument against direct reciprocity is that human societies are characterized by uneven relations, which do not circumscribe only to interactions among identical individuals. It often happens that return of altruistic costs does not take place directly but via somebody else. Therefore, Alexander [4] argues in favor of indirect reciprocity. Cooperation on the basis of indirect reciprocity assumes that individual will look for potential cues or signs of status or reputation. Nowak stands that indirect reciprocity can promote cooperation, only if recognize beneficiary's reputation [10].

In order that indirect reciprocity could function as the base of cooperation, in the first place it has to be assured that those who could profit from this system – free riders – have to be eliminated. Furthermore, the hypothesis of cooperation solves this as a result of officiating reputation [5, 6, 7, and 8]. Since this theory is part of my argument, at the moment I will leave aside its introduction and will return to it after the introduction of my fieldwork.

3.

In this part, the summary of fieldwork is provided via the description of the place, time period, survey questions and character of the acquired material.

The fieldwork took place in two phases, firstly from August till September 2007 and subsequently from November till December 2007. I have conducted the research in student dormitories on the grounds of *Mlynská dolina* in *Bratislava*. The area of the student residence consists of four main parts; two high-rise buildings make up the *Šturák,* then there are two sections consisting of lower panel-blocs buildings *Manželské internáty* and *Mladosť* and as well a complex of connected buildings forming atrium spaces called *Átriové domčeky.*

All participants in my research were accommodated in the complex *Átriové domčeky* and they were living in rooms available for three persons. The rooms were located in two parallel rows or halls between which the common sanitary facilities and bathrooms were placed one for each row or hall. The two rows of three beds rooms were connected with 2 shorter halls. There were 4 one person bedrooms on one of the short halls and the second short hall led to the staircase. From the outer side, every floor was surrounded by an interconnected balcony to which every 3 person bedroom had access. Thus, *Átriové domčeky* were not just connected internally but also from the outside by the balcony. The halls of the dormitories were also utilized as a space for social events well-known as *chodbovice* (hall parties), when students sharing the same room and the same hall organized small gatherings where admittance was free and open for all other inhabitants. Similar functions were attributed to the terraces on top of the highest floors of *Átriové domčeky*.

Each accommodated was assigned to his room on basis of the affiliation to certain faculty of Comenius University. In this way, the students of the same faculty utilized together the common spaces. However, the system envisaged the possibility of change and selection of rooms based on the preferences of individual students. Furthermore, the system of accommodation of students allowed the exchange of rooms for other rooms within the floor, buildings or the overall complex of all student houses.

The number of participants in the fieldwork was 33 students. More precisely, 18 female and 15 male students lived in 3 *átriové domčeky* buildings, on 5 different floors (or halls) and in 11 rooms. The rooms were occupied with students from either the same faculties, or from mixed faculties. None of the participants were first year students on the university.

Firstly, I have approached the actual participants via *a gatekeeper,* who was my former colleague during my undergraduate studies. Via the gatekeeper I have accessed to his environment of social network and he has also participated in my fieldwork as a respondent. By this way I got to half of the actual participants. Consequently, I had used their social contacts and this way I filled in the actual sample of student.

From the nature of this paper, it derives that the working questions used in the fieldwork concerned the manifestation of cooperation among students, strategies of cooperation, which they followed and as well as relations, which were not characterized by cooperation. Into the field I have entered with the assumption of a certain type of reciprocal behavior and on that based cooperation. The fieldwork was meant to verify this assumption.

As a method for collecting data, I have decided to use participant observation and an indirect informal interview. The gatekeeper has introduced me as a student working on the *project of students living in dormitories*. I have experienced a high degree of understanding coming from the respondents. In addition, I have attributed this eagerness to the fact that students are often used by their colleagues for their goals in their researches, though mostly quantitative in character.

The actual material which I have recorded relates to the relations between individuals within their room and relations, which they commonly share but these extend behind the borders of their accommodation. Relations extending behind the borders of the room concern the individual's social network. Social life of a student in dormitories is not just reduced to interactions with his roommates, but often involves interactions with students from the neighboring room(s) or from other parts of the residence complex, relatives, classmate, etc.

However the character of my research focused on the dormitories allowed me to concentrate on the individual's social networks within the framework of the dormitory complex. Accordingly, I have recorded relations outside the rooms but not outside the dormitory complex. Thus, under the notion of *own social network of an individual* I propose to call such a network, which refers to people, who are included in it by the individual himself. Correspondently, in the case of my participants, *social network of an individual* is limited to the area of his residence. Therefore, I address to such social networks as *dormitory social networks of an individual.*

The intention alone to examine the relations of cooperation emerged from the presumption of examining the ways of supplying and sharing objects of material consumption or services. I have begun to note down relations regarding this type of cooperative behavior. On the basis of the acquired information I shifted my attention to the manifestation of cooperation in another spheres. The primary interest was to record such types of relations, which would reflect the contributions and costs of individuals in reference to sharing a common room. In fact, this has made possible to shift towards the analysis of relations and material objects.

The conception of the individual representation was the core center of my interest and approach, and besides I have focused on the manner how individuals evaluate their position within the room and their social network. In addition, the above described has become the key element in my research. Subsequently, after the individual's conception of the relative costs and benefits expressed via his/her description of the situation I have confronted these with the conception of individuals within his/her room and social network.

4.

Individuals have the tendency to change their own ‘selfish’ behavior in favor of altruism on the basis of expected appraisal coming from others [15]. Experiments have revealed this tendency is most present when individuals’ behavior is observed [16]. Indeed, this supports the rise of a presumption that an individual behaves cooperative with respect to reputation [4]. An example showing that it depends on reputation is the usage of oblique phrasing in cases involving both cooperation and conflict [17].

An explanation taking into consideration reputation argues that cooperative individuals will relate to other cooperative individuals based on information on their own previous behavior [7]. Strategy for establishing cooperative networks proposed by Mari Rege assumes that cooperative individuals will spread such information, which concern experience with cooperation. Basically, those individual who will have the reputation of not cooperating will be excluded from the networks built on information about cooperative individuals. Mari Rege [7] explains why an individual decides to cooperate in situations of prisoner’s dilemma, even though he knows he will not play against his opponent in the future: *if a co-operator A gets information about another co-operator B in period t-1, then A will search for B and try pre-match with him in period t. If A succeeds in pre-matching B, then A and B will play the prisoners’ dilemma against each other in period t. A defector A will never succeed in such pre-matching because a cooperator B will only pre-match with A if A can send a signal by greeting from a cooperator against whom B has played in a previous period. Since a cooperator will never tell a defector about cooperators, a defector will never be able to send such signal (p: 2 [7]).*

Bergh and Engseld introduce into the model the noisy observation of reputation and seek such conditions of stable cooperative population, which would achieve lower benefits than current ones in case of being invaded by less cooperative strategies [8]. Their model exists on few assumptions. For them, the essentials to form cooperative networks is the selection of partners with who an individual gets into contact. Individuals within the society decide by themselves with whom they will or will not enter into relations of possible cooperation. Moreover, in this way their model is drawing closer to relations in real life. Furthermore, they distinguish propensity and reputation: *By propensity we mean a measure of the actual probability that an agent with a certain strategy will cooperate in a given population. By reputation we mean a noisy observation of an agent’s propensity (p: 7 [8]).*

According to them, reputation is the indicator of the probability that an individual will act cooperatively. Similarly this way they assume an individual perceiving only the opponent’s reputation. The important aspect of the model is the noisy observation of reputation. If reputation perfectly reflected the individual’s inclination then everybody would exactly know to differentiate between those who cooperate and those who not. If the observation is noisy then it means that the population formed purely and only by those cooperating would not be stable. Bergh and Engseld draw a conclusion that *Degree of cooperation is decreasing in the level of reputation noise (p: 12, [8]).*

Similarly as in the conclusions of the experiments [15], respondents of my research have also given a verbal evaluation of experiences from the games. During the analysis of a series of interviews with students concerning relations in the rooms I came to realize that students often formulate the evaluations as they were descriptions of repeating games of prisoner's dilemma. It means, that for example to my question concerning habits of cleaning the room, one respondent described the situation that how he has a subjective feeling since he contributes more than the others. In addition, this exact respondent further compared the contribution to cleaning the room of his two other roommates. While conducting further interviews, in most cases I observed the prevailing tendency to utter relations in the room and consequently evaluation of students sharing the same room as in the above described case. Furthermore, the recorded interviews I have compared in cases regarding students sharing a common room or belonging to a mutual social network.

This technique to classify data, when I regarded the individual's evaluation as the criteria of cooperation made it possible to avoid hasty conclusion on the character of cooperation. Forby, I have verified the recorded declarations by participating observation in the rooms of the individuals. For instance, study of the content of refrigerators (component of every room) was also part of the survey. By this way, I have attempted to verify the correlation between cooperation declared by students and the manner of sharing the content of refrigerators.

In cases, when among the students were not evident discrepancies in how they viewed their mutual relations in the room, I discovered that they have a greater tendency to share items from the refrigerators. On the contrary, there were also cases when students criticized how a roommate or the roommates behave. In such instances, the fridge was divided into sections each for every accommodated student in the room. Similar example was the utilization of other items, such is the toilette paper. In the first case, when there were not any conflicts between the roommates the toilette paper was purchased collectively and was part of the common space, but in the second case the accommodated students bought the toilette paper individually and it was located in their private closet. An interesting case was of three students, from which two were roommates for already two years. The third student happened to be accommodated in the room, after the departure of the previous roommate as well as friend of the other two. In this room, the refrigerator was divided into a common part and a part which belonged to the newcomer, and also the toilette paper was a content of the common space as the two original inhabitants have declared it, but the newcomer referred to it as 'their' space.

By the above described examples I wanted to point out two crucial issues in the recorded material. One of the issues is the verbal evaluation of the situation in the room and the second is the social network of each individual. In comparison to observation I regard the verbal evaluation as a method, which allowed me to classify relations of students with respect to cooperation. However, when speaking about social network then it is rather like pondering about a paradox which calls for explanation. On the basis of the data analysis I claim that potential partners for cooperation are defined by the social networks of the individual student. In fact the paradox is that some students, which for example were not content with their roommate or roommates they did not included these persons into their network. Nevertheless, they share a common room, meet everyday, sometimes even during the weekends, but they still do not consider them as members of their own social network.

For the own inhabitant of the room it holds, that only individuals belong to it who share the dormitory, thus it would be suitable a determinant ‘dormitory‘. But why are there cases when from it members of the own room are excluded? To this conclusion I have arrived following an extended tendency not just among my respondents to speak about ‘roommates’ or ‘roommate’ instead of calling them by their names as it is common practice. In addition, this was a common usage even though these students lived together day by day. This phenomenon as well as the items in the refrigerator was in correlation with how the particular students evaluated their own relations. In the earlier mentioned case of the two long-term inhabitants of one room, they did not refer to the third roommate by his real name (Martin) but talked about him as ‘he’.

Based on the introduced model of reputation [8], I ascribe the establishment of individual’s social network to the work of an evolutionary evolved mechanism considering propensity of individuals with regard to cooperation. As it was proved, each student forms a network of possible partners for subsequent cooperation, and he also spreads the information about those, who are sufficient for his own requirements. My research has shown that this network of individuals is not restricted to only student friends, but there are also classmates within the network or older or recent acquaintances. Paradox of the dormitory social network is grounded in the exclusion of own roommates in cases when they are considered unfit for cooperation.

In connection to verbal evaluation of others and establishment of own networks, my research points out the specific processes for partner selection aimed at sharing a common room. During the interview with the participants, I have attempted to request from them a reconstruction of a method how they happened to live in the room with their recent roommates. In the cases, when students were not satisfied with the situation prevailing in the room, in almost all instances these rooms were randomly allocated to them without any previous information with whom they will share one living space. Furthermore, in some instances newcomer roommates were assigned to rooms where already two roommates have been living together for a longer period of time. For example, when a student graduated or left the dormitories his/her place had to be filled in by a new inhabitant.

In both cases, it was a result of the administrative department of the student dormitories. On the contrary, rooms with content inhabitants were results of previous mutual agreement. These cases included also such students, which were allocated to certain room positions in accordance to the earlier mentioned principle (by the random choice of the administrative staff). Information attained in this way assumed the dormitory social network of a student, from what an individual chooses the suitable partners for sharing a room and this has been processed according to reputation. The interviews have confirmed the evaluation of reputation, since the students while reconstructing the method of accommodation had the tendency to talk about the possibilities they could chose from.

5. Conclusion

Social life in dormitories takes various forms. One of them is how students share they rooms. In some cases the explicit interests are fulfilled through the process of searching for ‘better’ roommates. In my fieldwork I have observed and interviewed individuals in this process.

Reputation models for evolution of cooperation have been employed to understand students’ behavior. My aim was to give recorded data an evolutionary explanation. I put stress on two main problems. The first problem was of mechanism of creating students’ social networks. The second was the mechanism of choosing potential partners for cooperation from this network. In this paper I argue for the reputation based scenario. Potential partners of one’s social network are being assessed with respect to their reputation and consequently chosen as roommates.

The problem of cooperation in my material was explained via indirect reciprocity approach. On this assumption, reputation based choices are product of possible evolutionary scenario when individual seeks for cues and signs of cooperating behavior.

References

1 Hamilton WD. The genetic evolution of social behaviour I & II. Journal Theoretical Biology 1964; 7: 1-52.

2 Dawkins R. The Selfish Gene. Oxford. Oxford University Press. 1989 [1976].

3 Trivers R. The evolution of reciprocal altruism. The Quarterly Review of Biology 1971; 46: 35-57.

4 Alexander RD. The Biology of Moral Systems. New York. Aldine de Gruyer 1987.

5 Panchanathan K & Boyd R. A Tale of Two Defectors: The Importance of Standing for the Evolution of Indirect Reciprocity. Journal of Theoretical Biology 2003; 224:115–26.

6 Nowak M & Sigmund K. Evolution of Indirect Reciprocity by Image Scoring. Nature 1998; 393: 573-577.

7 Rege M. Networking Strategy: Cooperate Today in Order To meet Cooperator Tomorrow. Discussion Papers 282. Research Department of Statistics Norway 2000.

8 Bergh A & Engseld P. The problem of Cooperation and ReputationBased Choice. Working papers 27. Lund University. Department of Economics 2005.

9 Darwin C. The Origin of Species. Oxford. Oxford University Press 1998 [1859].

10 Nowak M. A. Five rules for the evolution of cooperation. Science 2006; 314: 1560-63.

11 de Waal F. Chimpanzee Politics: Power and Sex among Apes, Baltimore. John Hopkins University Press 1989.

