14TH ANNUAL EUROMEDIA CONFERENCE

2008

PORTO, PORTUGAL

APRIL 9-11, 2008

Organized by

ETI

Sponsored by

EUROSIS

TTVI

EU-DG INFSO

BELGACOM

GHENT UNIVERSITY

IDMEC

INEGI

HOSTED BY

UNIVERSITY OF PORTO

EUROMEDIA'2008

FEATURING

FOURTEENTH ANNUAL SCIENTIFIC CONFERENCE
ON WEB TECHNOLOGY, NEW MEDIA
COMMUNICATIONS AND TELEMATICS THEORY
METHODS, TOOLS AND APPLICATIONS
MEDICAL IMAGING
AND D-TV

João Manuel R. S. Tavares

and

Renato Natal Jorge

APRIL 9-11, 2008
PORTO, PORTUGAL

A Publication of EUROSIS-ETI

Printed in Ghent, Belgium

EXECUTIVE EDITOR

PHILIPPE GERIL
(BELGIUM)

Editors

João Manuel R. S. Tavares
FEUP, University of Porto
Porto
Portugal

Renato Natal Jorge
FEUP, University of Porto
Porto
Portugal

D-TV Workshop Editor

Dr. H. Joachim Nern
TTVI
Dusseldorf, Germany

Programme Committee

WEBTEC Programme Committee

Sameh Abdel-Naby, University of Trento, Trento, Italy
Dr. Paul Dowland, University of Plymouth, Plymouth, United Kingdom
Dipl.-Inf. Steffen Harneit, CUTEC-Institute GmbH, Clausthal-Zellerfeld, Germany
Ass. Prof. Qingping Lin, Nanyang Technological University, Singapore
Prof. Dr. Jörn Loviscach, Hochschule Bremen, University of Applied Sciences, Bremen, Germany
Assoc. Prof. Wenji Mao, Institute of Automation, Chinese Academy of Sciences, Beijing, P.R. China
Lorenzo Motta, Ansaldo Segnalamento Ferroviaro s.p.a. Genova, Italy
Dr. H. Joachim Nern, Aspasia Knowledge Systems, Dusseldorf, Germany
Dr. Carlos E. Palau, Universidad Politecnica de Valencia, Valencia, Spain
Prof. Paola Salomoni, Universita di Bologna, Bologna, Italy
Dr. Elpida Tzafestas, National Technical University of Athens, Athens, Greece
Dr. Matthew Warren, Deakin University Geelong, Victoria, Australia

MEDIATEC Programme Committee

Assoc. Prof. Vincent Charvillat, IRIT-ENSEEIHT, Toulouse cedex, France
Dr. Fernando Boronat Segui, Universidad Politecnica de Valencia, Gran de Gandia, Spain
PhD. Jens Mueller-Iden, University of Münster, Münster, Germany
Dr. Ana Pajares, Universidad Politecnica de Valencia, Valencia, Spain
Jehan Francois Paris, University of Houston, Houston, USA
Prof. Marco Roccetti, Universita' di Bologna, Bologna, Italy
Dr. Leon Rothkrantz, Delft University of Technology, Delft, The Netherlands

COMTEC Programme Committee

Prof. Dr. Marwan Al-Akaidi, De Montfort University, Leicester, United Kingdom
Boguslaw Butrylo, Bialystok Technical University, Bialystok, Poland
Dr. Nathan Clarke, University of Plymouth, Plymouth, United Kingdom
Dr. Steven Furnell, University of Plymouth, Plymouth, United Kingdom
Prof. Chris Guy, The University of Reading, Reading, United Kingdom
PhD Mohammad Riaz Moghal, Ali Ahmad Shah-University College of Engineering and Technology, Mirpur, Pakistan
Ph. D. Oryal Tanir, Bell Canada, Montreal, Canada
Ass. Prof. Vassilis Triantafillou, Technological Educational Institution of Messolonghi Applied, Greece

INTERNATIONAL PROGRAMME COMMITTEE

APTEC Programme Committee

Prof. Dr. J. Broeckhove, RUCA-UA, Antwerp, Belgium
Dr. Juan Carlos Guerri Cebollada, Universidad Politecnica de Valencia, Valencia, Spain
Dr.ir. Johan Opsommer, Belgacom - BUS, Brussels, Belgium
Prof. Matthias Rauterberg, Eindhoven University of Technology, Eindhoven, The Netherlands.
Francisco Reinaldo, FEUP, University of Porto, Porto, Portugal
Prof. Jeanne Schreurs. Hasselt University, Diepenbeek, Belgium
Ass. Prof. Ramiro Velázquez, Universidad Panamericana, Aguascalientes, Mexico
Dr. Charles van der Mast, Delft University of Technology, Delft, The Netherlands

E-TEC Programme Committee

Dr.Steven Furnell, University of Plymouth, Plymouth, United Kingdom
Dr. Paul Dowland, University of Plymouth, Plymouth, United Kingdom

Knowledge Management and E-Mobility

Prof. Ricardo Chalmeta, Universidad Jaume I, Castellon, Spain
Prof. Dr.-Ing. Stephan Kassel, University of Applied Sciences Zwickau, Germany

Medical Imaging Systems

General Chair
João Manuel R. S. Tavares, FEUP, University of Porto, Porto, Portugal
General Co-Chair
Renato Natal Jorge, FEUP, University of Porto, Porto, Portugal

International Programme Committee

Alberto De Santis, Università degli Studi di Roma "La Sapienza", Italy
Arrate Muñoz Barrutia, University of Navarra, Spain
Behnam Heidari, University College Dublin, Ireland
Bernard Gosselin, Faculte Polytechnique de Mons, Belgium
Chandrajit Bajaj, University of Texas, USA
Christos E. Constantinou, Stanford University School of Medicine, USA
Daniela Iacoviello, Università degli Studi di Roma "La Sapienza", Italy
Dinggang Shen, University of Pennsylvania, USA
Djemel Ziou, University of Sherbrooke, Canada
Gerald Schaefer Aston University, United Kingdom
João Krug Noronha, Dr. Krug Noronha Clinic, Portugal
João Manuel R. S. Tavares, Faculty of Engineering of University of Porto, Portugal
João Paulo Costeira, Instituto Superior Técnico, Portugal
Jorge M. G. Barbosa, Faculty of Engineering of University of Porto, Portugal
Lyuba Alboul, Sheffield Hallam University, United Kingdom
Manuel González Hidalgo, Balearic Islands University, Spain
Maria Elizete Kunkel, Universität Ulm, Germany
Mário Forjaz Secca, Universidade Nova de Lisboa, Portugal
Miguel Angel López, Faculty University of Ciego de Avila, Cuba
Miguel Velhote Correia, Faculty of Engineering of University of Porto, Portugal
Patrick Dubois, Institut de Technologie Médicale, France
Reneta Barneva, State University of New York, USA
Renato M. Natal Jorge, Faculty of Engineering of University of Porto, Portugal
Sabina Tangaro, University of Bari, Italy
Valentin Brimkov, State University of New York, USA
Yongjie Zhan, Carnegie Mellon University, USA

D-TV Workshop

Dr. Hans-Joachim Nern, TTVI, Germany

All articles published in these Proceedings have been peer reviewed

EUROSIS-ETI Publications are ISI-Thomson and INSPEC referenced

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

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

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

EUROSIS-ETI Publication

ISBN: 978-9077381-38-0
EAN : 978-9077381-38-0

Preface

The EUROMEDIA conference is an annual meeting for dissemination of the state-of-the-art in multimedia research, technology, management and art. As in previous years, the conference seeks to bring together researchers and practitioners in academia and industry, who are interested in exploring and exploiting new and multiple media to create new capabilities for human expression, communication, collaboration, and interaction. EUROMEDIA intends to cover a broad of aspects of multimedia computing: theory to practice, from underlying technologies to applications. The present event is no exception, providing the technical programme an ideal forum for the presentation and exchange of research relating to the design and use of state-of-the-art multimedia and networked systems.

The EUROMEDIA 2008 conference was held, in the University of Porto, Portugal, during the period of 9-11 April 2008, concurrently with the ECEC and FUBUTEC conferences, being structured with three main tracks (WEBTEC - which deals with web technology, MEDIATEC - which covers multimedia technology, and APTEC - which provides an overview of media-integration) and two special workshops: 1st Workshop on Medical Imaging Systems and 2nd Workshop on Digital Television & Digital Special Interest Channels.

The EUROMEDIA 2008 conference brought together several researchers representing several fields related to web technology, multimedia technology, media-integration, communications technology, medical imaging systems, digital television, digital special interest channels and human computer interaction.

This book contains the 35 full papers presented at EUROMEDIA 2008 conference that came from 13 countries: Bulgaria, France, Germany, Greece, Iran, Japan, Malaysia, Netherlands, Poland, Portugal, Russia, Tunisia and United Kingdom.

We would like to thank to Philippe Geril, whose continued dedication and hard work as the conference organiser has enabled us to maintain the standard expected of EUROMEDIA 2008 conference, to The European Multidisciplinary Society for Modelling and Simulation Technology for the opportunity to be involved in the organization of EUROMEDIA 2008, to Faculdade de Engenharia da Universidade do Porto for hosting EUROMEDIA 2008, to all members of the Scientific Committee for their reviews and significant contribution for the high quality standards of EUROMEDIA 2008, to all sessions chairs for their effort for the smooth running of all scientific sessions of EUROMEDIA 2008, to our Invited Lecturer and to all Authors for sharing their excellent works during EUROMEDIA 2008 and to all attendees that enrich and validate the purposes of EUROMEDIA 2008.

João Manuel R. S. Tavares
Renato M. Natal Jorge
Faculty of Engineering, University of Porto, Portugal
General EUROMEDIA 2008 Chairs

CONTENTS

CONTENTS

DATA MANIPULATION

APPLIED MEDIA TECHNOLOGY

MEDICAL IMAGING

CONTENTS

D-TV

SCIENTIFIC PROGRAMME

WEB-BASED SERVICES AND VIRTUAL WORLDS

A CASE STUDY OF USING WEB-BASED SERVICES IN HIGHER EDUCATION

Hani Alers and Charles van der Mast
Section of Man Machine Interaction
Delft University of Technology
Mekelweg 4, 2628 CD
Delft, The Netherlands
E-mail: hani.alers@gmail.com | c.a.p.g.vandermast@tudelft.nl

KEYWORDS
E-learning, podcasts, collaboration, screencasts, web-lectures, educational tools.

ABSTRACT

This paper describes the motivation and research results of the use of new web-based technology to enhance the learning and teaching processes in higher education. The basic idea is to exploit the ubiquitous nature of the Internet to assist in instructing the students and interacting with them. This is accomplished by presenting them with rich media content and rapid communication channels that help them achieve the required educational goals while providing them with an enhanced learning experience. Our pilot experiment, discussed here, has shown that both students and instructors can benefit from using such educational tools.

INTRODUCTION

The field of higher education has produced many new Internet based tools and methods aimed at supporting teaching and learning. It is expected that younger students who grew up with the Internet are well trained in using many of the interactive tools and services it offers. Educators have therefore tried to exploit the experience they have by providing web-based tools to support students in learning as individuals and in groups (Van Aalst and Van der Mast 2003, Koppelman et al. 2000, Repenning et al. 2001). However, the development and the deployment of such educational software have proven to be a complex affair (Van der Mast. and Van Aalst 2002). It is not easy to use such tools and methods without a structured model of the content and didactic strategies, a careful selection of the used media, and the ways to optimize the engagement and attitude of the learners (Van der Mast 1995, Salomon et al. 1991). As a result, educators today tend to ignore most of the potential offered by the Internet, and mainly rely on traditional teaching approaches (Gagnon and Krovi 1999).

When developing a new course on human-computer interaction (HCI), we attempted to build a support structure for the course using web-based tools. We felt that such an approach was necessary since research has suggested that traditional methods of teaching HCI and usability are not as successful as generally thought (Kotzé and Oestreicher 2002). The course is called Intelligent User eXperience Engineering (IUXE), and teaches design and evaluation methodologies for developing user interfaces.

The goal of this paper is to describe the way these tools were implemented and how they were used by the students as individuals and in teams. We present subjective (questionnaires) and objective (recordings and loggings) measures to evaluate the way the tools were used (Karoulis and Pombortsis 2000, Russell 2001). The results of the experiment show how the tools were used by the students, and how students evaluated the approach used in this course compared to traditional educational methods.

The paper starts by giving an overview of the IUXE course and explain the main educational goals that the tools are meant to achieve. This is followed by a look at how the educational tools are implemented and the technology they are based on. Next, the results of the experiment are discussed by looking at the gathered data. The paper then ends with a discussion and some final conclusions.

THE COURSE IUXE

The IUXE course is given within the International Master program of the Delft University of Technology. It is one of the Human-Computer Interaction Specialization Courses offered by the Media & Knowledge Engineering program. The course provides a coherent engineering approach to human-computer interaction. It discusses guidelines necessary for the design of software tools or other interactive systems and gives the student a general framework to extend and apply usability knowledge. Based on this framework, available theories, methods, and technologies are discussed which can be used for validating user interfaces. In addition, the course handles the evaluation process of interactive technological devices. It concentrates on the traditional usability aspects (effectiveness, efficiency, and user satisfaction) as well as modern user-experience issues.

The course handles the theories discussed in the 'Designing Interactive Systems' book (Benyon et al. 2005), explained in the course's weekly lectures (10 lectures in total). The students are also required to read and present a number of research articles that discuss the latest advances in the field of human-computer interaction. After completing the course, the student should have:

- a coherent approach for developing interactive systems that allow users to accomplish their goals effectively and efficiently, and with a high level of satisfaction;
- knowledge of new theories and research-approaches for improving the user experiences in the development of intelligent systems;
- practical experience in the application of theories and methods for the generation and testing of intelligent user interfaces.

In the course, the instructors deal with the students on two separate levels. Individual students follow the course lectures, complete several homework assignments and are finally evaluated with an oral examination. Additionally,

while following the course the students will also take part in a lab project in teams of 4 students. There they develop an interactive system interface while observing the concepts taught in the course. A usability evaluation test is then performed on the developed interface by the students to measure the quality of their design. The students are given access to the Morae evaluation system. Using Morae gives the students a similar experience to using a full scale usability evaluation lab. The group's work is assessed by the quality of their designed system, and the quality of the project's reports and presentations. Simply said, the IUXE course is designed to handle the entire design process of system interfaces. It includes a lab project that takes the students through both development and evaluation stages. It also adds the new dimension of user experience to the development of interfaces.

As explained above, the course handles a comprehensive amount of theory from several sources. It also requires students to perform different types of course work (i.e., homework, reporting, presentations). One of the challenges in designing the course is that the students need to rapidly learn how to use several design and evaluation tools, which were necessary for completing different stages of the lab project. The goal of the implemented educational tools is to reduce the workload of the students and streamline their learning experience, without requiring the constant attention of the course instructors. It is also important to note that while the course is given in English, the absolute majority of the students who follow the International Master program are not native English speakers. They can therefore benefit from the ability to review course lectures and project instructions. That being the case, the course incorporated a number of educational tools specifically chosen to address each of the mentioned issues. The main requirements for the tools are to:

- help students in learning the discussed course theory and prepare for the course examination;
- allow them to review lectures and instructions;
- demonstrate how to use certain programs and services used in the course;
- make it easier to complete course work (e.g., homework, reports);
- reduce the time and effort needed from the instructors to administrate the course.

IMPLEMENTATION

The tools implemented for the course were aimed at assisting the students in both the theoretical and the project portions of the course. These tools utilized many open-source software packages and free online services (e.g., Audacity, Google Docs), as well as facilities provided by the university (e.g., PC's, Internet servers). Below is a description of the main tools used in the course.

Lecture Screencasts

The weekly lectures given by the instructors contain course theory. Generally, students are provided with a copy of the slides used in the lectures. However, students following the IUXE course were also provided with a screen recording (called a screencast) of the lectures. Using the Powerpoint plugin of the Camtasia software package, it was possible to make a recording of the instructors voice, the on screen mouse movements, videos, and any on screen activity (e.g., software demonstrations) performed during the lecture.

Creating the recordings did not call for any extra work for the presenter, and only involved using the Camtasia plugin to start and end the presentation instead of the standard Powerpoint interface. A simple restriction requires the presenter to stay within about 5 meters from the PC which is recording the presentation. This is necessary in order to ensure an acceptable voice level for the recording. After each lecture, the recordings were placed on the internet, giving the students access to the entire content of the lectures. Students were able to view the recording within any Internet browser with the flash player plugin (version 7 or higher). The lecture screencast plays within the Internet-browser as shown in Figure 1. It is possible to navigate through the recording using the slide titles (on the left) or the progress bar (at the bottom).

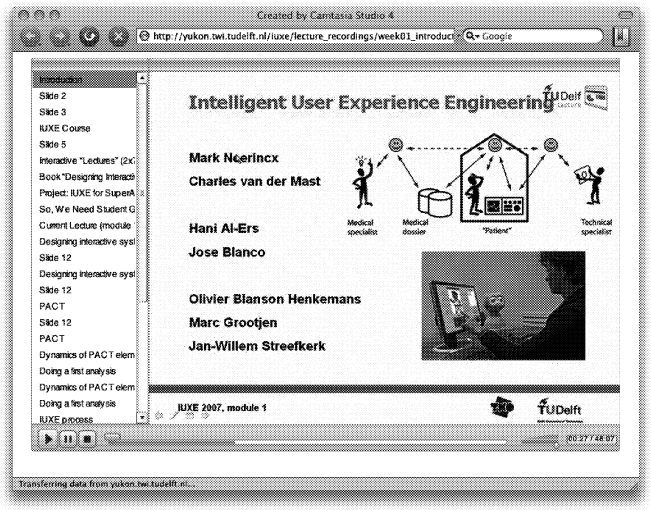


Figure 1: Screencast playback in an Internet browser

Screencast recordings were made of all 10 lectures given for the course. The aim of providing the students with these screencasts was to:

- allow the students to review information they do not understand while attending the lectures;
- provide them with reference information when they are performing the course project;
- assist them in preparing for the course examination.

Video Tutorials

Students are provided with tutorials that illustrate how to use specific software packages and perform specific tasks necessary to complete the course project. Traditionally, such instructions would have been provided to the students using written manuals. However, with video tutorials, students are able to see exactly how to complete each process with the voice of the instructor explaining the steps. Some videos contain short web-lectures of approximately 10 minutes, which discussed additional subjects related to the course project. In total, the course provided the students with 11 video tutorials and 4 web-lectures. The video tutorials were also constructed using the Camtasia software package. All tutorials were made available on the Internet where students can view them within the Internet-browser in a similar manner to the lecture screencasts. The aim of providing the video tutorials is to help the students in learning how to use specific tools and complete certain stages of the course project.

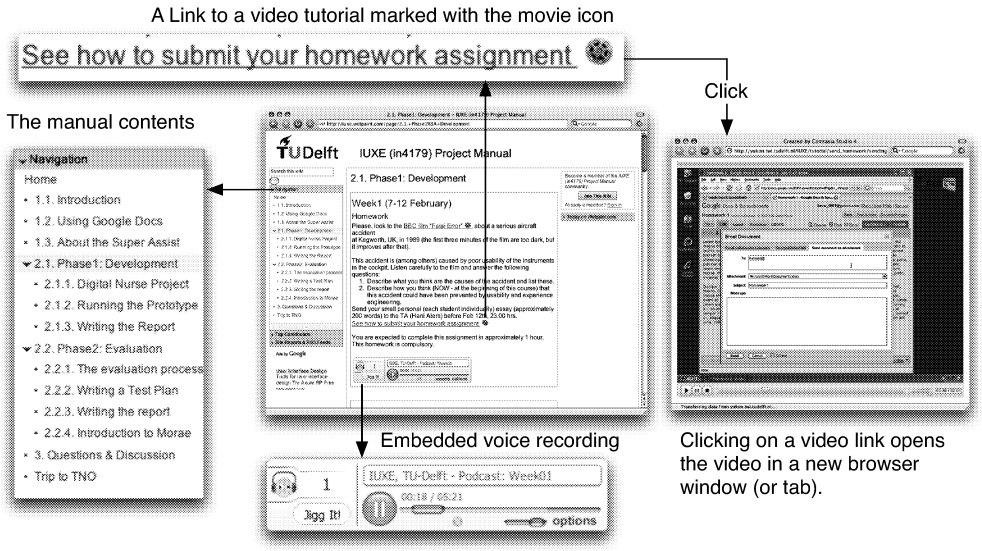


Figure 2: The online manual gives students access to the other tools from within the Internet browser

Voice Recordings

To provide the students with additional information, guidelines, and tips about the project as well as other course activities, students were provided with instructions in the form of voice recordings. Some recordings contained a summary of the activities for a specific week, while others gave tips and guidelines on what to write in the project's evaluation report. Students were provided a total of 13 voice recordings ranging between 2-5 minutes in length. The recordings were created using an open-source application called Audacity. They were initially provided to the students in the form of a podcast feed, which they could subscribe to using any podcast feed aggregator program such as iTunes. Halfway through the course, we stopped the podcast feed and decided to provide the recordings as embedded online players which work within the Internet-browser in a similar manner to the video tutorials. This was done in an effort to make the recordings more accessible for the students, as explained in the results section.

The Online Manual

Instead of providing the students with a printed project manual, the IUXE course used an online manual (reachable at http://iuxe.wetpaint.com) in order to incorporate the provided rich multimedia tools such as the video tutorials and the embedded sound recordings. The online manual was constructed on a wiki system which enables anyone to edit the content of the manual pages from within the Internet browser. With access restricted to the instructors which were involved with the course, this allowed rapid online collaboration among the instructors to build the content of the manual, and perform corrections or improvements whenever necessary. The manual included a built in message board system which allowed the students to discuss the project or other course activities with each other and with the course instructors. A free online wiki service called Wetpaint was used to host the course's online manual.

Collaborative Reporting

The instructors used an online collaborative authoring system in the form of Google-Docs. Using this system, any student can create a document and invite other team members as co-authors. This allows all team members to work on the same document either in terms or simultaneously. When course instructors are also invited as co-authors, they can monitor the progress of the teams and provide feedback and support if needed. The aim of using this system is to make it easier for the students to cooperate on course work. It also assists the instructors in communicating with the students and keep track of their progress.

RESULTS

In this section we discuss some of the results from our pilot experiment. The course was given for the first time in the spring semester of 2007. We collected subjective data using two interviews and two surveys throughout the semester. The course was attended by 20 students, none of which were native English speakers. The feedback results given in this paper are collected from 19 of the students. We asked the students about their perceived level of usefulness and enjoyability for each of the used tools. Figure 3 represents the average responses to the questions with a 0 to 4 Likert scale where higher numbers are better.

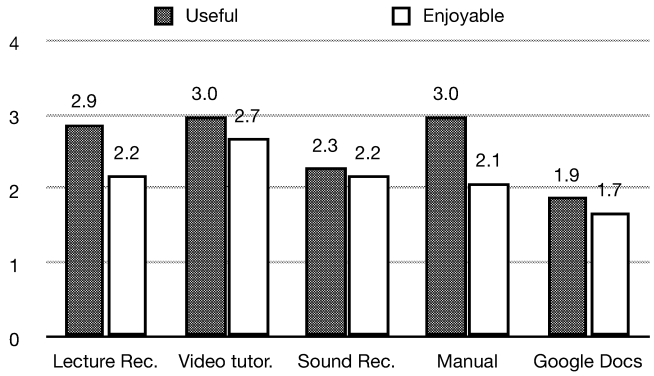


Figure 3: The Students' Perceived Usefulness and Enjoyability of Each of the Tools Used in the IUXE Course

One can note from the figure that the online manual, the video tutorials, and the lecture screencasts were found to be the most useful tools by the students. Table 1 shows the

number of times the students visited the lecture screencast database. The table shows that the students did review the screencasts while following the lectures and continued to do so even after the lectures stopped in April. During the interviews, students mentioned that following the screencasts was a very helpful alternative when they were unable to attend the real lectures. Nonetheless, they did not find that having access to the screencasts made them feel less motivated to attend the real lectures, which supports earlier findings in similar research efforts (Brotherton and Abowd 2004). Table 1 also indicates that the highest usage of the screencasts was in the examination period, where students pointed out that they were very helpful to prepare for the examination, and often expressed their desire to see similar screencasts in other courses.

It is worth noting that when the number of viewed screencasts was measured for the examination period in June, it turned out that the students have viewed 139 individual screencasts. This is more than double the 68 visits logged in Table 1, indicating that the number of views can be significantly higher than the visits listed in the table. One should also take into consideration that the screencasts only contain recordings of the course's weekly lectures (since extra information and instructions were only provided through the video tutorials and the voice recordings). As a result, they were not expected to be used heavily by the students.

Table 1: The Number of Times the Screencast Database was Accessed Throughout the Course Duration, and How They Correspond to the Different Stages of the Course.

Activity	Month	Visits
Lectures + Project	February	32
	March	24
Only project	April	9
	May	6
Examination	June	68

The Video tutorials were also well received (as shown in Figure 3). Students pointed out that following a video of the task they needed to perform left no room for guessing, and made it much easier to learn how to use the different tools they encountered while completing the course's lab project. Students also mentioned that watching the video tutorials and listening to the sound recordings gave them as close an experience as possible to having the instructor explaining things for them in person, which was helpful in working on the project. The sound recordings provided other interesting results. Initially, usage statistics showed that very few students were listening to the sound recordings when they were offered as podcasts. When asked, students pointed out that they did not want to spend the effort in learning how to get and configure the needed software to download them. However, once the recordings became available within the online manual, they were used by most students. This indicates how important it is for educational tools to function seamlessly within the students' work flow. If learning how to use the tool becomes an extra task to learn, then students will prefer not to use it.

The online manual proved to be a very successful method to provide students with instructions. Thanks to its open nature, it was continuously edited and improved throughout the course duration, involving a total of 436 entries and revisions by the course instructors. It was also used heavily by the students throughout the course (see Figure 4), with a total of 3516 pages viewed in 982 separate visits. Students pointed out that having the manual online made it very accessible. Students also appreciated the fact that the manual was quickly improved if any errors or ambiguity was pointed out to the instructors. In fact, together with the Video tutorials, the online manual received the highest score for usefulness .

Figure 4 also shows the average amount of time visitors spend on the site. This type of information gave the instructors the ability to observe the usage trends of the students, something which is impossible to achieve with printed media. It is, for example, possible to see that the average visit time was the highest in April, which is the period when the students were using the manual to prepare for their evaluation experiments. This showed the instructors that the course material was being used as intended.

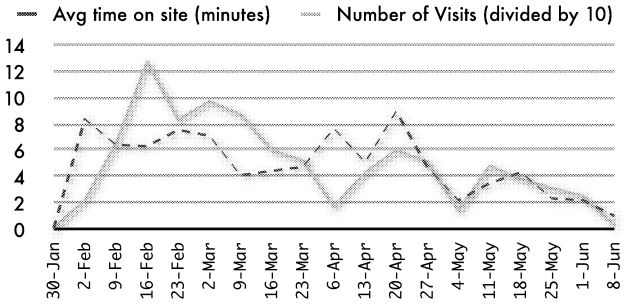
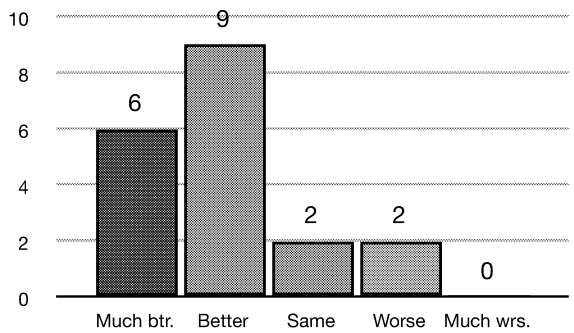


Figure 4: The number of visits and the average amount of time spent on the online manual each week throughout the duration of the course

For the collaborative reporting, students were introduced to the Google-Docs system during the first two weeks of the course where they used it to hand in homework assignments. Interviewed students pointed out that the ability to get rapid feedback through the system is a great feature. On the other hand, some students did not desire to spend the effort on learning how to use the system, and were frustrated with the limited features it offered. As a result, they opted to use Microsoft Word instead. This result supports the earlier conclusion (regarding the podcasts) in that educational tools have to be easy to use in order to be successful. The eventual conclusion was that the system offers very useful features, however it needed to be more powerful and user friendly in order to be fully adopted by the students.



Figures 5: Students' Opinion of the New IUXE Setup Compared to the Traditional Course Format

The students were also asked questions regarding their opinion of the general approach of the coarse. Figure 5 shows that the majority of the students favored the approach of IUXE to the traditional course format. In the interviews students pointed out that the tools used in the course made it interesting and engaging. They even expressed strong desire to see similar tools implemented on other courses.

Students also pointed out that using the tools made course work seem less of a chore. This is reflected in Figure 6, where the majority of the students found IUXE to be more enjoyable than other courses. For example, students mentioned that getting spoken instructions allowed them to carryout other activities while preparing for course work.

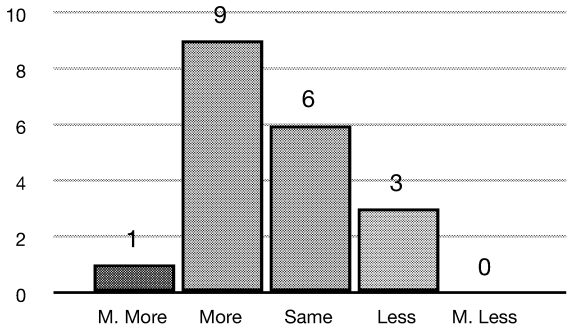


Figure 6: Students' Survey Answers to the Question: "All Else Being Equal, in Comparison to other Courses you Have Taken, How Much Did You Enjoy This Course"

CONCLUSIONS

In this paper, we have presented our use of internet based technology to enhance the learning and teaching processes in higher education. A pilot study in Spring 2007 served as a testbed for many of these tools as educators were looking for new educational approaches to teaching the subject of HCI. As mentioned in Section 2, the instructors had specific goals that they intended these tools to achieve.

The heavy usage of the lecture screencasts throughout the course and particularly during the exams period shows that they played a significant role for learning the course theory. This was also confirmed by subjective student opinions. Considering the (relatively small) cost and effort that goes into creating them and how useful they proved to be, screencasts are a good place to start when implementing educational tools in higher education. Similarly, the video tutorials played a significant role in helping the students to work on the course's lab project. Students were able to use all the different tools needed to complete the project without the need for direct supervision by the course instructors.

By looking at the tools which were not widely utilized in the course (i.e., podcasts and Google-Docs), one can see that both required some effort by the students to setup and use. This shows that educational tools need to be based on technology that is already familiar to the students in order to be accepted. This, in term, means that it is important for educational institutions to keep track of new technological devices (e.g., portable multimedia players) and services (e.g., social networks) which are starting to get wide acceptance by the students. This allows course designers to target such technologies when designing educational tools for their courses.

Using an online manual built on a wiki system allowed the instructors to effectively collaborate on creating the course material and provide the students with rich content. It facilitated the use of the rest of the educational tools and allowed the students to complete the course with little direct supervision of the instructors. The instructors were also able to track the students' progress and their usage of the course material with the help of the usage statistics gathered by the implemented tools.

These results, together with the positive opinion the students had towards the general new approach of the course, are very encouraging signs for this type of research. We plan to further develop the tools used in the IUXE course and implement similar tools in other courses.

REFERENSES

Alers, H. 2007. "Internet and multimedia for teaching and learning" Master's thesis, Computer Science, Delft University of Technology, Available at http://mmi.tudelft.nl/pub/halers/hani_alers_msc_thesis.pdf.

Benyon, D.; P. Turner; and S. Turner. 2005. "Designing Interactive Systems. People, Activities, Contexts, Technologies" Pearson Education Limited, England. ISBN 978-0321116291.

Brotherton, J. A.; G. D. Abowd. 2004. "Lessons Learned From eClass: Assessing Automated Capture and Access in the Classroom" *ACM Transactions on Computer-Human Interaction*, Vol. 11, No. 2.

Gagnon, R. J. and Ravi Krovi. 1999. "Internet Usage in Undergraduate Management Science and Operations Management Courses" *The Internet and Higher Education,* Volume 2, Issues 2-3, 107-118.

Karoulis, A. and A. Pombortsis. 2000. "Evaluating the Usability of Multimedia Educational Software for Use in the Classroom Using a «Combinatory Evaluation» Approach" *Eden 4th Open Classroom Conference*. Barcelona, Spain.

Koppelman, H.; E. Van Dijk; C Van der Mast and G. C. Van der Veer. 2000. "Team projects in distance education: a case in HCI design" *5th annual ACM SIGCSE/SIGCUE conference on Innovation and technology in computer science education*,, Helsinki Finland. 97-100.

Kotzé, P. and L. Oestreicher. 2002. "Teaching Human Computer Interaction: Qualitative Support for an Alternative Approach. Usability: Gaining a Competitive Edge" *IFIP World Computer Congress*, Montreal, Canada, 267–281.

Quintana C.; E. Soloway; C. A. Norris. 2001. "Learner-Centered Design: Developing Software That Scaffolds Learning" *ICALT*. 499-500.

Repenning, A.; A. Ioannidou; M. Payton; Ye Wenming and J. Roschelle. 2001. "Using components for rapid distributed software development" *Software, IEEE*. Volume: 18, Issue: 2, 38-45.

Russell, M. 2001. "Framing technology program evaluations" In W. Heineke &L. Blasi (Eds.), Methods of evaluating educational technology. Greenwich, CT: Information Age.

Salomon, G.; D. N. Perkins, and T. Globerson. 1991. "Partners in cognition: Extending human intelligence with intelligent technologies" *Educational Researcher*, 20, 2-9.

Van Aalst, J.W. and C. Van der Mast. 2003. "Performer: an instrument for multidisciplinary courseware teams to share knowledge and experiences" *Computers & Education,* 41, 39-48.

Van der Mast, C. 1995. "Professional Development of Multimedia Courseware" *Machine-Mediated Learning*, 5 (3 & 4), 269-292.

Van der Mast, C. and J.W. van Aalst. 2002. "A knowledge management instrument to support learning of highly interactive systems development" *M. Roccetti (Ed.), Proceedings of the 7TH Annual Euromedia Conference*, Modena, Italy, 105-112.

BIOGRAPHY

HANI ALERS received the M.S. degree in Computer Science from the Delft University of Technology in September 2007, where he is currently pursuing the Ph.D. degree.
His research interests include modeling human image and video quality perception, as well as employing internet technology in educational systems.

CHARLES VAN DER MAST has a PHD Computer Science from Delft University of Technology where he is employed at the Man-Machine Interaction group. He teaches HCI at several levels. His research includes using various media to improve teaching and VR therapy for phobia treatment.

PROVIDING MULTIMEDIA RECOMMENDATIONS BASED ON A MARKOV DECISION PROCESS INSIDE E-LEARNING PLATFORMS

Mihaela Brut
Alexandru Ioan Cuza University
Iasi, Romania
mihaela@infoiasi.ro

Romulus Grigoras
IRIT-ENSEEIHT
Toulouse, France
romulus.grigoras@enseeiht.fr

Vincent Charvillat
IRIT-ENSEEIHT
Toulouse, France
charvi@enseeiht.fr

Florence Sedes
Institut de Recherche en Informatique
Toulouse, France
sedes@irit.fr

KEYWORDS
Personalized recommendation, data mining, multimedia.

ABSTRACT

In this paper we provide a solution for recommending multimedia materials inside an e-learning platform, which considers the binding of user profile information with document annotations via the exploitation of domain ontologies. The user profile is structured on three levels – competences, interests and fingerprints – all expressed through ontological constructs, as also the multimedia materials annotations are. The user current activity (his fingerprints) is supervised in terms of conceptual ontology navigation (which ontology concepts are related to the currently accessed materials). For establishing the suitable topics for being recommended at each moment, a Markov decision process is used.

INTRODUCTION

Inside an e-learning platform, there are a lot of materials into multimedia formats. Because of their high production cost, an efficient multiple usage is desired, and recommending each of them to the suitable users, in the suitable moments is a possible approach.
This paper provide a such solution based on the exploitation of ontology based modeling of users and documents. In the beginning, the existing approaches for adopting the ontologies and semantic Web techniques into e-learning and multimedia management fields are discussed. Then, the paper exposes a model of ontology-based annotating the e-learning multimedia materials. Further, a three layer user modeling approach is presented, also expressed by ontology constructs. The recommendation system is modeled as a Markov decision process: the user profile is developed by supervising his current activity in terms of locating the ontology concept related to the currently accessed document and approximating the next concept which will be focused by the user. The improvement of the recommendations accuracy in time is mentioned in final, together with conclusions and further research directions.

ONTOLOGIES FOR MULTIMEDIA E-LEARNING MATERIALS ANNOTATION

The main goal of the existing e-learning standards is to increase the accessibility and the reusability of the e-learning materials. An improvement in this respect could be acquired by combining the e-learning standards with the Semantic Web technologies, whose aim is to make the electronic content comprehensible for the computers.
There are many researches in the field of integrating Semantic Web technologies into e-learning environments.
In the case of multimedia e-learning materials, there should be also considered the particularities of the semantic Web technologies adoption into the multimedia field.
The main issue of this problem is that, alongside with the Semantic Web activity of the Web Consortium, the main goal of transforming multimedia materials into machine-processable content is also assumed by the ISO's efforts in the direction of complex media content modeling, in particular the Multimedia Content Description Interface (MPEG-7).
In order to delimitate a semantic oriented approach of multimedia managing, there were discussed the differences between the two directions, as well as some modalities of combining them.
The differences are encountered at multiple levels (Nack et al. 2005):

- *Syntactically*, the difference could be reduced at those between XML representation (adopted by the MPEG) and RDF (used by the semantic Web): because one RDF assertion could be XML serialized in many ways, it is hard to process RDF using generic XML tools , and reverse;
- *Semantically*, the semantic Web approach makes use of different layers that define semantic structures and ontologies as third parties, while MPEG is a monolithic specification, including a large number of schemata.

A solution for unifying the two directions could consists into a semantic Web approach which to make use of the schemata developed in MPEG-7 as third-party specifications.

Another difference emphasizes that semantic Web technology is still mainly text oriented, while MPEG is dedicated to multimedia content description (Nilsson et al. 2002). A solution for unifying the two directions is provided by SMIL (*Synchronized Multimedia Information Language*) (Michel et al. 2005), which facilitates a textual serialization of temporal and spatial aspects for multimedia presentations.
Another distinction points out that the desiderate of semantic Web approaches to make explicit the semantics of media units is challenged by finding techniques for automatically metadata association. MPEG could be a useful aliat: it uses low-level features for semantic based descriptions, which constitutes one of the few mechanisms available for the automatic annotation of media.
In order to gain a unified view of the two directions, there were developed multiple MPEG-7 translations into RDF and OWL, as well as translations of the MPEG visual part into RDF and OWL (Hausenblas et al. 2007). Moreover, there were developed tools enabling to extract the visual features of multimedia materials (as MPEG proposes) and to associate them with domain ontology concepts (as a semantic web approach requires). PhotoStuff, AKTive Media, Vannotea, M-OntoMat-Annotizer, SWAD are such examples (Obrenovic et al. 2007).
In the educational field, the MIR (*Multimedia Informaion Repository*) project (Schmidt and Engelhardt 2005) provides a solution of adaptive facilities inside an e-learning platform which manage the multimedia information: it includes a user modeling component (MUMS - *Massive User Modeling System*), a component for managing and annotating the learning objects (HYLOS - *Hypermedia Learning Object System*), and a component for defining the adaptation model (MIRaCLE – *MIR Adaptive Linking Environment*).
We will expose a solution with a similar architecture which provides personalized recommendations to learners. Alongside with defining the multimedia annotations model and the user competences model, our focus will be in developing the user model by supervising his conceptual navigation activity in order to provide him the most relevant materials for his currently focused topic.

MULTIMEDIA INDEXING TECHNIQUES

The semantic contents of multimedia data can be indexed by two main ways:

- *Manually annotate the multimedia documents* (image, video, audio) with textual description (ideally, ontology concepts). It's not so comfortable for large multimedia databases, but could be a solution for multimedia learning objects, since the teachers spend long time for their development. The main issue remains still the difficulty to keep the consistency of the annotations.
- Provide the system with a *rule base or a knowledge base* where knowledge or rules are used to extract features from the raw data, to match content, to analyze queries, and so forth.

This second type of semantic analysis of the multimedia content involve a first step of multimedia automatic indexing. The indexing algorithms are applied successively in order to obtain (Chen et al. 2002):

- Feature extraction
- Clustering/Segmentation
- Object descriptor extraction
- Object recognition.

The results are then processed by using the knowledge database.
For example, in (Hsu et al. 1993) there is developed a hierarchy for the radiology domain, called Type Abstraction Hierarchy (TAH). The radiological shapes and their semantics are conceptualized through hierarchy concepts detailed into sets of attribute values. The knowledge base is targeted to the evaluation of shapes and spatial relationships of the objects (e.g., a tumor). The main goal is to improve search efficiency in radiological databases.
The maintenance of a knowledge base raise also the problem of keeping it semantically consistent with the database schema. (Yoshitaka and Ichikawa, 1999) provide a possible solution for this issue: to make the knowledge base and the database schema semantically dependent on each other by integrating them together with rules that prescribe semantic association of one with the other.
Despite the time consuming inconvenience, the manual annotation solution could provide a very good semantic level of metadata which to facilitate other further operations such as information retrieval or access personalization.

MULTIMEDIA ANNOTATIONS MODEL

We adopt a very simple model of multimedia materials semantic annotation, by using:

- an ontology already available inside the e-learning platform: that which is used for structuring the competences assigned to the existing courses (and to the corresponding certificates);
- an existing tool - *M-OntoMat-Annotizer* – which enables the automatic visual features extraction, as well as the semantic ontology-based manual annotations.

M-OntoMat-Annotizer provides support for extracting the MPEG-7 specific multimedia descriptors into XML format and for transforming them into RDF descriptor instances (which will be finally linked with the appropriate concepts of the selected ontology). Such descriptors are dominant color, scalable color, color layout, color structure, texture browsing, edge histogram, region shape, contour shape, homogenous texture. In addition, for video documents M-OntoMat-Annotizer enables to select a certain frame or frame sequence in order to associate its descriptors with ontology concepts. The only restriction is that each domain concept prototype instance could be linked with only one extracted visual descriptor of each type.
For our purposes, we consider that the teachers (educational multimedia materials developers) will use the M-OntoMat-Annotizer in order to annotate the video frames/sequences and image regions with the ontology concepts.
For example, let's suppose that we load the iswc.daml ontology (http://annotation.semanticweb.org/iswc/), we create the E-Learning, Modern Teaching and Blended Learning instances for the Application_Domain concept. We could select, for instance, the Blended Learning instance in order to associate a certain region of the image displayed in the right frame. As result, there will be generated a RDF description, including the following fragment:

<vdoext:Prototype rdf:about="http://www.acemedia.org
/ontologies/VDO-EXT#**Blended Learning**">
<rdf:type rdf:resource="http://annotation.semanticweb.org
/iswc/iswc.daml#**Application_Domain**"/>
…
</vdoext:Prototype>

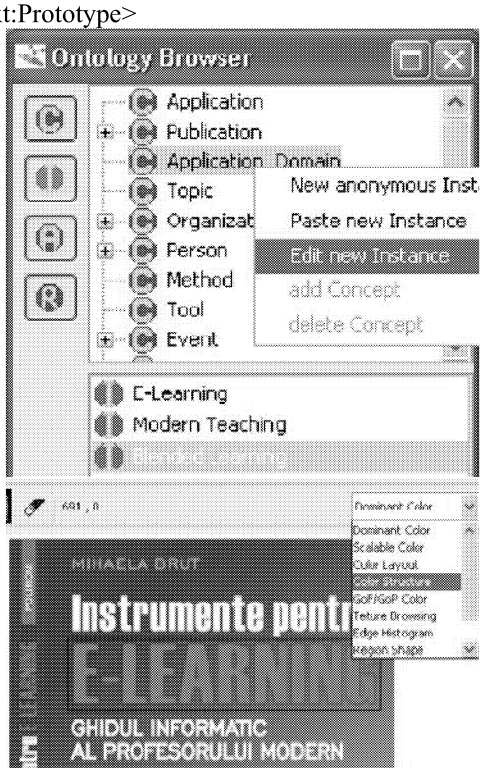


Figure 1: Using M-Onto-Annotazier for multimedia annotation

The effective values of the created instances (referred through their URLs) are stored in separate XML files, in MPEG-7 standard format. For example, the RegionShape.xml file contains:

<?xml version="1.0" encoding="ISO-8859-1" ?>
<Mpeg7 xmlns="http://www.mpeg7.org/2001
/MPEG-7_Schema" xmlns:xsi="http://www.w3.org
/2000/10/XMLSchema-instance">
<DescriptionUnit
xsi:type="DescriptorCollectionType">
<Descriptor xsi:type="RegionShapeType">
<MagnitudeOfART>
14 15 11 12 14 14 11 12 13 8 9 15 9 12 12 4
10 12 10 3 11 9 8 11 5 5 10 7 4 12 8 5 12 8 5
</MagnitudeOfART>
</Descriptor>
</DescriptionUnit>
</Mpeg7>

This description could be parsed in order to be integrated into an educational metadata repository. Among the established e-learning standards, the IEEE/LOM is considered as the most enabling for semantic extensions (Al-Khalifa and Hugh, 2006). It is structured into many categories, and the Classification enables to specify that the learning object belongs to a certain classification system, which could be also an ontology. It's possible to specify the information for identifying the ontology and its particular concept which want to be refferred:

<Classification>
<Purpose> competency </Purpose>
<TaxonPath>
<Source> ("en", "ISWC") </Source>
<Taxon> <id>I.2.6.5</id>
<entry> ("en","Application Domain")
</entry>
<instance>Blended Learning </instance>
</Taxon>
</TaxonPath>
</Classification>

The spatial and/or temporal description of the multimedia element could be referred through the General IEEE/LOM category.

Thus, by using the M-Onto-Annotizer tool and a conversion operation, a multimedia learning object could be described according to IEEE/LOM standard and also could gain ontology-based semantic annotations.

THE USER MODEL

The main purpose of creating and maintaining an user model is to provide each user with the most suitable information (Brusilovsky and Millán 2007). The five most popular features considered for modeling an user as an individual are: the user's knowledge, interests, goals, background, and individual traits.

All these features (especially the first three) could be expressed through the overlay model, which represents an individual user's profile as a subset of the domain model [7]. We adopt this model considering an ontology as domain model – following in this respect another actual direction of research, as in the (Dolog et al. 2004) or (Kay and Lum 2004). Rather, we consider the same ontology as for multimedia annotations described above.

In our approach, the user model regards the first three features, and, accordingly, our ontology-based user competences profile is split into three layers:

1. *Competences* – the actual, acquired, competences (e.g. through previous acquired certificates and qualifications), expressed through ontology concepts.

For the automatically construction of this layer, a rule-based approach could be applied (Lin et al. 2002). For example, a certain certificate should have assigned some competences (ontology concepts). In function of the score mentioned by the certificate, the user detains those knowledge at a certain level (beginner, intermediate, advanced).

2. *Interests* – the desired, foresighted, competencies, according to the courses in which the students is currently enrolled: each course aims at developing a set of competences, expressed as concepts; these concepts will constitute the user interests profile. The particularity of this layer is that of being common to multiple users: all the students enrolled into the same course have a same part of their interest profile.

For developing this profile layer, a rule-based approach could be also adopted: when a user is enrolled to a certain course, the topics (ontology concepts) assigned to these are included automatically in his long term interests profile.

3. *Fingerprints* – representing the currently visited concepts (through associated materials). These concepts illustrate the particular goals encountered into the current moment (these are driven by a specific task

which has to be accomplished – for example an homework preparation or a project development).

These fingerprints could be automatically inferred from the user navigation activity, and should be correlated with his competences and interests for providing him with the appropriate recommendations. In the next section we will describe the Markov decision process based approach for developing the fingerprints user model layer. Because the fingerprints are modeled through ontology concepts, we will trace the conceptual navigation through ontology instead of the site navigation.

USER MODEL DEVELOPMENT TOWARDS PERSONALIZED RECOMMENDATION

The most general method for developing the user profile in order to provide him with suitable recommendations (or, morel general, with personalization) is by analyzing the user's navigational activity. For this purpose, there were developed techniques such Clustering, Classification, Association Rule Discovery, Sequential Rule Discovery, Markov models, or hidden (latent) variable models (Mobasher 2007).

Markov models are support for another type of sequential modeling, based on stochastic methods: the navigational activity in the Web site is modeled as a Markov chain, used for predicting subsequent visits. There could be predicted the next user choice, based on his last action (using a first order Markov model) or based on his last k actions (using a k order Markov model) (Deshpande and Karypis 2004).

Instead of tracing the user's site navigation, there were recently developed some approaches for modeling users' navigation behavior at "higher" abstraction levels. In the adaptive InterBook system, the concept-based navigation was introduced (Brusilovsky et al. 1998): each concept used to index documents constitutes also a navigation hub - providing links to all content pages indexed with this concept; also, from each page, all its related concepts are accessible.

In (Gutiérrez et al. 2006), Web documents are first clustered based on users' navigational data, and then user behavior models are built at this document cluster level. In (Antonioletti et al. 2006), an aggregated representation is created as a set of pseudo objects which characterize groups of similar users at the object attribute level.

(Jin et al. 2005) adopts the proposed task-oriented user modeling approach. The relations between the common navigational "tasks" and Web pages or users are characterized through the *Probability Latent Semantic Analysis (PLSA)* model. The user model development is accomplished through an algorithm based on Bayesian updating, and the Web recommendation technique is based on a maximum entropy model.

In our approach, the domain ontology used for modeling documents annotations and user knowledge and interests provides the abstraction level for user conceptual navigation modeling. A Markov Decision Process is used for this modeling, to predict the next focused ontology concept by the user in order to provide him with personalized recommendations.

More precisely, at a certain moment, our system will display to the user:

- the chronological list of the already reached concepts in the current session – $c'_1, c'_2, c'_3, \ldots, c'_i$;
- the currently chosen concept, c_0, accompanied by a list of recommended documents $d_{01}, d_{02}, d_{03}, \ldots, d_{0j}$;
- the recommended concepts for being further accessed – $c_1, c_2, \ldots, c_k$, displayed in the predicted importance order;
- a link to "Other" concept list, for the case user is not satisfied by the recommended concepts.

Two separate modules contributes for accomplishing this functionality:

1. *Document selection module*, characterized by:
- input: user profile (the competences and interests layers);
- output: the list of recommended documents $d_{01}, d_{02}, d_{03}, \ldots, d_{0j}$, as well as the total amount of the time spent by the user in reading these documents until the moment he makes another concept choice.
2. *Markov decision Process module*, characterized at each step by:
- input: currently chosen concept + time spent by the user at the previous concept, as reward for this concept;
- output: the list of recommended concepts for being further accessed – $c_1, c_2, \ldots, c_k$, displayed in the predicted importance order.

We will describe bellow how the Markov Decision Process is modeled and his functioning manner. We considered in our MDP a certain user interests profile as a join of stereotypes: all the students enrolled into the same course have in common this course interest concepts, which could be viewed as a stereotype. Thus, the behavior of all the students is relevant in establishing the recommendations provided to each of them. The customization will be influenced by the each student competences and fingerprints: student particular competences will be used by the *Document selection module* and student fingerprints – by the *Markov decision Process module*.

A *Markov Decision Process* (Puterman 1994) is defined by a quintuple (S, A, T, P, R) where:

- S is the set of all possible *states* $\sigma \in S$ of the system;
- A - set of all the *actions* $a \in A$ that can be applied to it;
- T is the ordered set of *instants* at which decisions can be made, that is the *global temporal horizon*;
- P defines the *transition probabilities* between any pair of states in S after executing an action in A;
- R defines the *function of local rewards* $r \in R$ associated with these transitions $R : S \times A \rightarrow \Re$.

At each step, the learner algorithm (called the agent) selects an action, and then as result is given a reward and transitions to a new state

Our basic idea is to consider the currently visited concept as the main information of the current MDP *state*, and to predict the next possible states as corresponding to the recommended concept list. This list defines in fact the set of possible *actions* which user could accomplishes (transitions from the current state to each state corresponding to the concepts from list).

Supposing that the competences and the interest part of the user profile are already available, a MDP current state σ_0 will consist in:

- co_0 – the currently visited concept;

- i_0 – the user interest regarding the concept, $i_j \in \{0, 1, 2, \ldots\}$: the number of the courses (interest paths) in which the students is enrolled and concern the concept co_0; it is a fixed parameter for a certain user;
- f_0 – the user fingerprints in using the concept (expressed in number of minutes, as average of previous periods spent by the current user at co_0).
- cf_0 – the collective fingerprints in using the concept (expressed in number of minutes, as average of previous periods spent by the current user at co_0)
- r_0 constitutes the reward for reaching the current state, corresponding to the concept co_0, which will be updated in function of the f_0 and cf_0.

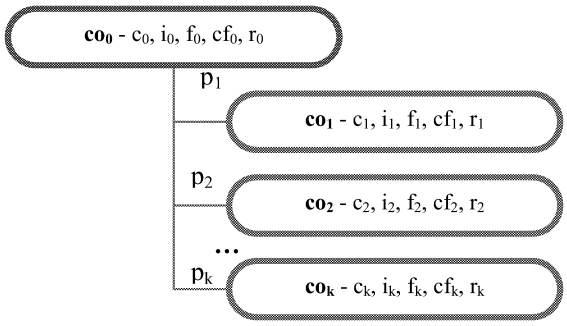


Figure 2: The set of possible actions that could be applied to a current state σ_0

From the current concept co_0, there could be considered for recommendation the concepts which are related to co_0 inside the ontology - c_1, c_2, …, c_k -, with respectively the probabilities p_1, p_2, …, p_k and the rewards r_1, r_2, …, r_k.

The goal of a MDP is to compute the so-called *optimal policy* - a function π^* that associates with any state $\sigma \in S$ and any time $t \in T$ the *optimal action* $\pi^*(\sigma, t)$, namely the action that maximizes the expected global reward on the remaining temporal horizon.

In our case, the policies possible to be applied into the state σ_0 (characterized by the co_0, i_0, f_0, cf_0) are characterized by the total reward that is expected to be gained over the temporal horizon as result to each action of choosing co_1, co_2, …, co_k respectively.

For estimating this optimal policy, a common adopted approach is the reinforcement learning (Sutton and Barto 1998), which consists in learning an optimal policy by estimating iteratively the optimal value function of the problem on the basis of simulations. It differs from supervised learning in that the learner algorithm is never told the correct action for a particular state, but is told how good or bad the selected action was, expressed in a form of scalar "reward" (McCallum et al. 1999).

The reinforcement learning approach is used in MDP to learn a *policy*, a mapping from states to actions, $\pi : \mathbf{S} \rightarrow \mathbf{A}$, that maximizes the sum of its reward over time. The most common formulation of the "reward over time" is a discounted sum of rewards into an infinite future. It is currently used an infinite-horizon discounted model where reward over time is a geometrically discounted sum in which the discount, $0 \leq \gamma < 1$, devaluates the reward received in the future. The discount γ could be constant, or could be decreased progressively: $\gamma = 1/(t + 1)$ or $\gamma = (1 / 2^t)$.

Accordingly, when following policy π, we can define the *value* of each state o be:

$$\mathbf{V}^{\pi}(\sigma) = \sum_{t=0, \infty} \gamma^t r_t$$
$$= R_1(\sigma, \pi) + \gamma R_2(\sigma, \pi) + \ldots + \gamma^{n-1} R_n(\sigma, \pi) + \ldots$$

where $r_n(\sigma, \pi)$ is the reward obtained at time n if the learner agent begins at the state σ at the time 0 and follows the policy π.

In order to learn the optimal policy, it must be learn its value function, $\mathbf{V}^*$, which maximize the $\mathbf{V}^{\pi}(\sigma)$ for all the states σ.

The previous equation could be written recursively:

$$\mathbf{V}^{\pi}(\sigma) = R_1(\sigma, \pi) + \gamma \, V^{\pi}(T(\sigma, \pi(s)))$$

Among the algorithms provided by the reinforcement learning approach, the on-step *Q-learning algorithm* (Watkins 1989) compute the optimal solution in a bottom-up manner, by *value-iteration* or by *policy iteration*.

In the value iteration variant, the Q-learning algorithm replaces the value function $\mathbf{V}^*$ with its more specific correlate, called $\mathbf{Q}$, providing the expected value of the criterion for the process that starts in σ, executes action a and thereafter follows policy π. Thus, $\mathbf{Q}^*(\sigma, a)$ will be the value of selecting the action a from state s, and thereafter following the optimal policy. This is expressed as:

$$\mathbf{Q}^*(\sigma, a) = \mathbf{R}(\sigma, a) + \gamma \, \mathbf{V}^*(\mathbf{T}(\sigma, a)),$$

where $\mathbf{T}(\sigma, a)$ denotes the state obtained as effect of applying action a to the state σ.

The optimal policy can be defined in terms of $\mathbf{Q}$ by selecting from each state the action with the highest expected future reward:

$$\pi^*(s) = \arg\max_a \mathbf{Q}^*(s, a)$$

(Bellman 1957) shows that the optimal policy can be found straightforwardly by dynamic programming.

The principle of the Q-learning algorithm consists in updating iteratively the values of the V^* function we search, observing each transition reward. Because the value function is expressed in terms of reward, we will update in fact the reward assigned with each state.

We provide the Q-learning algorithm below:

Initialize Q_0
for $n = 0$ **to** $N_{tot} - 1$ **do**
σ_n =choseState
a_n =choseAction
(σ'_n, r_n) =simulate(σ_n, a_n)
/* update Q_{n+1} */
$Q_{n+1} \leftarrow Q_n$
$d_n = r_n + \gamma \max_b Q_n(\sigma_n, b) - Q_n(\sigma_n, a_n)$
$Q_{n+1}(\sigma_n, a_n) \leftarrow Q_n(\sigma_n, a_n) + \alpha_n d_n$
end for
return QN_{tot}

We will illustrate this algorithm application for our particular situation, discussing as generic the case of the state σ_0.

The initialization part of the Q-learning algorithm

If we denote f_j – the current user fingerprints, and cf_i – the collective fingerprints in using the concept co_i, then:

- at the beginning of each working session: f_j=0;
- at the beginning of the first working session: cf_i=0; then, cf_i will represent the average time period spent by all the users over the concept co_i.

The initial values of p_j will depend by the current global interest profile, that is the stereotype profile constituted by the concepts assigned to all courses. We mention that a concept could be found in more than one course, meaning that concept could have a higher degree of interest associated. The algorithm for computing p_j first evaluates the total number of interest degrees associated with the all the concepts c_1, c_2, …, c_k, and then distribute probabilities proportionally with the interest degree of each concept. The total sum of probabilities is 1. Also, the initial positive rewards r_i are associated only with the states corresponding to the concepts of interests (as interest degree values), in rest the rewards are initialized with zero:

{ ni = 0; //the total number of interests
for j=1 to k { ni += i_j; $r_i = i_i$; rc_i=0; }
for j=1 to k $p_j = i_j$ / ni; }

The values of r_i and rc_i will be updated as average with the product $coef_i$ * tsr_i, where $coef_i$ represents a coefficient for estimating the accuracy of our current recommendation, and tsr_i representsa the time spent the user at the concept co_i.

So, for beginning, in the Markov chain, the states corresponding to the interest profile concepts have assigned a positive reward, and the actions which conduct to their selection have associate positive probabilities.

Choosing a state through an action

We denote by σ_0 the currently chosen state, characterized by the co_0, i_0, f_0, cf_0, r_0, rc_0. The state σ_0 was reached from σ_0' by applying a certain action a_0, and was obtained a reward r_0. The next step will be made the transition to the state (which, in our case, is one of the states σ_1, σ_2, …, σ_k). The reward r_0 will be added to the value of the specific correlate function Q_0, and the next reward which will be further added depend by which action will be chosen at the current step. Of course, the dynamic programming principle of the algorithm try to choose the action that will further lead to a maximum reward.

Updating the policy function as result of simulation

In order to improve the rewards pertinence, we will re-evaluate the current step rewards r_1, r_2, …, r_k (which establish the current step possible actions) according to the user choice and his reward to this choice (expressed through the time spent at the corresponding concept).

Suppose that the current user choice the concept c_l after the system provided him recommendations. The following algorithm re-compute the above mentioned rewards r_j, by increasing r_l and de-creasing the other rewards. We made the following notations: f_0 – the total amount of time spent by a user at c_0, representing his reward for the concept c_0; ni - the total number of interests; nf - the total number of the current user fingerprints; ncf - the total number of the all users fingerprints. The algorithm multiply r_l according the user interest and the personal/collective fingerprints relative to the concept c_l. We use some coefficients for assuring an harmonic medium quality; if we want to provide a greater importance to a certain thing, we could change the coefficients. We denote the increased quantity as nt, and we will decrease this quantity, proportionally, from the others rewards:

{ update f_0 with the amount spent by user at c0;
ni = 0; nf = 0; ncf=0;
for j=1 to k
{ ni += i_j; nf += f_j; ncf += cf_i}
r_l *= 0.4* i_l / ni + 0.3* f_l / nf + 0.3*cf_l / ncf;
n_t = 0.4* i_l / ni + 0.3* f_l / nf + 0.3*cf_l / ncf;
for j=1 to k
if j != l
r_j /= nt * (i_j + f_j) / (ni + nf – i_l – f_l); }

The sum of (i_j + f_j) will be exactly (ni + nf – i_l – f_l), so the total decreased value is nt. The next step will be performed when the user makes another choice, selecting another concept, after c_l. As an observation, if the user changes the documents, but remains at the same concept, there is considered no movement.

At this moment we dispose of the time period that user spent at the concept c_l, We denote this by tsr_l (time spent reading the concept c_l). In order to avoid the explicit feed-back request to the user, we consider the tsr_l as the user reward regarding the concept c_l (the spent time illustrates his interest), and its value be added both to f_l (user fingerprints for the concept c_l) and to nf (the total number of user fingerprints for all concepts c_1, c_2, …, c_k).

In order to evaluate the user reward (positive, negative, zero), we have to compare tsrl with nf. We will adapt the algorithm exposed in [16], used in the Knowledge Sea II system, for a certain page is the time spent reading (TSR) that page: the effective value of this time is normalized by using an algorithm which ignore the very short and very long page visits (the first are not enough for really reading a page, the second are probably caused by something like a coffee break), and consider the page length (for a short page it's enough a small time, but not for a long one). Because we are focused on the concepts, we do not measure the document length, and our adaptation considers: if the time is very short (ex., < 30 s), then the reward is negative; if the time is short (ex., 30 s < tsrl < 3min), then the reward is 0: the user considered the concept in concordance with his global interests, but no with the current ones; if the time is longer (tsrl > 3 min – it could be very long as one concept could correspond to a lot of documents), then the reward is positive, and we re-evaluate the previous reward according the algorithm below. We will adjust the reward r_l according the entire time f_l spent in the current session at the concept c_l (we consider it more relevant than the current tsr_l), denoting ap the adjusted quantity. We will decrease this quantity from the others rewards, proportionally to the corresponding fingerprints.

{ f_l += tsr_l; nf += tsr_l;
r_l = r_l + r_l * (f_l / nf); ap = f_l / nf;
for j=1 to k
if j != l
r_j = r_j – r_j * (ap * f_j / (nf - f_l)); }

The sum of f_j will be (nf - f_l) so, the total decreased value is exactly ap. In case of negative reward, the algorithm for adjusting current rewards is similar, just the ap quantity will be decreased from the r_l and proportionally increased to the other ones.

The simulations which we made in order to evaluate the prototype of our proposal illustrated a growth in the recommendations accuracy over the time, when the user profile is developed. The few user total unexpected actions lead to the necessity of taking into account also some off-topic recommendations.

CONCLUSIONS

The main advantage of our approach consist in its independence of the navigational structure of a particular site. Moreover, the conceptual navigation support the user in receiving recommendations according to his current main interests. The consideration of both personal and collective fingerprints define our solution as a mixture between navigation mining and collaborative filtering. The Markov Decision Process assures a good recommendations pertinence, increasing while the system is more and more used. Comparing to Collaborative filtering recommendations methods, our approach avoid the difficulty of the cold start problem, and provide a better recommendation accuracy due to the optimization process performed after the user reward.

REFERENCES

Al-Khalifa H.S., Hugh D. 2006. "The Evolution of Metadata from Standards to Semantics" in E-Learning Applications. Proceedings of Hypertext'06, ACM Press

M. Antonioletti, M. Atkinson, S. Malaika, S. Laws, N. W. Paton D. Pearson, G. Riccardi. 2006. "Web Services Data Access and Integration (WS-DAI)", OGF Grid Final Documents (GFDs): http://www.ogf.org/gf/docs/?final

Bellman, R. E. 1957. Dynamic Programming. Princeton University Press, Princeton

Brusilovsky, P., Eklund, J., Schwarz, E. 1998 "Web-based education for all: A tool for developing adaptive courseware". In: Ashman, H., Thistewaite, P. (eds.) Proc. of Seventh International World Wide Web Conference. Vol. 30. Elsevier Science B. V., 291-300

Brusilovsky, P., Millán, E. 2007. "User Models for Adaptive Hypermedia and Adaptive Educational Systems", in P. Brusilovsky, A. Kobsa, W. Nejdl (Eds.): The Adaptive Web, LNCS 4321, Springer-Verlag Berlin Heidelberg, 3 – 53.

Chen, S.-C., Kashyap, R. L., Ghafoor, A. 2002. Semantic Models for Multimedia Database Searching and Browsing, Kluwer Academic Publishers, NY, ISBN 0-792-37888-1

Deshpande, M., Karypis, G. 2004. "Selective Markov Models for Pre-dicting Web-page Accesses". ACM Transactions on Internet Technology 4(2) (2004) 163–184

Dolog, P., Henze, N., Nejdl, W., Sintek, M. 2004. "Personalization in distributed e-learning environments". In: Proc. of The Thirteenth International World Wide Web Conference, WWW 2004 (Alternate track papers and posters). ACM Press (2004) 161-169

Gutiérrez, M.E., Gómez-Pérez, A., García, O.M. 2006. "Ontology Access in Grids with WS-DAIOnt and the RDF(S) Realization", Proceedings of the 3rd Annual European Semantic Web Conference, LNCS, vol 4011, Springer

Hausenblas, M., Troncy, R., Halaschek-Wiener, C., Bürger, T., Celma, O., Boll, S., Mannens, E. 2007. "Multimedia Semantics on the Web: Vocabularies", W3C technical reports, Boston: http://www.w3.org/2005/Incubator/mmsem/wiki/Vocabularies

Hsu, C.C., Chu, W.W., Taira, R.K. 1993. "A Knowledge-Based Approach for Retrieving Images by Content," *IEEE Trans. Knowledge and Data Engineering*, vol. 8, no. 4, pp. 522-532

Jin , X., Zhou, Y., Mobasher, B. 2005. "Task-Oriented Web User Modeling for Recommendation" In *Proceedings of the 10th International Conference on User Modeling* (UM'05), Edinburgh, Scotland, July 2005, LNAI 3538, Springer, 109-118

Kay, J., Lum, A. 2004. "Ontologies for Scrutable Learner Modeling in Adaptive E-Learning". In: Aroyo, L., Tasso, C. (eds.) *Proc. of Workshop on Application of Semantic Web Technologies for Adaptive Educational Hypermedia*. Technische University Eindhoven, 292-301

Lin, W., Alvarez, S.A., Ruiz, C. 2002. "Efficient adaptive-support association rule mining for recommender systems". *Data Mining and Knowledge Discovery* 6 (2002) 83–105

McCallum, A., Nigam, K., Rennie, J., Seymore, K. 1999. "Building Domain Specific Search Engines with Machine Learning Techniques", *Proc. AAAI-99 Spring Symposium on Intelligent Agents in Cyberspace*

Michel, T. et al. 2005 "Synchronized Multimedia Integration Language (SMIL 2.0)" - Second Edition, W3C Recommendation, Boston, 2005: http://www.w3.org/TR/smil20

Mobasher, B. 2007 "Data Mining for Web Personalization", In Brusilovsky, P., Kobsa, A., Nejdl, W., eds.: *The Adaptive Web: Methods and Strategies of Web Personalization*. Volume 4321 of Lecture Notes in Computer Science. Springer

Nack, F., Van Ossenbruggen, J., Hardman, L. 2005 "That Obscure Object of Desire: Multimedia Metadata on the Web (Part II)", In: *IEEE Multimedia* 12(1), 54-63

Nilsson, M., Palmér, M., Naeve, A. 2002. "Semantic Web Metadata for e-Learning - Some Architectural Guidelines", *Proceedings of WWW Conference*, ACM

Obrenovic Z., et al. 2007 "Multimedia Semantics: Overview of Relevant Tools and Resources", Web Consortium, 2007: http://www.w3.org/2005/Incubator/mmsem/wiki/Tools_and_Resources

Puterman, M.L. 1994 *Markov decision processes: discrete stochastic dynamic programming*, Wiley-Interscience, NY

Schmidt, T.C., Engelhardt, M. 2005 "Educational content management", in García, F.J., García, J., López, M., López, R., Verdú, E. (Eds*), Educational Virtual Spaces in Practice*, Ariel Inter-national, Barcelona

Sutton, R. S., Barto, A. G. 1998. *Reinforcement Learning: An Introduction*, MIT Press, Cambridge, Massachusetts.

Watkins, C., 1989. *Learning from Delayed Rewards*, Thesis, University of Cambidge, England.

Yoshitaka A., Ichikawa, T. 1999. "A Survey on Content-Based Retrieval for Multimedia Databases," *IEEE Transactions on Knowledge and Data Engineering*, vol. 11, no. 1, 81-93

The Automatic Identification of the Emotion Status of Web Pages

Dr. David John
Bournemouth University
School of Design, Engineering & Computing
Fern Barrow, Poole
Dorset, BH12 5BB, UK
E-mail: djohn@bournemouth.ac.uk

Professor Anthony C. Boucouvalas
University of Peloponnese
Department of Telecommunication Science and Technology
Terma Karaiskaki
Tripoli, Greece, 22100
E-mail: acb@uop.gr

KEYWORDS
Development Tools for Internet Explorer, HTML Converters and Editors, Intelligent Analysis and Interpretation of Multimedia Data, Real Time Interactive Systems

ABSTRACT

A system has been developed that analyses the emotional content of web pages. Text is extracted from an embedded Internet browser and sent to a text-to-emotion engine that identifies the emotive content. An experiment was conducted that examines the effect the emotional content of the web page has on the reader. Our Online Emotion Stock Analyser application was used to assess the emotional content of web pages displaying articles that comment on the previous day's Stock Market share prices. The relevant emotive words relate to the up and down movement of share prices. The relative movement of the positive and negative emotions expressed in the articles over time was compared with the relative movement of the value of the London Stock Exchange. The experiment results support the loop effect between published articles and the future movement of share prices. A number of significant relationships have been observed between the emotion contained in articles about the Stock Market and the next day's market value. The output of the Online Emotion Analyser was shown to be a more accurate predictor of the next day's movement of the Stock Market share index than viewing the previous movement of the index.