12 Axelrod R. The Complexity of cooperation. Princeton. Princeton University Press 1997.

13 Axelrod R & Hamilton W D. The Evolution of Cooperation. Science 1981; 211: 1390–96.

14 Smith MJ. Evolution and the Theory of Games. Cambridge. Cambridge University Press 1982.

15 Ellingsen T & Johannesson M. Anticipated Verbal Feedback induces altruistic behavior. Working papers 668. Stockholm School of Economics 2007.

16 Bateson M, Nettle D, Roberts D. Cues of being watched enhance Cooperation in a real – world setting. Biology Letters 2006; 2: 412–14.

17 Pinker S, Nowak M, Lee JJ. The Logic of Indirect Speech. Proceedins of the National Academy of Sciences of USA 2008; 105: 833–38.

Perceived Work-Related Benefits of Smoking and Attitudes towards Smoking Restrictions at Work

M. Zlatev[1], *A. Rodafinos*[2], *R. J. Eiser*[3]

[1]*SEERC, 17 Mitropoleos Str, Thessaloniki, Greece, mzlatev@seerc.org*

[2]*City College, Dept. of Psychology, 24 Prox. Koromila Str., 54635 Thessaloniki, Greece, rodafnos@city.academic.gr*

[3]*The University of Sheffield, Dept. of Psychology, Sheffield, S10 2TP, UK, J.R.Eiser@sheffield.ac.uk*

Work-site smoking harms the health of both smokers and non-smokers. One of the most successful approaches to curb the risks of smoking is the introduction of strict smoking restrictions. However, many smokers would not be satisfied with such measures, since smoking bans not only prevent exposure to passive smoking, but they also restrict the perceived benefits of smoking (e.g., perceived relaxation and concentration) to smokers. Whereas much of the research on smoking behaviour in the field of health psychology has concentrated on the perceived risks, the aim of the current research is to study the link between perceived benefits of smoking and attitudes towards smoking bans in the workplace. Finding such association would mean that 'replacing' smoking at work with other behaviours that achieve the same goals as smoking (e.g., relaxation or concentration trainings) may help to increase the acceptance of smoking bans on part of the smokers. The current paper describes work in progress, therefore only the rationale of the study and the methods used to investigate the relationship between the perceived work related benefits of smoking and the support of workplace smoking restrictions are reported. The practical importance of the present research and possible limitations are also discussed.

Keywords
attitudes towards smoking restrictions, perceived benefits of smoking, smoking in the workplace.

1. Introduction

Smoking in the workplace is an important facet of the global problem of tobacco smoking, the major preventable cause of death in the world [1]. Smoking at work affects not only the health of smokers, but also the health of non-smokers, since even short-term exposure to tobacco smoke could be extremely harmful, e.g. [2], [3]. In fact, non-smokers exposed to tobacco smoke take more sick days than those in smoking-free offices [4]. Additionally, workplace smoking is linked to financial losses to companies and institutions due to health problems and health-related absents of the employees [5]. In line with this, the productivity of current smokers has been shown to decrease as compared to former smokers or non-smokers due to smoking breaks and sick leaves [6], [7], [8]. Considering the general context of anti-smoking policies, the workplace is a setting where large groups of affected by the tobacco smoke persons can be reached, where the legislation does not invade the private life and last but not least, where multiple positive effects of smoking restriction policies have been already demonstrated, e.g. [9], [10].

The most promising way to curb smoking at work is to impose strict smoking bans [11]. Smoking restrictions not only prevent non-smokers from second-hand smoke [12], but also affect the smoking behaviour of smokers. After the introduction of smoking bans, a significant increase in successful quit attempts and a decrease in the cigarette consumption per day have been observed, e.g. [11], [13], [14]. In addition, smoking restrictions seem to have positive effects on the perceived social norms and attitudes relevant to both active and passive smoking [15]. Thus, banning smoking in the workplace is a strong signal that smoking is not acceptable and cannot be tolerated.

However, there are some barriers against the implementation of smoking restrictions in the workplace. Besides legislation issues, many smokers would not be satisfied with policies restricting smoking at work, e.g. [16], [17]. Although smokers in general cope with existing bans on smoking, e.g. [18], [19], [20], not surprisingly, the support for smoking restrictions before their implementation is significantly lower among smokers than among non-smokers [21]. Achieving stronger approval of smokers might be important because smoking bans seem to be most accepted when they are enforced by or with the strong support from the local community and not when they have been dictated from above [22]. Moreover, the acceptance of smoking bans before their implementation might influence the compliance with these rules after their implementation. Thus, sophisticated approaches should be found to persuade smokers to approve smoking restrictions.

To date, the main approaches to promote supportiveness of smoking restrictions involved messages on the health-related harms of active and passive smoking. Whereas there is evidence that these approaches are effective to a certain extent, cf. [23], they have also some limitations. For instance, using the argument of the harms caused by passive smoking would prevent smokers from smoking in the presence of non-smokers and vulnerable ones, but, still might not alter their active smoking behaviour and would not reduce the smoking breaks during work and the health related absents. In addition, persuasive messages on the risks of active smoking could be even discounted by smokers [24]. Thus, adequate approaches should be developed to address smokers in order to achieve higher support for smoking bans and to affect to a greater extent active smoking. The current research pays special attention to the perceived benefits of smoking – a variable that has been often ignored by previous studies. The perceived benefits of smoking represent short-term gratifications, whereas the most serious negative consequences of smoking appear rather as long-term risks: smokers experience the “positive” effects of the cigarette immediately after lighting up whereas lung cancer or heart disease may happen somewhere in the future[1]. Therefore, the experienced positive consequences of smoking (e.g. decreased anxiety, relaxation) function as positive reinforcements building the difficult to treat psychological dependence of smoking, e.g. [25], [26]. In line with this, the Transtheoretical Model of Change (TTM) [27] posits that the perceived benefits of smoking, termed as the pros of smoking, and the perceived risks of smoking or the cons of smoking contribute to the so called decisional balance that determines the transition through different stages of health behaviour change [27]. According to TTM, smokers who do not intend to quit smoking report high perceived benefits of smoking. While moving through the different stages of change towards fully abandoning of the smoking habit, the perceived benefits of smoking decrease whereas the perceived risks of smoking increase [28]. Thus, the perceived benefits of smoking represent important factors for starting and maintaining smoking.

Particularly in the workplace context, the perceived benefits of smoking might have an important regulatory function. For instance, recent studies showed that job strain is positively associated with smoking prevalence [29], and smoking intensity [30]. Accordingly, work stress was found to be associated with nicotine dependence [31]. Moreover, smoking intensity was positively related to particular stressors such as the perceived organizational justice [32]. These findings suggest that smokers might derive positive consequences of smoking that are related to the perception of their work context, and that the perception of these positive consequences might be linked to the attitudes towards anti-smoking regulation in the workplace. There is some empirical evidence suggesting that the perceived benefits of smoking are strong predictors of the attitudes towards smoking restrictions. A recent study, for example, found that, among other variables (such as the perceived risks of smoking, nicotine dependence, or demographic variables), the perceived relaxation and pleasure of cigarette smoking were the best determinants of the attitudes towards smoking bans in public places [33]. Another study in the context of mobile phone technology found that the perceived benefits of mobile phones predicted best the approval of regulations of the mobile phone use [34]. Thus, it appears that the perceived work-related benefits of smoking could be an important factor for predicting restrictive tobacco smoke policies in the workplace. The present study aims to investigate this suggestion. It is part of a broader survey exploring work-site smoking in Bulgaria.

The current study suggests that smokers perceive particular work-related benefits of smoking. Specifically, it is hypothesized that the perceived work-related benefits of smoking are related to other variables related to smoking behaviour. These variables include nicotine dependence, risk perceptions of active and passive smoking, perceived and self-reported violation of smoking restrictions, intentions to quit smoking, and quit attempts in the past 30 days. It is expected that the perceived work-related benefits of smoking will be positively associated with nicotine dependence, risk perceptions of active and passive smoking, and intentions to quit smoking. Positive correlations are also expected between the perceived work-related benefits of smoking and perceived and self-reported violation of smoking restrictions, and quit attempts in the past 30

[1]Note, that the perceived positive effects of the cigarettes do not outweigh the objective negative consequences.

days. The third hypothesis posits that the perceived benefits of smoking are significant negative predictor of the support for smoking bans at work. Smokers who report high perceived positive effects of smoking at work should least approve or even disapprove smoking bans.

2. Method

2.1 Setting and Participants

The target setting is one without smoking restrictions, with partial smoking restrictions or with smoking restrictions that are not strictly followed. Settings with already established and strictly followed smoking restrictions are not target of the current research. Two appropriate institutions were contacted and expressed willingness to participate in the research. The first setting represents a state institution and the second is a media company. Both settings employ approximately 200 persons. There is no data from Bulgarian surveys that allows prediction how many of the employees will agree to participate in the study.

Data will be collected from both smokers and non-smokers. In this manner there would not be any bias through providing only smoking employees with the test instrument. Smoking status will be determined through self report of the participants.

2.2 Procedure

The participants will be provided with description of the survey, explaining the general aims of the study, the right of the respondents to withhold information or to withdraw from participation at any time during completion of the questionnaire. Further, it reassures the participants that their answers will be treated anonymously. After reading the general information about the study, the participants will be asked to fill in the test instrument. After this, the participants will be requested to sign a participation consent form, stating that they received all of the needed information about the study. Through signing the form the participants also agree that their results will be used in statistical analyses and in the subsequent presentation of the results. Finally, if interested in the results, the respondents will have the opportunity to submit their e-mail on a separate sheet of paper, in order to receive a brief review of the findings.

2.3 Questionnaire

The test instrument is paper-and-pencil questionnaire. Its items were derived from the International Tobacco Control Survey (ITC), [35], and studies conducted in the framework of TTM [28].

First, *smoking status* is measured according to common definition distinguishing between smokers, non-smokers and smokers [36] using two items. The first item represents self-report of the current smoking status including the following options: "I do not smoke" (non-smokers), "I used to smoke, but I have given up" (ex-smokers), "I smoke but not on a daily basis" (occasional smoker), "I smoke at least one cigarette a day" (daily smokers). The second item reflects smoking history: "Have you smoked 100 cigarettes (about 5 packs) or more throughout lifetime?" These items will help to identify the regular smokers in the sample and thus to use only their data in order to test the hypotheses.

Perceived work-related benefits of smoking are measured using a scale of seven items derived from TTM and ITC and were defined relating particularly to the work context (e.g., "smoking helps me concentrate better at work"). Six items reflect particular perceived benefits, namely: pleasure, social acceptance by colleagues, easier socialization, relaxation, concentration, avoidance of boredom. One additional item measures the perceived effect of smoking on work: **"**smoking helps me work better**"**). The responses to all of these items are coded on a 5-point scale from 1 – "strongly disagree" to 5 – "strongly agree."

Nicotine dependence is assessed on the base of three items: 1) the amount of cigarettes smoked per day with open response format; 2) the time after waking up prior to first cigarette in the day, divided into four response alternatives: "within 5 minutes," "within the first half hour," "within the first hour," "later than an hour;" and 3) the perceived difficulty to go without smoking for a whole day, measured on a 4 point scale ranging from 1 – "not difficult at all" to 5 – "almost impossible."

Perceived harms of passive smoking are assessed regarding five risks: asthma, lung cancer, heart disease, bronchitis, breathing problems. All of them are measured on a 5 point scale ranging from 1 – "definitely not" to 5 – "definitely yes." Three additional items reflect the perceived harm due to passive smoking for three vulnerable groups: pregnant women, people suffering from illness and young children, assessed on a scale ranging from 1 – "definitely no" to 5 – "definitely yes." *Perceived self risk* is measured with two questions regarding the risk due to active smoking and due to passive smoking: "How likely do you think it is that you will seriously damage your health through being a smoker?" respectively "How likely do you think it is that you will seriously damage your health by being exposed to other people's smoke?" A further item regarded the *perceived harm of the own smoking to other people's health*: "How likely do you think it is that smoke from your own cigarettes can seriously damage other people's health?" The responses of all three items were coded on a 5 point scale ranging from 1 – "extremely unlikely" to 5 – "extremely likely." Four items assess the self risk assessment regarding the chances of developing lung cancer and heart disease compared with the chances of a non-smoker (*perceived relative risk*). These items include two time perspectives for the estimations of the relative risk: if smokers continue to smoke the amount they currently do and if smokers assume they will completely quit smoking in the following 6 months: "Let's assume [that you continue to smoke the amount you currently do] / [that you completely quit smoking in the following 6 months]. How would you compare your chances of developing [lung cancer] / [heart disease] in the future to the chances of a non-smoker?" The four resulting items measured on a 4-point scale from 1 – "extremely higher" to 4 – "lower".

Perceived violation of smoking restrictions in the working place is formulated regarding own behaviour and perception of the behaviour of the others: "In your company, do [you] / [other colleagues] smoke in areas where smoking is not allowed?” This items use 5-point Likert scale ranging from 1 – “never” to 5 “almost always.”

Item reflecting *intentions to quit* requests the following self descriptive options to be selected: “I intend to quit smoking within the next month”, “I intend to quit smoking within the next 6 months”, “I intend to quit smoking in the future, but later than 6 months”, “I do not intend to quit smoking”. This item was adopted from TTM [27].

The question on *attempts to quit smoking* uses open response format: “In the past 12 months, how many times did you try to quit smoking and succeed in not smoking for at least 24 h.?". This item was derived from [37].

Approval of existing smoking regulations (“Do you approve of the current smoking policy in your company?”) is measured with a single 5-point Likert scale ranging from 1 – “definitely not” to 5 – “definitely yes.” A second item reflects the *approval of proposed future smoking restrictions*. It consists of a stem proposition “Which of the following policy options would you like to be implemented in your company?” and four response alternatives: 1) “smoking should be allowed everywhere,” 2) “smoking should be allowed only in certain areas” and 3) “smoking should not be allowed in any area.” The fourth alternative offers additional possibility indicating satisfaction with the existing policy, in case the participants see no need for future smoking restriction: “I feel satisfied with the existing policy and do not endorse any changes”. Additional item asks about possible problems that could emerge after the implementation of a total smoking ban: The options using “yes – no” format were as follows: “smokers will not comply,” “clients/visitors will not comply,” “smokers colleagues will seek for job elsewhere,” “it would be extremely hard to implement the policy,” “there will be no problems.” One additional option with open ended question was added giving the opportunity to name options that are not listed as response alternatives.

Sociodemographic measures that will be obtained are age, gender, job description and working area.

2.4 Translation of the Questionnaire

According to the guidelines for translation of psychological tests by [38], the questionnaire was translated from English into Bulgarian and then it was back translated by a bilingual person from Bulgarian into English. Both versions in English – the original and the back translation – were compared and some minor discrepancies were found. The majority of the discrepancies referred to the translation of synonyms or identical expressions, thus this did not lead to revision of the items wording. However, two items in the Bulgarian version were revised due to incorrect translation from English into Bulgarian. After these corrections, it could be concluded that the Bulgarian version of the questionnaire is identical with the English version.