INTRODUCTION

This paper describes the latest developments in our research project that aims to enhance communication over the Internet. In particular it presents the results of an experiment that investigated the link between emotions contained in web pages and the behaviour of the readers. The project contributes to the research into expressive communications and affective computing (Picard 2000; Paiva et al. 2007).
We have developed a prototype Emotion Analyser, which can analyse the emotive content within textual messages (John et al. 2006). The system enhances communication over the Internet by automatically identifying the emotions in the text and presenting appropriate emotive images. This system has been adapted to analyse the emotional content of web pages. The text to be analysed is automatically extracted from an embedded Internet browser. In order to examine how web pages presenting articles about the Stock Market affect the reader, the emotion extraction rules have been customised to identify words that relate to the up and down movement of share prices.
The Online Emotion Stock Analyser analyses the contents of web pages displaying objective-dominated stock articles and searches for the emotive content within them. The output of the engine is the Emotion Status of the article. In order to test whether the Online Emotion Stock Analyser can correctly identify the emotional information, the output of application was compared with the movement of share prices. The link between published articles and future movements of share prices was examined to determine whether the system could be used as a tool to predict future movements of the Stock Market share index.
Firstly, we define the term "stock market emotions"; secondly, we describe the operation of the Online Emotion Stock Analyser application; thirdly, we describe the experiment that compared the Emotion Status of the online articles with the Stock Market share index; fourthly, we examine the results of the experiment; and fifthly, we present our conclusions.

STOCK MARKET EMOTIONS

We define Stock Market emotions as the different states of expectation investors derive from their perception of news and events that influence their behaviour in the Stock Market. Stock Market emotions are directly reflected in the movement of the Stock Market share index and the volume of trading (Rutterford 2007; Bergen 2003). The positive state matches in the situation where share values are rising, and the negative state occurs when share values are falling.
Sensationalism and emotive news influence readers and shareholders behaviour. A loop effect (figure 1) has been observed between newspaper articles and the behaviour of investors (Gaughan 1986; Nelson 1991; Evatt 1997; Lo and Repin 2002).

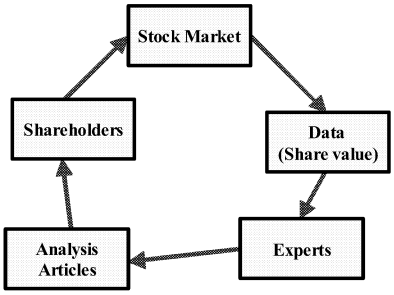


Figure 1: The Loop between Published Articles and the Behaviour of Investors

1. Shares are bought and sold in the Stock Market by shareholders.
2. Data is generated, such as the share value and the volume of trading.
3. Stock analysis experts analyse the data.
4. The Stock analysis experts publish articles.
5. The articles may be read by the shareholders, influencing their future buying or selling behaviour, starting the loop again.

The Emotion Analyser was adapted to test whether the mechanism for the loop effect can be detected. Information about the movement of share prices contained in published articles about Stock Market prices were compared with the future movement of Stock Market share index. If a link was detected, then the Online Emotion Stock Analyser could be used as a tool to predict the future movement of the Stock Market share index. The next section describes how the Emotion Analyser was adapted to suit the Stock Market environment.

ONLINE EMOTION STOCK ANALYSER ARCHITECTURE

There are a number of existing tools and research prototypes for the measurement of Stock Market emotions that are based on the analysis of market values (Arps 2007; CBOE 2007). For our Online Emotion Stock Analyser, the input is the text contained in web pages displaying Stock Market articles instead of the numerical data.

We developed a prototype Emotion Analyser, which can analyse the emotive content contained within textual messages passed in Internet communications (John et al. 2006). This system was adapted to create a new system called the Online Emotion Stock Analyser that evaluates the emotive content of online Stock Market analysis articles. The new system enhances communication over the Internet by automatically assessing the emotional state of a web page and indicating the "mood" of the information expressed.

For Internet communications a range of six emotions are identified, while for the Stock Market only two conditions are considered; positive (where stock prices are rising), and negative (where stock prices are falling).

The Online Emotion Stock Analyser is a rule-based system that uses key-word tagging in order to analyse text. Sentences can be broken down into two groups of words: function words and content words, which in turn have sub-classes. According to what groups the words belong to and the preceding and subsequent words, sentences can be decomposed and analysed (Robinson 1975; Gordon 1996). Rewrite rules are used to classify each word into different categories (Russell 1995). The Emotion Status of each sentence can be assessed by identifying the emotional words and analysing their interaction with the other words in the sentence and determining whom the emotion refers to and what the intensity of the emotion is. Different weights and categories are assigned to individual words. The overall Emotion Status of the article is defined as the absolute difference between the number of positive and negative sentences.

The Online Emotion Stock Analyser consists of two layers: the interface layer (the input from the Internet browser, and the output that indicates the emotional state of the web page) and the emotion extraction layer (the text-to-emotion engine). The engine includes three parts: input analysis, the tagging system and the parser (figure 2).

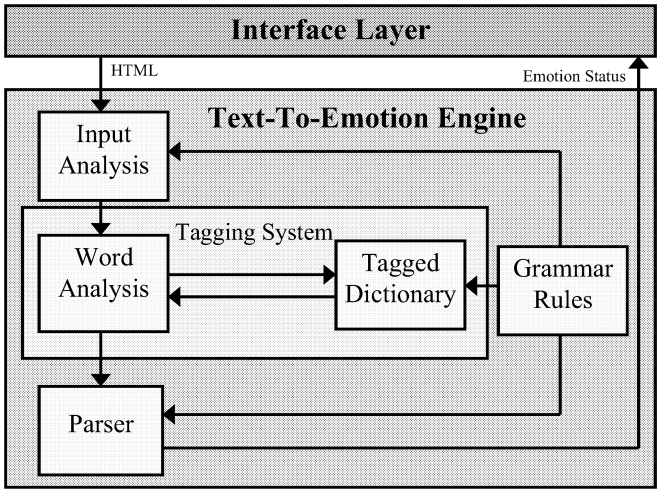


Figure 2: Online Emotion Stock Analyser program flow

Input

Input to the system is through an embedded Internet browser. Web pages are selected by following the hyperlinks in the browser interface. When a web page finishes loading, the text of the page is automatically extracted from the raw HTML by removing the text within HTML tags. This text is passed to the Text-To-Emotion engine for analysis, when the user clicks on the "Analyse" button.

There is a danger that the text of the article will be contaminated by other text displayed on the web page, such as links, or by badly formed HTML code. To verify the results of this experiment, the Emotion Status of the whole web page was compared against the Emotion Status of the article text, and no significant differences were found.

Input Analysis

The text of the web page is sent to the input analysis function for initial assessment. The tagging system can only handle one sentence a time therefore the input analysis function divides the text into individual sentences before sending them to the tagging system. To correctly identify the end of sentences, an analysis was carried out to test whether a full stop character is the sentence terminator, or is used for another purpose such as a decimal point.

The Tagging system

The tagging system contains two components; the word analysis mechanism and a tagged dictionary containing 22,000 words. In order to identify key-words, we manually searched through articles from The Financial Times and extracted the words related to the movement of stock prices. The stock word tag ("STO_W") was assigned to each

possible emotional word relating to the Stock Market. The tag-set uses numbers 0 and 1 to represent positive and negative price movements.
The tagging system searches through the tagged dictionary to find the corresponding tag category for each word in the sentence. A suffix and prefix examination is carried out to find words that have not been identified.
A special tag "NDP" (negative data point) is assigned to the words that give sentences the opposite meaning, for example, the word "halted" in the sentence "The market halted its slide", will overturn the meaning of the sentence. In addition, some phrases have opposite meaning compared to the words in it, for example, in the sentence "The market gave up yesterday's gains", the phrase "gave up" includes the positive word "up" but the phrase itself is negative. A special ambiguity tag is assigned to indicate these situations.
The output of the tagging system (the sentence words and the corresponding word category tags) is sent to the parser for further analysis.

Parser

The parser analyses the output from the tagging system using rewrite rules and tree representations (Russell 1995) to find possible combinable phrases and possible sentence structures. Any ambiguity is resolved within the context of the whole sentence, and previous sentences. If a word with the opposite meaning is found, the emotional state will be inverted.
The parser recognises conditional emotional sentences e.g., "the Stock Market will be stronger if the war finishes quickly" and interprets this as an emotional sentence. Sentences with more than one emotional word that are connected by a conjunction word, will be treated as an emotional sentence with two states, e.g. the sentence "The FTSE 100 was down but FTSE 500 was up" is recognised as an emotional sentence with both states.

Output

When the article analysis is complete, the interface displays the overall Emotion Status according to each sentence's emotion state and intensity (figure 3).

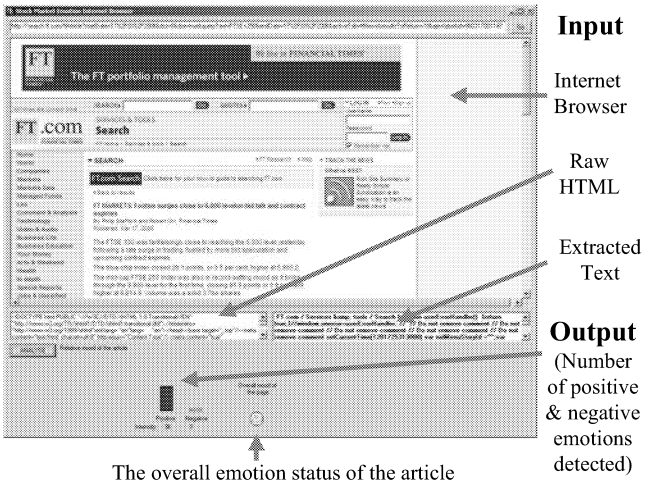


Figure 3: Online Emotion Stock Analyser interface

EXPERIMENT DESIGN

An experiment was carried out to test the effectiveness of the Online Emotion Analyser as a tool to predict the future movement of share prices in the London Stock Exchange.
A sample of 311 separate web pages were selected from the Financial Times web site. The sample contained one article for each day that the Stock Market was open during the 15 month period of 20 December 2004 to 17 March 2006. The chosen articles were written by different authors, but were from the same regular newspaper column that objectively summarised the movement of share prices and the value of the London Stock Market during the previous day.
The web pages were analysed by the Online Emotion Stock Analyser individually to derive an Emotion Status for each page (the number of positive emotions minus the number of negative emotions). The Emotion Status was compared against the value of the Stock Market as indicated by the Financial Times Stock Exchange index (FTSE). Calculations were performed for both the 100 share index (FTSE 100) and the 250 share index (FTSE 250).
Correlation calculations were carried out to compare:

1. The Emotion Status and the FTSE index for 8 days (from the day before, to one week after the day described in the article).
2. The Emotion Status and the value of the next day's FTSE index.
3. The Emotion Status and the movement of the next day's FTSE index.
4. A comparison of:
   a) The Emotion Status and the next day's movement of the FTSE index.
   b) The previous movement of the FTSE index and the next day's movement of the FTSE index.

Comparing the Emotion Status and the value of the FTSE index for 8 days

Correlation calculations were carried out over the 15 month period, comparing the Emotion Status of the articles with the FTSE index for a period of eight days, ranging from the day before the day described in the article, to six days in the future. This calculation assessed whether the Online Emotion Stock Analyser could accurately assess the "mood" of the market for the day described, and evaluates the best matches for the current mood in the following days. The results are shown in figure 4.
The closest correlation was for the day the article described (Day 0). For the FTSE 100 each of the next 6 days were above the correlation critical value for P=0.05 (0.113) which indicates the correlation was not by chance. For the FTSE 250 only the next day was above the correlation critical value. The correlation coefficient remains high for the FTSE 100 for the next three days (Day 1 - Day 3) but Day 1 is the only day that both the FTSE 100 and FTSE 250 are above the critical level.
This may reflect the fact that the articles concentrated their commentary on the companies in the FTSE 100 and not the companies in the wider FTSE 250. The articles focus on the companies that had most activity in the Stock Market.

The closest match was for the day that the article described which validates the use of the Online Emotion Analyser as a tool for identifying the emotion contained in Stock Market articles.

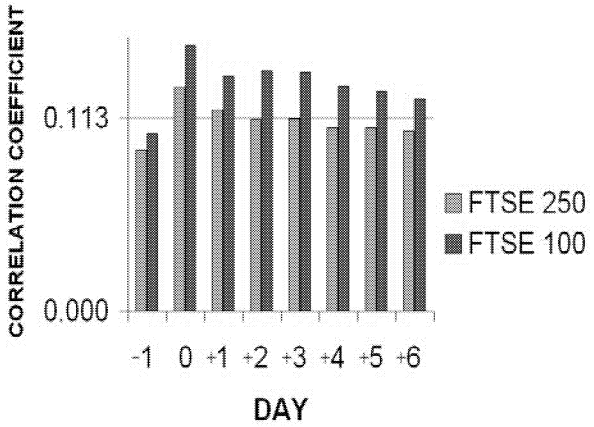


Figure 4: Daily Correlation between the Emotion Status of articles and the FTSE index over 8 Days

The best match with the future value of the Stock Market was the next day as it was the only day that there was a significant correlation for both the FTSE 100 and FTSE 250. This indicates that the Analyser is best suited to predict the value of the Stock Market the next day rather than longer periods of time in the future, and greater accuracy is achieved in the FTSE 100, than the FTSE 250.

The Emotion Status and the value of the next day's FTSE index

The results of the correlation calculations between the Emotion Status and the next day's share value as discussed above are shown in figure 5 and 6. A significant correlation was found between the Emotion Status and the FTSE 250 index, with a correlation coefficient of 0.117 (P=0.05) (figure 5).

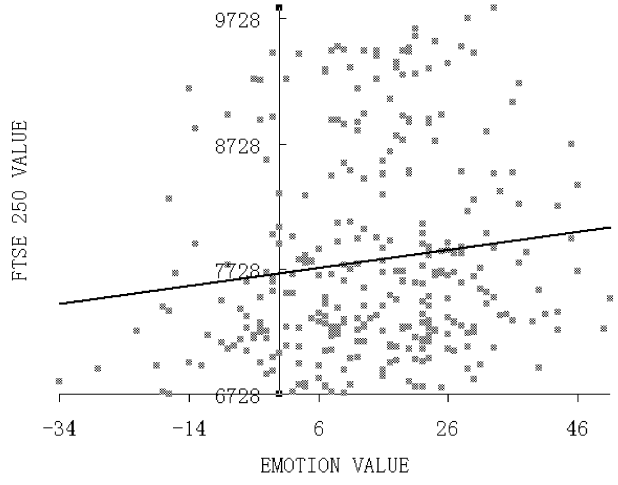


Figure 5: Correlation between Emotion Status and the FTSE 250 index

A significant correlation was also found for the FTSE 100 index, with a correlation coefficient of 0.137 (P=0.05) (figure 6).

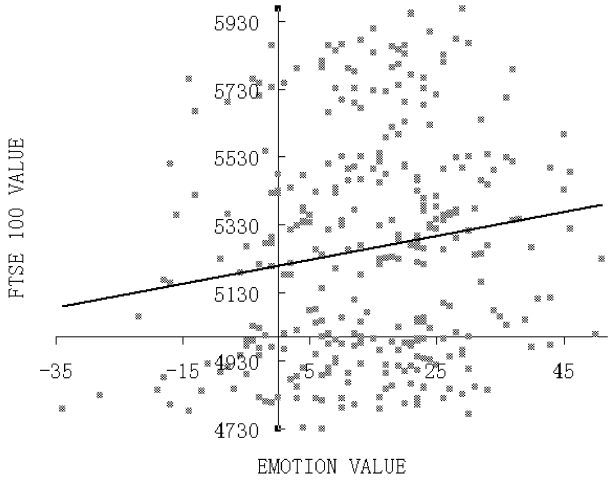


Figure 6: Correlation between Emotion Status and the FTSE 100 index

A link between the published articles and the behaviour of the readers has been observed, but this does not prove that their behaviour is solely based on the information obtained from this web site as the same information is available from other sources. This experiment verifies that the correct emotions are being detected, and that there is a link between the detected emotions and the future movement of the value of the Stock Market.

The Emotion Status and the movement of the next day's FTSE index

Correlation calculations were carried out between the movements of the Emotion Status of articles from one day to the next, with the movement of the next day's FTSE index value.
No significant correlation was found for the FTSE 250, but a significant correlation was found for the FTSE 100 index, with a correlation coefficient of 0.153 (P=0.01) (figure 7).

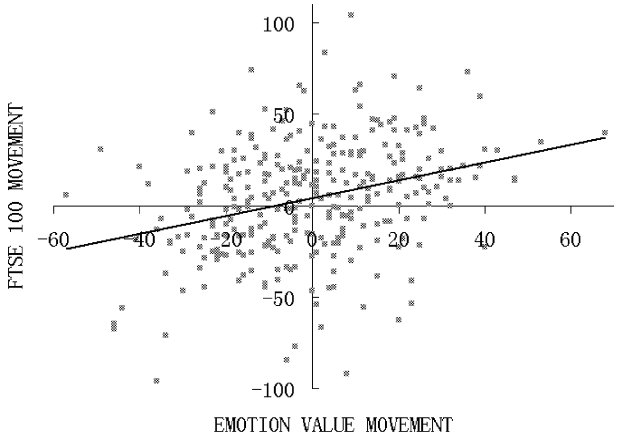


Figure 7: Correlation between the Movement of Emotion Status and the Movement of the FTSE 100 index

The result of this calculation again indicates a stronger correlation with the output of the Online Emotion Stock Analyser with the FTSE 100 than with the FTSE 250.

Emotion Status versus the previous movement of the FTSE index

It is possible to base predictions of future Market value on the previous values of the FTSE index, i.e. if the Market value is rising it may be reasonable to expect it to continue rising the next day. The final set of correlation calculations assessed whether the output of the Online Emotion Stock Analyser was a closer match to future value of FTSE index than the previous value of FTSE index.

The results of the correlation calculations for the Emotion Status and the next day's FTSE index movement were compared with the previous day's movement of FTSE index and the next day's movement of the FTSE index.

The results of the calculations are shown in figure 8. The first two columns show the results for the FTSE 250 index. Column one shows the correlation coefficient for the previous day's movement of the 250 index with the next day's movement of the FTSE 250 index, while the second column the correlation coefficient for the movement of the Emotion Status with the movement of the next day's FTSE 250 index.

The closest correlation is found using the Online Emotion Stock Analyser; however, both results are below the critical significance level. The next two columns show the corresponding results for the FTSE 100 index. There is a negative correlation between the previous movement of the FTSE 100 with the next day's movement of the FTSE 100, while there is a significant correlation between the movement of the Emotion Status and the movement of the next day's FTSE 100 index (0.153, P=0.01).

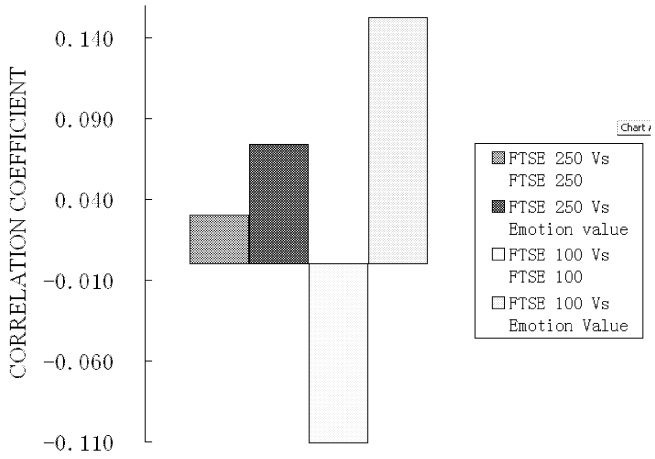


Figure 8: The Online Emotion Stock Analyser Versus the previous FTSE index

For both the FTSE 250 and the FTSE 100 indexes the correlation is stronger using the Online Emotion Stock Analyser than the values of the previous day's FTSE indexes. This shows the Emotion Analyser can be used as a more accurate predictor of future movements of the FTSE index than using the previous movement of the market.

CONCLUSIONS

We have developed an Online Emotion Stock Analyser that analyses the "emotions" contained in online articles about the movement of shares in the stock market.

Experiment results support the loop effect that has been observed between published articles and the behaviour of investors. A number of significant relationships have been observed between the emotion contained in published articles about the Stock Market and the next day's market value. There is a closer correlation between the Emotion Status of articles and the FTSE 100 index, than with the FTSE 250 index. The Online Emotion Stock Analyser was shown to be a more accurate predictor of the next day's movement of the FTSE index than using the previous day's movement.

Further research in this project could include the development of an application to assess the general emotions presented in a web page, with an experiment to determine how different combinations of emotions affect the perceptions of users. Authors of web pages could use this information to ensure the response of readers' match their intentions.

REFERENCES

Arps, J., 2007, *Arps Fear/Greed Index (Radar1)*, http://www.esignalcentral.com/university/esignal/addons/arps/fear.asp (accessed 11 Jan 2007).

Bergen J.V., 2003, *Investors Intelligence Sentiment Index*, http://www.investopedia.com/articles/trading/03/100103.asp (accessed 19 Jan 2007).

CBOE, 2007, *CBOE Market Volatility index (VIX)*. http://www.cboe.com/framed/ IVolframed.aspx (accessed 11 Jan 2007).

Evatt, D.L.S., 1997, *The influence of emotion-evoking content of news on issue salience*, PhD thesis, University of Texas At Austin.

Gaughan, P.A., 1986, *An Analysis Of The Impact Of Published News On Stock Prices And Stock Trading Volume*, PhD thesis, The City University of New York

Gordon, W.T., 1996, *Saussure for Beginners*, Harper Collins Publishers.

John, D., Boucouvalas, A.C. and Xu, Z., 2006, *Text-to-Emotion Analysis Engines - Theory and Practice*, Euromedia 2006.

Lo, A.W. and Repin, D.V., 2002, The Psychophysiology of Real-Time Financial Risk Processing, *Journal of Cognitive Neuroscience* 14, pp323-339.

Nelson, D., 1991, *How Business and Economic Coverage Changed In United States Daily Newspapers from 1970-1990*, PhD thesis.

Paiva, A., Prada, R. and Picard, R.W., 2007, Proceedings of the Second International Conference ACII 2007, Lisbon, Portugal 12-14 September 2007, Springer.

Picard, R.W., 2000, *Affective Computing*, MIT Press, Cambridge MA.

Robinson, I., 1975, *The New Grammarians' Funeral*, Cambridge University Press.

Rutterford, J., 2007, *Introduction to Stock Exchange Investment*, Third Edition, Macmillan Press.

Russell, R., and Novig, P., 1995, *Artificial Intelligence: A Modern Approach*, Prentice Hall.

DIGIMEM: REPRESENTING MEMORIES IN DIGITAL

Georgios D. Styliaras
Art Sciences Department
University of Ioannina, 45110, Greece
E-mail: gstyl@uoi.gr

KEYWORDS
Web application software, multimedia application, multimedia content management

ABSTRACT

This paper describes DigiMem, a web-based tool for producing a multimedia representation of an event's memory. An event can be a holiday; cultural events; conferences; exhibitions; social and sport events; and personal moments. Unlike other generic multimedia authoring tools and device specific software bundles, the main design paradigm behind DigiMem was to put the focus on the memory itself, rather than on the material that is representing it and the devices that captured the material. DigiMem comprises of two modes: the authoring mode where the content for a memory is prepared and structured properly and the playback mode where the memory is projected based on this content as a Dynamic HTML page. During both modes, the effort was to infuse seamlessly into the tool all the habits we everyday use in the real world for conserving our memories and define some new features that are only possible in the digital medium. Most features of DigiMem have been implemented by exploiting absolute positioning features. The representation is eventually published as a Dynamic HTML page.

INTRODUCTION

The paper introduces DigiMem, a web-based multimedia tool for representing memories of past events as multimedia documents. The aim of the tool is to represent a memory as vivid as possible with all digital material that is available at a specific moment. DigiMem tries to enumerate and support all features that we consider being necessary for such a representation and are currently scattered in different tools and technologies. In this way, DigiMem may be considered as a reference tool for such a representation. Memory representation may be enriched from time to time, as new material becomes available. Digital material may include images, sketches, video and audio. The events that DigiMem could be suitable for include holiday; cultural events; conferences; exhibitions; social and sport events; and personal moments. DigiMem comprises of two modes: the authoring mode where the content for a memory is prepared and the playback mode where the memory is projected based on this content as a web page. The tool presented may be used either for personal use, or for broadcasting a memory representation for a larger audience, which can annotate and discuss the event the memory is based or enrich and relate the memory representation to other memories. All these operations may be carried out either locally or online, by using, though, the same web-based interface.

As also discussed in Section 2, the effort during the design of DigiMem, unlike other generic multimedia authoring tools has been to put the focus on the memory itself, rather than on the material that is representing it and the devices that captured the material. Another innovative feature of DigiMem is that it permits the authoring of the memory representation through a web-based interface that requires neither technical, nor programmatic skills. The user is only required to place items on screens and set properties by using drag and drop and other simple interface elements, then save the representation online or locally and continue working at a later time, even from another machine. Additionally, all of the resources that are inserted in a representation are uniquely identified and have properties bundled to them, which follow them from authoring to presentation. The notion of flexibility is also present throughout the authoring mode of the tool: A lot of features are provided, but at any time only a set of them is suffice for certain presentation needs. For example, one memory representation may require a simple enumeration of images, whereas another may take advantage of the linking and overlay features. Another innovative feature is revisions that store several snapshots of a memory representation as a whole and independently for selected of its resources, as these evolve over time. What's more, instead of allowing a single resource (e.g. an image) to be placed in a specific area as in most multimedia authoring systems, in DigiMem a set of resources may be defined for a specific position and alternatively used. This feature permits the generation of several representations of the same memory depending on the exact combination of resources that are finally selected for every position before every publication. Multiple representations may address the needs of different audiences, such as for personal use, for the group of people that were present at the event on which the memory is based, or other people. Finally, the implementation of the playback mode as a simple Dynamic HTML page permits the viewing of the memory representation through a simple web browser.

The paper is organized as follows: Section 2 discusses the rationale behind the design of DigiMem and compares it with other work. Then, Section 3 describes the authoring mode and the playback mode of the tool, whereas Section 4 analyzes some interface elements of the tool. Then, Section 5 discusses implementation details, whereas Section 6 presents briefly some evaluation results. Finally, Section 7 concludes the paper by presenting some intended future work.

DISCUSSION AND RELATED WORK

People always have wanted to capture, store and reproduce their memories throughout their lives. A memory can be a travel, a visit to a museum or an archaeological site, a house or a neighborhood one has lived, a working environment, a concert, the company of friends, a strong emotion, a social or a sport event, a sudden event and everything some consider significant during their lives. In the film Until The End Of The World by Wim Wenders, Solveig Dommartin has a device that projects a past world's images. She watches over and over again these images, as she is afraid of forgetting close persons and landscapes that were displayed on them. Similar films that deal with the representation of illusive or past environments are The Eternal Sunshine Of A Spotless Mind and Vanilla Sky.

Ages before the appearance of digital technology, people used to keep up memories by writing, taking notes on personal diaries, sketching, painting and, later, by publishing books, by taking photographs and by video cameras. After gathering this material, people used to organize and further edit it. For example, they note the shooting date and the persons that were present in a photograph on its rear side or they store them in a photo album. They also label films and edit them. Personal diaries and journals are accompanied by dates and other information.

Digital technology offered the appropriate hardware and software that simplified the above processes with the use of a personal computer. Digital photo and video cameras help people shoot photos and films and transfer them to personal computers for further editing. It's a commonplace for everyone to take pictures during the holiday or shoot a video during a marriage. Once loaded on a computer, device dependent software bundles as well as generic multimedia editing packages (such as Macromedia Flash) help further editing the content. Even the simplest photo editors (for example AcdSee) permit the insertion of annotating tags. Accordingly, video editors (such as Adobe Premiere) allow selecting and moving specific film parts and creating navigation menus that point to specific film parts. Electronic blogs (as BlogSpot) permit the creation of personal journals over the World Wide Web. They are based on text and images and they can be available publicly and commented by others.

Apart from commercial tools, a lot of research has been conducted in the area of capturing the user's experience, but it mainly focuses on museums and tourist sites. With the use of handheld devices and mobile phones, one may grab images and video throughout a walk in a site and take notes at the same time, as in (Scherp and Boll 2004), (Hansen et al. 2004) and (Counsell 2002). These papers propose capturing content and interlink it with photos taken on the move. (Souza et al. 2005) deals with digital recording of information but the paper is focused on archaeological content. (Watkins and Russo 2005) is focused on the cultural domain and introduces digital cultural communication. This communication allows users to become co-creators of knowledge by providing tools and methods which enable the co-construction of creative artifacts. The concept in (Blue 2004) is to design a combination Book/DVD-ROM/Website that allows the reader direct interaction with the narrative. This project expands on the idea that interactivity is a dialogue between viewer and story by encouraging the book's reader to become a DVD-ROM user and website co-author. (Rocchi et al. 2004) and (Callaway et al. 2005) address the issue of the seamless interleaving of interaction with a mobile device and stationary devices during a museum visit. TOSCA (Salgado et al. 2002) provides user orientation with the capability to automatically show multimedia information (audio, static images, panoramic images and VRML). (Jacucci et al. 2004) exploits gesture-based interfaces as a means to navigate to previous multimedia recordings . Flickr is a program with which different users may upload photos to a common virtual place, but more concern is given on network collaboration, rather than on representation. A digital storytelling system is presented in (Fujita and Arikawa 2007), which is complementary to DigiMem.

However, regarding both analog and digital media capture devices, the representation of a memory tends to be device and media specific and, thus, fragmentary. For example, for a wedding someone may have stored photos in a photo-album; digital photos in the folder of a photo-editing program; a DVD disc in a bookshelf; a song of the wedding's reception in an audio CD; the invitation card; the description of the event in a personal blog accompanied with photos; and a photo or a video snapshot in the memory of a mobile phone. On the other hand, research related work shares some features with the tool presented, sometimes by going beyond a desktop computer, but the issue of gradual representation of a memory is not covered adequately.

Unlike photo and video editors or generic multimedia tools, DigiMem is a flexible multimedia environment that allows the definition of a certain memory representation in steps. Every step may be enriched with interconnected digital material such as a photo, a sketch, a video clip, a song, a label or a bigger text and a hyperlink.

It is true that the process of keeping a memory can be only a task of secondary priority among someone's activities and thus it may not be completed right after the principal event. For example, in the case of an analog photo camera, someone could take pictures, then develop the film, organize and reproduce some of the photographs, annotate them and place them in a photo album. It is not necessary that these tasks be performed simultaneously. Therefore, in the digital metaphor, it is necessary to be able to save the status of a memory's material and gradually enrich and edit it.

The nature of a memory is such that it must be stored as detailed as possible. Otherwise, it tends to fade out during time. So, while saving a memory, users should be able to attach to it all material that they are afraid of forgetting, even in an unorganized way at first, as they drop items in a bag. Even more, they should be able to come back to a saved memory and enrich it with new material, e.g., a better photograph of a landscape that they have revisited.

Regarding usage, most device specific software bundles and multimedia editors require some knowledge and technical expertise from end-users. Accordingly, generic multimedia authoring tools, even the most simple of those such as Microsoft Powerpoint, require programmatic and technical skills and the ability to find the right feature over a large set of options. Furthermore, it is difficult to deviate from the sequential, flat presentation type they are based on.

Finally, in most authoring and editing multimedia tools, imported multimedia content is usually anonymous or may be accompanied by a name. In DigiMem, multimedia content is organized as grouped and interlinked assets that may be reused or moved as a single unit. These relations are not applied just visually but are stored as reusable structures that are available to both modes of DigiMem. For example, as it will be described in the next section, for a certain photograph, the user may define on a layer above the photograph the participating persons along with the areas in the photograph that correspond to a certain person. This layer is related to the photograph and accompanies it during every usage of it. Furthermore, more layers may be defined for a certain multimedia object and more multimedia objects may be alternatively used for a certain position. Alternative objects may differ in quality, size, presentation purpose and time.

DIGIMEM PRESENTATION

In this section, the features of DigiMem will be described. DigiMem operates in two modes: the authoring mode and the playback mode. During the authoring mode, the user may create a new memory, analyze it in steps, import, annotate, organize and associate content for every step and define interactivity. At any time, the user may navigate among steps and enrich them with new content or make other corrections. During the playback mode, the memory is presented according to its steps. Playback may be paused and resumed and parts of it may be annotated, linked and enriched. The two modes are summarized in Figure 1.

Authoring mode

When the tool starts in authoring mode, the user may load an already saved memory or create a new one. In the latter case, a wizard screen appears that encourages the user to complete the memory's identity, i.e. title, date and some notes. Then the memory's steps should be defined, in parallel with the supporting multimedia pool, which concentrates the content that will be used in the presentation of the memory steps. More specifically the multimedia pools includes texts, acting as labels, persons, objects, places and narrations; images and videos representing persons, landscapes, cities, buildings, sites, background and memorabilia; panoramas consisting of a set of images that are stitched together and show an area around a certain viewpoint. maps on which some active areas may be defined and associated to short-sized explanations or point to another multimedia pool item or memory step; overlays consisting of a set of images, a basic one and the rest that may be transparently imposed over the basic one; guided tours defined by the selection of previously defined multimedia pool items that appear in a predefined order; and audio files of personal narrations, music or actual recordings that were heard during the event on which the memory is based.
When a new step is created, a blank screen appears. Around the screen, lay a set of thumbnails that correspond to multimedia pool items and thumbnails of previously defined steps. By using drag and drop, the user may define or set the following items: the step's background that may be an image from the multimedia pool, a solid or gradient color; squares on the step's screen and on every square, more than one multimedia pool items may be associated, one of which being the visible; freeform sketches that the user may draw in order to represent objects (e.g. landscapes, buildings, persons) for which no visual material (photos, videos) is available; background music chosen from the audio pool; and a map where the scene took place.
The user may keep going on in order to define all the steps of the memory. It should be noted that neither all features, nor all content must be used at once, e.g. some squares may remain blank, some steps may not be accompanied with music and sketches may be drawn later. During authoring, the following supporting features are available for the memory author: an identity viewer, a status bar showing the options the user has and shows what action is currently performed; synchronization that enables a time-based projection of items during the playback mode; and some zoom and transparency effects.
Having defined the steps, the user may navigate among them at any time and enrich them with more content or make corrections to existing content. More specifically, the user may perform the following maintenance operations that allow replacing sketeches with actual content, as the latter becomes available; linking among steps both sequentially and randomly; linking to World Wide Web resources or other memories; storing revisions of a memory's representation; publishing the representation on the World Wide Web; and making a rpoert on the content pool's items udage.

Playback mode

The web page that is produced by the tool's authoring mode may be viewed offline or through a web site by a standard web browser. Firstly, the memory's identity appears along with thumbnails that link to its steps. When a step is chosen, its objects appear with all supporting material and according to the synchronization and effects features. Apart from navigating, the user may perform the following operations that unfold from the corner of the step's screen: In case of images related to **imagemaps**, the user may choose to click on the image's segments and view the texts or follow the links that are possibly related to the image's segments. The most common usage of this operation would be to reveal the names of the persons that appear in a photograph. Regarding **thumbnail squares**, when a user hovers over such a thumbnail, the cursor turns into a magnifier tool. If such a thumbnail is clicked, the full version of the item that corresponds to the thumbnail is revealed. The user may choose to start or stop the **background music** and/or the rest of the sounds from a step. The user may set all kinds of **sounds**, the objects defined as overlays and steps to loop during the step's playback. In every step, the user may initiate a **chat session** where the content of the step can be enriched with new content and discussed with other viewers. Also the user may define a node on a memory's step that can be **linked** to other memories. In this way, a graph of memories can be created than can tell the same story from different perspectives. Finally, a **search interface** is available in every step that permits retrieval of items based on their textual information. Thus, the end-user can search persons, cities, landscapes etc.

INTERFACE ELEMENTS

As DigiMem addresses the needs of users with low computer expertise, its interface is modal, thus only a small set of permissible operations is available at any time, which constitute the operations the user may perform depending on the previous actions. For example, if the synchronization feature is selected during the editing of a memory's step, then the only permissible operation is to drag images from the step's screen on the time ruler. A check button updates the feature's status and permits the further editing of the items in the step's screen.

Towards this direction, all interface operations in DigiMem during both of its modes are based on the following, widely employed, simple interface elements: drag and drop, sliders, selecting items from dropdown menus and popups. *Drag and drop* is used when choosing an image from the multimedia pool and place it in a square, when placing thumbnails on the time ruler and when drawing sketches. *Sliders* are used for choosing a color and zoom and transparency levels in the respective features. The color palette contains some basic nuances and effects such as degradation. *Dropdown menus* help configure link destinations and assign text properties to multimedia items. *Popups* project item identities. Finally, the *status bar* is always present on the bottom of the screen and reveals the action that is currently being performed, the items that this action affects and which are the further permissible actions. Consequently, the tool's environment deviates from standard form-based applications that use lots of menus, buttons and controls. Ordinary actions (even in handwritten form) blend with simple computer interface actions in order to emulate, as much possible, a natural interface environment. Representing a memory is a hobby; therefore it should employ a relaxing interface, far from typical human-computer interfaces found in working environments, where results must be turning up instantly. In every case, not all features of DigiMem must be used in order to produce a successful memory representation: According to the given material and the presentation needs, the appropriate features are selected. At a later time, a new revision of the memory representation may be produced by exploiting more content and, if necessary, features. Furthermore, the editing of a representation is not a sequential procedure. On the contrary, the user should feel free to navigate among memory steps and complete or change content, parameters and features. The representation will still remain consistent, owing to the underlying integrity rules.

Finally, although steps may be defined before the supporting pools, it is recommended that the user first fills the pools with an adequate amount of content, so that the steps' analysis is performed based on already existing content. Then, pools and steps may be further completed with new material. Figure 1 presents schematically the procedure that a user needs to follow for representing a memory.

IMPLEMENTATION ISSUES

Most of DigiMem's features have been implemented by using Flash scripting that exploits absolute positioning of images and database access routines and enables the tool to be run either locally or through a web server. DigiMem encodes the playback file in Dynamic HTML, so that it can be viewed by most web browsers. In this way, no special software should be installed in the end-user's machine in order to view a memory. During authoring mode, a database is used for storing the items' details such as textual content and linking with external files, whereas every item in the multimedia pool is assigned a unique identity. Relations and groupings among multimedia pool items are also stored by exploiting their identities. By exploiting the images' position on the screens, image resizing routines and simple mouse events, it was possible to build the authoring mode of DigiMem and implement all features that require absolute positioning, such as organizing and arranging pool items; placing items on the time ruler and calculating the time they will appear during playback relatively to the timer ruler total length; dropping and associating items with a square when the item's and the square's surfaces are well overlapped; supporting relations, groupings and imagemaps; and positioning squares on the step's screen. When publishing a memory, the appropriate content is retrieved from the database and embedded in the playback file. The content includes selected images, sketches, externally linked video and audio and details about placement, settings and effects. These parameters allow the authoring mode to construct the Dynamic HTML page that projects the memory. In the playback HTML file, multimedia files are pointed by using IMG tags; settings are expressed by HTML properties; effects are implemented as JavaScript routines.

EVALUATION

In its current implementation, DigiMem has been presented to a group of undergraduate students in the Art Sciences Department of University of Ioannina in order to be evaluated. A brief presentation of the evaluation procedure and the results follows: A group of 20 students participated in the evaluation, which comprised four tasks: Use of DigiMem and opinion expression through a questionnaire; Silent observation of students while working with the tool in order to discover practices employed by users that agree or disagree with initial goals; Assignment of collection making tasks to two groups of students, the first of which used traditional procedures, while the second worked with DigiMem in order to form a representation of a current trip; and Interviewing of participating students while working with the authoring mode and the playback mode. Primary evaluation results are encouraging and have shown some desired features, as these are presented in the next section.

CONCLUSIONS

In this paper, DigiMem has been presented, a web-based tool that tries to recompose an experience from a past event through multimedia. It gives end-users the flexibility to employ in this composition all available material, form the memory representation gradually and share the representation for further commenting and linking. The main effort was to infuse seamlessly into the tool all the habits we everyday use in the real world for conserving our memories and define some new ones that are only possible in the digital medium. Former actions include arranging photos in an album; projecting a video to friends; and keeping a ticket

from a sport event or an invitation card. Regarding the latter, taking notes on a photograph; the combination and linking of the material; the effects that may be set; and the online publication of the memory can be only available by using multimedia. Possible deficiencies or extra desired functions denoted by evaluation results will help refining the tool's features and possibly integrating new ones. For example, implement a version of the authoring mode that could be compatible with a handheld device. The effort is to capture and classify content in the multimedia pool while an event takes place. Towards this direction, the panorama feature will be combined with the orientation capability that some handheld devices provide, so that it will be easier to define a link in specific directions of the panorama. Other improvements, which derived from evaluation interviews, include more tight cooperation with Web 2.0 technologies in terms of chatting and sharing memory representations; use of touch pads instead of mice in order to perform copy and paste operations on images in a more natural way; and simplifying memory pool population and other similar tasks during the authoring mode.

REFERENCES

Scherp, A. and S., Boll. 2004. "Generic Support for Personalized Mobile Multimedia Tourist Applications", in *Proceedings of ACM Multimedia 2004 conference*, 178-179.

Hansen, F.; N., Bouvin; B., Christensen; K., Grønbæk, T., Pedersen; and J, Gagacj. 2004. "Integrating the Web and the World: Contextual Trails on the Move", in *Proceedings of ACM Hypertext 2004 conference*, 98-107.

Counsell, J.. 2002. "An Evolutionary Approach to Digital Recording and Information about Heritage Sites", in *Proceedings of VAST 2002 conference*, 33-42.

Souza, M.; Postula, D.; Bergmann, A.; and M., Ros. 2005. "A Multimedia Guidebook Implementation Using a Bluetooth Wireless Information Point Network", in *Proceedings of WMASH'05 conference*¸ 33-38.

Watkins, J. and A., Russo. 2005. "Digital Cultural Communication: Designing Co-Creative New Media Environment", in *Proceedings of C&C'05 conference*, 144-149.

Blue, C. . 2004. "The Dawn At My Back", in *Proceedings of ACM Multimedia 2004 conference*, 993-994.

Rocchi, C.; O., Stock; M., Zancanaro; M. Kruppa; and A., Krüger. 2004. "The Museum Visit: Generating Seamless Personalized Presentations on Multiple Devices", in *Proceedings of IUI'04 conference*, 316-318.

Callaway, C.; T., Kuflik; E., Not; A., Novello; O., Stock, and M., Zancanaro. 2005. "Personal Reporting of a Museum Visit as an Entrypoint to Future Cultural Experience" in *Proceedings of IUI'05 conference*, 275-277.

Salgado, L.; E., Rendón; and R., Artola. 2002. "System Architecture for Tourist Orientation: The TOSCA High-End System", in *Proceedings of VAST 2002 conference*, 285-294.

Jacucci G.; J., Kela; and J., Plomp. 2004. "Configuring Gestures as Expressive Interactions to Navigate Multimedia Recordings from Visits on Multiple Projections", in *MUM 2004 conference*, 157-164.

Fujita, H. and M., Arikawa. 2007. "Creating Animation with Personal Photo Collections and Map for Storytelling", in *Proceedings of the EATIS 2007 conference*.

GEORGIOS STYLIARAS (1974) is a computer engineer and is currently a lecturer in the Department at the University of Ioannina, Greece. He has implemented many multimedia applications and systems with educational and cultural content. His research interests include multimedia applications for the open space, as well designing multimedia platforms for art applications. E-mail: gstyl@uoi.gr.

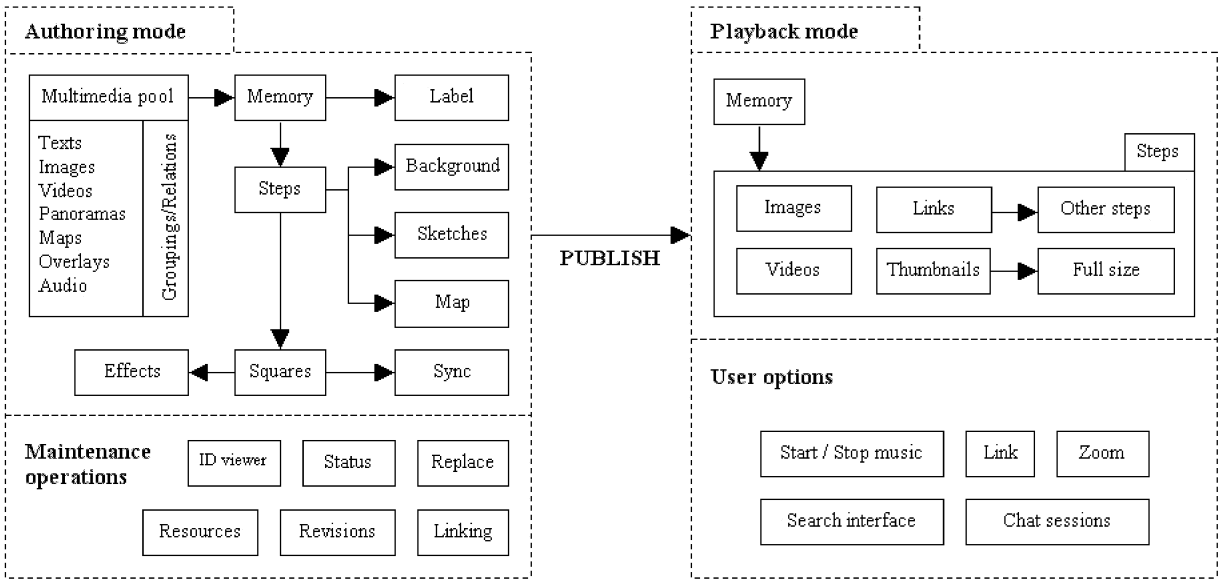


Figure 1: DigiMem usage

VIDEO CODING

HIGH-LEVEL PARALLEL H264/AVC ENCODER SPECIFICATION FOR MULTIPROCESSOR IMPLEMENTATION

Hajer Krichene Zrida
ENIS University
Computer & Embedded Systems Lab
3038 Sfax, Tunisia
Hajer_kri@yahoo.co.nz

Abderrazek Jemai
INSAT University
LIP2 Laboratory
BP676 – 1020 Tunis CEDEX Tunisia
abderrazek.jemai@insat.rnu.tn

Ahmed Chiheb Ammari
INSAT University
MMA Laboratory
BP676 – 1020 Tunis CEDEX Tunisia
ac.ammari@yahoo.fr

Mohamed Abid
ENIS University
CES Laboratory
3038 Sfax, Tunisia
mohamed.abid@enis.rnu.tn

KEYWORDS
H264 video encoder, coding performance, implementation complexity, task-level parallelization, Kahn Process Networks, YAPI, CAST and concurrency optimization

ABSTRACT
H264/AVC (Advanced Video Codec) is a new video coding standard developed by a joint effort of the ITU-TVCEG and ISO/IEC MPEG. This standard provides higher coding efficiency relative to former standards at the expense of higher computational requirements. This paper presents first a high-level complexity analysis of a H264 video encoder allowing for complexity reduction at the high system level. The coding performance is reported in terms of PSNR and bit rate. The complexity of the obtained cost-efficient configuration outlines the potential of using multi processor platforms for the execution of a parallel model of the encoder. For this, a YAPI-level parallel Kahn process network (KPN) model is proposed and implemented using the YAPI library Programming Interface. Finally, the system-level software CAST tool is used for a concurrency analysis of the implemented YAPI model to identify the potential concurrency bottlenecks to be resolved using "Task-splitting", "Data-parallelism", and "Task-merging" forms of parallelism.

INTRODUCTION

The H264/AVC has been designed with the goal of enabling significantly improved compression performance relative to all existing video coding standards (Joch et al. 2002). Such a standard uses advanced compression techniques that in turn, require high computational power (Alvarez et al. 2005). For a H264 encoder using all the new coding features, more than 50% average bit saving with 1–2 dB PSNR video quality gain are achieved compared to previous video encoding standards (Saponara et al. 2004). However, this comes with a complexity increase of a factor 2 for the decoder and larger than one order of magnitude for the encoder (Saponara et al. 2004).

Implementing a H264/AVC video encoder represents a big challenge for resource-constrained multimedia systems such as wireless devices or high-volume consumer electronics since this requires very high computational power to achieve real-time encoding. While the basic framework is similar to the motion compensated hybrid scheme of previous video coding standards, additional tools improve the compression efficiency at the expense of an increased implementation cost. For this, the exploration of the compression efficiency versus implementation cost design space is needed to provide early feedbacks on the standard bottlenecks and select the optimal use of its coding features.

In a previous study (Krichene Zrida et al. 2007), we performed a high-level performance analysis of a H264 video encoder to evaluate its compression efficiency versus its implementation complexity and to highlight important properties of the H264/AVC framework allowing for complexity reduction at the high system level. In this study, the complexity analysis covered major H264 encoding tools. Each new tool has been tested independently comparing the performance and complexity of a complex configuration to the same configuration minus the tool under evaluation. The coding performance is reported in terms of PSNR (Peak Signal-to-Noise Ratio) and bit rate, while the complexity is estimated as the total computational processing time of the application.

Absolute complexity values of the obtained cost-efficient configuration of the H264 encoder confirmed the big challenge of its cost-effective implementation. Given this, we will motivate the use of a multiprocessor approach to share the encoding time between several embedded processors. For this purpose, the sequential encoder reference code has to be distributed using a parallel programming model over a multiprocessor architecture. Prior to task-level decomposition and parallelization of the sequential H264 reference code (JM10.2. 2005), a computational profiling shall be considered to identify the most computationally-expensive tasks and to give a clear picture of the critical code parts candidate for parallelization. Based on the obtained profiling results and using the two predominant forms of parallelism (task and data levels), a first parallel model of the encoder will be proposed

and validated using Kahn process networks (KPN) models (Kahn. 1974) implemented by the Y-chart Applications Programmers Interface (YAPI) library of the multi-threading environment (Kock et al. 2000).

The paper is organized as follows. The next section presents the performance and complexity of the H264 major encoding tools. Section 3 discusses main aspects and issues for getting a high level parallel model of the H264 video encoder. A parallel YAPI-KPN model is proposed and the associated functional simulation results are presented in section 4. Section 5 provides task-level concurrency analysis and optimization using the CAST (Concurrency Analysis Software Tool) tool (Stuijk et al. 2003) and concluding remarks are given in the final section.

PERFORMANCE AND COMPLEXITY ANALYSIS

In a previous work (Krichene Zrida et al. 2007), we performed a complete performance and complexity analysis of a H264 video encoding application. In this section, the performance and complexity of the H264 major encoding tools are reviewed.

When combining the standard new coding features, the implementation complexity accumulates, while the global compression efficiency becomes saturated (Saponara et al. 2004). To find an optimal balance between the coding efficiency and the implementation cost, a proper use of the H264/AVC tools is needed to maintain the same coding performance as the most complex reference configuration while considerably reducing complexity. For this, a parametric influence of major encoding tools of this standard on performance and computing time complexity has been evaluated (Krichene Zrida et al. 2007).

In comparison with the most complex configuration, a complexity reduction of more than 80% has been achieved with less than 10% average bit rate increase for all the CIF and QCIF used test sequences. However, even with this configuration offering an optimal trade-off between coding efficiency and implementation complexity, we are still very far from a real time performance of 25 frames per second. Implementing this configuration of the encoder represents a big challenge for resource-constrained multimedia systems since this requires very high computational power to achieve real-time encoding.

PARALLEL KPN-BASED H264 MODELING ISSUES FOR MULTIPROCESSOR IMPLEMENTATION

In the previous section, we motivated the implementation of H264/AVC encoder application on a multiprocessor platform. The availability of many processing cores in such a multiprocessor system speeds-up the processing by introducing several types of parallelism, which are not envisioned for the implementation on single processor systems (Jemai et al. 1997), (Dwivedi et al. 1995), (Namneh at al. 1995). The predominant forms of parallelism are data-level parallelism (DLP) and thread-level parallelism (TLP). DLP is perhaps the most commonly used form of parallelism, implemented through vector or SIMD architectures. The benefits of thread-level parallelism (TLP) are achieved by distributing the workload of a single high performance processor among a number of slower and simpler processor cores. These two forms of parallelism should be used to get an optimal parallel model of the H264 video encoder. Any formal parallel model will be represented as a Kahn process network.

Typically, an optimal parallel khan process network model is obtained while extracting the concurrence available in the sequential reference code. For this, a computational profiling is considered to identify the most computationally-expensive tasks of the cost-efficient encoder configuration. Based on the profiling results, the task and data parallelism are exploited to maximize the parallelism between processes. Finally, the communication granularity must be determined for an optimal trade-off between the execution and the inter tasks communication time.

Computational profiling the execution of the cost-efficient encoder configuration

To identify the most computational-expensive tasks and to give a clear picture of the critical code parts candidate for parallelization, a computational “Gprof” GNU (Graham et al. 1982) profiling is performed for the 300 frames QCIF “Container” sequence using the cost-efficient H.264/AVC encoder configuration. The obtained results are reported in figure 1 in terms of the CPU time percentage spent in the execution of each module.

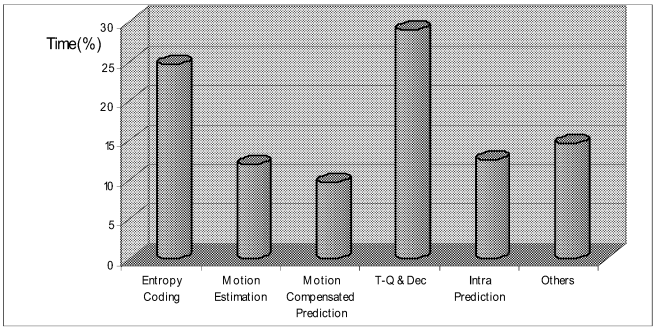


Figure 1: Computational Profile of H264 video encoding

The obtained profile shows that the motion estimation and compensation (MEC), DCT transform module, the entropycoding, and the intra-prediction modules are the most time-consuming modules. These tasks constitute the major bottlenecks of the encoder.

It is shown from figure 1 that the most computationally intensive task is the “T-Q & Dec” module (with about 30% of the total encoding time). For our case, this module represents the DCT transform T, the quantization Q, and the decoder rescaling and inverse transform blocks. In depth complexity exploration of this module has shown that the DCT transform is the major computational block of this module.

Task and Data parallelism

The application block diagram always serves as a starting point for extracting the task-level parallelism (Pastrnak et al. 2006). This type of parallelism may be achieved by decomposing the whole application into separate blocks. Each block defines one single task that runs a separate stage of an algorithm. For this case, the sequential H264 encoding algorithm is first split into concurrent tasks that may be executed at the same time, and then the necessary inter-task communication is established using message passing KPN primitives (Youssef et al. 2004). Each task represents one single block among the remaining blocks of the standard H264 diagram. Generally, such a parallel task execution is limited by data dependency between tasks. A data dependency means that one task needs the result of another one to be processed therefore limiting the ways for parallelization (Pastrnak et al. 2006).

In addition, using the obtained profiling results, as the "T-Q & Dec" module represents the most time-consuming module with about 30% of the total encoding time, data-level parallelism is also considered to increase the throughput and decrease the latency of the H264 optimal encoder application. The idea behind data parallelism is to perform the same transform on different data elements in parallel. For instance, it is possible to perform the "T-Q & Dec" Transform, quantization and decoding in parallel for the luminance plane (Y), the chrominance plane (U) and the chrominance plane (V). However as the combined size of the two chrominance planes is smaller than the size of the luminance plane for the used 4:2:0 coding format, it is possible to process the two chrominance planes consecutively in one chain.

Communication granularity

The third aspect to consider when extracting parallelism is the granularity at which data is communicated. The optimal communication granularity between tasks must be correct determined to prevent tasks from waiting and avoid spending too much time on synchronization.

Typically, the video processing in the H264 standard can be conducted either in Group of Pictures (GoP), slice, frame, or Macro-Block (MB) levels. This is made because of the hierarchical structure of the video streams, as described in the figure 2. Given this, the optimal level of granularity, at which each processor can operate independently, yet simultaneously with other processors, should be selected.

Many previous works elaborated on the task-level parallelization of different coding applications like those presented by *Shen* in (Shen et al. 1995) and *Bozoki* in (Bozoki et al. 1996) showed that the GoP-level parallelization for such a distributed system provides encoding performances better compared to those obtained with the others communication levels. However, this is not suitable for embedded System-on-Chip (SoC) implementations seeing that a substantial on-chip and shared memory capacity are required for the parallel compression of a group of pictures (Jacobs et al. 2006). The better granularity for such a SoC is thus threading at the fine grain level, i.e. at MB level, since only the current and reference frames needs to be stored. Each frame is considered as the current workload, and the encoding process of each frame is divided between the processors.

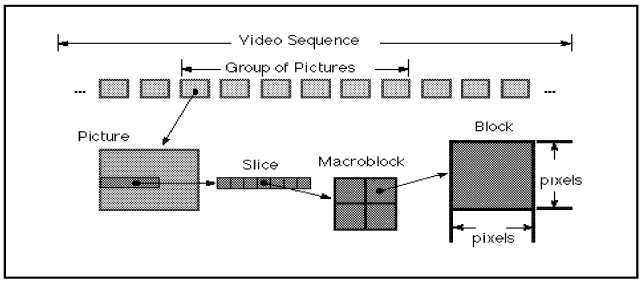


Figure 2: Layer Structure of a video stream

The proposed parallel KPN-based model

In our approach, Kahn Process Networks (KPNs) are used as a parallel programming model of computation. The KPN model assumes a network of concurrent autonomous processes that communicate in a point-to-point fashion over unbounded FIFO channels, using a blocking-read and write synchronization primitives. Read actions from these FIFOs block until at least one data item becomes available. The execution of a Kahn Process Network is deterministic and independent of process interleaving, meaning that for a given input always the same output is produced and the same workload is generated, irrespective of the execution schedule (Kock and Essink. 2001). The key characteristic of the KPN model is that it specifies an application in terms of distributed control and distributed memory which allows us in a future work to map the application onto a multiprocessor platform in a systematic and efficient way.

Given the functional blocks diagram of the H264 video encoder, the sequential C-code specification, and the results of the computational profiling, we propose the first parallel Kahn process network model of figure 3.

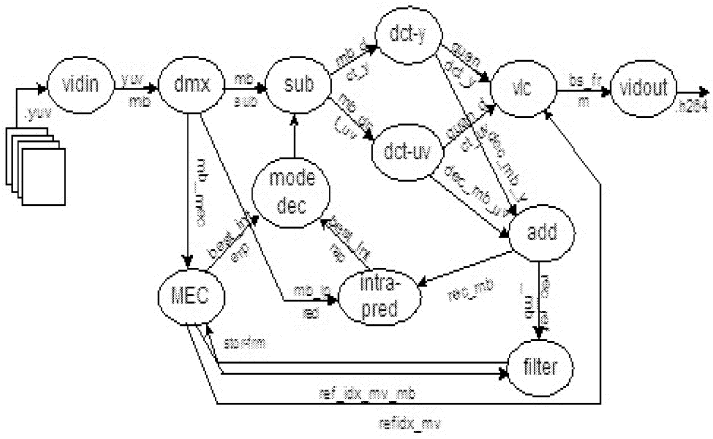


Figure 3: Proposed Parallel KPN-based H264 model

The "video_in" process shown in figure 3 represents the input of the encoder. It is responsible for collecting the video data (YUV frames) from the input file (video sequence with YUV format), the frame width and height dimensions, the total frames number, and the frame rate information. Each frame is divided into macro-blocks of 16x16 pixels. The "dmx"

process forwards the YUV-MBs to the "substract", "mec", and "intra prediction" processes. The "substract" process reads the predicted MB, subtracts it from the current MB, and sends the luminance residual data into the "dct_y" task and the chrominance data into the "dct_uv" task. The "dct_y" and the "dct_uv" processes perform associated transforms, respectively on the Y luminance and the UV chrominance macroblocks (MB). These macroblocks are received via dedicated data channels. Each 4x4 block is first transformed into block of DCT coefficients using an integer transform, then Q quantized, and finally decoded via a rescaling and an inverse transform. The "vlc" process receives the quantized DCT coefficients via two channels, applies on the CAVLC entropy coding method (Ostermann, 2004) and transmits the resulting compressed bit stream into the "video_out" process. Finally, the "video_out" process sends the H264 compressed data bit stream to the output file (.h264).

The "adder" process uses the decoded chrominance and luminance macroblocks to reconstruct the previously encoded macrobock. Using the current MB "mb_ipred" output of the "dmx" process and the reconstructed (but un-filtered) previously encoded MB "rec_mb" from the "adder" process, the "intra_prediction" process performs an intra-prediction on each macro-block using 9 prediction modes for 4x4 luma blocks, 4 prediction modes for 16x16 luma blocks, and 4 modes for 8x8 chroma blocks. The best intra-prediction mode cost obtained and the associated predicted MB are sent to the "mode_decision" process.

Parallel to the intra-prediction, each "dmx" output "mb_mec" current macroblock is inter-predicted using one or more reference frames in the "mec" process. The reference frames are received from the "loop_filter" process that is applied on each reconstructed frame (before storing the macroblock for future predictions) to reduce blocking distortion. The "loop_filter" process receives via two channels the reconstructed decoded MBs from the "adder" task and information about the references indexes and the motions vectors of the reconstructed MB from the "mec" task. The best inter-prediction mode cost obtained and the corresponding predicted MB are sent to the "mode_decision" process. Using the best intra-prediction and inter-prediction modes, the "mode_decision" process selects the best optimal predicted macroblock between them.

YAPI-LEVEL FUNCTIONAL SIMULATION OF THE PARALLEL KPN-BASED IMPLEMENTATION

The parallel H264 KPN model of figure 3 is implemented using the YAPI multi-threading environment. The YAPI implemented parallel KPN model is then validated by high level functional simulation. Typically, such a parallel thread-level implementation could not be performed in an automatic way only for small range of applications (Lange. 2004). For this case, we started with the sequential C reference code of the fixed optimal configuration defined in section 2. The sequential code is modified and structured by hand to describe the KPN in C++. Each Kahn process is described by a set of associated functions got from the original code. The inter process communication is performed using solely the YAPI I/O primitives. Using global variables for this purpose is not allowed. Thus, to ensure inter process communication, all of the shared variables used in the sequential reference code are grouped into associated data structures for communication over FIFO channels.

YAPI run-time environment

YAPI is a C++ library used to model signal processing applications as process networks based on the Kahn computation model. A process network is composed of processes and FIFOs. A process represents a computing node and a FIFO represents a communication channel. Each process interacts with its environment through input and output ports. It can be in running state**,** blocked state, or ready state**.** A process gets blocked when it tries to read from an empty FIFO or it tries to write into a FIFO which is full**.**

There are three primitives provided by YAPI for communication over channels. The "*read*" primitive is used to read from one channel. The "*write*" primitive is used to write into the channel. The "*select*" primitive is used in *non* deterministic applications (Kock et al. 2000). A validated YAPI application model is platform independent and hence, a validated specification is used as a starting point of any multi core platform implementation.

YAPI programming constraints

To describe the KPN using the YAPI C++ library, the application programmer has to take care of the following points:

- With the YAPI environment, a separate stack space is allocated for each process of the network. This stack is used to store the intermediate results, the local variables, and all function calls from the main member functions. For the H264 video coding application, there are a lot of local large video data structures that are allocated on the stack. As the total amount of stack space of each process is fixed to 64 Kilo bytes, this may be insufficient and results in a "*stack overflow*" (Kock and Essink. 2001). Such a stack overflow will leads to access violations that causes the program to be killed and a core dump to be generated. In our case, the adopted solution consisted in replacing the stack allocated local large video data structures by using the "*malloc"* and "*new*" routines to allocate these data structures dynamically on the heap.

- Typically, the processes could be blocked by a "*read*" operation if the number of produced tokens is smaller than the number of tokens to be consumed. On other hand, the YAPI run-time environment allocates 8 Kilo bytes of memory as a maximum size for each FIFO (Kock and Essink. 2001). So, if the maximal FIFO size is reached, the "*write*" operation can be also blocked and deadlock is appeared. To avoid such a writing deadlock for a given FIFO, the programmer has to study the tokens number being able to circulate in this channel before starting the "*read*" operation. This can be performed using two

optional arguments while declaring one FIFO, namely the minimum size and the maximum size.

Workload analysis

The proposed parallel model of figure 3 has been validated at YAPI system level. At this level, when this model is executed, the YAPI "*read*", "*write*", and ""*execute*" functions generate information on computation and communication workload of the application. For a QCIF "Bridge close" of 13 YUV frame sequence, the communication workload analysis is obtained, and is shown in the figure 4.

	size	Tsize	Wtokens	Wcalls	T/W	Rtokens	Rcalls	T/R
h264.YUVMB	1	768	1287	1287	1	1287	1287	1
h264.YUVMbToSub	1	768	1287	1287	1	1287	1287	1
h264.PredYUVMbToSub	1	640	1287	1287	1	1287	1287	1
h264.PredYUVMbToAdd	1	640	1287	1287	1	1287	1287	1
h264.LumaMbToDCT	1	38604	1287	1287	1	1287	1287	1
h264.ChromaMbToDCT	14	36	1287	1287	1	1287	1287	1
h264.QuanLumaMb	1	38604	1287	1287	1	1287	1287	1
h264.QuanChromaMb	14	36	1287	1287	1	1287	1287	1
h264.DecLumaMb	1	38604	1287	1287	1	1287	1287	1
h264.DecChromaMb	14	36	1287	1287	1	1287	1287	1
h264.AddedMbToReconsPic	1	768	1287	1287	1	1286	1286	1
h264.YUVMbToIntraPred	1	768	1287	1287	1	1287	1287	1
h264.BestIntraPred	1	648	1287	1287	1	1287	1287	1
h264.AddedMbToFilter	1	768	693	693	1	693	693	1
h264.RefIdxMvToFilter	42	12	594	594	1	594	594	1
h264.RefIdxMvToVlc	1	304	1188	1188	1	1188	1188	1
h264.FiltredStorableFrm	1	3284	7	7	1	7	7	1
h264.YUVMbToMotionEst	1	768	1188	1188	1	1188	1188	1
h264.BestInterPred	1	648	1188	1188	1	1188	1188	1
h264.BitStreamMb	12	40	13	13	1	13	13	1

Figure 4: Communication workload of the implemented H264 YAPI/KPN model

This figure describes the total number of tokens communicated between tasks via dedicated data channels. In the validated parallel model, the number of tokens per call is equal to 1 for all the "*reading*" and "*writing*" operations. Each QCIF frame consists of 99 macroblocks of 16x16 pixels. Every MB contains information about the Luminance (Y) and the Chrominance (UV) in a 4:2:0 format. That means that one MB contains two 8x8 blocks of Chrominance data, and one 16x16 pixels block of Luminance data. Given this, it is shown in figure 4 that 1287 (99*13frames) intra and inter macrobloks received by the "dmx" process via the "YUVMB" FIFO from the "video_in" process. The same MBs number is sent from the "dmx" task into the "substract" and "intra_prediction" tasks over respectively the "YUVMbToSub" and "YUVMbToIntraPred" channels. However, there are only 1188 inter predicted and bidirectional MBs sent from the "dmx" process into the "mec" process using the "YUVMbToMotionEst" channel.

The macroblock data structure is represented by one single token. But in the "substract" process, this structure is divided into two different data structures representing the Y Luminance MB and the UV Chromimance MB. These two tokens structures are received respectively from the "dct_y" and "dct_uv" processes with a total tokens number equal to 1287 via the "LumaMbToDCT" and "ChromaMbToDCT" FIFOs. On the other hand, the "vlc" process receive this tokens number twice times via the "QuanLumaMb" and "QuanChromaMb" channels to entropy code and to finally generate 13 bit streams sent into the "video_out" process.

Among 13 encoded YUV frames, there are 7 references frames (one I-frame, 2 P-frames, and 4 B-frames) received by the "mec" process from the "loop_filter" process via the "FiltredStorableFrm" FIFO. These references frames are reconstructed from 693 previously encoded macrobocks communicated between the "adder" and the "loop_filter" processes.

SYSTEM-LEVEL CONCURRENCY OPTIMIZATION ORIENTATIONS WITH THE CAST SOFTWARE TOOL

In this section, we will focus on the task-level independent target-architecture concurrency optimization of the implemented YAPI Kahn Process Network model of figure 3. The goal of this optimization is to transform this YAPI/KPN model in a structured way into a computational network that has a balanced workload and good communication behavior while maximizing the parallelism between tasks. We are thus seeking for an optimal parallel implementation of the H264 encoding application with respect to their actual concurrency measures introduced by *Sander Stuijk* in (Stuijk and Basten. 2003). Such measures would give a good indication whether the initial computational network design is optimal with respect to concurrency. To compute these measures, the system-level software CAST tool (Stuijk et al. 2003) is used. The input of the CAST tool is a YAPI/KPN network specification. The concurrency measures output shall provide insight to the potential bottlenecks and a global guidance when optimizing concurrency. Network bottleneck nodes have low values of concurrency measures. When a potential bottleneck is found, the programmer should modify the original specified network to resolve this bottleneck. For this, the three forms of parallelism "task-splitting", data-parallelism", and "task-merging" have to be explored to get high concurrency measured values.

Concurrency measures

The execution of a computational network can be described as a sequence of events (Stuijk and Basten. 2003). These events are ordered in time and per node. Based on these events, several measures may be derived to analyze the concurrency of the network under consideration. For an optimal parallel network model, these concurrency measures should be optimized. Typically for an optimal model, concurrency measure values are as close to 1 as possible (Stuijk and Basten. 2003). These measures are listed as follows:

- ***Computational load***: this measure describes the ratio between the amount of time that a compute node spends on computation with respect to the time spent in computation and communication. With a measured ratio of 1, a node of a given network is only doing computation.

- ***Execution load***: this measure considers the ratio between the execution time (computation plus communication time) and the run-time (execution plus idle time). This measure deals with the workload balance distribution over

the individual nodes of a network. If the execution load of all nodes is close to 1, this means that all nodes have the same amount of work to do.

- ***Restart***: is the third measure of concurrency. The throughput of a node or network is covered by the restart. This measure is inverse dependent on the runtime of the node. The inverse runtime means that a low restart for a node indicates that this node has a low throughput, and vice versa. For an optimal concurrency design, the throughput of all the nodes should be the same. No node or network should wait too much for another node or network to acquire or transmit data. Therefore all nodes or networks should have approximately the same restart values. The height of these restart values is dependent on the runtime of the system, and is therefore not able to reach 1 in useful designs. It is important however that the restart is as high as possible.

- ***Synchronization***: this is a measure of the speedup of a parallel network model in comparison with its sequential execution. Thus, if the synchronization value is close to 0, the communication and synchronisation overhead of the network are too high and the network is even slower than the sequential system. When the synchronization measure is close to one, executing the network takes almost no more time in comparison of the time it takes to execute the sequential system. Therefore this measure will be as close as possible to 1 in an optimal design.

- ***Structure***: the number of different data streams is considered in this last concurrency structure measure. When this measure is close to one, all nodes in the computational network take part in only few data streams. The parallelism is high in that case. If the structure measure is close to 0, there is more data parallelism available in the network which should be explored.

Concurrency bottlenecks analysis of the proposed model

For the proposed model of figure 3, the "video_in" and "video_out" processes are used for communication with the external environment. The "video_in" process is responsible for getting the input video from the input file. The "video_out" process stores the H264 compressed data to the output file. These tasks are very platform dependent and should not thus be used to compute concurrency measures. In addition, the "dmx", "substract", and "adder" processes are not also considered in our concurrency measures exploration.