2.5 Statistical Analyses

First of all, analyses of the statistical properties of the particular scales (distribution of the variables, extreme values, intercorrelations between the items, internal consistency) are planned in order to conduct appropriate further examination of the three postulated hypotheses.

Regarding the first hypothesis, the means of the reported work related benefits of smoking would indicate whether smokers perceive work-related benefits of smoking. However, the first step of preliminary analyses would suggest if there are items that could be deleted from this scale due to lack of statistical quality.

The second hypotheses regarding the connection between the perceived related benefits of smoking and other variables will be examined computing correlation analyses between the perceived work-related benefits of smoking and the variables related to smoking behaviour: nicotine dependence, number of smoked cigarettes per day, risk perceptions of active and passive smoking, perceived and self-reported violation of smoking restrictions. This step will include the examination of possible relationship between the perceived work-related benefits of smoking and the sociodemographic variables.

Since the attitudes towards smoking bans are measured on a nominal scale, the third hypotheses will be examined performing two multinominal logistic regressions with dependent variables 1) the attitudes towards the existing anti-smoking regulation and 2) the attitudes towards a future anti-smoking regulation. The independent variables in this model are the perceived benefits of smoking, as well as the other variables mentioned with regard to the second hypothesis. Significant negative β-weight of the perceived benefits of smoking would indicate conformation of the hypothesis.

3. Practical Implementations and Limitations

The present research focuses on the suggested association between the perceived benefits of smoking with work behaviour. At this point prior to data collection, it could be only speculated about the importance of the future findings; however, it is possible to sketch the possible practical implementations that could be derived from this research. It is supposed that if smokers perceive work-related benefits of smoking, they should be linked with other smoking related variables and most importantly with the attitudes towards smoking bans. Seeing the implementation of smoking bans from the perspective of the smoker is important, because it allows to understand the common opposition on part of the smokers against smoking restrictions and to find better arguments for the introduction of anti-smoking policies. It also allows finding positive ways for combating health damaging behaviour. Changing job characteristics and reducing job strain may have the effect of promoting health behaviour, cf. [30]. Policy makers could develop approaches of "replacing" the perceived benefits persons derive from smoking with healthier strategies. If smokers for example report that smoking helps them concentrate better at work, it would mean that the introduction of concentration trainings could be a promising technique to argue for implementation of smoking restrictions. An important positive effect of these approaches could be curbing the smoking breaks when smokers leave their workplace in order to smoke outside the premises they work in, since the smoking breaks could increase following the implementation of smoking bans [39]. Thus, smokers should not be left "coping alone" with their withdrawal symptoms and craving for smoking, but instead they would be provided with more adequate arrangements, cf. [17], [40].

However, the focus on the perceived benefits of smoking as predictors of attitudes towards smoking restrictions does not imply that other factors do not matter and should be therefore ignored. It means rather that the perceived benefits of smoking are studied in the present research as a possible inspiration for finding additional approaches against smoking in the workplace that complement existing practices.

There are some important limitations that can be anticipated at this stage of study execution. First, since the most studies on smoking behaviour were conducted in Western populations, there are maybe some unexpected influences in the Bulgarian sample. For instance, Bulgaria has smoking rates above the average of the European Union (36%) and only a tiny minority of Bulgarians (11%) think that the smoking restrictions in their country are respected by smokers [16], thus making the sample different than the Western samples involved in the majority of studies regarding smoking behaviour. Second, concerning the test instrument it should be noted that there will be no variation of the order of the questions, so it is possible that some of the items affect the responses to other items. However, at this point, it is not possible to determine exactly the possible influences of the items' order. Additional limitation emerges from the cross-sectional design: cross-sectional data does not allow causal interpretations of the findings.

References

1 World Health Organization. WHO Report on the Global Tobacco Epidemic, 2008 - The MPOWER package. WHO: Geneva 2008.

2 Barnoya J, Glantz S A. Cardiovascular effects of secondhand smoke nearly as large as smoking. Circulation 2005; 111: 2684-2698.

3 Raupach T, Schäfer K, Konstantinides S, Andreas S. Secondhand smoke as an acute threat for the cardiovascular system: a change in paradigm. European Heart Journal 2006; 27: 386-92.

4 McGhee S M, Adab P, Hedley A J, Lam T H, Ho L M, Fielding R, Wong C M. Passive smoking at work: the short-term cost. Journal of Epidemiology and Community Health 2000; 54: 673-676.

5 U.S. Department of Health and Human Services. The Health Consequences of Involuntary Exposure to Tobacco Smoke: A Report of the Surgeon General. U. S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health, 2006.

6 Halpern M T, Shikiar R, Rentz A M, Khan Z M. Impact of smoking status on workplace absenteeism and productivity. Tobacco Control 2001; 10: 233-238.

7 Kirkcaldy B D, Cooper C L, Brown J M, Athanasiou J A, Ma L B. Job stress and health profiles of smokers, ex-smokers and non-smokers. Stress Medicine 1994; 10: 159-166.

8 Lundberg P. Does smoking increases sick leave? Evidence using register data on Swedish workers. Tobacco Control 2007; 16: 114-118.

9 Chapman S, Borland R, Scollo M, Brownson R C, Dominello A, Woodward S. The impact of smoke-free workplaces on declining cigarette consumption in Australia and the United States. American Journal of Public Health 1999; 89: 1018-1024.

10 West R. Banning smoking in the workplace. Smoking bans work: so what is the government going to do about it? British Medical Journal 2002; 325: 174-175.
11 Fichtenberg C M, Glantz S A. Effects of smoke-free workplaces on smoking behavior: systematic review. British Medical Journal 2002; 325: 188-191.
12 Haw S J, Gruer L. Changes in exposure of adult non-smokers to secondhand smoke after implementation of smoke-free legislation in Scotland: national cross sectional survey. British Medical Journal 2007; 335: 549-552.
13 Bauer J E, Hyland A, Li Q, Steger C, Cummings K M. A longitudinal assessment of the impact of smoke-free worksite policies on tobacco use. American Journal of Public Health 2005; 95: 1024-1029.
14 Braverman M T, Aarø L E, Hetland J. Changes in smoking among restaurant and bar employees following Norway's comprehensive smoking ban. Health Promotion International 2008; 23: 5-15.
15 Tang H, Cowling D W, Stevens C M, Lloyd J C. Changes of knowledge, attitudes, beliefs, and preference of bar owner and staff in response to a smoke-free bar law. Tobacco Control 2004; 13: 87-89.
16 European Commission. Attitudes of Europeans towards tobacco 2006. http://ec.europa.eu/health/ph_information/documents/ebs_239_en.pdf March 2008.
17 Parry O, Platt S, Thomson C. Out of sight, out of mind: workplace smoking bans and the relocation of smoking at work. Health Promotion International 2000; 15: 125-133.
18 Fong G T, Hyland A, Borland R, Hammond D, Hastings G, McNeill A, Anderson S, Cummings K M, Allwright S, Mulcahy M, Howell F, Clancy L, Thompson M E, Connolly G, Driezen P. Reductions in tobacco smoke pollution and increases in support for smoke-free public places following the implementation of comprehensive smoke-free workplace legislation in the Republic of Ireland: findings from the ITC Ireland/UK Survey. Tobacco Control 2006; 15 (Suppl. 3): iii51-iii58.
19 Hilton S, Semple S, Miller B G, MacCalman L, Petticrew M, Dempsey S, Naji A, Ayres J G. Expectations and changing attitudes of bar workers before and after the implementation of smoke-free legislation in Scotland. BMC Public Health 2007; 7: 206.
20 Trinidad D R, Gilpin E A, Pierce J P. Compliance and support for smoke-free school policies. Health Education Research 2005; 20: 466-475.
21 Poland B D, Cohen J E, Ashley M J, Adlaf E, Ferrence R, Pederson L L, Bull S B, Raphael D. Heterogeneity among smokers and non-smokers in attitudes and behaviour regarding smoking and smoking restrictions. Tobacco Control 2000; 9: 364-371.
22 Batra V, Patkar A, Weibel S, Pincock G, Leone F. Public health implications of voters' attitudes regarding statewide tobacco policy. Chest: 2002; 122: 295-298.
23 King K A, Vidourek R A, Creighton S, Vogel S. Smokers' willingness to protect children from secondhand smoke. American Journal of Health Behavior 2003; 27: 554-63.
24 Brown S L, Smith E Z. The inhibitory effect of a distressing anti-smoking message on risk perception in smokers. Psychology and Health 2007; 22: 255-268.
25 Benowitz N L. The nature of nicotine addiction. In Slovic P. (Ed.) Smoking. Risk, Perception and Policy (pp. 159-187). Thousand Oaks, CA: Sage Publications 2001.

26 Glautier S. Measures and models of nicotine dependence: positive reinforcement. Addiciton 2004; 99 (Suppl. 1): 30-50.
27 Prochaska J O, Velicer W F. The transtheoretical model of health behavior change. American Journal of Health Promotion 1997; 12: 38-48.
28 Velicer W F, DiClemente C, Prochaska J O, Brandenburg N. Decisional balance measure for assessing and predicting smoking status. Journal of Personality and Social Psychology 1985; 48: 1279-1289.
29 Brisson C, Larocque B, Moisan J, Vezina M, Dagenais G R. Psychosocial factors at work, smoking, sedentary behavior, and body mass index: A prevalence study among 6995 white collar workers. Journal of Occupational and Environmental Medicine 2000; 42: 40-46.
30 Kouvonen A, Kivimäki M, Virtanen M, Pentti J, Vahtera J. Work stress, smoking status, and smoking intensity: an observational study of 46.190 employees. Journal of Epidemiology and Community Health 2005; 59: 63-69.
31 John U, Riedel J, Rumpf H J, Hapke U, Meyer C. Associations of perceived work strain with nicotine dependence in a community sample. Occupational and Environmental Medicine 2006; 63: 207-211.
32 Kouvonen A, Vahtera J, Elovainio M, Cox1 S J, Cox1 T, Linna A, Virtanen M, Kivimäki M. Organisational justice and smoking: the Finnish public sector study. Journal of Epidemiology and Community Health 2007; 61: 427-433.
33 Zlatev M, Pahl S, White M P. Perceived risk and benefit for self and others as predictors of smokers' attitudes towards smoking restrictions. Psychology and Health (under review).
34 White M P, Eiser J R, Harris P R, Pahl S. Who reaps the benefits, who bears the risks? Comparative optimism, comparative utility, and regulatory preferences for mobile phone technology. Risk Analysis 2007; 27: 741-753.
35 Fong G T, Cummings K M, Borland R, Hastings G, Hyland A, Giovino G A, Hammond D, Thompson M E. The conceptual framework of the International Tobacco Control (ITC) Policy Evaluation Project. Tobacco Control, 2006, 15 (Suppl 3): iii3-iii11.
36 Weinstein N D, Marcus S E, Moser R P. Smokers' unrealistic optimism about their risk. Tobacco Control 2005; 14: 55-59.
37 Etter J F, Sutton S. Assesing 'stage of change' in current and former smokers. Addiction 2002; 97: 1171-1182.
38 Hambleton R K. The next generation of the ITC test translation and adaptation guidelines. European Journal of Psychological Assessment 2001; 17: 164-172.
39 Borland R, Cappiello M, Owen N. Leaving work to smoke. Addiction 1997; 92: 1361-1368.
40 The Lancet. Want employees to quit smoking? Editorial. The Lancet 2007; 369: 1491.

Breaking Bad News - A Romanian Experience

Bigiu Nicusor Florin[1], *Moga Marius Alexandru*[3], *Mihaela Badea*[2]

University of Transilvania, Faculty of Medicine, Department of Fundamental and Prophylactic Sciences, 29 Eroilor Blvd, Brasov, Romania, {nicusorbigiu[1], badeamihaela[2]}@yahoo.com, moga.og@gmail.com[3]

Breaking bad news encounters serious problems either for receiver, unprepared to cope with the emotional content, or for deliverer as trained staff, but especially for untrained staff in which the impact is tremendous. Thus, the aim of this paper is to inform and investigate on this problem, by the assessment of physicians' conception on the procedure of breaking bad news to patients. In this insight two approaches will be considered: that of a professional and that of a patient. In this respect, the authors have selected a focus group of 75 physicians of various medical specialties which imply breaking bad news to patients. These physicians were administered a pretested questionnaire, containing 18 questions. The following aspects have been assessed: the frequency, location and time allotted to this process; the attitude of the physician towards breaking the bad news to the patient and the attitude during the unfolding of the process; the choice made by the physician regarding breaking the news either directly to the patient or to the family. The majority of the physicians questioned (62%, n= 46) frequently get to communicate bad news to patients. The study of L.Grassi, T.Giraldi shows that 44,8% of physicians always communicate bad news whereas 46,6% do it only depending on the case. The impersonal and pitiful reaction is a dangerous extreme that can and will negatively affect the quality of the physician-patient relationship. What is considered to be the ideal method is the use of a sufficient period of time (41%) or breaking the news during several visits (16%), thus being able to assure a prolonged support. The communication of the ad vitam prognostic may be a source of errors. Communicating a large amount of information creates anxiety. The most frequent reaction on finding out the bad news is the reversible nervous breakdown (in 59% of the cases). 90,6% of the study group members have never participated in courses focused on the process of breaking bad news to patients.

Keywords

bad news, chronic and terminal disease.

1. Introduction

The bad news are any information delivered by health staff that produces a negative alteration to a person's expectations about their present and future. Bad news promotes a gap between the patient expectation and medical reality shown by diagnosis and prognostic.

A history of telling the bad news in Western Medicine has to consider the basic tradition widespread in its roots. In the Hippocratic tradition there is a stress on prognostic, on foretelling the course of the disease supposedly based upon ancient oracular practices and interpretation of dreams. Hippocratic medicine was humble and passive, allowing criticism for its feature of expectance of death. The Hippocratic treatise on prognosis was a workmanlike guide for the doctor who wanted to achieve therapeutic success, but it said nothing about considering the patient's feelings. In the Galenic tradition, complex intellectual exercises were introduced to escape blame for failures and to win universal admiration in which the physician need to cultivate the art of making his diagnoses and prognoses seem like acts of divination. A clever application of this tactic was to predict the worst possible outcome while reluctantly agreeing to accept the case. If the patient died, the physician's prediction was vindicated; if the patient recovered, the physician appeared to be a miracle worker.

In medieval times the Galenic tradition continued; the doctor who believed that the prognosis was hopeless could say so boldly, perhaps losing his fee, but building his reputation; however, he should incline to the bright side when possible. One interesting example is a postal bad news breaking from 1693, when a young Italian doctor wrote, unasked, to a senior colleague to tell him that he was suffering from a lethal disorder, although the younger doctor had probably never seen the elder one during the illness. The young doctor wrote grimly to his elder colleague: „Only autopsy can show you graphically the specific characteristics of your polypous excrescence. . .", and from the moment of receiving the letter the symptoms of the last grew much worse. He died soon and the postal diagnostic was confirmed by necropsy. The name of the Italian doctor who deliver the news was Alexander Knips Macoppe and later in his career wrote advice to his juniors on prognosis: "Use ambiguous phrases"; "Smile with those who are going to be cured, but weep with the dying; such tears sell well"; "If the prognosis is bad, hide it from the patient, or merely hint at it." [1]

In the 1950s and 1960s, doctors, as a rule, felt it inappropriate to tell their patients the whole truth on the grounds that what someone does not know cannot harm them. The common thought was the patient doesn't want to know. Recent studies show that 50% to 97% of patients want to know the truth about their illness, even when dealing with bad news. [2]

Censoring information continued this traditional paternalistic attitudes which still can be seen despite moves towards increased autonomy and empowerment for patients. Since the 1970s, patient activists in the USA and Europe have aggressively asserted their claims to be regarded as experts on their own illnesses and to play a more active part in health-care decision-making. In spite of this "laicisation of the Temple of Aesculapius" and the patient-centred medicine, in the last decade, refusal to communicate on request the survival estimates or prognostic errors made in optimistic directions are common practices. [3] [4] Patient-centred medicine brought the recognition that doctors may require further preparation in order to communicate better the bad news and this allowed the development of countless number of courses and guidelines at undergraduate and postgraduate level.

Delivering the bad news to a patient implies negative consequences for both participants : the receiver and the deliverer. In the case of receiver (patient), response to loss viewed as an orderly progression through different stages, was proposed by a large number of authors: Lindemann 1944, Kubler-Ross 1969, Parkes 1975 Bowlby 1981, Worden 1991, Jacobs 1993. [5] [6] [7] [8] [9] [10] Initially these theories encircled issues regarding grief and bereavement, then, some of them were proposed for response of terminally ill patients to awareness of their impending death(Kubler-Ross stages of death and dying: denial-dissociation-isolation, anger, bargaining, depression, and acceptance). Many of the theories or models of grief (**Table 1**) share common themes, categories or tasks(shock-numbness) and in all of them the last stage(acceptance-recovery-adjustment) is mandatory to be achieved in order to insure a decent death. For the purposes of this study Kubler-Ross stages will be used because of the heavy reliance of medical education on this model. [11]

Lindemann 1944	Kubler-Ross 1969	Parkes 1975	Bowlby 1981	Worden 1991	Jacobs 1993
Shock	Denial	Numbness	Numbness	Acceptance of loss	Numbness, disbelief
Acute mourning	Anger	Searching	Searching	Working with pain	Separation distress
Resolution	Guilt	Depression	Disorganisation-despair	Adjustment with loss	Depression, mourning
	Depression	Recovery	Adjustment	Reinvestment	Recovery
	Acceptance				

Table 1 Grief Theories

In the case of the impact of the bad news on the deliverer (health staff) the research done in different parts of the world was reviewed by Fallowfield et all in 2004. The disparity between physicians in opinion about truthful disclosure, the stress experienced when giving bad news, and the desire for more training is notable. [4] Excepting the study of Grassi et all, 2000 which included 675 Italian physicians the rest of studies are small scale studies. [12] Recent empowerment tendencies allow other members of the heath staff like nurses not only to be part of the team who is delivering the bad news but also to deliver effectively bad news concerning caring. [13]

In this context we decided to asses the attitudes of doctors from Romania, Brasov city towards delivering bad news. Although emerging research reveals some signs of change in southern Europe in the traditional medical practice of non-disclosure of cancer diagnosis and prognosis to the patients affected, paternalistic attitudes still remain.

One observations is needed, delivering bad news does not refer only to medical specialities. Efforts had been made to improve, in many corporate organisations, their internal communications lines by encouraging employees to report important information, including bad news, up to the management ranks. A survey suggests that more than a third of employees believe that the top management in their own companies does not encourage the reporting of vital information—including bad news. [14]

Expressing bad news in medicine is not a narrow subject, limited let's say to oncology, since in almost every medical specialty bad, sad, and difficult information must be given to patients and their families

2. Material and Methods

The participants in the study were a focus group of 75 physicians of various medical specialties which imply breaking bad news to patients. Were included medical specialities like: internal medicine (oncologists), general surgery, emergency medicine and intensive therapy, neurology, obstetrics and paediatrics.

Authors have used as sampling frame a list of doctors activating in the city of Brasov (284,596 inhabitants) from which based on previous studies in international literature regarding delivering bad news were selected only specialities which have to deal with communicating bad news. A list of 178 doctors was constituted and each physician it was sent a packet that included a letter of explanation, detailed instruction, the questionnaire and a stamped return envelope. One week following this mailing, a reminder/thank-you card was sent to. Eighty-one packets were returned (response rate = 45,5%) and because of inconsistencies 6 packets were eliminated. A focus group of 75 resulted (42,1% of those sent).

The questionnaire, contained 18 items (17 multiple choice questions and one open ended question). The core of the self-administered questionnaire was based on the Grassi et all, 2000, ten items questionnaire used in North Italy. [12]. We added to this core another 8 items in order to asses other problems of breaking the bad news. The following aspects have been assessed:

- the frequency, location and time allotted to this process;
- the attitude of the physician towards breaking the bad news to the patient and the attitude and techniques used during the unfolding of the process;
- the choice made by the physician regarding breaking the news either directly to the patient or to the family;
- the attitude of the physician regarding communication of the survival estimates
- the most common reaction encountered by physicians

3. Results and Discussion

The majority of the physicians questioned (62%, n=46) frequently (weekly, monthly) get to communicate bad news to patients.(**Figure 1**) When we choose the sample we opted for specialities in which delivering was a frequent process and it is interesting that among these specialties there are individuals (4%, n=3) who refuse do the disclosure probably by referring the patient to GP or other specialist.

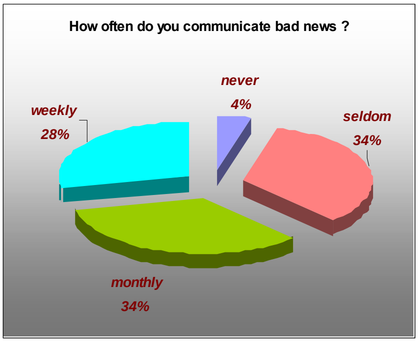


Figure 1 How often do the doctors communicate bad news?

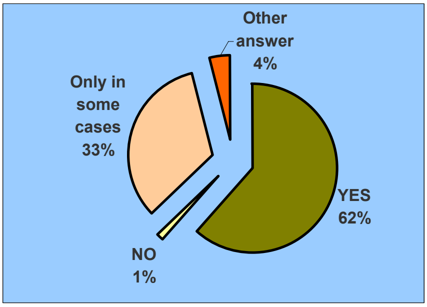


Figure 2 Do you think a patient should be delivered the bad news?

Regarding the situation in which they communicate the bad news we noticed that 62% (n=46) tell the patient the bad news since 33% (n=25) choose to tell the truth depending the case.(**Figure 2**) The study of Grassi et all, 2000 shows that 44,8% of physicians always communicate bad news whereas 46,6% do it only depending on the case. The percents of refusal to deliver the bad news is low in both studies: 1% in our study and 0,8% in the Italian study. [12] In one study realized in Peru (45% of internists) were in favour of not telling patients their diagnosis and informing the relatives instead in detriment of patient's autonomy. [15] The situation in other countries from South-east Europe is the following: Turkey (45% of oncologists generally tell the truth since 7% do it always, Ozdogan M. et all, 2006), [16] French (majority of respondents opted for prognosis disclosure only at patients' request, very few opted for systematic disclosure without patient's request, Peretti-Watel et all, 2006), [17] Japan(13% of doctors inform cancer patients of their disease, Tanida 1994). [18] The situation described above shows diversity in the attitude regarding communicating bad news with trends toward autonomy and trends toward paternalism.

The physicians from focus group deliver bad news to different kinds of patients: surgical patients (n=28), patients from oncology services (n=24), victims of accidents (n=19), neurological patients (n=12), terminally ill patients (n=12), others (n=11). The physicians that we investigated presented

In order to clarify the situation we asked the doctors in which circumstances they deliver the bad news. (**Figure 3**)

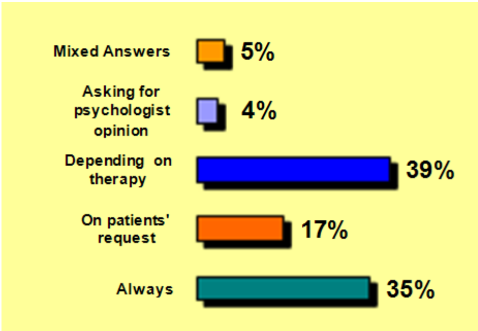


Figure 3 In which circumstances do you deliver the bad news?

We found out that the initial percent of 62% of those who consider that the patient should be delivered the bad news decreased to 35% (n=26) referring to those who are expressing the bad news, always, not thinking of circumstances. In this case we are thinking that some physicians are waiting either for direct request from the patient, either for cues that the patient is ready for the disclosure, or for psychologist opinion (4%, n=3).

In the case of the survival estimates (**Figure 4**) the percent of those who always allow the patient to know them is lower (27%, n=21)

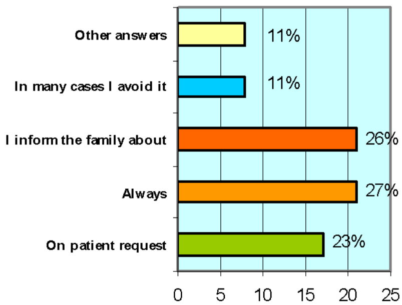


Figure 4 Do you deliver together with the bad news the survival estimates?

The communication of survival estimates could lead to errors of estimation. In one US study of physicians working in five hospices, even if patients requested survival estimates, physicians said that they would provide frank disclosure only around 37% of the time, favouring instead either no disclosure or a conscious overestimate of survival. (Fallowfield et all, 2002) [18]

We have asked the respondents how much time do they allotted for breaking the bad news and the answers were the following: 16% (n=12) few hours sectioned in few meetings, 41% (n=31) more than 15 minutes, 31% (n=23) less than 15 minutes and 9% (n=7) in few minutes. Our sample comprised busy doctors: 26% (n=19) have more than twenty medical appointments, 38% (n=28) have between 10 and 20 medical appointments and 36% (n=27) have less than 10 medical appointments. The discussion in which the disclosure is made take place usually in the doctors bureau (79%, n=62) but there are in some cases special facilities for this disclosure (15%, n=12).

When we questioned them about whom they are giving the bad news (**Figure 5**) we registered that 75% (n=56) imply the family in this disclosure, 5% (n=4) tell the truth only to patient and 17% (n=13) tell the bad news only to the family.

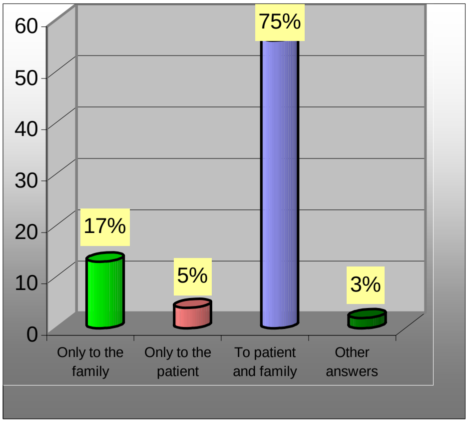


Figure 5 Whom are you giving the bad news?

The attitude of the doctors toward the patient during disclosure was empathic, 57% (n=41) but we found out other reactions like impersonal, technical (17%, n=12) and pitiful (26%, n=19) which can and will negatively affect the quality of the physician-patient relationship. In this conditions 90,7% (n=68) admitted that they never take part at courses to help them deliver the bad news. Considering international literature, in a ten years review on bad news literature, is notable the disparity between physicians in opinion about truthful disclosure, the stress experienced when giving bad news, and the desire for more training. [4]

We were interested about the most common reactions (**Figure 6**) that they did encounter in patients after the bad news disclosure: 59% (n=53) reversible nervous breakdown with support, 19% (n=17) fury, denial or requesting a second medical opinion, 11% (n=10) calm, cold, with interest more on technical details, 9% (n=8) hostility and anger directed towards the doctor and 2% (n=2) indifference, destructive negativism. The doctors report us that they found difficult to answer this question because they encounter these classes but it was hard for them to tell which is the most common.

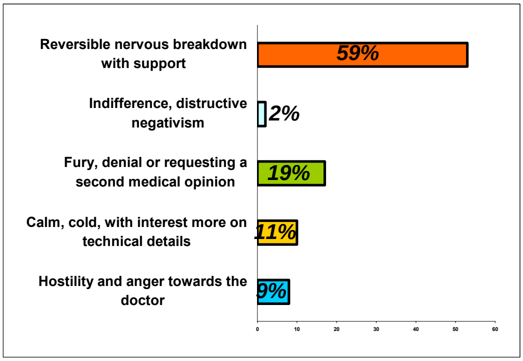


Figure 6 Which is the most frequent reaction do you encounter at hearing the news?

4. Conclusions

The delivery of sad, bad, and difficult news will always bean unpleasant but necessary part of medicine. As our focus group clinicians stated guidelines and courses could help them in developing a strategy for minimizing the stress experienced by the patient during the disclosure.

We found out that that 62% (n=46) consider that the patient should be delivered the bad news and the percent of those who always allow the patient to know the survival estimates is lower (27%, n=21). In any case we consider this trend as a move towards greater patient autonomy with some fear of errors in case of survival estimates. Certain caveats should be mentioned in relation to this study. First, physicians involved in this research are not representative of all Romanian physicians and thus generalizability of the results is not permissible. The study is limited by its small sample and by assigning, based on previous research in the field, of certain specialities which are believed to be in close contact with the disclosure of bad news. The nature of this study was exploratory and the authors intend to develop in the next year after some statistic calibration of the psychometric tool a larger sample study. Despite these limitations, the study indicates that communication of bad and sad news is still a complex problem in Romania and that further efforts are necessary to modify physicians' attitudes towards patients and to enhance their communication skills in clinical practice.

The involving of the family in breaking bad news was preferred by physicians (we registered that 75% (n=56) imply the family in this disclosure) which suggests that the patient social support system is important and effective. The most frequent reaction encountered after the delivery of bad news was reversible nervous breakdown (in 59% of the cases) and empathy was the attitude adopted by doctors during the process in 57% (n=41). Further limitation of the study shall be considered resulting from the fact that the meaning of death in Occident is different than in Eastern Europe and this differences could influence the psychometric tool.