We started initially this exploration using only the "execution load" and "restart" measures. Low values of these properties are obtained. The mean "execution load" value obtained of all the nodes of the YAPI/KPN model of figure 3 is about 0.1. This is very low indicating that many processes are blocked waiting for data or have fast finished their execution while others did not have done so. Thus, this model is not optimal balanced and more available concurrency should be explored. To improve the concurrency of this model, we propose to use "task-splitting" and "task-merging" mechanisms to get better execution loads.

The obtained restart measure values are presented in figure 5 for all the computation nodes of the proposed model. As shown in this figure, the restart values are too low. The mean for the network is about *3.74E-04*. This arise the need to split some nodes into possibly several nodes. The first node candidate for such a split is the motion estimation and compensation (MEC) node which has the lowest restart value. For this, we propose first to separate the motion estimation (ME) from the motion compensation (MC) modules. Second, as the computation of the ME module represents the highest part, we propose also to separate the part that performs storing the reference frames into the decoded pictures buffer (DPB) from the ME module.

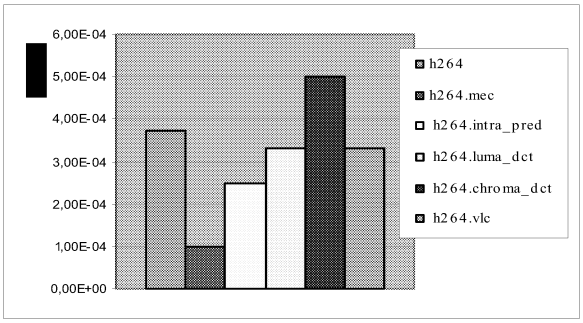


Figure 5: Concurrency Restart measures of the proposed H264 YAPI/KPN model

On other hand, the intra-prediction of the luma 4x4 blocks, and the intra-prediction of the luma 16x16 and chroma 8x8 blocks should be processed independently. More task splitting of the Luminance DCT quantization process need also to be explored by extracting the Luma decoder part that contains the rescaling and the inverse transform modules.

CONCLUSION

The H264/AVC has been designed with the goal of enabling significantly improved compression performance relative to all existing video coding standards. Implementing a H264 video encoder represents a big challenge for resource-constrained multimedia systems such as wireless devices or high-volume consumer electronics since this requires very high computational power to achieve real-time encoding. In this paper, a high-level performance analysis of a H264 video encoder is first performed to find an optimal balance between the coding efficiency and the implementation cost allowing for a complexity reduction at the high system level.

The obtained results have provided the best parameter configuration for an optimal use of the AVC tools. For this cost-efficient configuration, the obtained absolute complexity values confirmed the big challenge needed for its effective implementation. Given this, we proposed the use of a multiprocessor approach to share the encoding time between several processors. For efficient parallel code decomposition, a "Gprof" execution profiling and a study of the predominant forms of parallelism have been explored. Based on this, a first parallel model of the H264 encoder was proposed and validated using khan process networks (KPN) models

implemented by the YAPI multithreading environment library.

For an optimal parallel specification of the H264 encoding application, the system-level software CAST tool is used. The input of the CAST tool is a YAPI/KPN network specification. Outputs are concurrency measures providing insight to the potential concurrency bottlenecks of the proposed model. Given the obtained concurrency measure results, we motivated the optimization of the proposed YAPI/KPN model to get one with more balanced workload, better communication behavior, and maximum parallelism between tasks.

REFERENCES

Joch, A.; F. Kossentini; P. Nasiopoulos. 2002. "A Performance Analysis of the ITU-T Draft H. 26L Video Coding Standard". In *Proceeding of the 12th International Packet Video Worksho.,* Pittsburg, Pa, USA (April).

Alvarez, M.; A. Salami; A. Ramirez; M. Valero. 2005. "A Performance Characterization of high Definition Digital Video Decoding using H264/AVC". *Proceeding of the IEEE International, Symposium on Workload Characterization.* pp. 24–33, 6-8 (Oct).

Saponara, S.; K. Denolf; G. Lafruit; C. Blanch; J. Bormans. 2004. "Performance and Complexity Co-evaluation of the Advanced Video Coding Standard for Cost-Effective multimedia communication", *EURASIP Journal on Applied Signal Processing*. pp. 220-235, (Feb).

Krichene Zrida, H. ; A.C. Ammari; A. Jemai; M. Abid. 2007. "Performance/Complexity Analysis of a H264 Video Encoder". *International Review on Computers and Software (IRECOS)* – (July).

Kahn G. 1974. "The semantics of a simple language for parallel programming". *Proceeding of the IFIP Congress 74.* North-Holland Publishing Co.

Kock, E.A.; G. Essink; W.J.M. Smits; P. van der Wolf; J.Y. Brunel; W.M. Kruijtzer; P. Lieverse; and K.A. Vissers. 2000. "YAPI: Application modeling for signal processing system". In *Proceeding 37th Design Automation Conference (DAC'2000)*. Los Angeles, CA. pp. 402–405, (June 5-9).

H264 Reference Software Version JM 10.2. 2005 (Nov). http://iphome.hhi.de/suehring/tml/.

Krichene, H.; A. C. Ammari; A. Jemai; M. Abid; S. Maalej. 2007. "Parametric Complexity/Performance Analysis of a H264 Video Encoder", *Fourth International Multi-Conference on Systems, Signals & Devices (SSD07), Communication & Signal Processing Conference*. Vol. 3 (March 19-22). Hammamet – Tunisia.

Jemai, A.; P. Kisson; A. A. Jerraya. 1997. "Combining Architectural Simulation and Behavioral Synthesis", *IEICE Transaction Fundamentals*. Vol. E80-A, No 10 (October).

Dwivedi, B. K.; A. Kumar; and M. Balakrishnan. 2004. "Synthesis of application specific multiprocessor architectures for process networks*". Proceeding 17th International Conference on VLSI Design (VLSI-2004)*. Mumbai, India (January).

Namneh, R.A.; W.D. Pan; and S.M. Yoo. 1995. **"**Parallel implementations of 1-D fast Fourier transform without interprocessor communication", *International Journal of Computers and Applications*. Volume 29, Issue 2.

Graham, Susan L.; Peter B. Kessler; and Marshall K. McKusick. 1982. "Gprof: A Call Graph Execution Profiler". *Proceedings of the SIGPLAN '82 Symposium on Compiler Construction*. http://www.gnu.org/software/binutils/manual/gprof-2.9.1/

Pastrnak, M.; P.H.N. de With; S. Stuijk; and J. van Meerbergen. 2006. "Parallel Implementation of Arbitrary-Shaped MPEG-4 Decoder for Multiprocessor Systems"*, Visual Communications and Image Processing (VCIP'06)*. pp 60771I-1 - 60771I-10.

Youssef, M.; S. Yoo; A. Sasongko; Y. Paviot; and A.A. Jerraya. 2004. "Debugging HW/SW interface for MPSOC: Video Encoder System Design Case Study", *proceedings of the 41st Design Automation Conference*.

Shen, K.; L.A. Rowe; and E.J. Delp. 1995. "A Parallel Implementation of an MPEG1 Encoder: Fater than Real-Time", *Proceedings of SPIE Conference on Digital Video Compression: Algorithms and Technologies*. San Jose (Feb).

Bozoki, S.; S.J.P. Westen; R.L. Lagendijk; and J. Biemond. 1996. "Parallel algorithms for MPEG video compression with PVM". In *EUROSIM: Delft.* The Netherlands 315-326.

Ostermann, J.; J. Bormans; P. List; D. Marpe; M. Narroschke, F. Pereira; Th. Stockhammer; and Th. Wedi. 2004. "Video coding with H.264/AVC: Tools, Performance, and Complexity". *IEEE Circuits and Systems Magazine*.

Jacobs, T.R.; V.A. Chouliaras; D.J. Mulvaney. 2006. "Thread-parallel MPEG-4 and H.264 coders for system-on-chip multi-processor architectures", *International Conference on Consumer Electronics (ICCE 06)*. Pages: 91-92 (7-11 Jan).

Kock, E.A.; and G. Essink. 2001. "Y-chart Application Programmer's Interface. Application programmer's giude version 1.0.1". *Philips Research*. Eindhoven.

Lange, M. De. 2004. "ACE Associated Compiler Experts, Will the Software Dinosaurs Step Aside or Step on MPSoC?". In *MPSoC'04, "Saint-Maximin la Sainte Baume"*. France (July).

Stuijk S.; and T. Basten. 2003. "Analyzing Concurrency in Computational Networks (Extended Abstract), Formal Methods and Models for Codesign". *1st ACM & IEEE International Conference, MEMOCODE'2003, Proceedings*. Mont Saint-Michel, France (24-26 June). *IEEE Computer Society Press*. Los Alamitos, CA, USA.

Stuijk, S.; T. Basten; and J. Ypma. 2003. "CAST - A Task-Level Concurrency Analysis Tool, Application of Concurrency to System Design". *3rd International Conference, ACSD 03, Proceedings*. Guimarães, Portugal (18-20 June). *IEEE Computer Society Press.* Los Alamitos, CA, USA, 2003.

COMPLEXITY CONSTRAINED VIDEO CODING FOR DECODERS WITH LIMITED RESOURCES

Paulo J. Cordeiro[1,2,3]
[1]Inst. Politécnico de Leiria - ESTG
Campus 2, Morro do Lena,
2411-901 Leiria, Portugal
email: cordeiro@estg.ipleiria.pt

Juan Gomez-Pulido[2]
[2]Dept. Tecn. Comp. y Comunic.
Universidad de Extremadura
10071 Cáceres, España
email: jangomez@unex.es

Pedro A. Assunção [1,3]
[3]Inst. de Telecomunicações
Campus 2, Morro do Lena,
2411-901 Leiria, Portugal
email: amado@co.it.pt

KEYWORDS
Constrained video coding, Decoding complexity, Power-aware coding.

ABSTRACT

This paper describes an efficient method to generate compressed video streams with low complexity decoding requirements. This is particularly targeted at portable media players, where the use of such video streams leads to reduced power consumption thus extending the battery life. The proposed method is based on the fact that motion compensation is the most complex operation in standard H.264/AVC video decoders, mainly because of the operations involved in computing sub-pixel predictions. By using a measure of the computational complexity needed to decode a stream and including it in the encoder cost function used to select the type of sub-pixel motion vector, it is shown that the decoding complexity of compressed video streams can be reduced with negligible penalty in rate-distortion performance. The experimental results show that such optimized constrained coding method is capable of achieving significant decoding complexity reduction at the expense of negligible PSNR loss within the range of bit rates with practical interest.

INTRODUCTION

Current portable devices such as media players, mobile phones, personal digital assistants, palmtops, etc are increasingly used to access, decode and render multimedia content in which compressed video plays a major role. The limitation imposed by battery life and computational constraints of processing equipment have been a strong motivation for research on complexity issues of video coding and decoding systems in the last few years (Pouwelse et al. 2001, Poellabauer and Schwan 2002, Lin et al. 2004). Extending the time that multimedia content can be delivered and consumed in portable equipment is a desirable feature for both users and service providers which may be achieved through different approaches. Reduction of the computational complexity (Lu et al. 2007) and power saving mechanisms (Adams 2007) are among the most popular approaches to deal with this issue.
Although power saving in portable devices is dependent on the characteristics of several different functional components of the device itself such as hardware, operating system, processing software implementation, communication protocols, etc, in video decoders this is highly related to the computational complexity of the decoding process. It is particularly relevant in the case of H.264/AVC video coding standard where complexity, at both the encoding and decoding sides, is a major concern in any practical real time implementation. Besides all implementation optimizations that can be considered for the purpose of minimizing decoding complexity, there is still room for further reduction by producing compressed video streams such that their inherent decoding complexity is made lower (Wang and Chang 2006, Kalva and Furht 2005). This can be done by constraining the encoder in order to reduce the use of the most complex coding tools and options which lead to higher decoding computational complexity. The challenge is how to achieve this goal without compromising too much the signal quality when compared with unconstrained encoding.

Decoding complexity of H.264/AVC streams is mainly due to motion compensation with sub-pixel accuracy computations, which demand a great deal of processing power. This is because computation of predictions from sub-pixel motion vectors is done by interpolation filters, which need a different amount of filtering operations according to each sub-pixel location. In this paper we propose a measure of decoding computational complexity based on the amount of filtering operations needed to compute half and quarter-pixels. Such a measure allows to account for both rate-distortion and decoding complexity in the process of selecting the best motion vector for each block. Then the rate-distortion-complexity performance is evaluated and compared with the normal rate-distortion optimized coding.

The next section of the paper is a brief review of relevant issues in the H.264/AVC standard, with particular emphasis on motion compensation and

half/quarter-pixel interpolation, then the proposed rate-distortion-complexity optimization procedure is described in third section, the experimental results are presented and discussed in fourth section, and finally fifth section concludes the paper.

MOTION COMPENSATION IN H.264/AVC

As in previous video coding standards from the Moving Picture Experts Group (MPEG) / International Telecommunication Union Telecommunication Standardization Sector (ITU-T), the H.264/AVC video codec defines a video signal structure as a series of groups of pictures (GOPs) comprising slices of luminance (Y) and two chrominance (Cb and Cr). Each slice is divided into non-overlapping blocks i.e. macroblocks (MBs) and is typically encoded as either Intra (I), Predictive (P), or Bi-Predictive (B). In general, I slices are encoded using prediction from the same slice and they are independent of all others, P slices mainly use inter-frame prediction but can also use intra coding with or without intra-prediction. Finally B slices use an expanded set of forward and/or backward inter-prediction modes compared to P-slices. In P and B slices, motion estimation (ME) is carried out for each MB to find the best match from a given reference frame. In H.264/AVC, each MB can be further partitioned into sub-blocks and multiple reference frames can also be used. An integer transform (IT), with similar characteristics to those of the discrete cosine transform (DCT), used in previous standards, is applied to the residual block, concentrating most of the block energy into the low frequency region. Quantization and entropy coding of the remaining coefficients are then applied to further reduce irrelevancy and redundancy of the video signal.

ME is among the most complex operations in H.264/AVC video coding because of the high number of operations involved in finding the best match for each block. Also on the decoder side, previous studies on H.264/AVC decoder complexity have shown that motion compensation (MC) is the most computationally complex functional block, followed by the deblocking filter process (Horowitz et al. 2003, Lappalainen et al. 2003). High complexity in motion compensation is mainly due to interpolation needed to decode motion vectors with half or quarter pixel accuracy. In the case of the baseline H.264/AVC decoder, it was shown that interpolation takes around 39% of the execution time on average, and it can go up to 44% for some sequences. Figure 1 shows the breakdown of the complexity of a typical H.264/AVC decoder implementation as reported in (Lappalainen et al. 2003). Since this is an important topic in the context of this paper, a more detailed description of sub-pixel interpolation in H.264/AVC and the origin of its computational complexity is provided in the next subsection.

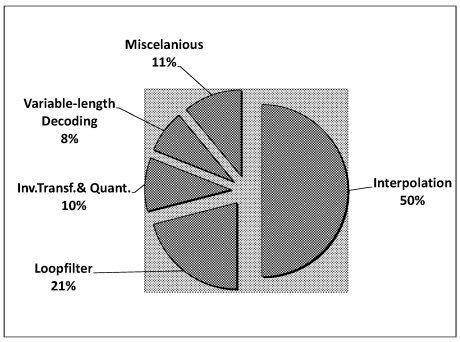


Figure 1: Computational Complexity Distribution in a Typical H.264/AVC Decoding Process
(Lappalainen et al. 2003)

HALF/QUARTER-PIXEL INTERPOLATION

When motion vectors received by the decoder point to reference image locations which are not coincident with the pixel sampling grid of the image, interpolated sub-pixel values must be computed at the decoder in order to build exactly the same prediction that was previously used in the encoder. Such an interpolation process is implemented through filters defined in the standard which lead to either half or quarter pixel accuracy. H.264/AVC can use up to quarter pixel precision during interpolation (Wiegand et al. 2003). Figure 2 illustrates the details of sub-pixel interpolation where the shaded squares with capital letters are the integer locations, corresponding to the sampling grid, and the white squares with lowercase letters correspond to interpolated pixels with sub-pixel accuracy.

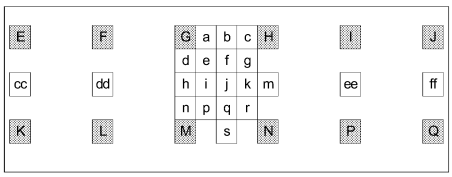


Figure 2: Notations for Integer and Fractional Pixels Locations in H.264/AVC

The sample values at half-pixel locations (b, h) are computed with a 6-tap FIR filter applied in both horizontal and vertical directions. All quarter-pixel values are computed by a 2-tap average filtering using both integer and half-pixels. For example, the following formulae are used to calculate sub-pixels b and p.

$$b = \frac{(E - 5F + 20H - 5I - J) + 16}{32} \quad (1)$$

$$p = \frac{h + s + 1}{2} \quad (2)$$

The total number of filtering operations required to determine each sub-pixel depends on the exact location of each one. In this work, this is closely related to the computational complexity and therefore with power saving in portable decoding devices.
Table 1 lists the number of taps of the interpolation filters, which is used as a measure of the corresponding complexity. From the table, it is obvious that interpolation of different sub-pixels implies quite different computational complexity. Since this operation must be done at the decoder whenever sub-pixel accuracy motion vectors are received in the coded stream, a strong dependence between both the number and type of such sub-pixels and decoding computational complexity of P or B slices is expected.

Table 1: Interpolation Filtering according to Sub-pixel Location

Pixel/Sub-pixel		Interpolation filtering
Type	Location	
integer	G	No
in-line half-pels	b,h	6-tap
center half-pel	j	7*6-tap
in-line quarter-pels	a,c,d,n	6-tap + 2-tap
corner quarter-pels	p,e,g,r	2*6-tap + 2-tap
center quarter-pels	f, q, i,k	7*6-tap + 2-tap

According to Table 1, one can define a cost associated with each type of motion vector corresponding to the associated computational complexity. The underlying idea for reducing the decoder complexity is to select those motion vectors that involve less interpolation operations, especially 6-tap filtering, while keeping the video quality high. Since in general a tradeoff between only rate and distortion is used to select the best motion vector at the encoder, by introducing the decoding complexity cost in the selection process, one may optimally reduce this type of complexity without penalizing too much rate-distortion performance.

Therefore, the basic approach to complexity reduction is to change motion vectors from high complexity sub-pixel positions into the ones with low complexity, or even to integer-pixel positions, by jointly optimize quality, bitrate, and computational complexity. This will lead to lower energy consumption in portable video players without significantly affecting the signal quality.

RATE-DISTORTION-COMPLEXITY OPTIMIZATION

In video coding, the bit allocation and rate control mechanisms dynamically adjust the encoding parameters to achieve a target bit budget at the highest possible video quality. This is done through rate-distortion (RD) Lagrangian optimization in two coding stages: motion estimation and mode decision. In motion estimation, for each block B with a block mode M, the motion vector associated with the block is selected through a rate-distortion joint cost function (Wiegand et al. 2003):

$$J_{Motion}^{R,D} = D_{DFD} + \lambda_{Motion} R_{Motion} \quad (3)$$

In Equation (3), D_{DFD} is the prediction error, computed as either the sum of absolute differences (SAD) or the sum of squared difference (SSD), R_{Motion} is the estimated bit rate to encode the corresponding motion vector, λ_{Motion} is the Lagrange multiplier to control the weight of the bit rate cost relative to the prediction error. $J_{Motion}^{R,D}$ is the rate-distortion joint cost comprising both R_{Motion} and D_{DFD}. Since in general the search space for motion vectors is very large and SAD has lighter computation cost than SSD, the former is used more often.

In this work, in order to favor those motion vectors with less interpolation complexity and penalize the ones with higher complexity, the conventional cost function presented in Equation (3) is added a specific Lagrange term to model the complexity cost. This allows selecting the motion vectors based on a joint cost function which takes the three parameters into account, i.e., a rate-distortion-complexity joint cost function as given by Equation (4),

$$J_{Motion}^{R,D,C} = J_{Motion}^{R,D} + \lambda c_{Motion} C_{Motion} \quad (4)$$

where C_{Motion} is the complexity cost associated with the selected motion vector and λc_{Motion} is the Lagrange multiplier for the complexity term. J_{Motion}^{RD} is the rate-distortion defined in Equation (3) and $J_{Motion}^{R,D,C}$ is the rate-distortion-complexity cost function.

The macroblock mode is directly related with the computational complexity because it defines which motion vector should be associated with each of its sub-blocks. For each motion vector, a prediction block must be computed by the decoder in order to reconstruct a predicted macroblock from all sub-blocks. After all inter mode candidates have an associated motion vector, the coding results of the modes are compared and the one which minimizes the Lagrangian cost given by Equation (5) is chosen.

$$J_{Mode}^{R,D,C} = J_{Mode}^{R,D} + \lambda c_{Mode} C_{Mode} \quad (5)$$

Where C_{Mode} is the complexity cost associated with the selected block mode and λc_{Mode} is the Lagrange multipier for the complexity term. $J_{Mode}^{R,D}$ is the rate-distortion for the selected block mode and $J_{Mode}^{R,D,C}$ is the rate-distortion-complexity cost function. Considering two extreme cases of $\lambda c_{Motion} = \lambda c_{Mode} = 0$ the solution for Equations (4) and (5) is identical to that of Equation (3), i.e., the complexity cost is not considered, the motion vectors and block modes are chosen based only

Table 2: Complexity Cost of Half/Quarter-pixel Interpolation

Pixel location	MV complexity cost
G	0
b,h	6
a,c,d,n	8
p,e,g,r	14
j	42
f, q, i,k	44

on the rate-distortion cost without taking into account the inherent decoding complexity of half and quarter-pixel accuracy motion. At the other extreme, $\lambda c_{Motion} = \lambda c_{Mode} = \infty$ results in minimum decoding complexity, because complexity cost is dominant. However, in this case the rate-distortion performance drops, thus the best solution is to tradeoff between rate, distortion and complexity using $0 < \lambda c_{Motion} = \lambda c_{Mode} < \infty$ in Equations (4) and (5).

COST MODEL FOR COMPLEXITY OF HALF PIXEL INTERPOLATION FILTERING

As pointed out before, the computational complexity is heavily influenced by the type of the motion vector (integer, half-pixel, or quarter-pixel) which defines the interpolation filters to be used in the motion compensation process. Therefore, a cost model was defined to associate the decoding complexity to each candidate motion vector. Since interpolation is implemented by the predefined filters, the complexity cost was chosen to be directly related to the number of filtering operations needed in each type of interpolation. Table 2 shows the complexity costs used for each type of half and quarter-pixel location.

EXPERIMENTAL RESULTS

The efficiency of the proposed constrained coding method was experimentally evaluated in order to assess how much quality drop is obtained for reducing the decoding complexity. Two sequences (Container and Foreman) with different types of motion were used in order to assess the influence of the video signal characteristics on the performance of the proposed method. The baseline mode of H.264/AVC was used since this is more appropriate to be delivered to mobile devices. The experimental setup was defined according to the recommended simulation conditions for coding efficiency experiments (Tan et al. 2007) and also followed the complexity evaluation methods proposed in (Horowitz 2007).
Figures 3 and 4 show the rate-PSNR performance obtained from sequences Container and Foreman for different values of λc and Table 3 shows the decoding complexity measured as the processing time

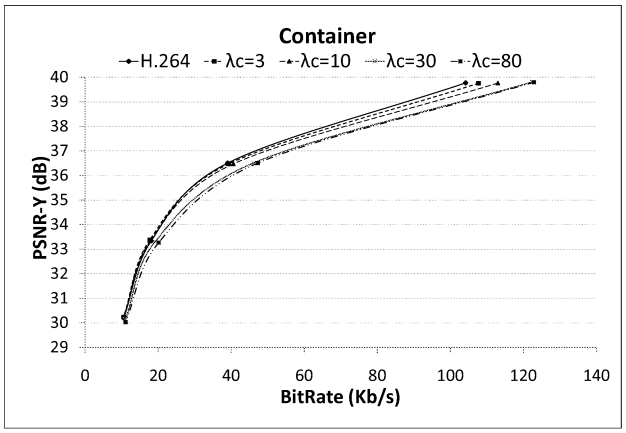


Figure 3: Rate-PSNR Performance for Different Values of λc (Container)

Table 3: Decoding Computational Complexity

Seq.	R-D (ms)	R-D-C (λc=10)		R-D-C (λc=30)		R-D-C (λc=80)	
		ms	ΔC	ms	ΔC	ms	ΔC
For.	1174,6	996.6	15,2%	941	19,9%	891,1	24,1%
Con.	520,6	423,2	18,8%	411	21,1%	406	22%

consumed by the motion compensation function in the H.264/AVC decoder. The proposed method is compared with H.264/AVC RD optimization, referred to as "H.264" in both figures and Table 3.

As one can see in the Figures 3 and 4, the proposed method produces a very small drop in PSNR for a wide range of bit rates. For higher values of λc the choice for integer motion vectors increase while sub-pixel interpolation decreases because the complexity weight is higher in the Lagrangian cost function. This causes lower RD performance but significant savings in decoding computational complexity. At PSNR values close to 40 dB, i.e., at higher bit rates, the quality drop is slightly higher but the computational complexity saving is also higher. This is because at such high values of PSNR the picture quality is very high and there is almost no coefficient quantization noise to mask the small extra distortion. Nevertheless the bit rate range of interest in this type of applications is below 100 kbps, where quality drop is about 0.5 db or less and decoding complexity can be reduced up to 25%.
Figure 5 shows the histogram with all selected motion vectors (mv occurrences) in the Foreman sequence encoded by H.264/AVC without any modification. Coordinates (1,1) corresponds to integer G pixel in Figure 2. Position complexity can be differentiated by gray level in the figure where darker bars identify more complex (filter operations) positions.
Figure 6 shows the results for the same sequence encoded by using λc = 80. It shows a significant reduction

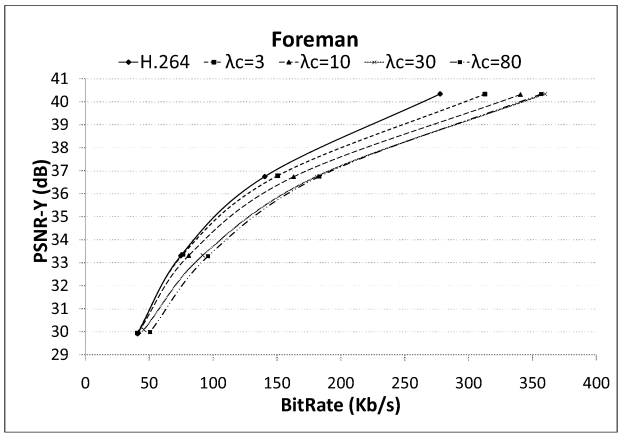


Figure 4: Rate-PSNR Performance for Different Values of λc (Foreman)

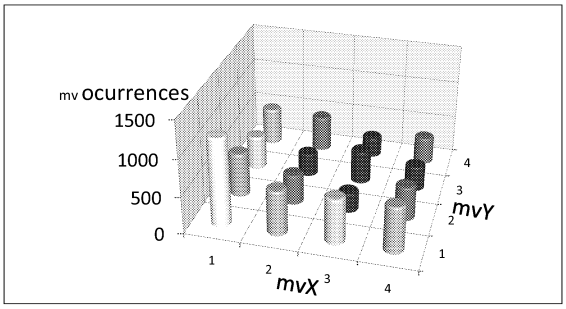


Figure 5: Foreman Sequence Histogram with no Complexity Control

of the more complex interpolations and an increase on the number of motion vectors pointing to integer locations, thus reducing the decoding complexity.

CONCLUSION

The results presented in this paper show that decoding complexity can be significantly reduced by constraining the video encoder in the choice of complex operations such as sub-pixel interpolation in the motion compensation process. By including the complexity cost in the selection procedure of the best motion vector, the rate-distortion-complexity tradeoff is optimized such that the quality loss in PSNR is perceptually negligible while the decoding complexity is significantly reduced. Therefore, the proposed method is particularly useful in video services and applications which deliver H.264/AVC video to portable devices. Future work includes application of this method in media adaptation proxies and reduction of the computational complexity of the constrained process itself.

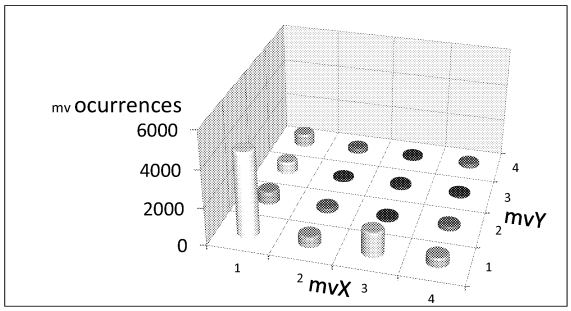


Figure 6: Foreman Sequence Histogram with Complexity Control

REFERENCES

Adams J., 2007. *Adaptive Buffer Power Save Mechanism for Mobile Multimedia Streaming.* Master's thesis, School of Electronic Engineering, Faculty of Computing and Engineering Dublin City University.

Horowitz M., 2007. *Towards useful complexity evaluation methods.* Document VCEG-AG19, 33rd meeting: Shenzen, China.

Horowitz M.; Joch A.; Kossentini F.; and Hallapuro A., 2003. *H.264/AVC baseline profile decoder complexity analysis. IEEE Trans Circuits Syst Video Techn*, 13, no. 7, 704–716.

Kalva H. and Furht B., 2005. *Complexity Estimation of the H.264 Coded Video Bitstreams. Comput J*, 48, no. 5, 504–513.

Lappalainen V.; Hallapuro A.; and Hämäläinen T.D., 2003. *Complexity of optimized H.26L video decoder implementation. IEEE Trans Circuits Syst Video Techn*, 13, no. 7, 717–725.

Lin S.S.; Tseng P.C.; Lin C.P.; and Chen L.G., 2004. *Multi-mode content-aware motion estimation algorithm for power-aware video coding systems. Signal Processing Systems, 2004 SIPS 2004 IEEE Workshop on*, 239–244.

Lu M.T.; Yao J.J.; and Chen H.H., 2007. *A complexity-aware video adaptation mechanism for live streaming systems. EURASIP J Appl Signal Process*, 2007, no. 1, 215–215. ISSN 1110-8657.

Poellabauer C. and Schwan K., 2002. *Power-Aware Video Decoding using Real-Time Event Handlers.* In *5th International Workshop on Wireless Mobile Multimedia (WoWMoM).*

Pouwelse J.; Langendoen K.; Lagendijk R.; and Sips H., 2001. *Power-aware video decoding.* In 22nd Picture Coding Symposium, Seoul, Korea.

Tan T.; Sullivan G.; and Wedi T., 2007. *Recommended Simulation Common Conditions for Coding Efficiency Experiments.* Revision 1. ITU-T Video Coding Experts Group (ITU-T SG16 Q.16), Document VCEG-AE010,31st Meeting: Morocco.

Wang Y. and Chang S.F., 2006. *Complexity Adaptive H.264 Encoding for Light Weight Streams.* In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).* Toulouse, France.

Wiegand T.; Sullivan G.J.; Bjntegaard G.; and Luthra A., 2003. *Overview of the H.264/AVC video coding standard. Circuits and Systems for Video Technology, IEEE Transactions on*, 13, no. 7, 560–576.

DATA DETECTION

On Dual View Lipreading Using High Speed Camera[1]

Alin G. Chiţu and Leon J.M. Rothkrantz
Faculty of Information Technology and Systems
Delft University of Technology
Mekelweg 4, 2628CD Delft,
The Netherlands
E-mail: {A.G.Chitu,L.J.M.Rothkrantz}@ewi.tudelft.nl

KEYWORDS: Lipreading, audio-visual speech recognition, audio-visual data corpus, multimodal data corpus, high speed camera, dual view recording, frontal view, side view, active appearance models, optical flow and lipreading, emotional speech.

ABSTRACT

Lipreading gets increasingly attention from the scientific society. However, many aspects related to lipreading are still unknown or poorly understood. In the current paper we present the entire process used for engineering the data for building a lip recognizer. Firstly, we provide detailed information on compiling an advanced multimodal data corpus for audio-visual speech recognition, lipreading and related domains. This data corpus contains synchronized dual view acquired using high speed camera. We paid careful attention to the language content of the corpus and the affective state of the speaker. Secondly, we introduce several methods for extraction features from both views and detail the problem of combining the information from the two views. While the information of the frontal view processing is more like a state of the art, we bring as well valuable new information and analysis for the profile view.

INTRODUCTION

Lipreading is getting more and more importance in the scientific community. There are however, still, many unknowns about what aspects are important when doing lipreading (i.d. what features hold the most useful information or how accurate must the sampling rate be and of course how do people do lipreading). There is an increasing belief, common sense but also based on scientific research (see McGurk 1976), that people use context information acquired through different communication channels to improve the accuracy of their speech recognition. This is the case for almost everything people do throughout their existence.

During the years there were quite a few methods developed to extract features relevant for lipreading. These methods are shape based (Essa and Pentland 1994, Yoshinaga et. al 2003 and 2004, Rothkrantz et. al 2006), appearance based (Bregler and Konig 1994, Duchnowski 1995, Li et. al 1997, Hong et al. 2006, etc.), or mixed, searching to digitize static or dynamic aspects of the input information, using more or less complex machine learning techniques. A special case is optical flow analysis methods which lies somewhere in between the two approaches (Mase and Pentland 1991, Iwano 2001, Tamura et. al 2002, Chitu et. al 2007). Although some level of agreement was achieved, there is still large space for improvement. Data corpora are an important part of any application related scientific study. The data corpus should provide the means for understanding all the aspects of a given process, direct the development of the techniques toward an optimum solution by allowing for the necessary calibration and tuning of the methods and also give good means for evaluation and comparison. Having a good data corpus, (i.e. well designed, capturing both general and also particular aspects of a certain process) is of great help for the researchers in this field as it greatly influences the research results.

Having this in mind we decided to build such a data corpus. A good data corpus should have a good coverage of the language such that every speech and visual item is well represented in the database. The audio and video quality is also an important issue to be covered. A thorough study of the existing data corpora can be found in Chitu and Rothkrantz 2007. An open question is however, what is the optimum sampling rate in the visual domain? Current standard for video recording frame rate ranges from 24 up to 30 frames per second, but is that enough? There are a number of issues related with the sampling rate in the visual domain. A first problem and the most intuitive is the difficulty in handling the increased amount of data, since the bandwidth needed is many times larger. A second problem is a technical problem and is related with the techniques used for fusing the audio and video channels. Namely, since it is common practice to sample the audio stream at a rate of 100 feature vectors per second, in the case when the information is fused in an early stage, we encounter the need to use interpolation to match the two data sampling rates. A third issue, that actually convinced us to use a high speed camera, is related to the coverage of the visemes during recording, namely the number of frames per visemes. In the paper Chitu and Rothkrantz 2007b it was showed that the visemes coverage becomes a big issue when the speech rate increases.

Figure 1 shows the poor coverage of the visemes in the case of fast speech rate as found in the DUTAVSC (Wojdeł et. al 2002). Hence, in the case of fast speech rate the data becomes

[1] The research reported here is part of the Interactive Collaborative Information Systems (ICIS) project, supported by the Dutch Ministry of Economic Affairs, grant nr: BSIK03024

very scarce, we have a mean of 3 frames per viseme which can not be sufficient. Therefore, during the recordings we asked the speakers to alternate their speech rate, in order to capture this aspect as well.

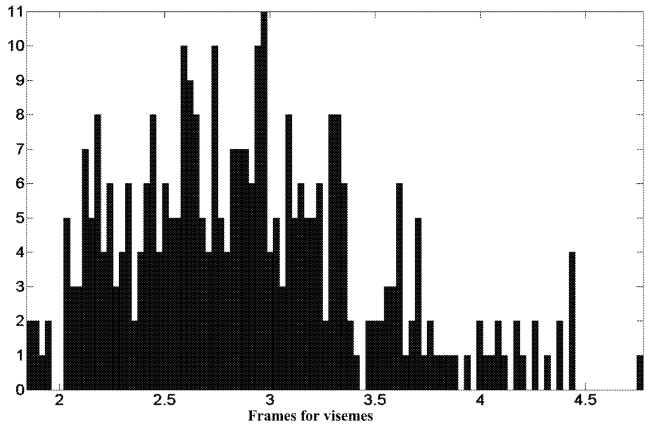


Figure 1. Viseme coverage by data in the case of fast speech rate in DUTAVSC data corpus.