References

1 Forrester J. Postal diagnosis: breaking the bad news in the 17th century. BMJ 1995;311:1694-1696.

2 Buckman R. Breaking bad news: why is it still so difficult? BMJ 1984;288:1597–1599

3 Tomes N. Patient empowerment and the dilemmas of late-modern medicalisation . The Lancet 2007;369:698 – 700.

4 Fallowfield L., Jenkins V. Communicating sad, bad, and difficult news in medicine. The Lancet 2004; 363: 312–19.

5 Lindemann E. SYMPTOMATOLOGY AND MANAGEMENT OF ACUTE GRIEF. Am J Psychiatry 1944 101: 141-148

6 Kübler-Ross E. On Death and Dying. New York, NY: Macmillan; 1969

7 Parkes C. Psycho-social Transitions: Comparison between Reactions to Loss of a Limb and Loss of a Spouse. The British Journal of Psychiatry (1975) 127: 204-21

8 Bowlby J. Processes of mourning. Int J Psychoanal. 1961;42:317-339

9 Kim K, Jacobs S. Pathologic grief and its relationship to other psychiatric disorders. J Affect Disord. 1991;21(4):257–263

10 Maciejewski P K, Baohui Zhang, Susan D. Block, Holly G. Prigerson, An Empirical Examination of the Stage Theory of Grief. JAMA. 2007;297:716-723.

11 Downe-Wamboldt B, Tamlyn D. An international survey of death education trends in faculties of nursing and medicine. Death Stud. 1997;21:177-188

12 Grassi L, Giraldi T, Messina EG, Magnani K, Valle E, Cartei G. Physicians' attitudes to and problems with truth-telling to cancer patients. Support Care Cancer 2000; 8: 40–45

13 Peel N. The role of the critical care nurse in the delivery of bad news. British Journal of Nursing. 2003; 12: 966-971

14 Sirota survey www.sirota.com/pressrelease/4Employees_Reluctant_To_Report_Bad_News_905.pdf, September 2005

15 Monge E., Sotomayor R. Telling the truth to patients with cancer: what is the truth? . The Lancet Oncology.2004;7 :944 – 950.

16 Ozdogan M, Samur M, Artac M, Yildiz M, Savas B, Bozcuk HS. Factors Related to Truth-Telling Practice of Physicians Treating Patients with Cancer in TurkeyJournal of Palliative Medicine. October 1, 2006, 9(5): 1114-1119.

17 Peretti-Watel P., Bendiane M.K., Obadia Y., Lapiana J.M., Galinier A., Pegliasco H., Favre R., Moatti J.P. Disclosure of Prognosis to Terminally Ill Patients: Attitudes and Practices Among French PhysiciansJournal of Palliative Medicine. April 1, 2005, 8(2): 280-290.

18 Tanida N. Japanese attitudes towards truth disclosure in cancer. Scand J Soc Med. 1994 Mar;22(1):50-7

19 Fallowfield LJ, Jenkins VA, Beveridge HA. Truth may hurt but deceit hurts more: communication in palliative care. Palliat Med 2002; 16:297–303

Scoliosis severity diagnostics: the new application for automatic X-ray image analysis

Lucia Hutnikova[1], Radovan Hudak[2], Jozef Zivcak[3]

Technical University of Košice, Faculty of mechanical Engineering, Department of Biomedical Engineering, Automation and Measurement, Letná 9, Košice, Slovakia, {lucia.hutnikova[1], radovan.hudak[2], jozef.zivcak[3]}@tuke.sk

Scoliosis is a medical condition in which a person's spine is curved from side to side or front to back, and may also be rotated. On an X-ray, the spine of an individual with a typical scoliosis may look more like an "S" or a "C" than a straight line. An X-ray of the full-length standing spine is the standard method for evaluating the severity and progression of the scoliosis, and whether it is congenital or idiopathic in nature. The traditional medical management of scoliosis is complex and is determined mainly by the severity of the curvature and skeletal maturity. Consequently, the severity of the curvature is commonly represented by Cobb angle and Nash-Moe rotation of vertebrae. These parameters are manually obtained by an orthopedist using X–ray images, which is an inaccurate and time consuming process.

The new sophisticated image analysis software for scoliosis detection was developed. The patient's database module with digitalized X-ray images was included. The software calculates the spine curvature and evaluates the rotation level of the spine for optimization of the scoliosis diagnostics.

Keywords

scoliosis, spine, Cobb angle, X-ray, image analysis, software.

1. Introduction

Scoliosis is a medical condition in which a person's spine is curved from side to side, and may also be rotated. On an x-ray, the spine of an individual with a typical scoliosis may look more like an "S" or a "C" than a straight line (fig.1).

It is typically classified as:

- congenital (caused by vertebral anomalies present at birth),
- idiopathic (infantile, juvenile, adolescent, or adult),

- a secondary symptom of another condition.

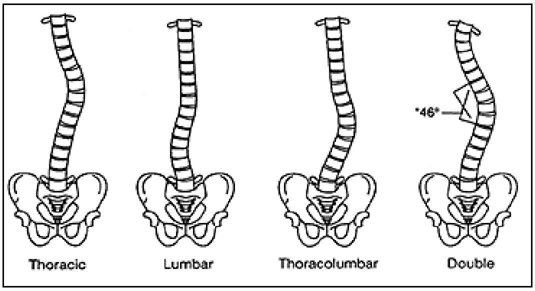


Figure 1 Type of spine curve by localization.

The appropriate management of scoliosis relies on the vigilance of clinicians: first to identify the presence of a scoliosis, and secondly to ensure an appropriate clinical examination is performed to identify those who require a more imaging strategy.

Non-invasive systems have evolved, such as the handheld ‘‘scoliometer’’ (Bunnel, 1984, Burwell et al., 1990), Moire-fringe mapping (Takasaki, 1970, Moreland et al, 1981, Willner et al, 1982, Idesawa, 1982, Karachalios et al, 1999), the raster-based systems like the ISIS system (Weisz et al, 1988, Theologis et al, 1997), or the Quantec system (Goldberg et al, 2001, Thometz et al, 2000) or the Ortelius (Dickman et al, 2001b) scanners, and devices that scan 360° torso profiles (Gomes et al, 1995, Sciandra et al 1995, Poncet et al, 2000, Schmitz et al., 2002), ultrasound systems (Suzuki, et al, 1989), Computer tomography (scoliosis was reconstructed in 3-Dimensional using a multiplanar digital radiographic technique allowing the visualization of the vertebral line of the spine in any projection using auto CAD software), Magnetic Resonance Imaging and last but not least stereo-photogrammetric systems (Frobin, et al, 1981 and 1983, Thomson, 1985, Hill et al, 1992, Sechidis et al, 2000), the FONAR images (coronal, sagittal and axial views of the entire spine - with no radiation) and 3-Dimensional Motion Analysis System.

Plain radiography remains the usual method of measuring curvature progression, the place of radiology in scoliosis management is straightforward and it is to confirm the diagnosis, identify any underlying cause, and monitor the degree of curvature.

2. The Standard Method for Assessing the Curvature

Patients with scoliosis are examined to determine if there is an underlying cause of the deformity. During a physical examination the patient is asked to bend forward (Adam's Bend Test). Neurologic examination includes reflexes, sensation, motor strength to ensure that there are no deficits or deterioration of baseline deficits. If scoliosis is a possibility the patient should be sent for an x-ray to confirm the diagnosis.

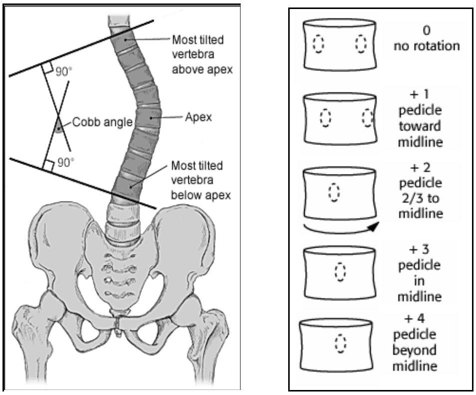


Figure 2 a) The Cobb angle b) Nash-Moe rotation.

The standard method for assessing the curvature on an anteroposterior radiographic projection of the spine quantitatively is the measurement of the Cobb angle. When assessing a curve the apical vertebra is first identified; this is the most likely displaced and rotated vertebra with the least tilted end plate. The end vertebra is the most superior and inferior vertebra which are least displaced and rotated and have the maximally tilted end plate. A line is drawn along the superior end plate of the superior end vertebra and a second line drawn along the inferior end plate of the inferior end vertebra. If the end plates are indistinct the line may be drawn through the pedicles. The angle between these two lines is measured as the Cobb angle (fig. 2a).

Rotation is an inherent structural change in scoliosis. Displacement of Pedicles: On anteroposterior view, one pedicle rotates toward the midline and the other rotates to the lateral border of the vertebra. Nash-Moe technique measures pedicle rotation by dividing the vertebral body into segments. The segment into which the pedicle is located quantifies rotation (fig. 2b). requirements from practise and missing image analysis software were the inspiration to develop a new practical too for analyzing.

3. Analysis of the New Software

In the present, the computer technique is used in the diagnosis. We decided to develop the software for simplifying the process of the scoliosis detection. The software was written in C++ language.

C++ is a general-purpose programming language. C++ is regarded as a mid-level language, as it comprises a combination of both high-level and low-level language features. It is a statically typed, free-form, multi-paradigm, usually compiled language supporting procedural programming, data abstraction, object-oriented programming, and generic programming. [8]

We developed new sophisticated image analysis software for scoliosis detection. The ScolioX software is used by orthopaedists for simplifying diagnostic spine deformity and also for easy data archiving and searching. The software consist of 2 parts, database system and part for X/ray analyzing. The patient's database module with digitalized X-ray images was included (fig. 3).

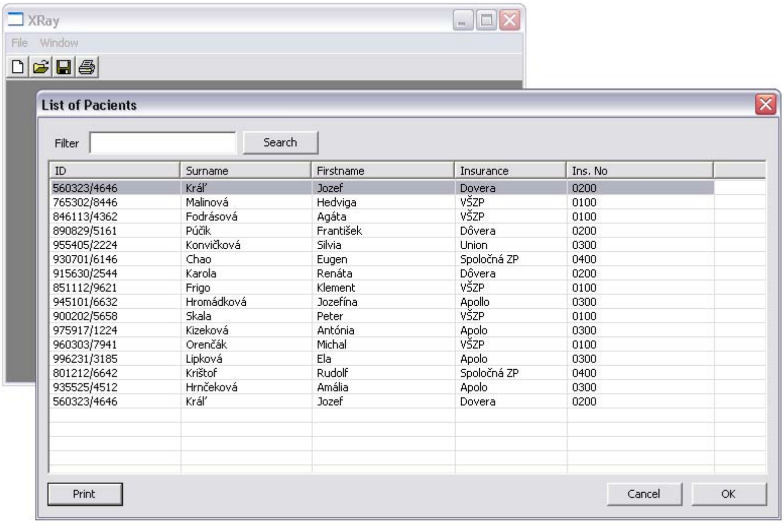


Figure 3 Main window with list of patients.

The software provides efficient tools for image processing, evidence X-ray images and patients, possibility insertion notes or remarks straight to pictures, printing ascending documents by personable computer (fig. 4).

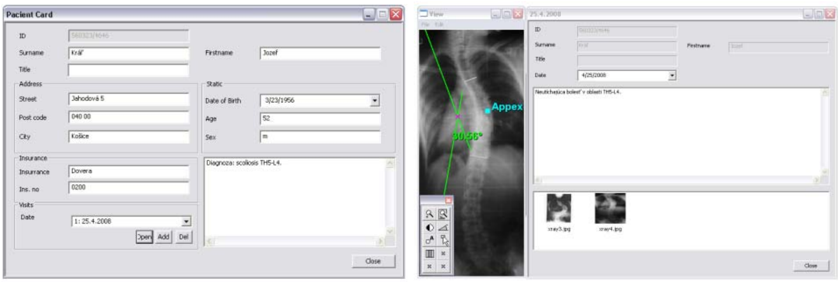


Figure 4 The Patient card with progress reports and x-ray attachments.

The software calculates the spine curvature (fig. 5) and evaluates the rotation level of the spine for optimization of the scoliosis diagnostics. It is controlled by user friendly graphical interface and supports multi-language (in English, Slovak). To work with this program it is necessary to use digitizing X-ray images.

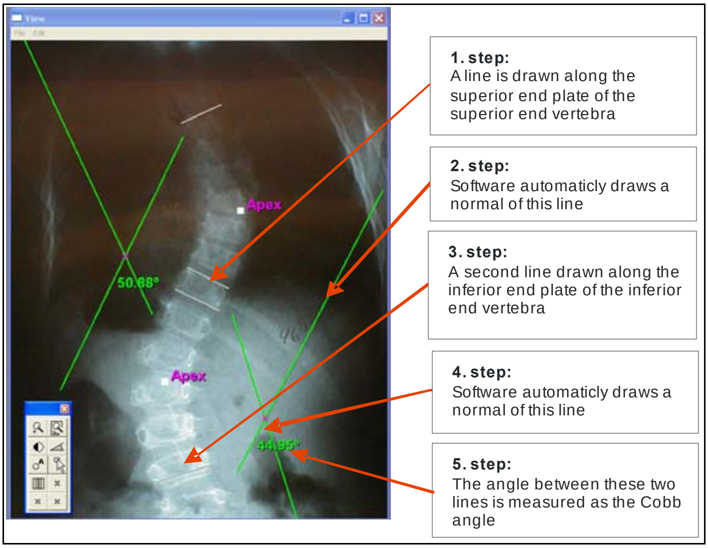


Figure 5 Calculation of Cobb angle.

3.1 Digitalization Radiographs

For the digitalization of roentgenograms we used ScanMaker 9800XL. This scanner captures details in highlight and shadow areas at a high optical density range of 3,7 Dmax, optical resolution 1600 dpi, and 16 bits of stored data. The 16" x 12" transparency scan area can scan the critical areas of almost all large format X-Rays. The ability to resolve huge amounts of highlight and shadow detail means that subtle X-Ray information will not be missed. This provides Medical Professionals a solution to digitize their roentgenogram for electronic data storage.

The software requires the following hardware, software configuration: Processor Pentium II PC or higher, Minimal resolution 1024x768x16bit, Windows XP or higher, DirectX9 compatible card, RAM 512MB.

4. Conclusions

To develop the efficiency, software output precision and repeatability actual clinical practice will be evaluated. In comparison to the previous manual method, a higher diagnostics efficiency is expected, which helps to assess an optimal treatment method. Future plans are clinical tests of ScolioX software and statistical evaluation of obtain results.

Presented paper is the objective of grant work VEGA 1/0829/08 – The Correlation of the Input Parameters Changes and the Output Thermograms at Infrared Thermographic Diagnostics.

References

1 Dungl P. a kol. Ortopedie. Grada Publishing, a. s., 2005, Praha, ISBN 80-247-0550-8
2 Kokavec, M., Novorolský, K. Skolióza a pohybová aktivita u detí. Pediatr. prax, 2007; 2: 70–74
3 Goldstein LA Wangh TR. Classification and Terminology of scoliosis. Clin. Orthop., 93, 1973, s. 30 – 41.
4 Andrews M. Programujeme v jazyce C++. Computer Press, 1997, ISBN 80-85896-91-5
5 Herout P. Učebnica jazyka C. Kopp, 1995, ISBN 80-7232-220-6
6 Patias P., Stylianidis E., Pateraki M., Chrysanthou Y., Contozis C., Zavitsanakis Th. 3D digital photogrammetric reconstructions for screening
7 Wright N. Imaging in scoliosis. Arch Dis Child 2000; 82:38–40
8 Perdriolle R., Le Borgne P., Dansereau J., De Guise J. Idiopathic scoliosis in three dimensions: A succession of two-dimensional deformities? Spine 2001, vol. 26, no24, ISSN 0362-2436, pp. 2719-2726
9 http://www.microtekusa.com/sm9800xl_fb.html
10 http://www.harms-spinesurgery.com
11 http://en.wikipedia.org/wiki/C%2B%2B

Formal Thought Disorder (FTD) Disrupts IOR in Schizophrenia

Fenia Kalogeropoulou[1], Ana B. Vivas[2],Peter W.R. Woodruff[3]

[1]SEERC, 17 Mitropoleos Str, Thessaloniki, Greece, fkalogeropoulou@seerc.org

[2]City College, Dept. of Psychology, 24 Prox. Koromila Str, 54622 Thessaloniki, Greece, vivas@city.academic.gr

[3]Academic Clinical Psychiatry & Sheffield Cognition and Neuroimaging Laboratory (SCANLab), Section of Neuroscience, Sheffield University School of Medicine & Biomedical Science, The Longley Centre, Norwood Grange Drive, Sheffield S5 7JT, United Kingdom, p.w.woodruff@sheffield.ac.uk

This study aimed at examining the relationship between Formal Thought Disorder –FTD- and inhibition of return -IOR- [1] in patients with schizophrenia. Two experiments were performed. In Experiment 1 (single-cue IOR procedure) participates 26 patients and 28 healthy adults. In Experiment 2 (double-cue IOR procedure) participated 24 patients -diagnosed with schizophrenia- and 21 healthy adults from Experiment 1. Both experiments involved seven SOA manipulations (140, 230, 350, 450, 600, 2000, and 3200 ms). Patients' symptoms were assessed using the SAPS and SANS scales. Patients with schizophrenia showed significant IOR effects at the double-cue and absence of IOR at the single-cue procedure. Also, there was a significant negative correlation between FTD scores and cueing scores (RTcued-RTuncued) at the 2000ms SOA value in both experiments. Patients who scored high for FTD displayed more facilitation; whereas patients who scored low for FTD tended to show inhibition. This finding is consistent with the presence of impaired inhibitory processing in the semantic domain, together with impaired inhibitory processes in the visuospatial domain in patients with FTD.

Keywords

formal thought disorder, IOR, schizophrenia.

1. Introduction

Deficits of visual processing in schizophrenia have been found in cognitive tasks measuring visuospatial selective attention [2]. According to Posner [3] orienting attention within the visual field involves three mental operations, disengaging attention from the current focus, moving attention to the location of the target and engaging the target. Studies with brain-injured patients have shown that these operations are localized in the posterior parietal lobe, the midbrain and the thalamus [4].

Orienting of attention within the visual field is exerted through the deployment of both facilitatory and inhibitory processes in tasks that require target detection [1]. [1] developed a widely used covert orienting of attention task (COVAT). In a typical COVAT task, a target can appear at the location of a previously presented spatial cue (valid location) or at the opposite non-cued location (invalid location). When the time interval between the onset of the cue and the target (SOA: stimulus onset asynchrony) is short (~<300ms), valid trials result in a reaction time (RTs) advantage over invalid trials, as participants detect the target more efficiently in the valid location. However, when the cue-target interval is longer than 300 ms, response times are higher for targets presented in the previously cued valid location as compared to the invalid location. This phenomenon has been called *inhibition of return* (IOR) and it has been proposed to reflect a bias of attention against re-orienting to previously explored locations [3]. Recent research suggests that early facilitation and late inhibition may be two independent processes that arise simultaneously when there is a non-informative cue on the periphery [5]. Although early studies [1] regarded both facilitation and IOR as automatic and reflexive phenomena, other researchers [5] suggested that, unlike facilitation, IOR is not purely automatic and is affected by experimental manipulations such as target-location predictability [6], task difficulty [7] and SOA range [8].

The interplay of facilitation and IOR has been tested in patients with schizophrenia, although the literature reflects contradictory evidence probably due to methodological differences and heterogeneity of the disorder. For instance, [2], [9] and [10] reported delayed onset or absent IOR in schizophrenia with the single-cue paradigm. On the other hand, [11], [12] and [13] found preserved IOR in schizophrenia patients when employing a second central cue to summon back attention to the centre. The different procedures used in the studies to measure IOR may account for the discrepant results.

In one of the above studies, [2] employed the single-cue IOR procedure in 11 medicated schizophrenia patients (6 females and five males) with four SOAs (100ms, 200ms, 700ms and 1200ms). The control participants of the study showed facilitation (shorter RTs in cued trials) at the 100ms SOA, but this facilitatory effect was no longer significant at the 200ms SOA, and instead a significant IOR effect was found at this interval value. [1]. However, schizophrenia patients exhibited longer lasting facilitation effect (it was significant at the 200ms SOA) and delayed onset of the IOR effect, which was significant only at the 700ms and 1200ms SOA. Furthermore, the magnitude of the effects was greater for facilitation, and smaller for the IOR effect in the group of patients as compared to the control participants. [2] concluded that the delayed IOR onset in schizophrenia patients might have been due to a difficulty to overcome the initial facilitatory effect of the cue. In line with [2], [9] employed the single-cue exogenous paradigm with two SOAs (100ms and 800ms) on 40 medicated schizophrenia patients and 34 healthy participants. The researchers examined the patients on two occasions, once during the acute state of the disorder and then during partial remission (12-16 weeks after). The results showed a disorder state-independent blunted IOR, that is patients had similar RTs for both cued and uncued targets in the long SOA (800ms) during both the initial and the follow-up examination after clinical improvement. [9] concluded that impaired IOR could be regarded as a trait characteristic of schizophrenia as therapy did not improve patients' performance.

Hence, the blunted IOR could reflect a biological vulnerability marker for the development of psychosis.

On the other hand [11] proposed an alternative explanation for the finding of blunted IOR in schizophrenia. They proposed that this deficit might reflect an impairment in the processes that reorient attention rather than a deficit in filtering information at the inhibited locations. [11] argued that certain psychiatric patients, such as patients with schizophrenia, fail to show IOR with the single cue procedure because they need a second central cue in order to reorient their attention back to the centre. The voluntary (endogenous) orientation of attention to the centre depends on the frontal eye fields and other frontal areas which are impaired in schizophrenia patients [11]. The researchers used the double cue procedure with two long SOAs (950ms, 1250ms) conditions and found normal IOR effects in a group of medicated schizophrenia patients. Similar effects were also reported by [14]. Taking into account the findings by [11], [12] and [14] that the employment of a second cue normalizes the differences between healthy and schizophrenic subjects in inhibitory effects, [13], argued that none of the above studies used different SOAs to probe the crosspoint (where facilitation changed to inhibition in the double cue paradigm). As a consequence [13] administered the double cue IOR procedure with 10 different SOAs (66,79, 106, 133, 159, 226, 305, 505, 705 and 1000ms) in order to measure facilitation, IOR and the crossover point where facilitation changes to inhibition in 14 medicated patients. By examining the intermediate SOAs, that had previously been excluded, the researchers demonstrated that patients showed prolonged facilitation relative to healthy subjects as the patients crossed from facilitation to inhibition at 450ms, while the crossover point of healthy subjects was at 226ms. These findings with the double cue paradigm lend support to previous results by [2] and [14] that point to the presence of increased facilitation in patients for short SOAs. Hence, mechanisms that lead to IOR are impaired in schizophrenia [13].

However, apart from the different experimental manipulations very few studies have considered potential differences between the patients in terms of symptomatology. [15] employed both cueing procedures (single vs double) in a group of medicated schizophrenia patients and assessed their performance taking into account their predominant symptomatology. In the single-cue condition the researchers reported greater facilitation effects in positive- relative to negative-symptom patients as well as reduced IOR effects in the negative-symptom group. In contrast, both positive- and negative-symptom patients showed intact IOR in the double-cue paradigm.

1.2 The Symptoms Approach in Chronic Schizophrenia

The clinical heterogeneity of schizophrenia has prompted many researchers to define types of the disorder. For example, [16] distinguished between positive and negative symptoms of schizophrenia. According to this model, positive symptoms (hallucinations, delusions, Formal Thought Disorder, bizarre behaviour) and negative symptoms (flat affect, apathy, alogia, avolition, anhedonia, inattention) represent two distinct pathological processes that are thought to be related to specific cognitive impairments. Recently, the validity of the positive/negative symptoms dichotomy in schizophrenia has been questioned in the sense that it did not explore whether or not particular symptoms were associated with external measures. To address this issue [17] examined correlations between symptoms in a group of 40 patients and found that symptoms segregated into three distinct clusters: the negative cluster, the reality distortion cluster (hallucinations, delusions) and the disorganization (FTD, bizarre behaviour) cluster.

Neuropsychological studies [18], [19], [20], [21], [22], [23] have shown that each cluster is characterized by a specific pattern of test performance and may even be related with a distinct pattern of brain impairments. Studies [19], [20] have suggested that: the negative-symptom cluster is associated with frontal lobe abnormalities such as functional hypofrontality (reduced metabolism or blood flow) [19], widening of the frontal interhemispheric fissure and abnomal frontotemporal connections, whereas the reality distortion cluster is associated with altered cerebral blood flow [20][22] in the temporal lobes. Most important, the neural correlates of the cluster appear to converge with the cognitive findings. That is specific psychotic symptoms are associated with specific cognitive deficits. Thus, the negative cluster is associated with more pronounced cognitive deficits in tests of verbal memory, language, executive function and visual-motor performance [20], while the disorganization cluster reflects a breakdown of inhibitory mechanisms in the semantic domain, as measured by executive tests such as the Stroop task and the Continuous Performance Test [24]. For instance, studies have linked the presence of formal thought disorder (FTD) to impaired processing of semantic information [25]. More specifically, [26] indicated that FTD is associated with increased semantic priming and lack of phonological inhibition. The failure of phonological inhibition, that is automatic in healthy participants, suggests that inhibitory processes in the semantic domain are dysfunctional in FTD. Furthermore, [27] administered the emotional Stroop test comprising conditions separated by both emotional valence (positive, negative) and arousal (low, high) to disorganized and non-disorganized schizophrenia patients and found that disorganized patients were unable to inhibit the influence of negative words in the incongruent condition ON WHAT?. [27] claimed that the above findings support the conclusion that the 'disorganization syndrome' is associated with disinhibition of cognitive control. Studies have not reported specific cognitive deficits in relation to the reality distortion cluster.

1.3 The Parietal Lobe in Schizophrenia: Relation with FTD Symptoms

The parietal lobe is involved in cognitive processes such as visuospatial attention (IOR) [28] and language processing [29]. [30] scanned 53 patients with schizophrenia and indicated significant grey matter volume reductions in the inferior parietal lobule which, together with frontal regions, subserve visuospatial processing. [28] concluded that impairments to inhibitory control are linked to processes of spatial attention, indicating dysfunctions of the frontoparietal network in patients with schizophrenia.

Apart from visuospatial attention, the parietal lobe (together with the planum temporale) contributes to the lexical-semantic network that controls semantic processing. The degree of reduction in planum temporale volume and surface area has been correlated with the extent of formal thought disorder symptoms [31]. In schizophrenia, there is evidence of reduced normal asymmetry (left bigger than right) in the planum temporale. [29] hypothesized that reduced or possibly reversed asymmetry would extend to the angular gyrus of the parietal lobe as both the parietal lobe and the angular gyrus are part of the lexical-semantic system that controls language and thought [29]. Indeed, [29] hypothesis was supported, as there was found to be an absence of normal left-greater-than-right asymmetry of gray matter in the parietal lobes of patients. The above findings provide evidence in support of a possible link between the manifestation of FTD with parietal lobe abnormalities.

1.4 The Present Study

The aim of the present study was to investigate whether schizophrenia patients' predisposition to Formal Thought Disorder would modulate performance on the IOR procedures. To our knowledge there are no previous studies that have examined the role of FTD in the manifestation of IOR.

We conducted two experiments –the single and the double cue IOR procedures, both with identical SOAs values. In the single-cue experiment we expected that schizophrenia patients will not show the IOR effect, in line with previous reports [2] as schizophrenia patients cannot disengage their attention voluntary from the cued location. In the double-cue experiment we expected the patients to show the inhibitory effect as the central cue summons back attention to the centre and elicits IOR [11].

Furthermore, we suggested that FTD global ratings would modulate patients' performance in SOAs values, in the sense that patients with high scores (above 2) on these items will show facilitation relative to patients with low scores that show IOR. As the disorganization cluster and more specificically FTD is related to a breakdown of inhibitory mechanisms in the semantic domain it was possible that FTD was also linked to impaired inhibitory processes in the visuospatial domain.

2. Method

2.1 Experiment 1 Single-Cue

2.1.1. Participants

24 male and 2 female patients with the diagnosis of schizophrenia participated. Patients were recruited from the psychiatric hospital 'Agia Fotini' and ranged in age from 24 to 54 years old (M= 36, SD= 12,02). All patients were medicated and signed an informed consent before their participation. Handedness was assessed through the Edinburgh Handedness Inventory. Except for one, all patients were right-handed. The inclusion criteria consisted of a diagnosis of schizophrenia according to the DSM-IV that was confirmed by the Mini International Neuropsychiatric Interview Greek version (M.I.N.I) (Papadimitriou et al, 2004) and documented evidence of regular antipsychotic drug therapy. Exclusion criteria were the history of head trauma, mental retardation, neurological disorder and drug or alcohol abuse. At the time of the assessment all patients had been stabilized on typical neuroleptic medication.

28 healhty adults participated in Experiment 1 (25 males, 3 females). Healthy participants were recruited from the staff of the Psychiatric Hospital 'Agia Fotini'. Although healthy volunteers were matched in age with the patients group (M=34, SD=10,07) they had more education years relative to patients (M=15, SD=5,04). Exclusion criteria were the history of mental or neurological disorder, head trauma, mental retardation and drug or alcohol abuse that were assessed through the M.I.N.I (Papadimitriou et al, 2004).