As we said in the beginning we aim to discover where the most useful information for lipreading lies. We also want to give the possibility for developing new applications for lipreading. Therefore we decided to include side view recordings of the speaker's face in our corpus. A useful application could be lipreading though the mobile phone's camera. While the idea of side view lipreading is not entirely new (Yoshinaga et. al 2003 and 2004), we do not believe it has yet received the attention it deserves as less than a hand full of papers have addressed the issue. Besides that, a data corpus with side view recordings is nowhere to find at this moment.

We present in this paper a detailed analysis of the process of building an advanced lipreading system starting with the settings used during recording and continuing with the state of the art in feature extraction methods from both frontal and profile views.

To get a synchronized frontal-profile view we used a mirror placed at 45 degrees behind the speaker. The next Section will detail the recording setup. In Section 3 we introduce the state of the art methods for extracting visual features from frontal view. However, the attention falls here on the profile view processing. Hence, we give there a thorough analysis on the problem of extracting features from profile view images. We introduce three methods for key point detection from side view images: the AAM method, a functional method and a heuristic method. We also discuss there the possibilities brought up by combining the information from the two views. For all methods we give experimental results on the spot. The preliminary conclusions and future work directions will be given in Section 5.

RECORDINGS' SETTINGS

This section presents the settings used when compiling the data corpus. Figure 3 shows the complete image of the setup. We used a high speed camera, a professional microphone and a mirror for dual view synchronization. The camera was controlled by the speaker, through a prompter like software. The software was presenting the speaker the next item to be uttered together with directions on the speaking style required. This provided us with a better control of the recordings.

Video device

When one goes outside the range of consumer devices, things become extremely more complicated and definitely more expensive. The quality of the sensors and the huge bandwidth necessary to stream high speed video to the PC makes high speed video recording very restrictive. Fortunately, lately, by the advance made in image sensors (i.d. CCD and CMOS technology), it is possible to develop medium speed computer vision cameras at acceptable prices. We used for recording a Pike F032C camera built by AVT. The camera is capable of recording at 200Hz in black and white, 139Hz when using the chroma subsampling ratio 4:1:1 and 105Hz when using the chroma subsampling ratio 4:2:2 while capturing at maximum resolution 640X480. By setting a lower ROI the frame rate can be increased. In order to increase the Field Of View(FOV), as we will mention later, we recorded in full VGA resolution at 100Hz. To be able to guarantee a fix and uniform sampling rate and to permit an accurate synchronization with the audio signal we used a pulse generator as an external trigger. A sample frame is shown in Figure 2.

In the case of video data recording there are a larger number of important factors that control the success of the resulted data corpus. Hence, not only the environment, but also the equipment used for recording and other settings are actively influencing the final result. The environment where the recordings are made is very important since it can determine the illumination of the scene, and the background of the speakers. We use mono-chrome background so that by using a "chroma keying" technique the speaker can be placed in different locations inducing in this way some degree of visual noise.

Figure 2. Sample frame with dual view.

Audio device

For recording the audio signal we used NT2A Studio Condensators. We recorded a stereo signal using a sample rate of 48kHz and a sample size of 16bits. The data was stored in PCM audio format. Special laboratory conditions were maintained, such that the signal to noise ratio (SNR) was kept at controlled level. We considered that it is more advantageous to have very good quality recordings and degrade them in a post process as needed. The specific noise can be simulated or recorded in the required conditions and later superimposed on the clear audio data. An example of such database is NOISEX-92 (Varga and Steeneken 1993). This dataset contains white noise, pink noise, speech babble, factory noise, car interior noise, etc. As said before special attention was paid to the synchronization of the two modalities.

Mirror

The mirror was placed at 45 degrees on the side of the speaker so that a parallel side view of the speaker could be captured synchronized with the frontal view. The mirror covered the speaker face entirely. Since the available mirror was 50cm by 70cm the holder gave the possibility to adjust the height of the mirror, thus tailoring it for all participants.

Figure 3. The setup of the experiment.

LANGUAGE CONTENT FOR LIPREADING

The language coverage is very important for the success of a speech data corpus. The language pool of our new data corpus was based on the DUTAVSC data corpus, however, enriched to obtain a better distribution of the phonemes. Hence, the new pool contains 1966 unique words, 427 phonetically rich unique sentences, 91 context aware sentences, 72 conversation starters and endings and 41 simple open questions (i.d. for these questions the user was asked to utter the first answer that they think of. In this way we expect to collect more spontaneous aspects of the speech). For each session the speaker was asked to utter 64 different items (sentences, connected digits combination, random words, free answer questions) divided in 16 categories with respect to the language content and speech style: normal rate, fast rate and whisper. The total recording time was estimated to lie in the range 45-60 minutes. The complete dataset should contain some 5000 utterances, hence a few hours of recordings, thus we target 30-40 respondents that should record 2-3 sessions. However, the data corpus is at this moment still under development.

EMOTIONAL SPEECH

When speaking humans communicate their emotional state to their interlocutor. They use facial expressions and prosody to enrich the semantic content of their speech. For speech recognition and lipreading the transformations occurred in the signal in this way is actually considered as noise. Training for such target, namely emotional rich speech recognition, can be extremely difficult since the variation increases very much. One solution would be to try to filter out the affective information. Of course the information about the speaker affective state can be extremely valuable if we could have access to it as context information, without the induced perturbation of the information channel. We therefore included in our data corpus a second section which deals with emotional speech. For this we asked the speaker to read a short story that transmits a certain feeling, then ask her to set her mind to that affective state. Then the speaker was asked to utter a set of 5 sentences as a possible reaction to the particular story.

The speakers were divided into two groups: professional actors and naive speakers. All speakers were native Dutch. This is very important for the case of emotional speech since the performance of the speaker could get less genuine and definitely less spontaneous as result of the speaker spending more time in preparing his speech. However, it could be very interesting to analyze the cultural effect on expressing ones emotions though facial expressions and prosody. We recorded 21 emotions which are listed in Table 1.

Table 1. List of emotions considered for recordings.

#	Emotion	#	Emotion
1	Admiration	**12**	Fear
2	Amusement	**13**	Fury
3	Anger	**14**	Happiness
4	Boredom	**15**	Indignation
5	Contempt	**16**	Inspiration
7	Desire	**17**	Interest
6	Disappointment	**18**	Pleasant surprise
8	Disgust	**19**	Sadness
9	Dislike	**20**	Satisfaction
10	Dissatisfaction	**21**	Unpleasant surprise
11	Fascination		

VISUAL FEATURES

There are two general approaches in digitizing the input images for performing lipreading: appearance based processing and model based processing. While in the first case some transformation of the image data is considered as input for the inference engine (e.g. eigen faces/lips), in the

second approach some key elements are extracted from the image, namely, key points or contours, or geometrical measures of anatomical elements of the human face (e.g. mouth height/width, lips size). The second approach also permits the description of motion patterns. However, by using optical flow analysis we can recover more accurate motion information. The first approach has the disadvantage that it results in very large feature vector dimensions. The second approach on the other side provides smaller dimensionality, however, to the expense of the information content that is lost this way.

In order to build a consistent 3D model of the speaker face we need to find a correspondence between the feature points found in the frontal view with the feature points found in the side view. Hence comparing the Figure 4 and Figure 8 we find that the points K and L are superimposed with the points I1 and I2, respectively. However it should be noticed that the point P6 is actually the same with the point L, while the point M is similar with the point P7. When the mouth is closed the point P7 is also superimposed with the point N, of course. Also point P and P8 have almost the same location on the speaker face.

While the use of a mirror for recordings has the advantage of simplifying the process of recordings and provides a synchronized dual view, it has the disadvantage that the side view is slightly smaller due to the projection on the mirror, as can be seen in Figure 2. Hence before trying to match the feature points extracted from the two views a metric must be computed based on the optics of the camera, or purely by statistical analysis, and scale the profile image accordingly. As can be seen in Figure 2 if the camera is somewhat tilted the metric might be more difficult to find.

The following sections detail some very useful feature extraction techniques.

FRONTAL VIEW FEATURES

The shape based approach starts with the detection of some key points on the speaker's face. Figure 4 shows the typical feature points used for facial expression recognition, lip reading and other related applications. For lipreading however only the lower half of the face is usually considered.

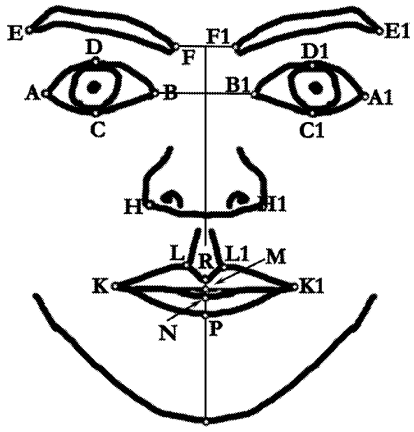


Figure 4. Frontal key points.

Mouth size features

Through the most common features used for lipreading from frontal images are the mouth width, mouth height and mouth aperture, defined as the distance between points K and K1, the distance between the points RP or the vertical displacement between points L and P or L1 and P and the distance between the points M and N, respectively. The vertical displacement between the points L and P is very important in the case of dual view processing, since it represents a strong correspondence measure.

Active Appearance Model(AAM) for key points detection in frontal view images

Active Appearance Models(AAM) (Cootes et. al 1998) is a generalization of the Active Shape Models(ASM) and combines both an appearance based approach and a model based approach. The AAM starts from a mean model, as shown in Figure 5, and then modifies the model parameters inside some learned range while minimizing the difference in appearance between the real image and the synthesized imaged based on the new model. The required number of parameters is computed in both cases by using PCA. To shorten the search and improve the solution a good initial guess is necessary. Hence, a face/mouth detection/tracking algorithm should be used. We use a Viola-Jones cascade classifier to do this job. This makes possible a real-time implementation of the technique. A search path sample is given in Figure 5 together with the final result.

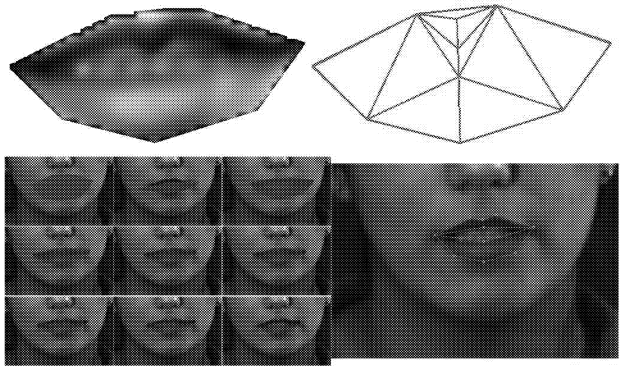


Figure 5. AAM Sample. First raw shows the appearance and the geometrical models, respectively. The second raw shows a sample search path and the final result.

Lip Geometry Estimation features

Lip Geometry Estimation method was introduced in the paper Wojdel and Rothkrantz 2000. The method starts with identifying the pixels on the lips. For this, any kind of image segmentation techniques can be used. Then by interpreting the filtered image as a two dimensional distribution, the method determines the mouth aperture and the thickness of the lips by analyzing the properties of the given distribution. Figure 6 shows a possible result of this method.

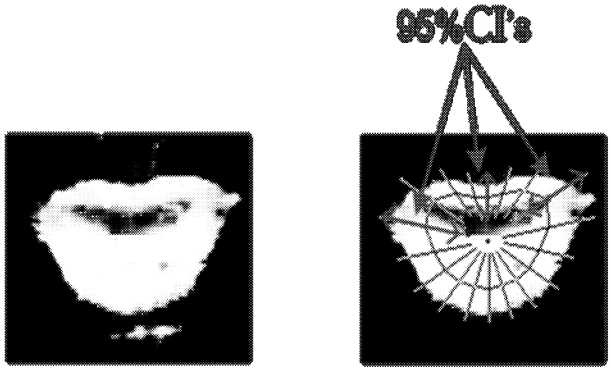


Figure 6. Lip Geometry Features.

Optical Flow features

Optical flow measures the apparent motion in a 2D clip. We employed optical flow analysis to directly recover the motion information, apparent around the speaker's mouth. The method was introduced in Chitu et. al 2007. The features extracted, represents quantitative measures of the muscular activity visible around the speaker's mouth. Hence we retain the horizontal and the vertical displacement in 18 sectors, centered in the mouth center. Figure 7 shows a possible result.

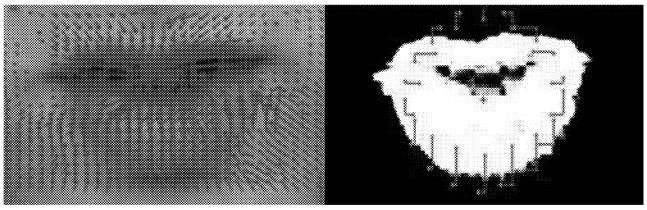


Figure 7. Optical Flow Features.

SIDE VIEW FEATURES

As in the case of frontal view Figure 8 shows the typical feature points used for encoding the speaker's profile posture in shape based approach.

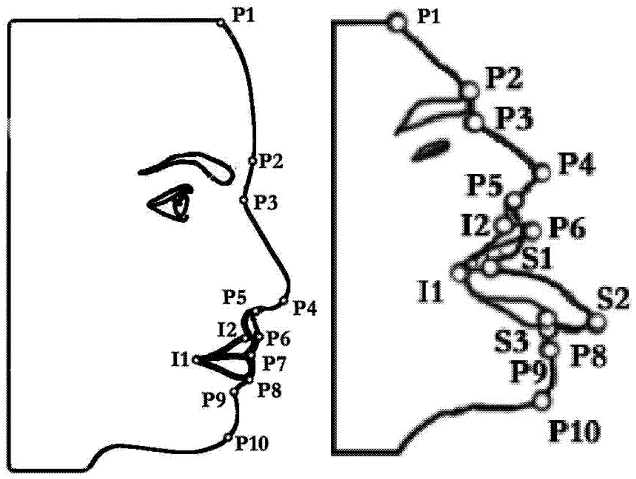


Figure 8. Side view feature points.

Two postures are shown because when the tongue is visible the point P7 is replaced with the group of points S1, S2, and S3 respectively, which describe the position of the tongue with respect to the lips. Again in the case of lipreading applications only the lower half of the face is important. With the exception of the points I1 and I2, all the other points are found on the silhouette. While the points that form the outline of the image can be found for instance by searching the extreme points of the function determined by the contour, the detection of the points I1 and I2 can be extremely difficult, and usually implies using some heuristic rules (e.g. Yoshinaga et. al 2003 considers the point I1 as the left most point in the dark area inside the mouth area). The AAM method is extremely valuable for this process, since it is not only detecting the contour but also the key points at once.

Mouth size features

In the case of profile view the first measure that comes into mind is the mouth height or mouth aperture, compared to the frontal view. However, increased attention should be paid to the definition of the two notions, since viewed from profile points P6, P7, P8, S1 and S3 are now differently projected. A consistent definition with the frontal view is to consider as mouth height only the vertical displacement between the points P6 and P8. However, this measure is vulnerable to head tilting.

A new feature that can classify as a mouth size feature is the mouth protrusion. This measure should describe the horizontal movement of the lips. There is however an issue related with the reference from which this displacement is computed. In Kumar et. al 2007 the reference was formed by the line linking the tip of the nose with the tip of the chin, namely the points P4 and P10. The protrusions features are then computed as the distances from the points P6 and P8 respectively to this line. However, the point P10 is not entirely independent of the movement of the mouth/lips. Therefore using the line <P4,P10> as a reference line introduces artificial correlation between the features, which is in fact hiding the speech related information. Hence, the reference point or line should be as much as possible uncorrelated with the feature we extract. Table 2 gives some correlation coefficients between the three protrusion features computed this way and the key points P4 and P10 on the profile. It is visible that P10 is more correlated with the protrusion features than P4. While this is not a clear proof that P10 is not good to be taken as reference, it is a good indication that further analysis is needed.

Table 2. Correlation coefficients between protrusion features and different point on the profile.

Quantity 1	Quantity 2	Correlation coeff. (x,y)
Protrusion in P6	P4	(0.05, -0.19)
Protrusion in P6	P10	(0.49, -0.44)
Protrusion in P8	P4	(-0.02, -0.06)
Protrusion in P8	P10	(0.46, -0.56)
Protrusion in I1	P4	(-0.08, -0.38)
Protrusion in I1	P10	(0.15, -0.17)

One solution would be to use only the tip of the nose as reference point for computing the protrusion. While the displacement from the tip of the nose is disturbed by head

tilting, the Euclidian distance is not. We could also find another reference point which is not moving too much when the speaker tilts his/her head. Such a point can be the tip of the ear. The distance from this point to P6, I1 and P8 can be a good measure for protrusion. Figure 9 shows the possible protrusion features.

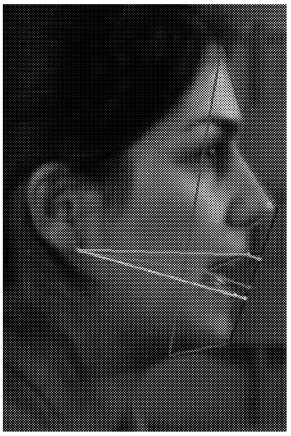


Figure 9. Possible protrusion features.

One more possible feature that could be used to express the aperture of the mouth is the angle <P6,I1,P8> (Yoshinaga et. al 2004). However, we need again to detect the off contour point I1. The AAM method for key points detection from profile view images given in the next subsection is very useful for detecting the point I1. This measure is correlated with both protuberance and mouth aperture.

AAM for key points detection in profile view images

Similar with the frontal view we can use here AAM. The appearance and the shape models are shown in Figure 10. Figure 10 also shows an example of a search path. We included the entire contour for detection since it gave more stabile results. Moreover, it gives extra information. For instance we noted that in the case of mouth protrusion we need some extra reference points in order to compute the features.

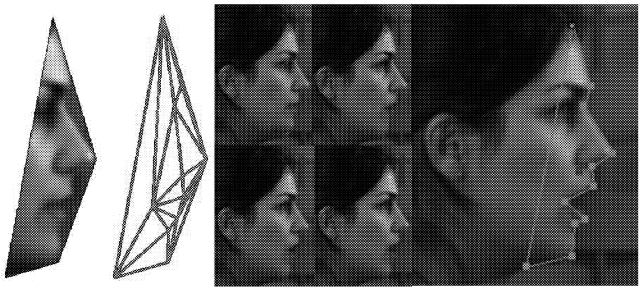


Figure 10. AAM Sample for profile view processing. At left are shown the appearance and shape mean models.

CONCLUSIONS

We presented in this paper in detail the entire process of making data available for building a lip recognizer. We started with the setup of the recordings for acquiring an advanced data corpus, that includes multimodal information. This is the starting point of any lipreading research and therefore is extremely important. Our new data corpus consist of high speed recordings of synchronized dual view of speaker faces while uttering both language rich and emotional speech.

We then give a detailed overview on the possible approaches for extracting visual features suitable for lipreading. While in the case of frontal view we portrayed the state of the art, in the case of profile view we introduced the use of Active Appearance Models for feature extraction and detailed to computation of mouth size like features.

The next step in completing a lipreading is to choose the inference engine for training and doing the actual recognition. In our work we used Hidden Markov Models for this job, being the most successful mathematical model for speech recognition up to date. When audio-visual speech recognition is considered, using HMMs is still possible, making the fusion approach to fall in feature fusion category. However, when more context information is used, information made available through the multimodal corpus, we consider using an inference engine based on Dynamic Bayesian Network(DBN). For future work we plan to evaluate different feature extraction methods and analyze and compare the performance of all approaches. We also envision a complete lip recognition tool that could exemplify real time the different methods.

Acknowledgments

We would like to thank Karin Driel (K.F.Driel@student.TUDelft.NL), Pegah Takapoui (pegahtak@gmail.com) and Martijs van Vulpen (mathijs@ch.tudelft.nl) for their valuable help with building the language corpus and setting up and conducting the recording sessions.

REFERENCES

(Bregler and Konig 1994) C. Bregler and Y. Konig, “”Eigenlips” for robust speech recognition”, in Acoustics, Speech, and Signal Processing, 1994. ICASSP-94 IEEE International Conference on, 1994.

(Chițu and Rothkrantz 2007) Alin G. Chițu and Leon J.M. Rothkrantz, "Building a Data Corpus for Audio-Visual Speech Recognition", *Euromedia2007*, ISBN 9789077381328, pp. 88-92, April 2007.

(Chițu and Rothkrantz 2007b) Alin G. Chițu and Leon J.M. Rothkrantz, "The Influence of Video Sampling Rate on Lipreading Performance", *12-th International Conference on Speech and Computer (SPECOM'2007)*, ISBN 6-7452-0110-x, pp. 678-684, Moscow State Linguistic University, Moscow, October 2007.

(Chițu et. al 2007) Alin G. Chițu, Leon J.M. Rothkrantz, Jacek C. Wojdel, Pascal Wiggers, "Comparison Between Different Feature Extraction Techniques for Audio-Visual Speech Recognition", *Journal on Multimodal User Interfaces*, pp. 7-20, Springer, March 2007.

(Cootes et. al 1998) Cootes, T.F., Edwards, G.J., Taylor, C.J., Active Appearance$_{th}$Models, In H.Burkhardt and B.Neumann, editors, 5 European Conference on Computer Vision, Vol.2, 484-498, Springer, 1998.

(Duchnowski et. al 1995) P. Duchnowski, M. Hunke, D. B¨usching, U. Meier, and A. Waibel, "Toward Movement-Invariant Automatic Lip-Reading and Speech Recognition", in International Conference on Acoustics, Speech, and Signal Processing, 1995 (ICASSP-95), vol. 1, pp. 109–112, 1995.

(Essa and Pentland 1994) I. A. Essa and A. Pentland, "A Vision System for Observing and Extracting Facial Action Parameters", in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 76–83, IEEE, June 1994.

(Hong et. al 2006) X. Hong, H. Yao, Y. Wan, and R. Chen, "A PCA Based Visual DCT Feature Extraction Method for Lip-Reading", iih-msp, vol. 0, pp. 321–326, 2006.

(Iwano et. al 2001) K. Iwano, S. Tamura, and S. Furui, "Bimodal Speech Recognition Using Lip Movement Measured By Optical-Flow analysis", in HSC2001, 2001.

(Kumar et. al 2007) Kshitiz Kumar, Tsuhan Chen, Richard M. Stern, "Profile View Lip Reading", IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP), 2007, Honolulu, Hawaii.

(Li et. al 1997) N. Li, S. Dettmer, and M. Shah, "Visually recognizing speech using eigensequences", Motion-based recognition, 1997.

(Mase and Pentland 1991) K. Mase and A. Pentland., "Automatic Lipreading by Optical-Flow Analysis", in Systems and Computers in Japan, vol. 22, pp. 67–76, 1991.

(McGurk and MacDonald 1976) McGurk, H. & MacDonald, J. Hearing lips and seeing voices *Nature,* 1976*, 264*, 746 - 748

(Rothkrantz et. al 2005) L. J. M. Rothkrantz, J. C. Wojdel, and P. Wiggers, "Fusing Data Streams in Continuous Audio-Visual Speech Recognition", in Text, Speech and Dialogue: 8th International Conference, TSD 2005, vol. 3658, (Karlovy Vary, Czech Republic), pp. 33–44, Springer Berlin / Heidelberg, September 2005.

(Rothkrantz et. al 2006) L. J. M. Rothkrantz, J. C. Wojdel, and P. Wiggers, "Comparison between different feature extraction techniques in lipreading applications", in Specom'2006, SpIIRAS Petersburg, 2006.

(Tamura et. al 2002) S. Tamura, K. Iwano, and S. Furui, "A Robust Multi-Modal Speech Recognition Method Using Optical-Flow Analysis", in Extended summary of IDS02, (Kloster Irsee, Germany), pp. 2–4, June 2002.

(Varga and Steeneken 1993) Varga, A. and Steeneken, H. 1993. "Assessment for automatic speech recognition II: NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems." *Speech Communication*, (vol. 12, no. 3, pp. 247-251, July)

(Wojdel and Rothkrantz 2000) J. C. Wojdel and L. J. M. Rothkrantz, "Visually based speech onset/offset detection", in Proceedings of 5th Annual Scientific Conference on Web Technology, New Media, Communications and Telematics Theory, Methods, Tools and Application (Euromedia 2000), (Antwerp, Belgium), pp. 156–160, 2000.

(Wojdeł et. al 2002) Wojdeł, J.C.; Wiggers, P. and Rothkrantz, L.J.M. 2002. "An audio-visual corpus for multimodal speech recognition in Dutch language" In *Proceedings of the International Conference on Spoken Language Processing (ICSLP2002)* (Denver CO, USA, September, pp. 1917-1920)

(Yoshinaga et. al 2003) T. Yoshinaga, S. Tamura, K. Iwano, and S. Furui, "Audio-Visual Speech Recognition Using Lip Movement Extracted from Side-Face Images", in AVSP2003, pp. 117–120, September 2003.

(Yoshinaga et. al 2004) T. Yoshinaga, S. Tamura, K. iwano, and S. Furui, "Audio-Visual Speech Recognition Using New Lip Features Extracted from Side-Face Images", in Robust 2004, August 2004.

AUTHORS BIOGRAPHY

ALIN GAVRIL CHIȚU was born on November 8, 1978 in Buşteni, Romania. He graduated in 2001 at the Faculty of Mathematics and Computer Science at University of Bucharest, which is one of the top universities in Romania. In 2003 he received the MSc. degree in applied computer science at the same university. Starting September 2003 he joined the Risk and Environmental Master Program at Delft University of Technology, Delft, The Netherlands which he graduated with honors in August 2005. Since then he is pursuing his PhD degree in the Man-Machine Interaction Group, Mediamatics Department at Delft University of Technology under the supervision of Dr. Leon J.M. Rothkrantz. His main interest is in data fusion as the means to build robust and reliable systems, audio-visual speech recognition being one of the case studies. He is also interested in robust computer vision, machine learning and computer graphics.
Email: a.g.chitu@ewi.tudelft.nl
Web address: http://mmi.tudelft.nl/~alin

LEON J.M. ROTHKRANTZ received the MSc. degree in mathematics from the University of Utrecht, Utrecht, The Netherlands, in 1971, the Ph.D. degree in mathematics from the University of Amsterdam, Amsterdam, The Netherlands, in 1980, and the MSc. degree in psychology from the University of Leiden, Leiden, The Netherlands, in 1990. He is currently an Associate Professor with the Man-Machine Interaction Group, Mediamatics Department, Delft University of Technology, Delft, The Netherlands, since 1992. His current research focuses on a wide range of the related issues, including lip reading, speech recognition and synthesis, facial expression analysis and synthesis, multimodal information fusion, natural dialogue management, and human affective feedback recognition. The long-range goal of his research is the design and development of natural, context-aware, multimodal man–machine interfaces. Drs. Dr. Rothkrantz is a member of the Program Committee for EUROSIS.
Email: l.j.m.rothkrantz@ewi.tudelft.nl
Web address: http://mmi.tudelft.nl/~leon

BEHAVIOUR DETECTION IN DUTCH TRAIN COMPARTMENTS

Z. Yang, A. Keur and L. J. M. Rothkrantz
Faculty of Electrical Engineering, Mathematics and Computer science
Delft University of Technology
Mekelweg 4, 2628CD Delft, The Netherlands
E-mail: {Z.Yang, L.J.M.Rothkrantz}@tudelft.nl

KEYWORDS

Aggression detection, Dutch train compartment, aggressive behaviour, Multi-modal cameras

ABSTRACT

Aggressive behavior in public places can cause great distress on the part of innocent bystanders. This paper describes research done to automatically detect forms of aggression by recognising the behaviour of people in a train. A dataset was gathered in a real train with semi professional actors performing aggressive and non-aggressive scenarios. We developed a system to recognize a number of predefined behaviours from features extracted from the sensor data.

INTRODUCTION

Safety in public places has gained a lot of attention in the past few years. The need for increased surveillance in public places as a guard against terrorist attacks and other forms of aggression, have made people more tolerant of cameras and microphones in public areas. With the increased number and complexity of these devices, people also expect a higher level of safety. Up until now, there has been limited success in living up to these expectations. The Dutch railway company (NS) for example, strives to decrease the number of incidents on Dutch trains by equipping them with cameras (e.g. the trains in the Zoetermeer Stadslijn). The primary role of these cameras is to increase the feeling of security of the passengers and to have a deterring effect on people with bad intentions. However, the camera images have to be inspected manually. With the growing number of camera images to process, the chances of detecting aggression manually becomes very small.

The goal of an ongoing project at the Man-Machine Interaction (MMI) group in Delft is to solve this problem by creating a system to automatically detect aggression as it is happening or is about to happen. In this paper we explore methods and techniques to describe normal and unusual behaviour in a train compartment. First we describe the train compartment and the situations we want to detect. We also specify the particular problems that we have to cope with in our environment such as varying light conditions and occlusion. Faced with these problems we present our solution which uses off-the-shelf classification algorithms.

The remainder of this paper is structured as follows. First we give an overview of the background and the related work in the area. Then we describe the data that was captured in the train. Next the classification of the behaviour that we want to detect. Afterwards come the detection methods and the results. We finish with a discussion and conclusions.

BACKGROUND

With the availability of inexpensive sensors and the ever increasing processing power at our disposal, the number of surveillance and surveillance related research projects has increased. The most commonly used modalities for this purpose are video (Foresti et al., 2005; Javed et al., 2003), audio (Clavel et al., 2005; Härmä et al., 2005) or a combination of both (Beal et al., 2002). We observe that in complex surveillance environments, such as in public transport systems, the combination of multiple modalities is more common, e.g. PRISMATICA for railways (Velastin et al., 2002) and ADVISOR for metro stations (Cupillard et al., 2004).

The usual approach to the surveillance problem is to view the individual events (e.g arm motions, gestures) as related parts of a bigger scenario e.g. fighting, ticket checking. A scenario is defined as a combination of states, events or sub scenarios. This means that in the representation of the scenario, the influence of the individual events on the outcome of the scenario is also included e.g. the occurrence of shouting might cause the ticket checking scenario to escalate. At runtime, the surveillance system tries to infer the consequences of the activity/scene recognized based on this prior knowledge. Bayesian networks can be used for the inference, but other approaches have also been proposed, including multi-layered HMMs (Zhang et al., 2006) and CHMMs (Oliver et al., 2000).

As suggested above, the surveillance system can be divided into two steps. A first step to detect the features and events in the incoming sensor data and a second step to combine these events (over time) into activities and scenarios. For surveillance/activity recognition in relatively controllable environments (e.g. rooms, offices)

data can be collected quite easily. Thanks to the controlled environment, feature extraction and event recognition can also be robustly performed.

In the train compartment however, we have to cope with more challenging circumstances. These include the varying (and unpredictable) light conditions throughout the course of the day, occlusion and echos as a result of the confined space of the compartment etc. Over the years however, huge improvements have been gained in classification algorithms. Technology evaluations, like the Face Recognition Grand Challenge (FRGC), have shown a huge progress in face recognition over the last 5-10 years (Phillips et al., 2005). When looking into the details, this progress was mainly driven by algorithmic innovations and improvement in sensor technology.

METHOD

In this paper our goal is to recognize specific behavior based on the recognition (and tracking) of some features in the input data over time. The surveillance system can be divided into two general steps. The first step detects the features and events in the incoming sensor data and the second step uses automated reasoning to combine these events (over time) into activities and scenarios (figure 1). The reasoning method is based on expert knowledge gathered after interviews with security experts from the Dutch Railways (NS).

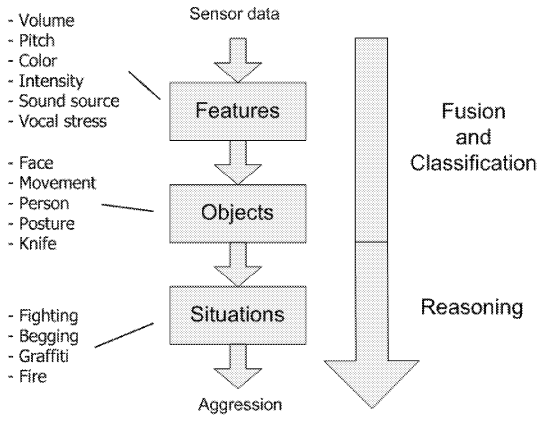


Figure 1: Overview of the aggression detection process

From interviews with experts we compiled a list of behaviours to detect and the features that the human experts use themselves to detect these behaviours. Next we gathered data of these behaviours in a train. Finally we used off-the-shelf classification algorithms to extract the features from the data and implemented our own algorithms to combine the detected the behaviours.

Aggressive behavior

The Dutch Railways (NS) has a system to classify incidents that occur in a train. The NS tailors this classification toward the procedures that should be taken when an incident of a certain category occurs (see table 1).

Table 1: Incident categorisation used by the NS.

Category	Description
A	Suspicious behavior
B	Robbery and theft
C	Violence
D	Serious public inconveniences
E	Small public inconveniences
F	Vandalism
G	Accident
H	Fire

Based on this classification, we created scenarios to be performed by actors in a real train, trying to get at least one scenario per category. In this paper we will focus on these scenarios (listed below).

1 Suspicious behaviour: a passenger prefers to stand in an empty compartment. Features to watch for are: the compartment is empty or almost empty, a passenger stands in hallway, passenger does not move forward or backward.
2 Small public inconvenience: a beggar enters the compartment and starts asking for money. Features to watch for are: a passenger walking along the hallway stopping periodically and speaking (with normal volume) to passengers. The passenger does not take a seat.
3 Serious public inconvenience: overcrowding. The most important feature is the number of people in the compartment.
4 Ticket checking: a conductor enters the compartment and checks the tickets of the passengers. Features to watch for are: a person dressed in blue with a blue hat walks along the hallway stopping periodically and speaking (with normal volume) to passengers. The person receives an object from a passenger and gives it back after a while. The person does not take a seat.
5 Enter train: one or more persons enter the train. Features: People come into the train from the entrance doors. Some take a seat if there is a free seat available.

Data

The aim of the data collection experiment is to gather data that can be used to test the aggression detection algorithms. Due to the scarcity of this kind of recordings and the privacy issues involved, we hired semi-professional actors to perform the scenarios described above in a real train. We used multiple microphones and cameras to record the actions. The location of the sensors in the train compartment and their orientation is shown in figure 2. Most scenarios were performed in the middle of the train, where the two cameras in the

middle have the largest overlap.

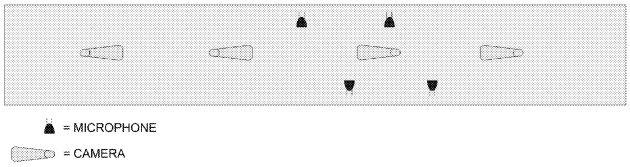


Figure 2: The locations of the sensors seen from a top view of the train compartment

The scenarios are recorded in sequences which total up to about one and a half hours of audio and video data. The data contains the scenarios as described earlier as well as recordings of normal and spontaneous situations. All the data of the sensors is stored in separate streams (four audio streams and four video streams). The four video cameras captured video at about 13 frames per second, at a resolution of 640x256 pixels (see figure 3).

Figure 3: Each scenes as captured by the four cameras

Each microphone captured sound generated by the actors performing the scenarios at a sample rate of 44100Hz with a 24 bit sample size. Each track is synchronized in hardware with sample accuracy. The audio data can be addressed in a single synchronized project consisting of the four streams of the four microphones, or as separate mono audio streams for each individual microphone (figure 4).