All participants had normal or corrected-to-normal eye vision and were naïve about the purpose of the experiment. Table 1 presents characteristics of the patient group.

	Mean	SD
Age	36	12,02
Education (in years)	11,66	3,97
Years of the disorder	12	12,59
Chlorpromazine-equivalent neuroleptic dose	8,66	4,87
Positive syndrome score	6,2	1,37
Negative syndrome score	8,32	5,04
Disorganization syndrome score	4,11	2,32

Table 1 Demographic and clinical data (N=26)

2.1.2. Materials and Stimuli

The scale for the assessment of Positive symptoms (SAPS) and the scale for the assessment of Negative symptoms (SANS) [16] were used to assess symptomatology. Positive symptoms used in the analysis consisted of global ratings of the severity of hallucinations, delusions, bizarre behaviour and formal thought disorder. Negative symptoms consisted of global ratings of the severity of alogia, apathy, affective flattening, anhedonia and attention. Global ratings were based on ratings of individual items which together reflect the severity of each symptom in a scale ranging from 0 (absent) to 5 (severe). Symptoms were evaluated by two psychiatrists of the hospital and inter-rater reliability for SAPS and SANS items was 0,79 and 0,82 respectively. The evaluation of symptoms was performed by the two psychiatrists at the same time and setting and the sample was consisted by 15 patients. In the correlation analysis we included both subscores and global ratings of SAPS and SANS items. The Neuropsychological function of patients was assessed through two executive-frontal tests: the verbal fluency and the Trial-making tests which both have been translated and validated in Greek population [32] [33]. The verbal fluency test consists of two parts, semantic and phonological. In the semantic part words have to be searched according to their semantic relationship through three categories (animals, fruits, objects) as compared to the phonological part where participants have to ignore semantic associations and select words on the basis of their first letter. The Trail-Making test consists of two parts (part A and part B). In Part A, the participant is presented with a sheet of paper with 25 scattered numbers from 1 to 25 enclosed in circles. The test requires participants to draw lines connecting circles in numerical order as quickly as possible. In Part B, the participant is presented with a sheet of paper which has 25 scattered circles containing numbers from 1 to 13 and letters from A to L. Participants are instructed to alternately connect numbers and letters in ascending order 1,A, 2, B, 3, C….as quickly as possible. The dependent variable is the amount of time the subject needs to complete the tests. The intellectual functioning of patients was assessed through the Raven Progressive Matrices test, that is an IQ test of non-verbal abilities and consists of 60 sequence of shapes and drawings with a missing part. Also we assessed patients' abnormal movements due to medication by using the AIMS scale (Abnormal Involuntary Movements Scale).

Stimuli were presented on a colour monitor (VGA) of an IBM/PC compatible computer, and responses were recorded through the computer keyboard. The software used for creating and running the experiment was E-Prime. The stimuli were the two boxes and the target was a white asterisk presented inside one of the boxes. Participants had to press the spacebar as soon as they detected the target.

Psychiatric symptoms and neuropsychological function were assessed first and participants performed the experiments during the following four days.

2.1.3. Procedure

Participants sat approximately 60cm from the computer and the experimenter explained the experiment verbally to them. Each trial began with a fixation point (a cross) presented in the middle of the screen for 1000ms. Participants were instructed to maintain their eyes at the fixation point during the experiment. The fixation point was followed by two white boxes were presented in the left and right of the cross for 1000ms. Then, one of the peripheral boxes became thicker for 50ms. This served as a cue to attract attention to the periphery. After a further time interval of 50ms, 90ms, 180ms, 300ms, 400ms, 3150ms, 3950ms (according to the SOA value) the target (a white asterisk) appeared inside one of the two lateral boxes. The target remained visible either until 2000ms or until a response was made (Figure 1).

Participants performed four identical ten-minutes sessions of each experiment. Each session included one practice block of 10 trials and two experimental blocks of 42 experimental trials and 15 catch trials (without target) to avoid anticipated responses. There were 12 trials of each SOA condition of which 6 were cued and six were uncued.

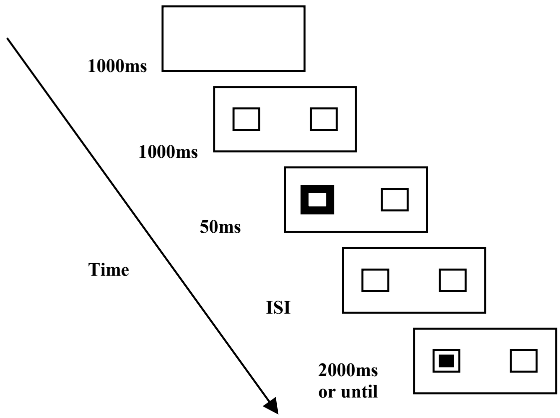


Figure 1 Sequence of events in a valid trial in Experiment 1

2.1.4. Results

The mean of the median correct response times and standard deviation are shown in Table 2. Response times were submitted to a repeated measures analysis of variance (ANOVA) with SOA (140ms, 230ms, 350ms, 450ms, 600ms, 2000ms, 3200) and target location (valid and invalid) as the within-subject factors.

Healthy Adults' Group

The analysis of the results showed a significant main effect of the SOA F(6,162)=3,371 p<.05. There was a significant two-way interaction between target location and SOA F(6,162)=10,863 p<.05.

Further analysis of the interaction using a t-test showed the following results: the mean number of RTs at cued locations (M=482,96, SD=92,49) and the mean number of RTs at uncued locations (M=500,27, SD=93,13) at the 140ms SOA value (t=-3.599, df=27, p<.05). The mean number of RTs at cued locations (M=474,46, SD=96,47) and the mean number of RTs at uncued locations (M=500,41, SD=102,52) at the 230ms SOA value differed significantly (t=-4,303, df=27,p<.05). So, there was a significant RT advantage for target detection to cued relative to uncued locations at the 140ms (-18ms) and at the 230ms (-26ms) SOAs values. Also, there was found a significant difference in the mean RTs between cued (M=500,74, SD=85,73) and uncued locations (M=489,78, SD=89,43) at the 600ms SOA value (t=2,313, df=27, p<.05). The mean number of RTs at cued locations (M=502,65, SD=80,22) and the mean number of RTs at uncued locations (M=490,16, SD=79,04) at the 2000ms SOA differed significantly (t=2,059, df=27, p<.05). Finally, the mean number of RTs at cued locations (M=503,58, SD=83,49) and the mean number of RTs at uncued locations (M=456,47, SD=79,16) at the 3200ms SOA value differed significantly (t=2,119, df=27, p<.05). So, there was a RT advantage for targets presented at uncued locations as compared to cued locations at the SOAs of 600ms (11ms), 2000ms (12ms) and 3200ms (7ms).

	140 ms	230 ms	350 ms	SOA 450 ms	600 ms	2000 ms	3200 ms
Cued	482 (92,49)	474 (96,47)	479 (90,62)	480 (91,09)	500 (85,73)	502 (80,22)	503 (83,49)
Uncued	500 (93,13)	500 (102,52)	483 (95,32)	485 (90,91)	489 (89,43)	490 (79,04)	496 (79,16)
Cue Effect	-18*	-26*	-4	-5	11*	12*	7*

Note: Cue effect = RTcued-RTuncued; * p < 0.05.

Table 2 Mean of the median correct RTs and SD as a function of target location and SOA value

Patients' Group

The analysis of the results showed significant main effects of target location F(1,25)=27. 603 p<,05 and SOA F(6,150)=2.601 p<.05. That is RTs were faster at valid

(504ms) relative to invalid (522ms) target locations. Also, there was a significant two-way interaction between target location and SOA F(6,150)= 5.582 p<.05.

Further analysis of the interaction using a t-test showed the following results: the mean number of RTs at cued locations (M=500ms, SD=105,11) and the mean number of RTs at uncued locations (M=530ms, SD=112,91) at the 140ms SOA value differed significantly (t=-4.421, df=25, p<.05). The mean number of RTs at cued locations (M= 482ms, SD=106) and the mean number of RTs at uncued locations (M=518ms, SD=113) at the 230ms SOA value differed significantly (t=-4.617, df=25, p<.05). Also, the mean number of RTs at cued locations (M=492,73, SD=105,28) and the mean number of RTs at uncued locations (M=521,65, SD=112,79) differed significantly at the 350ms SOA value (t=-4.264, df=25, p<.05). Finally, the mean number of the cued RTs (M=509,57, SD=109,97) and the mean number of the uncued RTs (M=527,23, SD=113,38) at the 450ms SOA value differed significantly (t=-2.051, df=15, p<.05). The above results suggest that there was a RT advantage for target detection of cued locations at the 140ms (-30ms), the 230ms (-36ms), the 350ms (-29ms) and the 450ms (-18ms) SOAs values.

	SOA						
	140ms	230ms	350ms	450ms	600ms	2000ms	3200ms
Cued	500 (105,11)	482 (106,59)	492 (105,28)	509 (109,97)	515 (95,54)	519 (83,55)	516 (78,82)
Uncued	530 (112,91)	518 (113,73)	521 (112,79)	527 (113,38)	526 (97,39)	513 (87,91)	519 (89,41)
Cue Effect	-30*	-36*	-29*	-18*	-11	6	-3

Note: Cue effect = RTcued-RTuncued; * p < 0.05.

Table 3 Mean of the median correct RTs and SD as a function of target location and SOA value

We correlated each symptom global rating individually with cueing scores (Cued RTs-Uncued RTs) at each SOA value. The results from the bivariate correlational analysis indicated a significant negative correlation between Formal Thought Disorder (FTD) rating and cueing effect at the 2000ms SOA, r =0.65 p<.05. We did not find any significant correlation between symptoms global ratings and neuropsychological test (verbal fluency, Trail Making, Raven Progressive Matrices).

2.1.5 Discussion

As compared to healthy adults, schizophrenia patients showed prolonged facilitation and absence of IOR in the single-cue procedure [2]. The lack of significant IOR for the longer SOA conditions may be due to impairments in the processes that reorient attention rather than to a deficit to attend to information that appear at the inhibited location. On the other hand, it may be due to increased facilitation in earlier SOA that would mask any potential IOR effects at those SOAs.

Taking into account the role of symptoms in the development of IOR, we observed a negative correlation between FTD ratings and cueing effects at the 2000ms SOA. That means that FTD patients continued showing facilitation at the 2000ms SOA, as compared with non-FTD patients who showed inhibition at 2000ms SOA. FTD probably modulated patients' performance at the 2000ms SOA because at all previous short SOAs schizophrenia patients showed significant facilitatory effects. As it seems, patients started overcoming the initial facilitation of the cue around the 2000ms SOA, so at this point the influence of FTD became evident.

We conducted experiment 2 (double-cue) IOR to replicate previous findings [11] that suggest that the use of a central cue helps patients to exhibit IOR. In addition, we aimed to see whether symptoms would modulate patients' performance despite the use of a central cue. To our knowledge all previous studies [11], [14] that have employed a central cue reported normal IOR effects in schizophrenia but none of them has looked the influence of specific symptoms. In experiment 2 we expected that patients would not show prolonged facilitation and would exhibit the IOR effect as the central cue attracts attention from the periphery. However, if we were to find a similar association between the cueing effect and FTD as in Experiment 1, we may conclude that, regardless of the central cue, the presence of these symptoms disrupt inhibitory processes in schizophrenia.

2.2 Experiment 2 Double-Cue

2.2.1 Participants

24 patients (M= 36, SD=11,09) with schizophrenia from Experiment 1 and 21 healthy adults (M=33, SD=12,02) from Experiment 1 participated in Experiment 2.

2.2.2 Materials and Stimuli

We used the same clinical (SAPS/SANS) and neuropsychological (Verbal fluency, Trail Making, Raven) measures employed in Experiment 1. The stimuli of the experiment were created through the E-Prime Software and were presented in a colour monitor compatible computer. Responses were recorded through the computer keyboard.

2.2.3 Procedure

Participants sat approximately 60cm from the computer and the experimenter explained them the experiment verbally. Each trial began with a fixation point (a cross) presented in the middle of the screen for 1000ms Participants were instructed to look at the fixation point during the duration of the experiment. Then two white boxes were presented in the left and the right of the cross for 1000ms. Then, one of the peripheral boxes became thicker for 50ms. After 40ms the second cue (a brightening of a central box) was presented for 50ms in order to summon back attention to the centre. After a further time interval of 0ms, 90ms, 210ms, 310ms, 460ms, 1860ms or 3060ms, according to the value of the SOA the target (a white asterisk) appeared inside one of the peripheral boxes. The target remained visible either until 2000ms or until response (Figure 2).

Participants run four identical ten-minutes sessions of each experiment. Each session included one practice block of 10 trials and two experimental blocks of 42 experimental trials and 15 catch trials (without target) to avoid anticipated responses. There were 12 trials of each SOA condition of which 6 were cued and six were uncued in each experimental block.

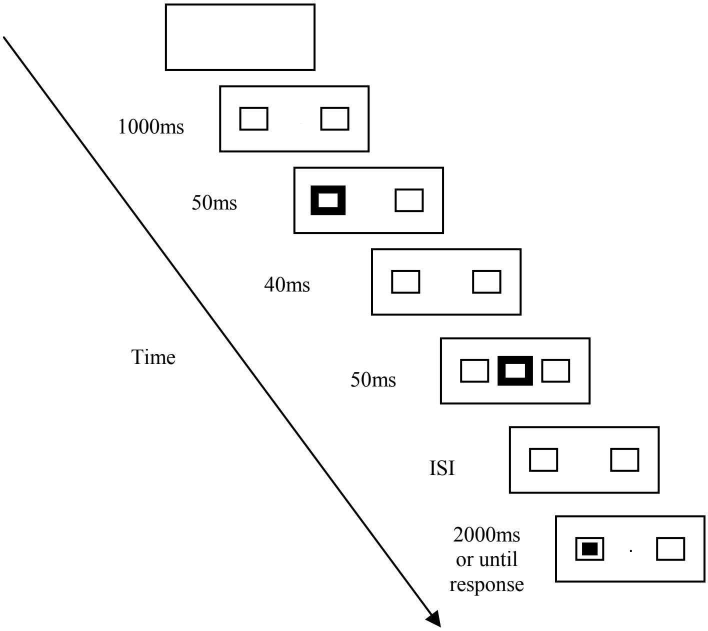


Figure 2 Sequence of events in a valid trial in experiment 2

2.2.4 Results

The mean of the median correct RTs and SD are shown in Table 3. Response times were submitted in a Repeated Measures Analysis of Variance (ANOVA) with SOA (140ms, 230ms, 350ms, 450ms, 600ms, 2000ms, 3200ms) and target location (valid, invalid) as within-subject factors.

Healthy Adults' Group

The analysis of the RTs showed significant main effects of target location $F(1,20)=12{,}292$ $p<.05$ and SOA $F(6,120)=32{,}760$ $p<.05$. That is RTs were faster to uncued (403ms) relative to cued (465ms) locations. There was observed a significant two-way interaction between target location and SOA $F(6,120)=3{,}187$, $p<.05$.

Further analysis of the interaction by using a t-test showed the following results: the mean number of RTs at cued locations (M=445, SD=63,77) and the mean number of RTs at uncued locations (M=434, SD=61,18) at the 350ms SOA value differed significantly (t=2.743, df=20, p<.05). Also, the mean cued RTs (M=454, SD=67,67) and the mean uncued RTs (M=445, SD=58,05) at the 450ms SOA value differed significantly (t=1.77, df=20, p<.05). Finally, there was found a significant difference between the mean cued RTs (M=476, SD=62,68) and the mean uncued RTs (M=460, SD=63,36) at the 600ms SOA value (t=4.595, df=20, p<.05). The above results suggest that there was a significant RT advantage for target detection at uncued as compared to cued locations at the SOAs values of 350ms (11ms), 450ms (11ms) and 600ms (16ms).

	SOA						
	140ms	230ms	350ms	450ms	600ms	2000ms	3200ms
Cued	464 (59,11)	449 (65,41)	445 (63,77)	454 (67,67)	476 (62,68)	481 (57,96)	486 (53,74)
Uncued	471 (71,48)	440 (71,26)	434 (61,18)	445 (58,05)	460 (63,36)	478 (59,95)	489 (56,16)
Cue Effect	-7	9	11*	11*	16*	3	-3

Note: Cue effect = RTcued-RTuncued; * p < 0.05.

Table 4 Mean of the median correct RTs and SD as a function of target location and SOA value

Patients' Group

The analysis of the results showed significant main effects of target location F(1,23)=3,584 p<.05 and SOA value F(6,138)=2,986 p<.05. That is RTs were faster to uncued (554ms) as compared to cued (560ms) locations. Also, there was found a significant two-way interaction between target location and SOA F(6,138)=1,352 p<.05.

Further analysis of the interaction by a T-Test showed the following results: the mean number of RTs at the cued location (M=554, SD=116,64) and then mean number of RTs at the uncued locations (M=542, SD=113,95) at the 450ms SOA value differed significantly (t=2.513, df=23, p<.05). Also, the mean number of cued RTs (M=575, SD=110,36) and the mean number of uncued RTs (M=560, SD=106,037) at the 600ms SOA value differed significantly (t=2.530, df=23, p<.05). The above results showed that there was a significant RTs advantage for uncued relative to cued locations at the 450ms (12ms) and at the 600ms (15ms) SOAs values.

	SOA						
	140ms	230ms	350ms	450ms	600ms	2000ms	3200ms
Cued	577 (137,33)	553 (122,75)	552 (109,23)	554 (116,64)	575 (110,36)	553 (87,43)	558 (103,06)
Uncued	580 (135,45)	556 (126,08)	545 (115,98)	542 (113,95)	560 (106,03)	550 (92,15)	549 (87,39)
Cue Effect	-3	-3	7	12*	15*	3	9

Note: Cue effect = RTcued-RTuncued; * p < 0.05.

Table 5 Mean of the median correct RTs and SD as a function of target location and SOA value

The bivariate correlation analysis between symptoms global ratings and cueing effects showed a significant negative correlation between FTD and cueing at the 2000ms SOA, *r*=-0.733, p < .05. We did not observe any significant correlation between symptoms global ratings and neuropsychological test (verbal fluency, Trail Making, Raven

Progressive Matrices)> We were interested in correlating these because we wanted to see whether specific symptoms are related to specific performance in these tests. More specific, verbal fluency and Trail Making tests are executive tests, mediated through frontal lobes, and as negative symptoms are related to frontal lobes we expected an association between these two.

2.2.5 Discussion

The findings from experiment 2 suggest that, the central cue was effective for patients to exhibit IOR effects [11]. Although, in comparison to healthy participants (350ms SOA), patients showed later onset of IOR (450ms SOA) the central cue enabled schizophrenia patients to overcome the initial facilitation and re-orient their attention to the centre. The different cross-over point between healthy participants and schizophrenia patients is in line with previous research [13] which had shown that even with the double-cue IOR procedure patients display later onset of IOR due to prolonged facilitation.

In line with experiment 1, FTD ratings modulated patients' performance at the 2000ms SOA value. That means that high scores of FTD were associated with facilitation while low scores were associated with IOR. These findings further support our hypothesis that FTD is associated with increased facilitation at the IOR tasks both with the single (voluntary disengagement) and the double-cue (automatic disengagement) procedures.

3. General Discussion and Conclusions

A new finding that emerged from our experiments is the different time-course of IOR in healthy individuals between single and double-cue procedures. There was found an earlier onset of IOR (350ms SOA) in the double as compared to the single-cue (600ms SOA) experiment but IOR diminished earlier with the double-cue (600ms SOA) relative to the single-cue (3200ms SOA). Although [34] indicated that the presence of a central cue does not influence IOR in detection tasks, our findings suggest that IOR is not as robust as it was thought at the double-cue procedure. It remains to be shown which factors determine the time course facilitation and IOR under specific experimental manipulations.

Schizophrenia patients showed prolonged facilitation and diminished or absent IOR in the single cue paradigm (Experiment 1), in line with previous findings [2] [14]. It seems with the single cue procedure patients cannot overcome the initial facilitation of the valid cue. Early facilitation and late inhibition are developed in parallel but at short SOAs the facilitatory effect may mask inhibition. However, in schizophrenia patients' facilitation is much more enhanced and prolonged [2] so that inhibition cannot be observed with the single cue procedure. On the contrary, the employment of a second central cue helped schizophrenia patients to disengage their attention from the cued location and re-orient to the centre which is a necessary condition for the development of IOR [11].

In both Experiments a negative correlation emerged between cueing effects at the 2000ms SOA and FTD. These results suggest that FTD modulated patients' performance in both single and double cue IOR tasks. Patients with high ratings in FTD item showed greater facilitation at the 2000ms SOA contrary to low FTD ratings patients who showed greater inhibition. The above results are in line with our hypothesis that Formal Thought Disorder in schizophrenia is related to prolonged facilitatory effects.

There are two possibilities why FTD disrupted IOR in patients. First of all, it could be that apart from the semantic domain [25] FTD is also associated with a breakdown of inhibitory mechanisms in the spatial domain. Evidence from the semantic priming paradigm and the emotional Stroop test [27] suggest that disorganized patients are impaired in measures of cognitive inhibition. If this is the case our results suggest that apart from diminished inhibitory function in measures of semantic memory or selective attention, FTD is also related to impaired visuospatial inhibitory processes. This would be in agreement with the evidence that support the role of the parietal lobe in mediating IOR; one of the cortical areas that has been suggested to be dysfunctional in FTD patients. However, another explanation for our findings could be that facilitation masks IOR at the 2000ms SOA in patients with high FTD. If this is the case, the lack of IOR for these patients would be due to increased facilitation and not to the breakdown of visuospatial inhibitory processes. Further research is needed in order to determine the relationship between disorganization symptoms, brain abnormalities in the parietal lobe and the absence of IOR in schizophrenia.

References

1 Posner M I & Cohen Y. Components of visual orienting. (in H. Bouma & D. Bouwhuis Eds.) Attention and Performance London: Erlbaum 1984

2 Huey E D &Wexler B E . Abnormalities in rapid automatic aspects of attention in schizophrenia: Blunted Inhibition of Return. Schizophrenia Research 1994; 14*:* 57-63

3 Posner M I. Orienting of Attention. Quarterly Journal of Experimental Psychology 1980; 32: 3-25

4 Posner, M.I., Early, T. S., Peiman, E., Pardo, P. & Dhawan, M. Asymmetries in hemispheric control of attention in schizophrenia, Archives of General Psychiatry, 45, 814-821

5 Maruff, P., Yucel, M., Dancert, Stuart, G. & Currie, J. (1999) Facilitation and inhibition arising from the exogenous orienting of covert attention depends on the temporal properties of spatial cues and targets, Neuropsychologia, 37, 731-744

6 Wright, R.D. & Richard, C.M. (2000) Location cue validity affects inhibition of return of visual processing, *Vision Research, 40,* 2351-2358

7 Lupianez, J., Milan, E.G., Tornay, F.J., Madrit, E. & Tudela, P. (1997)Does IOR occur in discrimination tasks? Yesit does, but later, Perception and Psychophysics, 59, 1241-1259

8 Cheal, M. & Chastain, G. (2002) Timing of facilitatory and inhibitory effects of visual attention, Visual Cognition, 9 (8), 969-1002

9 Gouzoulis-Mayfrank, E., Heekeren, K., Voss, T., Moerth, D., Thelen, B. & Meincke, U. (2004). Blunted inhibition of return in schizophrenia. Evidence from a longitudinal study. Progress in Neuro-Psychopharmacology & Biological Psychiatry, 28, 389-396

10 Nestor, P.G., Faux, S.F., McCarley, R.W., Penhune, V., Shenton, M.E., Pollak, S. & Sands, S.F. (2000) Attentional cues in chronic schizophrenia. Abnormal disengagement of attention, Journal of Abnormal Psychology, 101(4), 682-689
11 Fuentes, L.J. & Santiago, E. (1999). Spatial and semantic inhibitory processing in schizophrenia. Neuropsychology, 14(1), 134-140
12 Sapir, A., Henik, A., Dobrusin, M. & Hochman, E.Y. (2001). Attentional assymetry in schizophrenia. Disengagement and inhibition of return deficits. Neuropsychology, 15(3), 361-370
13 Larrison-Faucher, A., Briand, K.A. & Sereno, A.B. (2002). Delayed onset of inhibition of return in schizophrenia. Progress in Neuro-Psychopharmacology & Biological Psychiatry, 26, 505-512
14 Fuentes, L.J., Boucart, M., Alvarez, R., Vivas, A.B. & Zimmerman, M.A. (1999) Inhibitory processing in visuospatial attention in healthy adults and schizophrenic patients, Schizophrenia Research, 40(1), 75-80
15 Fuentes, L.J. & Santiago, E. (2002). Inhibitory mechanisms deficits in schizophrenics with either positive or negative symptoms. Paper presented at the Fourth Conference of Spanish Experimental Society, Oviedo, Spain
16 Andreasen, N.C. & Olsen, S. (1982) Negative vs positive schizophrenia, Archives of General Psychiatry, 39, 789-794
17 Liddle, P.F. (1987) The symptoms of chronic schizophrenia. A re-examination of the positive-negative dichotomy, The British Journal of Psychiatry, 151, 145-151
18 Basso, M.R., Nasrallah, H.A., Olson, S.C. & Bornstein, R.A. (1998) Neuropsychological correlates of negative, disorganised and psychotic symptoms in schizophrenia, Schizophrenia Research, 31, 99-111
19 Kaplan, R.D., Szechtman, H., Franco, S., Szechtman, B., Nahmias, C., Garnett, E.S., List, S. & Cleghorn, J.M. (1993) Three clinical syndromes of schizophrenia in untreated subjects: relation to brain glucose activity measured by positron emission tomography (PET), Schizophrenia Research, 11, 47-54
20 Cuesta, M.J. & Peralta, V. (1995) Cognitive disorders in the positive, negative and disorganization syndromes schizophrenia, Psychiatry Research, 58, 227-235
21 Schrooler, C., Neumann, E., Caplan, L.J. & Roberts, B.R. (1997) A time course analysi of stroop interference and facilitation comparing normal individuals and individuals with schizophrenia, Journal of Experimental Psychology: General, 126(1), 19-36
22 Liddle, P.F. & Morris, D.L. (1991) Schizophrenic syndromes and frontal lobe performance, British Journal of Psychiatry, 158, 340-345
23 Liddle, P.F., Friston, K.J., Ftith, C.D., Hirsch, S.R., Jones, T. & Frackowiak, S.J. (1992) Patterns of cerebral blood flow in schizophrenia, British Journal of Psychiatry, 160, 179-186
24 Jensterle, J., Mlakar, J., Vodusek, D.B. & Frith, C.D. (2000) Disorganisation in schizophrenia need not result from a failure to inhibit dominant response tendencies, Cognitive Neuropsychiatry, 5(2), 105-121
25 Kerns, J.G., Berenbaum, H., Barch, D.M., Banich, M.T. & Stolar, N. (1999) Word production in schizophrenia and relationship to positive symptoms, Psychiatry Research, 87, 29-37
26 Spitzer, M., Braun, U., Hermle, L. & Maier, S. (1993) Associative semantic network dysfunction in thought- disordered schizophrenic patients: Direct evidence from indirect semantic priming, Biological Psychiatry, 34, 864-877

27 Phillips, L.K., Deldin, P.J., Voglmaier, M.M. & Rabbitt, S. (2005) Emotional stroop performance predicts disorganization in schizophrenia, Schizophrenia Research, 77, 141-149
28 Danckert, J., Saoud, M. & Maruff, P. (2004) Attention, motor control and motor imagery in schizophrenia: Implications fort he role of the parietal cortex, Schizophrenia Research, 70, 241-261
29 Niznikiewicz, M., Donnino, R., McCarley, R.W., Nestor, P.G., Losifescu, D.V., O'Donell, B., Levitt, J. & Shenton, M.E. (2000) Abnormal angular gyrus asymmetry in schizophrenia, American Journal of Psychiatry, 157, 428-437
30 Zhou, S., Suzuki, M., Takahashi, T., Hagino, H., Kawasaki, Y., Matsui, M., Seto, H. & Kurachi, M. (2007) Parietal lobe volume deficits in schizophrenia spectrum disorders, Schizophrenia Research, 89, 35-48
31 McCarley, R.W., Salisbury, D.F., Hirayasu, Y., Yurgelun-Todd, D.A., Hohen, M., Zarate, C., Kikinis, R., Jolesz, F.A. & Shenton, M.E. (2002) Association between smaller left posterior superior temporal gyrus volume on Magnetic Resonance Imaging and smaller left temporal P300 amplitude in first-episode schizophrenia, Archives of General Psychiatry, 59(4), 321-331
32 Kosmidis, M.H., Vlachou, C.H., Panagiotaki, P. & Kiosseoglou, G. (2004) The verbal fluency task in the Greek population: Normative data, and clustering and switching strategies, Journal of the International Neuropsychological Society, 10, 164-172
33 Vlachou, C.E & Kosmidis, M.H. (2002) Η ελληνικη δοκιμασια οπτικο-νοητικης ιχνηλατησης, Προκαταρκτικες νορμες για κλινικη και ερευνητικη εφαρμογη, Ψυχολογια, 9(3), 336-352
34 Prime, D.J., Visser, T. A. & Ward, L.M. (2006) Reorienting attention and inhibition of return, Perception & Psychophysics, 68 (8), 1310-1323