BEHAVIOUR RECOGNITION

Automated surveillance systems require the ability to recognize scenarios and behaviour from data. It is not sufficient to extract features and recognize objects since these have to be put in the correct context to determine the correct situation. For the scenarios we have defined earlier in this paper we have a list of features that need to be calculated at each time frame.

At each time step we determine:

- Number of people in the compartment
- Total movement (compared to the previous frame)

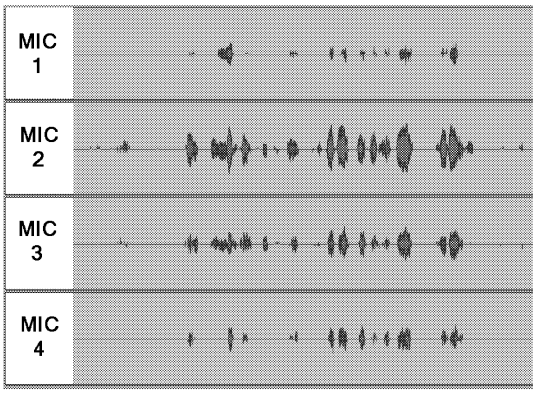


Figure 4: Four waveforms of a shouting scene recorded by the microphones. The waveforms are different in energy yet similar in form

- Total volume (over 100 ms)
- For each detected person position, pose and speed are determined.

By combining the feature vectors over time and using knowledge of the location of fixed objects in the train (such as seats), the behaviour of people in the train can be determined.

Preprocessing

Our work differs from others by the fact that our system has to work under a more problematic setting. The challenging circumstances we have to cope with in train compartments include the varying (and relatively unpredictable e.g. snow, rain, tunnels) light conditions. The preprocessing step consists of reducing noise in the video stream.

The raw video data consists of a sequence of jpeg frames with a resolution of 640x256 pixels interlaced. Therefore, the true resolution of the images should be 640x512 pixels (a 4:3 aspect ratio). As the Voila & Jones frontal face detection algorithm was trained for larger faces (larger window size) than the faces that normally occur in our video images, we further upscaled the images during preprocessing. (The scaling factor was obtained by trial and error until the classifier performed well for a number of preselected test images from our dataset.)

The raw camera images recorded during the experiments in the train are not directly usable in classification algorithms. The camera is somewhat rotated causing horizontal lines to be slanted in the recorded images. Finally, the camera faces downward with an unknown angle, so that the images recorded are a perspective projection of objects in a 3-D scene onto a 2-D image.

The method for image adjustment is based on a camera model called the Direct Linear Transformation (DLT). The DLT model describes a model for camera calibration using a linear transformation that takes into account the zoom, pan, and tilt of the camera. The DLT

method is a linear transformation so it is computationally cheap, but it is unable to compensate for non linear effects such as radial distortion.

The imaging process produced by the cameras can be interpreted as a sequence of three projective transformations. Given a point $p = (x_w, y_w, z_w, 1)$ in homogeneous world coordinates and a point $q = (f \cdot x_i, f \cdot y_i, f)$ in image coordinates corresponding to the projection of p onto the image, the mapping of p to q can be expressed as:

$$q = K \cdot \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \cdot M \cdot p \tag{1}$$

where K represents the intrinsic parameters of the camera and is given by:

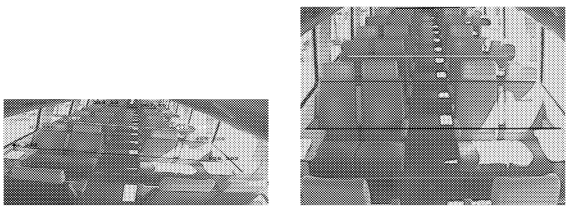
$$K = \begin{bmatrix} \sigma_x & \sigma_\theta & u_0 \\ 0 & \sigma_y & v_0 \\ 0 & 0 & 1 \end{bmatrix} \tag{2}$$

With (u_0, v_0) the coordinates of the principal point, and σ_x and σ_y the scale factors in image u and v axes. The parameter σ_θ describes the skewness of the two image axes. In practice it accounts for the skewness due to non-rectangular pixels. However, in most cameras the pixels nowadays are almost perfectly rectangular and thus σ_θ is very close to zero.

M represents the extrinsic parameters of the camera and is given by:

$$M = \begin{bmatrix} \ddots & \vdots & \iddots & \vdots \\ \cdots & R & \cdots & T \\ \iddots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{3}$$

Where R is the rotation and T the translation which relates the world coordinate system to the camera coordinate system. Figure 5 shows the images before and after adjustment.



(a) (b)

Figure 5: Comparison of an original image from a train camera (a) with the same image after preprocessing (b)

Face detection

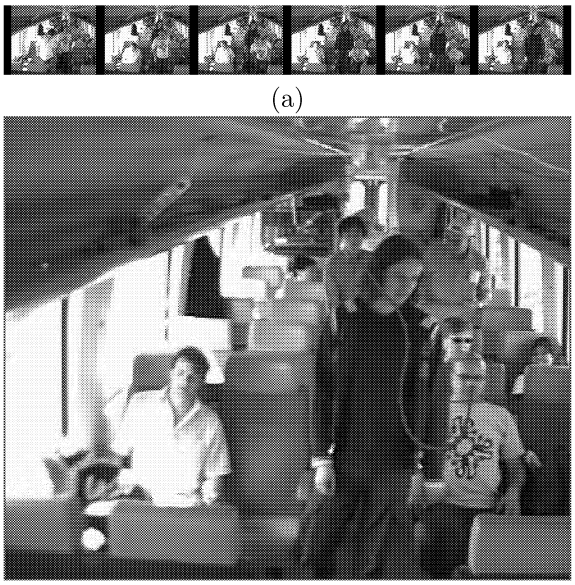
The method we considered for the purpose of person detection is face detection. When a face is detected in an image, obviously this means a person has been detected as well. We used the face detection method implemented in the OpenCV library, which is based on the method proposed by Viola & Jones (Viola and Jones, 2001).

The Viola & Jones method is capable of accurately detecting faces for which the classifier is trained, in reasonable time. It is however not very robust under noisy circumstances, and the frontal face classifier used is very susceptible to changes in orientation. In larger frontal faces of sizes around 100x100 pixels, we achieved a detection rate close to the rates reported in the literature. However, if the size of a face drops below this figure, detection rates fall dramatically, to a point where they hardly contribute to detection at all. Faces down to a resolution of 16x16 pixels can be detected, however the reduced size of detectable features and the higher signal to noise ratios in these smaller areas result in a very high rate of false negatives and false positives. From the detected faces in the data almost half were false positives.

Given the number of false positives, determining the number of people by counting the number of faces, is inaccurate at the least. In addition, we did not have enough actors to capture data of an overcrowded train. At the peak of occupation, the train compartment was fairly crowded at most. To determine the number of people in the current frame more accurately, we first filter all positives from areas where no faces are expected to be found using a mask. This excludes areas such as the windows and the ceiling, where false positives commonly occur. We analyze the measurements of a limited number of frames up to the current frame. The theory is that the number of people in a scene will not change abruptly, but instead change gradually. If for example, in one frame we detect 4 faces, and none in the next, a scenario not uncommon with a low detection rate, we assume the scene to still contain 4 people.

Action recognition

Our approach for activity recognition is by comparing the characteristics of the trajectory of people. We apply greedy nearest-neighbor matching to construct most probable tracks from the coordinates of detected faces. To guard against false positives, we apply a mask focused on the area around the corridor and the seats. To account for the low detection rate, we search over a maximum of 10 frames increasing the search area by 5 pixels every frame without a face found. This produces satisfactory results, due to the test data containing little occlusion of actors. Alternatively, other prediction methods, such as linear- and Kalman-filtering are widely used (Wang et al., 2003). The paths thus obtained (see figure 6) are compared to some predefined trajectory templates of actions such as entering the compartment and sitting, walking through the corridor, begging etc. The resulting measurement vector is compared to the template. The sum of the Euclidean distances between the current trajectory coordinates and each template trajectory coordinates is calculated and the action template with the smallest cumulative difference is selected.



(a)

(b)

Figure 6: Individual frames with detected faces (a) and calculated path (red) of a person overlayed against a video image (b)

The results of the action recognition algorithm are somewhat disappointing at the moment. There are just a handfull of trajectories correctly recognized. This is partly due to the low detection rate and the high number of false positives. More importantly, we think the bad performance is caused by the way people walk. People tend to wobble while they are walking, this effect is amplified when people are walking near the camera. These sideway movements corrupted the speed measurements to such an extend that they where left out of the trajectory recognition algorithm. A solution would be better smoothing of the tracks or applying a normalisation measure determined by the distance to the camera. The distance to the camera can be determined by the size of a person's face or body. An additional benefit of working with distances is that positions can be translated into 3-D coordinates instead of the currently used position on the 2-D projection plane of the image.

Behavior interpretation

Our goal is to define a simple set of rules based on interviews with experts, to recognize the predefined behaviours from observed data. Currently, the behaviour recognition is implemented as a rule based decision system that combines incoming features (number of people in the compartment, total volume, position, paths and speed of people in the scene etc.) into a conclusion.

The rule based system contains rules that describe the salient features of each scenario. As features are detected over time, these features are asserted into the rule base system as facts. Those rules with their features satisfied gain a higher score. If the score of a scenario reaches a certain threshold, that scenario is concluded to be the true scenario (figure 7).

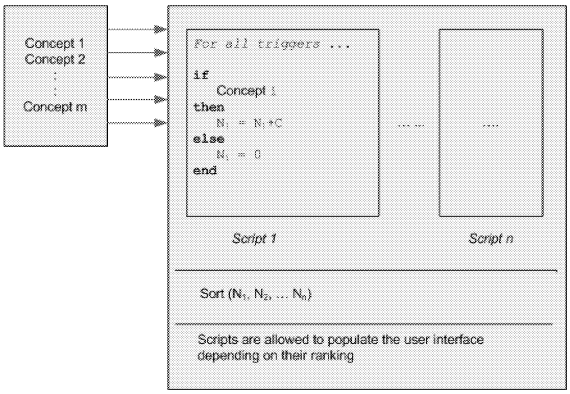


Figure 7: Reasoning scheme for behavior recognition

Features can be entered but also removed from the rule based system, making the system dynamic. To deal with uncertainty in the reasoning system, we will look into (dynamic) Bayesian networks. Since, it is not possible to model all scenarios and their particularities we plan to adopt techniques from emergent behaviour theory.

FUTURE WORK

To improve the performance of face recognition, background modeling techniques could be used. Faces can only be detected in foreground pixels. Being able to reduce the search area to only the foreground pixels will greatly reduce the running time of a face detection algorithm. The simplest background model is taking the difference between two frames, and considering the previous frame to be the background. By using the produced motion history image (MHI) we can limit our search to the foreground pixels only. Since we have to deal with varying lighting conditions, this method is not expected to work well. A more robust method that is widely used is the median filter. This method takes the median value of a pixel over all the frames in the stream that have been detected until the current time index, and constructs a background image from that. Alternatively, if an image of the scene without people in it is available, we can use this instead to perform an offline background subtraction. Other methods rely on statistical analysis to construct a probability model of a single pixel.

Detecting skin tones in an image will effectively allow us to localize persons in a scene as well. Skin detection can be done in several ways, most of which are computationally inexpensive. The most basic method is simply determining for each pixel whether it belongs to an certain empirically determined color range, in this case that of skin tones. We can take advantage of the fact that the RGB values of skin tones are highly correlated for

skin tones. The same holds for the YCbCr color space (Albiol et al., 2000). This correlation is quite specific, but not unique to skin tones. We experienced difficulties applying it in train compartments, specifically because the color of the upholstery was similar to skin color.

The key to differentiate between some scenarios depends on the recognition of certain salient objects or people. The conductor checking for tickets and the beggar scenarios for example can be differentiated by the detection of the conductor. Since conductors in the Netherlands wear specific uniforms, it is worthwhile to develop algorithms to detect the conductor specifically.

To deal with uncertainty in the reasoning system, we will look into (dynamic) Bayesian networks. Since, it is not possible to model all scenarios and their particularities we plan to adopt techniques from emergent behaviour theory. The idea is to have the scenario emerge as a completed puzzle from the detected features (puzzle pieces) instead of the fixed scenarios in the expert system approach.

DISCUSSION AND CONCLUSIONS

In this paper we presented our work so far in the development of an aggression detection system for train compartments. Most of the work is still in a preliminary stage and many tasks need still to be done.

Nevertheless, we have managed to develop a prototype for simple behavior recognition in a train. We used off-the-shelf algorithms to detect low level features from data and we developed a high-level rule based reasoning system that combines the features into recognized behaviors. Rules of thumb used by security expert have been translated into rules for the reasoning system.

Since most of the classification algorithms that we used are trained (and meant) to work in lab environments, better fine tuning of the algorithms to suit the train compartment might improve results. For example, for person detection we used face detection. The downside of this method however is that only specific views of faces, such as frontal or profile, will yield positives. We suggest therefore that this method be used in conjunction with other person detection methods to increase the detection rate, as well as decreasing the likelihoods of false positives.

Currently the reasoning system is only capable of recognizing predefined scenarios from facts. We need to expand this with uncertainty and reasoning about unanticipated scenarios.

ACKNOWLEDGEMENTS

This research was done as part of an ongoing project at the TUDelft funded by the MultimediaN project. We like to thank the NS/ProRail for providing us with a train to do recordings in and a train conductor to give advice.

REFERENCES

Albiol, A., Torres, L., Bouman, C. A., and Delp, E. J. 2000. "A Simple and Efficient Face Detection Algorithm for Video Database Applications". In *Proceedings of the IEEE International Conference on Image Processing*, Vol. 2, pp. 239–242.

Beal, M. J., Attias, H., and Jojic, N. 2002. "Audio-Video Sensor Fusion with Probabilistic Graphical Models". In *Proceedings of the 7th European Conference on Computer Vision*, pp. 736–752, London, UK. Springer-Verlag.

Clavel, C., Ehrette, T., and Richard, G. 2005. "Events Detection For an Audio-based Surveillance System". In *the IEEE International Conference on Multimedia and Expo (ICME 2005)*, pp. 1306– 1309.

Cupillard, F., Avanzi, A., Brémond, F., and Thonnat, M. 2004. "Video Understanding For Metro Surveillance". In *Proceedings of the IEEE International Conference on Networking, Sensing & Control*, Taipei, Taiwan.

Foresti, G., Micheloni, C., Snidaro, L., Remagnino, P., and Ellis, T. 2005. "Active video-based surveillance system: the low-level image and video processing techniques needed for implementation". *IEEE Signal Processing Magazine*, Vol. 22 No. 2 pp. 25–37.

Härmä, A., McKinney, M. F., and Skowronek, J. 2005. "Automatic Surveillance of the Acoustic Activity in our Living Environment". In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME 2005)*.

Javed, O., Rasheed, Z., Alatas, O., and Shah, M. 2003. "Knight^M: A Real-time Surveillance System for Multiple Overlapping and Non-overlapping Cameras". In *Proceedings of the International Conference on Multimedia and Expo (ICME 2003)*.

Oliver, N., Rosario, B., and Pentland, A. 2000. "A Bayesian Computer Vision System for Modeling Human Interactions". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22 pp. 831–843.

Phillips, P., Flynn, P., Scruggs, T., Bowyer, K., Chang, J., Hoffman, K., Marques, J., Min, J., and Worek, W. 2005. "Overview of the Face Recognition Grand Challenge". In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005)*, Vol. 1, pp. 947–954.

Velastin, S. A., Maria Alicia Vicencio-Silva, B. L., and Khoudour, L. 2002. "A Distributed Surveillance System For Improving Security In Public Transport Networks". *Special Issue on Remote Surveillance Measurement and Control*, Vol. 35 No. 8 pp. 209–13.

Viola, P. and Jones, M. 2001. "Rapid Object Detection using a Boosted Cascade of Simple Features". In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognitio (CVPR 2001)*, Vol. 1, pp. I–511– I–518.

Wang, L., Hu, W., and Tan, T. 2003. "Recent Developments in Human Motion Analysis". *Pattern Recognition*, Vol. 36 No. 3 pp. 585–601.

Zhang, D., Gatica-Perez, D., Bengio, S., and McCowan, I. 2006. "Modeling Individual and Group Actions in Meetings With Layered HMMs". *IEEE Transactions on Multimedia*, Vol. 8 No. 3 pp. 509–520.

Semantic Audio-Visual Data Fusion for Automatic Emotion Recognition

Dragos Datcu, Leon J.M. Rothkrantz
Man-Machine Interaction Group
Delft University of Technology
2628 CD, Delft,
The Netherlands
E-mail: {D.Datcu ; L.J.M.Rothkrantz}@tudelft.nl

KEYWORDS
Data fusion, automatic emotion recognition, speech analysis, face detection, facial feature extraction, facial characteristic point extraction, Active Appearance Models, Support Vector Machines.

ABSTRACT

The paper describes a novel technique for the recognition of emotions from multimodal data. We focus on the recognition of the six prototypic emotions. The results from the facial expression recognition and from the emotion recognition from speech are combined using a bi-modal multimodal semantic data fusion model that determines the most probable emotion of the subject. Two types of models based on geometric face features for facial expression recognition are being used, depending on the presence or absence of speech. In our approach we define an algorithm that is robust to changes of face shape that occur during regular speech. The influence of phoneme generation on the face shape during speech is removed by using features that are only related to the eyes and the eyebrows. The paper includes results from testing the presented models.

INTRODUCTION

The ability to replicate the human competence in naturally processing emotional clues from different channels during interpersonal communication has achieved an even higher role for the modern society nowadays. Smart human computer interfaces, affect sensitive robots or systems to support and coordinate people's daily activities are all about to incorporate knowledge on how emotions are to be perceived and interpreted in a similar way they are sensed by human beings.
While the study on human emotion recognition using unimodal information has considerably maturated for the last decade, the research on multimodal emotion understanding is still at the preliminary phase (Pantic and Rothkrantz, 2003). Sustained efforts attempt answering the question of what is the role and how information from various modalities can support or attenuate each other so as to get the smooth determination of human emotions. Recent research works have pointed to the advantage of using combinations of facial expressions and speech for correctly determining the subject's emotion (Busso et al., 2004; Zeng et al., 2007).
In the current paper we investigate the creation of a bimodal emotion recognition algorithm that incorporates facial expression recognition and emotion extraction from speech. Mostly we are interested on the design of a multimodal emotion data fusion model that works at the high, semantic level and that takes account of the dynamics in facial expressions and speech. Following recent comparable studies, we aim at obtaining higher performance rates for our method when compared to the unimodal approaches. The algorithms we use derive representative and robust feature sets for emotion classification model. The current research is a continuation of our previous work on facial expression recognition (Datcu and Rothkrantz, 2007; Datcu and Rothkrantz, 2005) and emotion extraction from speech signals (Datcu and Rothkrantz, 2006).

Acknowledgments. The research reported here is part of the Interactive Collaborative Information Systems (ICIS) project, supported by the Dutch Ministry of Economic Affairs, grant nr: BSIK03024.

RELATED WORK

The paper of (Wimmer et al., 2008) studies early feature fusion models based on statistically analyzing multivariate time-series for combining the processing of video based and audio based low-level descriptors (LLDs).
The work of (Hoch et al., 2005) presents an algorithm for bimodal emotion recognition in automotive environment. The fusion of results from unimodal acoustic and visual emotion recognizers is realized at abstract decision level.
For the analysis, the authors used a database of 840 audiovisual samples that contain recordings from seven different speakers showing three emotions. By using a fusion model based on a weighted linear combination, the performance gain becomes nearly 4% compared to the results in the case of unimodal emotion recognition.
(Song et al., 2004) presents a emotion recognition method based on Active Appearance Models – AAM for facial feature tracking. Facial Animation Parameters – FAPs are extracted from video data and are used together with low level audio features as input for a HMM to classify the human emotions.
The paper of (Paleari and Lisetti, 2006) presents a multimodal fusion framework for emotion recognition that relies on MAUI - Multimodal Affective User Interface paradigm. The approach is based on the Scherer's theory Component Process Theory (CPT) for the definition of the user model and to simulate the agent emotion generation.

(Sebe et al., 2006) proposes a Bayesian network topology for recognizing emotions from audio and facial expressions. The database they used includes recordings of 38 subjects which show 11 classes of affects. According to the authors, the achieved performance results pointed to around 90% for bimodal classification of emotions from speech and facial expressions compared to 56% for the face-only classifier and about 45% for the prosody-only classifier.
(Zeng et al., 2007) conducted a series of experiments related to the multimodal recognition of spontaneous emotions in a realistic setup for Adult Attachment Interview. They use Facial Action Coding System – FACS (Ekman and Friesen, 1978) to label the emotion samples. Their bimodal fusion model combines facial texture and prosody in a framework of Adaboost multi-stream hidden Markov model (AdaMHMM).
(Joo et al., 2007) investigates the use of S-type membership functions for creating bimodal fusion models for the recognition of five emotions from speech signal and facial expressions. The achieved recognition rate of the fusion model was 70.4% whereas the performance of the audio-based analysis was 63% and the performance of the face-based analysis was 53.4%. (Go et al., 2003) uses Z-type membership functions to compute the membership degree of each of the six emotions based on the facial expression and the speech data. The facial expression recognition algorithm uses multi-resolution analysis based on discrete wavelets. An initial gender classification is done by the pitch of the speech signal criterium. The authors report final emotion recognition results of 95% in case of male and 98.3% for female subjects. (Fellenz et al., 2000) uses a hybrid classification procedure organized in a two-stages architecture to select and fuse the features extracted from face and speech to perform the recognition of emotions. In the first stage, a multi-layered perceptron (MLP) is trained with the backpropagation of error procedure. The second symbolic stage involves the use of PAC learning paradigm for Boolean functions.
(Meng et al., 2007) presents a speech-emotion recognizer that works in combination with an automatic speech recognition system. The algorithm uses Hidden Markov Model – HMM as a classifier. The features considered for the experiments consisted of 39 MFCCs plus pitch, intensity and three formants, including some of their statistical derivates.
(Busso et al., 2004) explores the properties of both unimodal and multimodal systems for emotion recognition in case of four emotion classes. In this study, the multimodal fusion is realized separately at the semantic level and at the feature level. The overall performance of the classifier based on feature level fusion is 89.1% which is close to the performance of the semantic fusion based classifier when the product-combining criterion is used.

MULTIMODAL APPROACH

In our approach, the emotion recognition algorithm works for the prototypic emotions (Ekman and Friesen, 1978) and is based on semantic fusion of audio and video data. We have based our single modality data processing methods on previous work (Datcu and Rothkrantz, 2006; Datcu and Rothkrantz, 2007) we have conducted for the recognition of emotions from human faces and speech.

Facial Expression Recognition

In the case of video data processing, we have developed automatic systems for the recognition of facial expressions for both still pictures and video sequences. The recognition was done by using Viola&Jones features and boosting techniques for face detection (Viola and Jones, 2001), Active Appearance Model – AAM for the extraction of face shape and Support Vector Machines –SVM (Vapnik 1995; Vapnik 1998) for the classification of feature patterns in one of the prototypic facial expressions. For training and testing the systems we have used Cohn-Kanade database (Kanade et al., 2000) by creating a subset of relevant data for each facial expression. The structure of the final dataset is presented in Table 1.

Table 1: The structure of the Cohn-Kanade subset for facial expression recognition.

Expression	#samples
Fear	84
Surprise	105
Sadness	92
Anger	30
Disgust	56
Happy	107

The Active Appearance Model – AAM (Cootes et al., 1998) makes sure the shapes of the face and of the facial features are correctly extracted from each detected face.
Starting with the samples we have collected from the Cohn-Kanade database, we have determined the average face shape and texture (Figure 1).

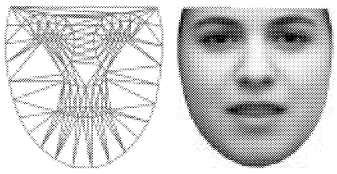


Figure 1: The mean face shape (left) and the mean face texture aligned to the mean shape (right).

According to the AAM model, the shape and texture can be represented as depicted in Equation 1, where the values of $\bar{s}$ and $\bar{t}$ represent the mean face shape and the mean face texture. The matrices Φ_s and Φ_t contain the eigenvectors of the shape and texture variations.

Equation 1

$$s = \bar{s} + \Phi_s b_s$$

$$t = \bar{t} + \Phi_t b_t$$

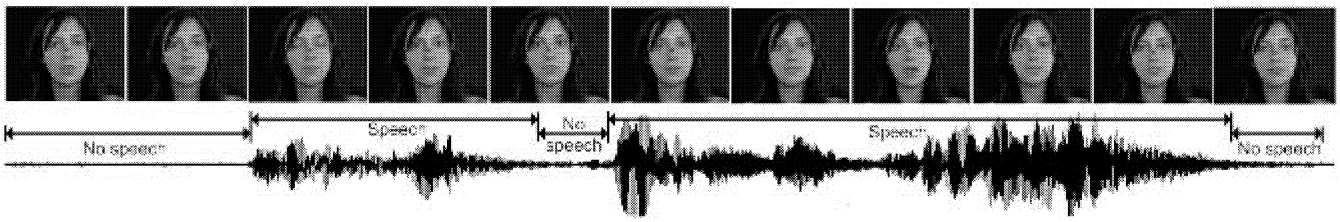


Figure 2: The silence/non-silence detection using multimodal data.
Sample taken from eNTERFACE 2005 database: "Oh my god, there is someone in the house!"

The final combined model contains information regarding both the shape and texture and is written as in Equation 2. The term W_s is a diagonal matrix that introduces the weighting between units of intensities and units of distances.

Equation 2

$$b = \begin{bmatrix} W_s b_s \\ b_t \end{bmatrix}$$

Based on the AAM face shape, the facial expression recognition algorithm generates a set of features to be used further on during the emotion classification stage. The features stand for geometric parameters as distances computed between specific Facial Characteristic Points – FCPs (Figure 3).
For the recognition of expressions in still pictures, the distances determined from one face form a representative set of features to reflect the emotion at a certain moment of time. In the case of recognition of facial expressions in video sequences, the features are determined as the variation of the same distances between FCPs as observed during several consecutive frames.

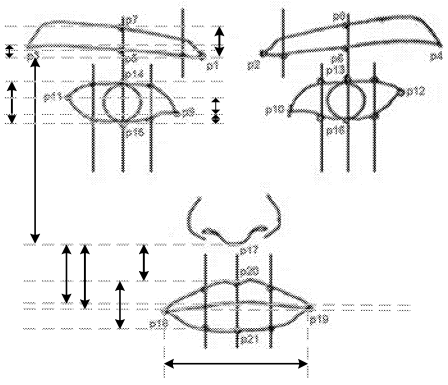


Figure 3: The Facial Characteristic Point FCP model.

Our algorithm for the recognition of emotions in videos implies an initial processing of the multimodal data. Firstly, the audio-video input data is rescaled by conversion to a specific frame-rate (Figure 4). This process may imply downscaling by skipping some video and audio frames. Secondly, the audio data is processed in order to determine the silence and non-silence segments. The resulting segments are correlated to the correspondent audio data and constitute the major data for the analysis.

In the case of facial expression recognition, within each segment an overlaping sliding window (Figure 5) groups together adjacent video frames. Based on the set of video frames, the recognition of facial expressions determines the most probable facial expression using a voting algorithm and a classifier trained on still pictures.

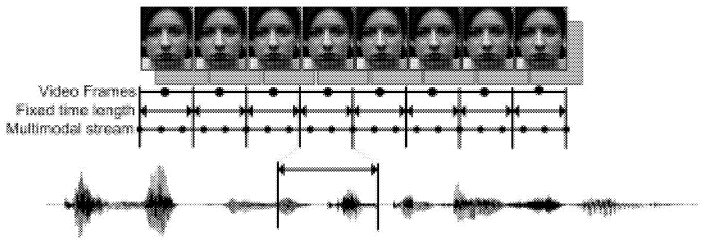


Figure 4: Multimodal frame rescaling algorithm.

For the video oriented classifier, the most probable facial expression is determined by taking into account the variation of the features extracted from all the video frames in the group.

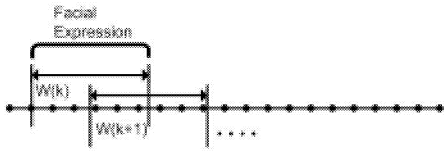


Figure 5: Video frame selection algorithm.

The indentification of silence and non-silence segments is realized by using both acoustic and video information (Figure 2). Apart from running acoustic analysis of the data, speech can be detected by tracking features from a simple FCPs based model that includes data from the mouth area.
The recognition of emotions is realized differently for silence segments and non-silence segments (Figure 6). For silence segments, the emotion is represented by the facial expression as detemined by the facial expression classification algorithm.
For the non-silence segments, the emotion recognition is based on the multimodal semantic fusion of the results of the emotion classification on single modalities. Additionally, the facial expression classification algorithm for non-silence segments determines the most probable facial expression by considering a different set of geometric features. The input features in this case relate to only FCPs from the upper part of the face.

Table 2: The geometric feature set for facial expression recognition for silence data segments.

		Visual feature			Visual feature			Visual feature
v_1	$(P_1, P_7)_y$	Left eyebrow	v_7	$(P_{14}, P_{15})_y$	Left eye	v_{13}	$(P_{17}, P_{20})_y$	Mouth
v_2	$(P_1, P_3)_y$	Left eyebrow	v_8	$(P_9, P_{11})_y$	Left eye	v_{14}	$(P_{20}, P_{21})_y$	Mouth
v_3	$(P_2, P_8)_y$	Right eyebrow	v_9	$(P_9, P_{15})_y$	Left eye	v_{15}	$(P_{18}, P_{19})_y$	Mouth
v_4	$(P_2, P_4)_y$	Right Eyebrow	v_{10}	$(P_{13}, P_{16})_y$	Right eye	v_{16}	$(P_{17}, P_{18})_y$	Mouth
v_5	$(P_1, P_{17})_y$	Left Eyebrow	v_{11}	$(P_{10}, P_{12})_y$	Right eye	v_{17}	$(P_{17}, P_{19})_x$	Mouth
v_6	$(P_2, P_{17})_y$	Right eyebrow	v_{12}	$(P_{10}, P_{16})_y$	Right eye			

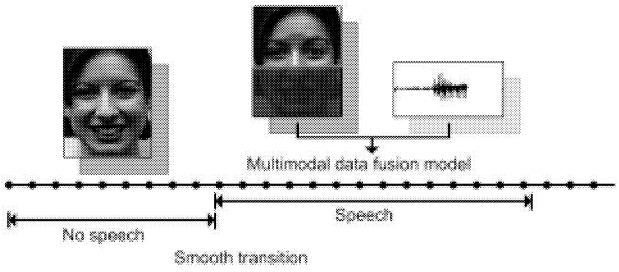


Figure 6: Emotion recognition regarding the transition between two adjacent segments.

The reason for not considering the FCPs of the mouth is explained by the natural influence of the phoneme generation on the mouth shape during the process of speaking.
The geometric features used for the recognition of facial expressions are illustrated in Table 2 for non-silence data segments and in Table 3 for silence data segments.

Table 3: The geometric feature set for facial expression recognition for speech-containing enhanced data segments.

		Feature			Feature
v_1	$(P_1, P_7)_y$	Left eyebrow	v_7	$(P_{14}, P_{15})_y$	Left eye
v_2	$(P_1, P_3)_y$	Left eyebrow	v_8	$(P_9, P_{11})_y$	Left eye
v_3	$(P_2, P_8)_y$	Right eyebrow	v_9	$(P_9, P_{15})_y$	Left eye
v_4	$(P_2, P_4)_y$	Right eyebrow	v_{10}	$(P_{13}, P_{16})_y$	Right eye
v_5	$(P_1, P_9)_y$	Left eyebrow	v_{11}	$(P_{10}, P_{12})_y$	Right eye
v_6	$(P_2, P_{10})_y$	Right eyebrow	v_{12}	$(P_{10}, P_{16})_y$	Right eye

All the FCPs are adjusted for correcting against the head rotation prior to computing the values of the geometric features used for the facial expression classification.

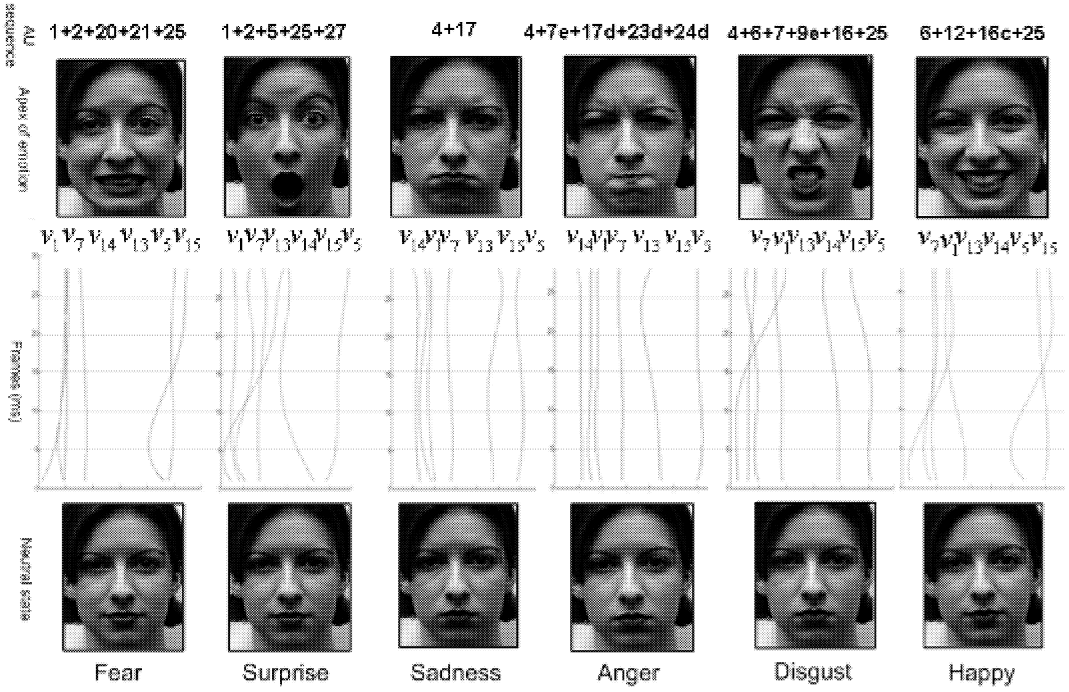


Figure 7: The dependency of temporal changes on emotion featured sequences (reduced parameter set).

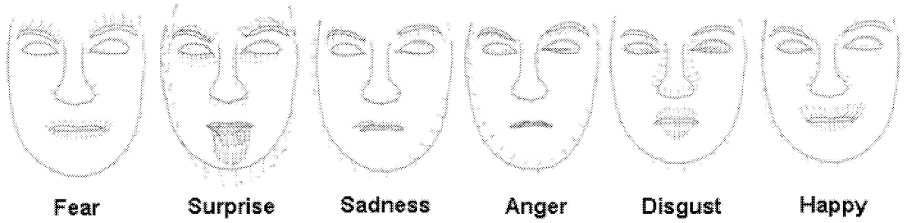


Figure 8: The emotion facial shape deformation patterns for the six prototypic emotion classes.

Moreover, another adjustement of the FCPs applies a correction against the varience of the distance between the subject and the camera. This is realized by scaling all the distance-oriented feature values by the distance between the inner corners of the eyes. The models that use the feature sets in Table 2 and Table 3 allow for the independent consideration of features from both sides of the face. The advantage of a facial expression recognition system that makes use of such a set of features is the ability to still offer good results for limited degrees of occlusion. For such cases, the features computed from the side that is not occluded can be mirrored to the features from the occluded side of the face.
The values of the geometric features over time may be plot for each facial expression (Figure 7).
An alternative to the previously described set of features is to take into account the dynamics of the features presented in Table 2 so as to determine the emotion given the relative deformation of the facial features in time (Figure 8).

Emotion recognition from speech

In the case of emotion recognition from speech, the analysis is handled separately for different number of frames per speech segment (Datcu and Rothkrantz, 2006). In the current approach there are five types of split methods applied on the initial audio data. Each type of split produces a number of data sets, according to all the frame combinations in one segment.

The data set used for emotion analysis from speech is Berlin (Burkhardt et al., 2005) – a database of German emotional speech. The database contains utterances of both male and female speakers, two sentences pro speaker. The emotions were simulated by ten native German actors (five female and five male). The result consists of ten utterances (five short and five long sentences). The length of the utterance samples ranges from 1.2255 seconds to 8.9782 seconds. The recording frequency is 16kHz.

The final speech data set contains the utterances for which the associated emotional class was recognized by at least 80% of the listeners. Following a speech sample selection, an initial data set was generated comprising 456 samples and six basic emotions (anger: 127 samples, boredom: 81 samples, disgust: 46 samples, anxiety/fear: 69 samples, happiness: 71 samples and sadness: 62 samples).
The Praat (Boersma and Weenink, 2005) tool was used for extracting the features from each sample from all generated data sets. According to each data set frame configuration, the parameters mean, standard deviation, minimum and maximum of the following acoustic features were computed: *Fundamental frequency* (pitch), *Intensity*, *F1*, *F2*, *F3*, *F4* and *Bandwidth*. All these parameters form the input for separate GentleBoost classifiers according to data sets with distinct segmentation characteristics.
The GentleBoost strong classifier is trained for a maximum number of 200 stages. Separate data sets containing male, female and both male and female utterances are considered for training and testing the classifier models.

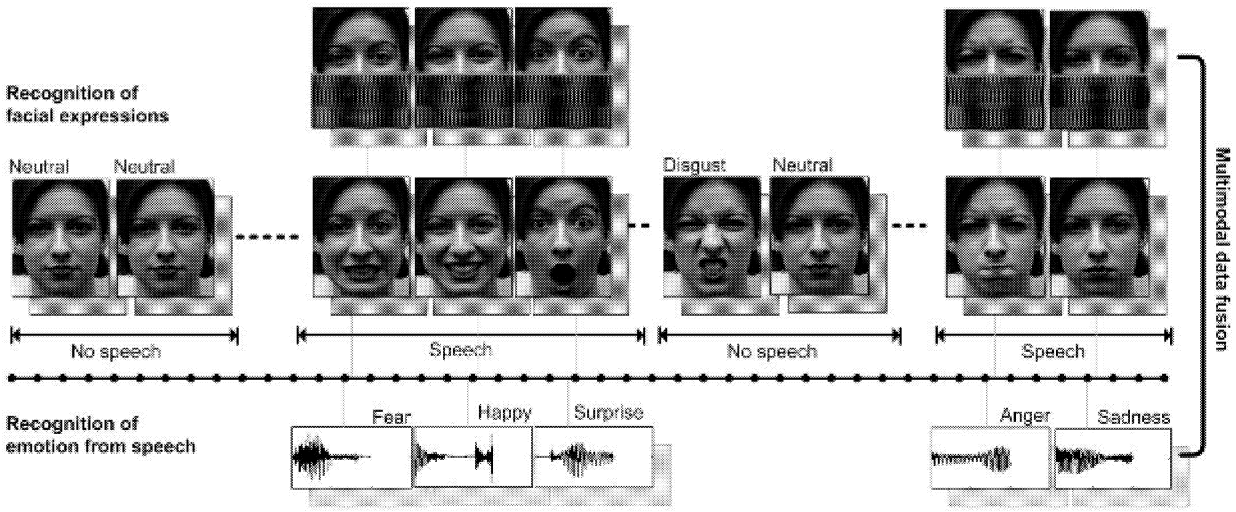


Figure 9: The sequencial recognition of human emotions from audio and video data.

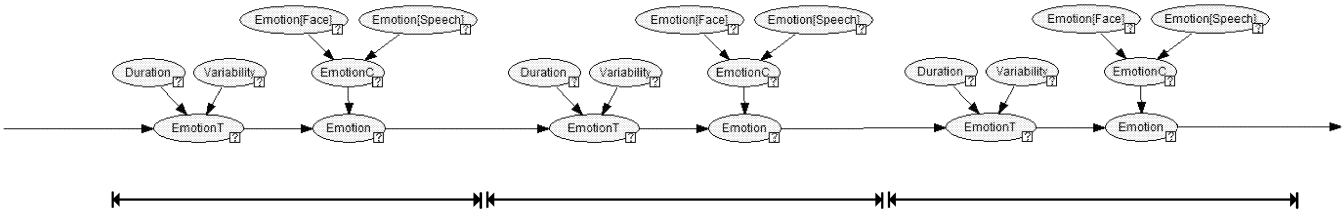


Figure 10: The DBN model for multimodal emotion recognition.

Bimodal emotion data fusion model

Figure 9 illustrates an example of the multimodal emotion recognition algorithm. High-level data fusion works only during the analyses of speech-enhanced multimodal segments.

The fusion model aims at determining the most probable emotion of the subject given the emotions determined in the previous frames. A data window contains the current and the previous *n* frames for the analysis.
Figure 10 depicts the Dynamic Bayesian Network - DBN for the emotion data fusion.
The variable *Duration* represents the stability of last determined emotion in consecutive frames in the analysis window. It has three states: *short*, *normal* and *long*. Each of the three possible situations are assumed to hold differently for each facial expression. Accordingly, it is assumed that for instance an emotion transition is likely to happen from one emotion to another after the former emotion has been shown during a number of consecutive frames.
The variable *Variability* represents the knowledge on the variability of previously determined emotions in the analysis window. It has three states: *low*, *medium* and *high*. Presumably the probability of emotion transition should be higher when the subject has shown rapid changes of emotions during the previous time.
The variable *EmotionT* represents the most probable emotion taking into account only the emotion of the subject determined at the previous frame in the analysis window and the probability of showing another emotion.
The variable *EmotionC* is the emotion of the subject as it is computed by the facial expression recognition and the emotion extraction from speech at the current frame. The variable *Emotion* is the emotion of the subject at the current frame to be determined.

RESULTS

For the classification of facial expressions, different models have been taken into account. In our experiments we have used 2-fold Cross Validation method for testing the performance of the models. For training, we have used Cohn-Kanade database for experiments on facial expression recognition and Berlin database for emotion extraction from speech.
We have partly used the eNTERFACE'05 audio-visual emotion database (Martin et al., 2006) for testing our multimodal algorithms for emotion recognition.

The partial results presented in the paper show the performance achieved by our algorithms for facial expression recognition in processing silence (Table 4 and Table 5) and speech segments (Table 6 and Table 7).
Table 5 shows the results of algorithms that use the dynamic behaviour shown by geometric features as input for the emotion classification process. Addionally we show the results in the case of emotion recognition from speech (Table 8).
Ongoing work is set to test the multimodal fusion model by using eNTERFACE'05 data set.
The results of the facial expression recognition clearly show that a higher performance is obtained by the models that make use of features computed from the entire face shape in comparison to the model that uses information regarding only the eyes and eyebrows.

Table 4: The results for facial expression recognition using SVM (polynomial kernel of degree 3) for still pictures.

(%)	Fear	Surprise	Sadness	Anger	Disgust	Happy
Fear	**84.70**	3.52	3.52	4.70	1.17	2.35
Surprise	12.38	**83.80**	0.95	0	0	2.85
Sadness	6.45	3.22	**82.79**	1.07	3.22	3.22
Anger	3.44	6.89	6.89	**75.86**	6.89	0
Disgust	0	0	7.14	10.71	**80.35**	1.78
Happy	7.54	8.49	2.83	3.77	4.71	**72.64**

Table 5: The results for facial expression recognition using SVM (polynomial kernel of degree 3) for sequence of frames.

(%)	Fear	Surprise	Sadness	Anger	Disgust	Happy
Fear	**88.09**	2.38	4.76	3.57	1.19	0
Surprise	0	**88.67**	2.83	8.49	0	0
Sadness	5.43	2.17	**85.86**	2.17	1.08	3.26
Anger	10.71	0	3.57	**85.71**	0	0
Disgust	5.35	5.35	3.57	1.78	**82.14**	1.78
Happy	4.62	0	7.40	2.77	5.55	**79.62**

Table 6: The results for facial expression recognition using SVM (polynomial kernel of degree 3) for still pictures using only the eyes and eyebrows information.

(%)	Fear	Surprise	Sadness	Anger	Disgust	Happy
Fear	**66.67**	6.67	13.33	0	13.33	0
Surprise	0	**63.64**	0	36.36	0	0
Sadness	0	0	**64.71**	0	35.29	0
Anger	20.00	20.00	0	**60.00**	0	0
Disgust	0	0	25.00	12.50	**62.50**	0
Happy	39.13	0	0	0	0	**60.87**

Table 7: The results for facial expression recognition using SVM (polynomial kernel of degree 3) for sequence of frames using only the eyes and eyebrows information.

(%)	Fear	Surprise	Sadness	Anger	Disgust	Happy
Fear	**70.59**	0	0	0	29.41	0
Surprise	15.00	**70.00**	15.00	0	0	0
Sadness	15.79	15.79	**63.16**	0	5.26	0
Anger	16.67	16.66	0	**66.67**	0	0
Disgust	0	21.22	2	11.11	**65.67**	0
Happy	0	36.36	0	0	0	**63.64**

Table 8: The optimal classifier for each emotion class, Berlin data set.

(%)	**ac (%)**	**tpr (%)**	**fpr (%)**
Anger	0.83±0.03	0.72±0.16	0.13±0.06
Boredom	0.84±0.07	0.49±0.18	0.09±0.09
Disgust	0.92±0.05	0.24±0.43	0.00±0.00
Fear	0.87±0.03	0.38±0.15	0.05±0.04
Happy	0.81±0.06	0.54±0.41	0.14±0.13
Sadness	0.91±0.05	0.83±0.06	0.08±0.06

IMPLEMENTATION

Figure 11 shows a snap shot of our software implementation (Datcu and Rothkrantz, Software Demo 2007) for the bimodal human emotion recognition system. Our system runs on Windows machines. For the detection of faces we have mainly used the implementation of Viola&Jones method from Intel's Open Source Computer Vision Library – OpenCV. We have used AAM-API (Stegmann, 2003) libraries for the implementation of Active Appearance Models. For the speech processing part we have built Tcl/Tk scripts in combination with Snack Sound Toolkit, a public domain toolkit developed at KTH.
Finally, we have built our facial feature extraction routines, the facial expression recognition system and the emotion extraction from speech in C++ programming language. For the classification component, we have used LIBSVM (Chang and Lin, 2001). On an Intel Core 2 CPU @2.00 GHz, 2.00 GB of RAM our software implementation works at speed of about 5 fps.

The AAM module of our initial facial expression recognition system requires the detection of faces for each frame in the incoming video sequence. We have obtained a considerable improvement in terms of speed by nearly doubling the frame rate with an algorithm that uses information regarding the face shape of one subject at the current frame as initial location for the AAM fitting procedure at the next frame in the video sequence (Figure 12). In this way the face detection is run only at certain time intervals comparing to the case when it is run for all the frames.

Figure 11: A snap shot of our software implementation of the bimodal emotion recognition algorithm.

The disadvantage of this algorithm is the higher probability of generating faulty results for the extraction of the face shape.
Moreover, the faulty such cases attract definitive errorneous results for the rest of the frames in the sequence. This happens because of the possibly sharp moves of the head, rather low speed of the implementation and because of the incapacity of the AAM algorithm to match model face shapes to image face shapes when the translation, rotation or scalation effects present high magnitudes. The case can be overcome by using an effective face tracking algorithm to anticipate the move of the head in the sequence of frames.

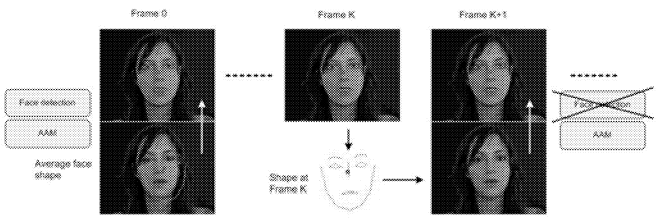


Figure 12: The shape at frame K is used as initial shape estimation for the AAM shape matching algorithm at frame K+1; the face detection is run only once every n frames instead of once for every frame.

CONCLUSION

In our experiments we have used Cohn-Kanade database for training and testing our facial expression recognition models. For the emotion recognition from speech we have used Berlin database that contains utterances in German language. A better approach is to use a unique multimodal database for running the full set of experiments on the algorithms detailed in the paper.

The use of additional semantic information regarding, for instance the emotion level from text or gestures - would greately increase the performance of the multimodal emotion recognition system. In such a situation, more advanced multimodal data fusion models may be developed.

Eventually, the fusion technique described in the paper focuses on information from only the upper part of the face. Instead, an efficient alternative would be to filter out the influence of phonemes and to run the same type of facial expression recognition models also for the speech-enhanced multimodal segments.

REFERENCES

Aleksic, P. S., A. K. Katsaggelos, "Automatic Facial Expression Recognition using Facial Animation Parameters and Multi-Stream HMMs", ISSN: 1556-6013, in IEEE Transactions on Information Forensics and Security, 2006.

Boersma, P., Weenink, D., "Praat: doing phonetics by computer (Version 4.3.14)" [Computer program], 2005.

Burkhardt, F., A. Paeschke, M. Rolfes, W. Sendlmeier, B. Weiss, "A Database of German Emotional Speech", Proceedings Interspeech, Lissabon, Portugal 2005.

Busso, C., Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, S. Narayanan, "Analysis of Emotion Recognition using Facial Expressions, Speech and Multimodal Information", Proceedings of ACM 6th International Conference on Mutlmodal Interfaces (ICMI 2004), ISBN: 1-58113-890-3, State College, PA, 2004.

Chang, C. C., C. J. Lin, "LIBSVM: a library for support vector machines", http://www.csie.ntu.edu.tw/~cjlin/libsvm, 2001.

Cootes, T. F., G. J. Edwards, C. J. Taylor, "Active appearance models", Lecture Notes in Computer Science, vol. 1407, pp. 484–98, 1998.

Datcu, D., L. J. M. Rothkrantz, "Multimodal workbench for human emotion recognition", Software Demo at IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR'07, Minneapolis, Minnesota, USA, 2007.

Datcu, D., L. J. M. Rothkrantz, "The recognition of emotions from speech using GentleBoost Classifier", CompSysTech'06, June 2006.

Datcu, D., L. J. M. Rothkrantz, "Facial Expression Recognition in still pictures and videos using Active Appearance Models. A comparison approach.", CompSysTech'07, ISBN 978-954-9641-50-9, pp. VI.13-1-VI.13-6, Rousse, Bugaria, June 2007.

Datcu, D., L. J. M. Rothkrantz, "The use of Active Appearance Model for facial expression recognition in crisis environments", Proceedings ISCRAM2007, ISBN 9789054874171, pp. 515-524, 2007.

Datcu, D., L. J. M. Rothkrantz, "Machine learning techniques for face analysis", Euromedia 2005, ISBN 90-77381-17-1, pp. 105-109, 2005.

Ekman, P., W. Friesen, "Facial Action Coding System", Consulting Psychologists Press, Inc., Palo Alto California, USA, 1978.

Fellenz, W. A., J. G. Taylor, R. Cowie, E. Douglas-Cowie, F. Piat, S. Kollias, C. Orovas, B. Apolloni, "On emotion recognition of faces and of speech using neuralnetworks, fuzzy logic and the ASSESS system", Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks, IJCNN 2000, ISBN: 0-7695-0619-4, pp. 93-98, 2000.

Go, H. J., K. C. Kwak, D. J. Lee, "Emotion recognition from the facial image and speech signal," SICE Annual Conference, Japan, pp. 2890-2895, 2003.

Hoch, S., F. Althoff, G. McGlaun, G. Rigoll, "Bimodal fusion of emotional data in an automotive environment", IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP '05, ISBN: 0-7803-8874-7, Vol.2, pp. 1085-1088, 2005.

Joo, J. T., S. W. Seo, K. E. Ko, K. B. Sim, "Emotion Recognition Method Based on Multimodal Sensor Fusion Algorithm", 8th International Symposium on Advanced Intelligent Systems ISIS'07, 2007.

Kanade, T., J. F. Cohn, Y. Tian, "Comprehensive database for facial expression analysis,". Proc. of the 4th IEEE Int. Con. On Automatic Face and Gestures Reco., France, 2000.

Martin, O., I. Kotsia, B. Macq, I. Pitas, "The eNTERFACE'05 Audio-Visual Emotion Database", Proceedings of the 22nd International Conference on Data Engineering Workshops (ICDEW'06), ISBN:0-7695-2571-7, 2006.

Meng, H., J. Pittermann, A. Pittermann, W. Minker, "Combined speech-emotion recognition for spoken human-computer interfaces", IEEE International Conference on Signal Processing and Communications, Dubai (United Emirates), 2007.

OpenCV: Open Source Computer Vision Library; http://www.intel.com/technology/computing/opencv/.

Pantic, M., L. J. M. Rothkrantz, "Toward an affect-sensitive multimodal human-computer interaction", Proceedings of the IEEE, Special Issue on human-computer multimodal interface, 91(9):1370–1390, 2003.

Paleari, M., C. L. Lisetti, "Toward Multimodal Fusion of Affective Cues", HCM'06, Santa Barbara, California, USA, pp. 99-108, October 27, 2006.

Sebe N., I. Cohen, T. Gevers, T. S. Huang, "Emotion Recognition Based on Joint Visual and Audio Cues", Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06), 2006.

Song, M., J.J. Bu, C. Chen, N. Li, "Audio-visual based emotion recognition-a new approach", Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004 - CVPR 2004, ISBN: 0-7695-2158-4, pp. 1020-1025, 2004.

Snack Sound Toolkit; http://www.speech.kth.se/snack/.

Stegmann, M. B., "The AAM-API: An Open Source Active Appearance Model Implementation", Medical Image Computing and Computer-Assisted Intervention - MICCAI 2003, 6th Int. Conference, Montréal, Canada, Springer, pp. 951-952, 2003.

Vapnik, V., "The Nature of Statistical Learning Theory", Springer-Verlag, New York, 1995.

Vapnik, V., "Statistical Learning Theory", John Wiley and Sons, Inc., New York, 1998.

Viola, P., M. Jones, "Robust Real-time Object Detection." Second International Workshop on Statistical and Computational Theories of Vision-Modeling, Learning, Computing, and Sampling, 2001.

Zeng, Z., Y. Hu, G. I. Roisman, Z. Wen, Y. Fu, T. S. Huang, "Audio-Visual Spontaneous Emotion Recognition", Artificial Intelligence for Human Computing, ISBN: 978-3-540-72346-2, Springer Berlin/Heidelberg, pp.72-90, 2007.

Wimmer, M., B. Schuller, D. Arsic, G. Rigoll, B. Radig, "LOW-LEVEL FUSION OF AUDIO AND VIDEO FEATURE FOR MULTI-MODAL EMOTION RECOGNITION", In 3rd International Conference on Computer Vision Theory and Applications (VISAPP), Madeira, Portugal, 2008.

DATA MANIPULATION

A TOOL FOR TURNING SERIES OF DIGITAL PHOTOGRAPHS INTO DYNAMIC VIDEO FLUX

Philippe Codognet
Keio University,
Research Institute for Digital Media and Content,
2-15-45 Mita, Minato-ku, Tokyo 108-8345, Japan
Email: philippe@dmc.keio.ac.jp

KEYWORDS
Presentation Software, Digital Video, Toolboxes for Moving Graphics

ABSTRACT

We propose a method for turning a slideshow of photographs into a media art installation consisting of a digital video flux of intermingled images. Our idea is to combine one photograph with another on a pixel-by-pixel basis and to apply Cellular Automata rules to further mix in real-time the images together. The Cellular Automaton, will, from a series of randomly chosen seed pixels, slowly merge an image with another. It is worth noticing that more than two images can overlap at the same time. The key point is that this transformation process is part of the artwork itself, bringing some impressionist-like aspect to the flux of images.

INTRODUCTION

The basic idea of this paper is to turn a static series of photographs into a continuous, ever-changing stream of dynamically created pictures composed of parts of basic images taken from a database of digital photographs. Indeed, the premises of our system lie in a simple observation in the psychology of perception: the so-called *subjective* or *illusory* contour. Following the tradition of *Gestalt* psychology from early 20th century in Germany, the Italian psychologist Gaetano Kanizsa pointed out in the 50's, with his well known illusion called "Kanizsa triange", that the human eye has a natural tendency to create lines even if they are not properly depicted, and thus to create forms by matter of continuity (Kanizsa 1955).

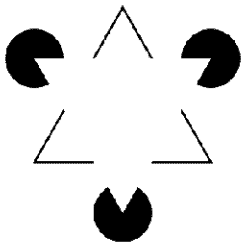


Figure 1: Kanizsa Triangle.

Therefore the basic idea is that we do not need all the pixels of an image to cognitively apprehend or "see" the forms depicted in it. Thus some pixels can be removed and used to present another photograph, creating some kind of pixel-based photomontage.

In the domain of computer graphics, most art-oriented works have been attempting to avoid the discrete (pixel-based) nature of digital images and to reach the (apparent) continuity of finer-grain image, such as silver-based photography or painting, trying thus to make the pixels disappear. Although this approach is perfectly valid from an engineering point of view (in the tradition of the "recreation of nature"), it has to be noted that in art history, opposite approaches would usually be found, in the sense that, when faced with a new technology or challenge, artists would rather tend to reduce their discipline to the core that cannot be handled by other means. This can be seen for instance in the 19th century with the beginning of impressionism after the birth of photography, as painters no more need to replicate reality (mimesis) as this could be better achieved by black and white photography, but they rather focused on colours and subjective "impressions". Another example is from the early 20th century when Russian avant-guard artists and especially the Suprematist and Constructivist painters, advocated that, in contrast to the cubist experiments and sculpture, painting should be flat and should not try to integrate the third dimension into the picture plane. These theories, developed in New York in the late 30's by the founder of New York Museum of Modern Art, Alfred Barr, and in the late 40's by art theorist Clement Greenberg (Greenberg 1961), will pave the way for the American "abstract expressionism" - Jackson Pollock, etc - where painting is reduced to a its pure 2D aspects.

Thus our approach is to treat the pixel as a first class aesthetics object and propose a pixel-by-pixel transformation of series of photographs into a seamless flux of changing pixels. Therefore our system has not to be seen as a slideshow tool such as the Tiling Slideshow (Chen et al. 2006) (Chen et al. 2007) which is used for presenting series of personal photographs or the Impressionism Slideshow (Li and Chan 2007), which is used to create a slideshow of Impressionist paintings together with musical soundtrack. In these tools, each image is kept intact, even if in (Chen et al. 2006) several photographs are tiled together within a single frame. Our approach is rather to merge photographs at the pixel level, creating thus new images. For instance in the Figure 2 below, elements of three photographs are merged together.

We therefore named our system Palimpsest, as when a manuscript is erased and re-written with new text, in the tradition of medieval copies of books.

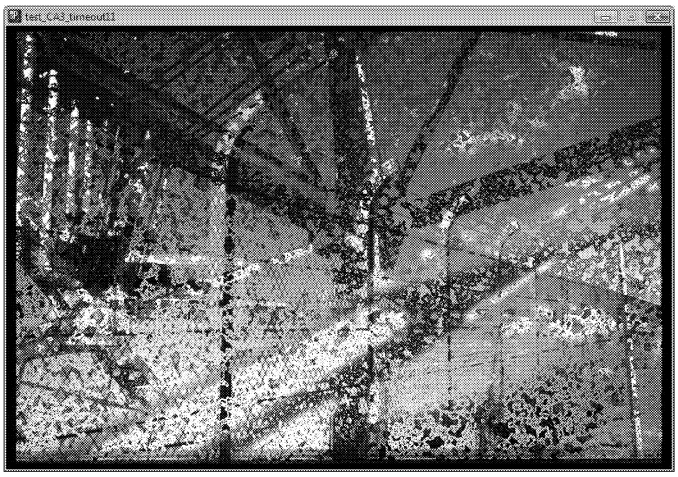


Figure 2: Image Produced by the Palimpsest System.

METHOD

The *Palimpsest* system works on a database of digitalized photographs and produces a single-channel visual stream consisting of pixel-based transformation of one image into another. It is thus a continuous flux which has to be appreciated as a digital art video rather than a slideshow.

Basic Idea

The main novelty in our approach is to use a Cellular Automaton (CA) to perform the transition effect in a smooth and seamless manner, mixing thus two or more images together. CA have been defined in the 50's by Von Neumann (von Neuman 1966) and have gradually gained success in various fields, see for instance (Wolfram 2002). In its basic form, a CA consists in a 2D matrix where cells can evolve over time according to a set of predefined rules and to the state of their neighbourhood (presence/absence of other cells). In the following, we will use the so-called *Moore-neighbourhood*, consisting of the 9 cells adjacent to a cell in the 2D matrix

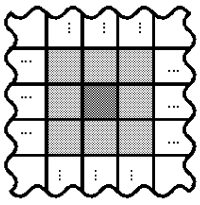


Figure 3: Moore-neighbourhood of a cell

There are two main components in our method for the transition / mixing effect between images:

1. Create "random seeds" in the current image
   That is, change a certain percentage of the pixels of the current image by pixels from next image
2. Apply Cellular Automaton rules
   This will result in extending the seed pixels to their neighbours, and thus propagate the transition to the next image for the seed pixels to the overall display.

Algorithm

The core loop of the algorithm can be stated easily below. It consists in a pixel-by-pixel replacement of the elements of the current image by the corresponding ones in a new image. However, the order in which these pixels are modified is of course not linear, but dynamic and based on the use of a Cellular Automaton. The algorithm can be tuned using the following parameters:

- *Step* : the number of seed pixels to be generated at each loop. Seed pixels are pixels from the next image that replace corresponding pixels in the currently depicted image. In general *Step* will be set to 1% of the total number of pixels of the image.
- *NChanged* : the percentage of pixels that have to be changed in the current image before introducing (mixing) another new image. In general *NChanged* will be set between 50% and 100%.
- *Trigger* : the number of pixels in the neighbourhood of a given cell that have to belong to the new image in order to convert this pixel - that is, to replace it by the corresponding pixel in the next image.

Core Algorithm:

```
img1 = load image from database
img2 = load another image from database
loop
/* 1. Generate seed pixels randomly */
for  k=0 to Step
    choose randomly coordinates (i,j);
        replace pixel (i,j) in img1
                by pixel (i,j) of img2;

/* 2. apply CA rules  */
for each pixel (i,j)of img1
    if  (number of pixels from img2 in neighbourhood of (i,j) ) > Trigger
        replace pixel (i,j) from img1
                by pixel (i,j) from img2

/* 3. load new image if  current transition finished */
if ( NChanged pixels changed since img1 is loaded)
   img1 = img2;
       img2 = load new image from database;
```

Parameter tuning

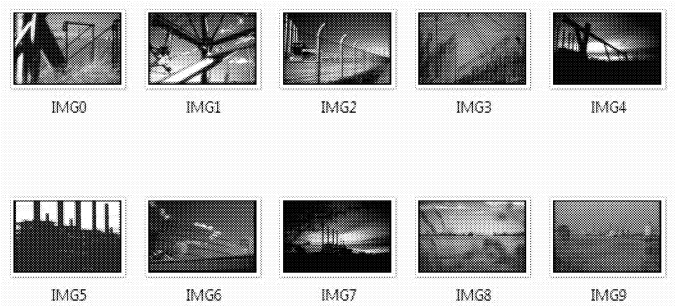
As seen in the algorithm above, there are three main parameters that could and indeed should be adjusted in order to produce interesting artistic effects. Section 3 will show the different types of effect that can be achieved.

Implementation

We have implemented the above approach in the Processing programming language (Reas and Fry 2007). Processing is an open source programming environment based on Java and aimed at easy manipulation of audiovisual objects. The language is simple and limited (as geared towards designers and artists rather than programmers) but is in fact open to the full power of Java language if needed. It is compiled to Java byte code and produces standalone executables.

EXPERIMENTS

We applied the *Palimpsest* system to a series of 10 photographs representing industrial landscapes, depicted below

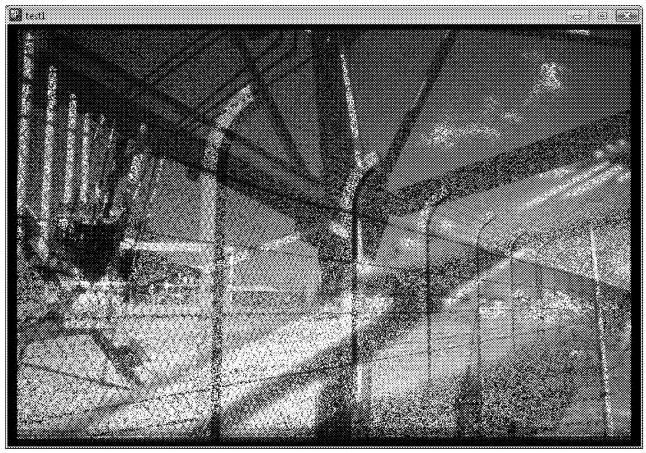


In the following experiments, we always ordered the photographs in the same sequential manner rather that choosing randomly in the database, in order to better show the different transition effects.

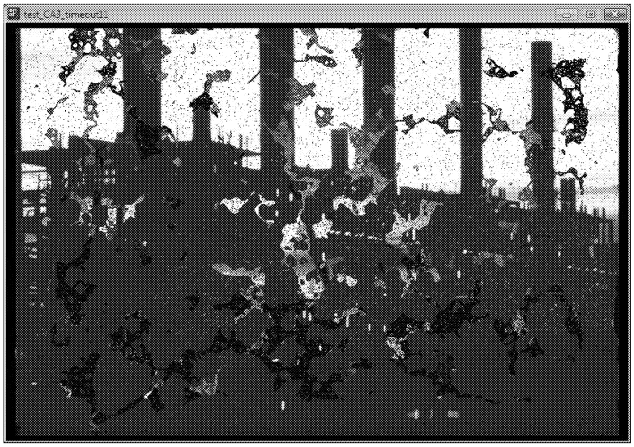
Also in the following examples *Step* is always set to 1%.

No CA

If we do not apply the CA rules and thus mix images by using only random pixels:



CA with *Trigger* =3, *NChanged* =0.7



CA with *Trigger* = 4, NChanged = 0.5

CONCLUSION

Palimpsest is a simple tool for transforming a static set of photographs into a mesmerizing dynamic flux of combined images that is better appreciated as digital video art rather than a simple slideshow. The algorithm is based on random seed pixels and cellular automaton rules to perform the transition between images. It is worth noticing that a simple real-time algorithm can produce interesting artistic effects. A key point however is the tuning of the parameters, that can engender a large variety of effects.

REFERENCES

CHEN J-C., CHU W-T., KUO J-H., WENG C-Y., WU J-L.: Audiovisual Slideshow: present your journey by photos. In *Proc. MM'06, 14th ACM International Conference on Multimedia*, ACM Press 2006, 955-956.

CHEN J-C., CHU W-T., WU J-L.: Tiling Slideshow: an audiovisual presentation method for consumer photos. *IEEE Multimedia,* 14, 3 (Jul-Sep 2007), 36-45.

GREENBERG C.: The Crisis of the Easel Picture. *Art and Culture Critical essays*, Beacon Press, 1961.

FLECKLES D.: Picturing Systems. *IEEE Multimedia,* 13, 2 (Apr-Jun 2006), 8-14.

KANIZSA G.: *Margini quasi-percettivi in campi con stimulazione omogenea. Rivista di Psychologia 49, 1, 1955.*

LI C-T., SHAN M-K.: Emotion-based impressionism slideshow with automatic music accompaniment. In *Proc. MM'07, 15th ACM International Conference on Multimedia*, ACM Press 2007, 839-842.

VON NEUMANN J.: *The Theory of Self-reproducing Automata,* edited by A. Burke, University of Illinois Press 1966.

REAS C., FRY B.: *Processing: A Programming Handbook for Visual Designers and Artists,* MIT Press 2007.

WOLFFRAM S.: History of Cellular Automata, In *A new Kind of Science,* Wolfram Media Inc, 2002

Efficiency of peer-to-peer overlays for content distribution

Gerhard Haßlinger [a], Halldór Matthías Sigurðsson [b], Úlfur Ron Halldórsson [c], Julian Schröder-Bernhardi [d]

[a] *T-Systems, Deutsche Telekom Allee 7, D-64295 Darmstadt, Germany*
[b] *Center for Information & Communication Technologies, Danish Technical University, 2800 Lyngby, Denmark*
[c] *Iceland Telecom R & D, Ármúla 25, 150 Reykjavik, Iceland*
[d] *Darmstadt University of Technology, D-64289 Darmstadt, Germany*

Abstract - **Peer-to-peer (P2P) networks have introduced a distributed communication paradigm, which is useful for many communication services. While P2P based file sharing pushes the Internet traffic growth, P2P approaches for VoIP and video streaming also have established popular golbal scale networks.**

We investigate P2P content delivery schemes to get insight into the structures and features required to distribute data volumes over a large community at about the same speed as for a single end-to-end data transfer. In addition, we briefly discuss alternative content delivery overlays with regard to the efficiency of transfer paths to avoid unnecessary traffic load and overhead.

I. INTRODUCTION: SPECTRUM OF PEER-TO-PEER APPLICATIONS

The Internet has seen a proliferation of multimedia content in the past decade indicating a paradigm shift away from text based HTML pages and images towards more resource demanding multimedia content [16]. The ongoing widespread deployment of broadband access to the Internet [4][11] has enabled a wide range of multimedia applications for residential users including audio/video content and TV. The consumption and demand for content has been reinforced by increasing access capacities for delivering high quality continuous media, although there are still limitations and demands for even higher bandwidths, since HDTV video transfers are in the range of Gbit/s per stream for optimum quality.

A controversial driver for this development are content distribution systems based on peer-to-peer overlay networks, which pose challenges to all actors in the value chain. P2P networking has shown its potential of increasing the efficiency and unleashing idle resources to form scalable and resilient overlay networks. Today, the potentials of peer-to-peer networks are reflected through hugely successful file sharing [5] and VoIP [1][15] applications which can offer stability, redundancy and scalability at a fraction of the cost of traditional server based approaches. Overlays on the end systems of the users give new opportunities to offer services worldwide at a minimum of own network and server infrastructure.

Peer-to-peer networks now stand at the brim of an evolution into a wide spectrum of mainstream applications. Content distribution in the convergent Internet-TV environment are a promising trend where video streaming and controlled P2P solutions are launched [2]. Application types that are most likely to benefit from peer-to-peer systems are gaming, eLearning, a support of smaller communities or enterprises as well as extensions into mobile and wireless networks:

- Online gaming and eLearning
  In both cases, there are similar demands to distribute and update large software packages and to support online activity in multi user games and interactive learning frameworks. The BitTorrent file sharing protocol is used for software downloads e.g. on the Blizzard gaming platform <www.blizzard.co.uk/wow/faq/bittorrent.shtml>.
- Support for small communities and enterprises
  P2P overlays can give access to content and multimedia services according to the specific demands of a community or company. Therefore secured and authorized access within closed user groups can be set up if required.
- Wireless and mobile networks
  Fixed mobile convergence is a challenging environment for P2P systems. Links in mobile and wireless networks have smaller bandwidths and are less reliable. There is a higher churn rate due to changing access conditions of the nodes. Nevertheless, P2P networks have been developed to adapt to unreliable and variable network conditions and provide a promising alternative for ad hoc and self-organizing networks.
- Internet-TV, streaming and multicast services
  Regarding future directions in the convergence of TV broadcasting with Internet applications, the BBC has launched an integrated media player [2]. They used a peer-to-peer networking approach to make a part of their TV and radio program available for watching and viewing for seven days after the broadcast transmission date including software for digital rights management and enforcement. A currently started EU project also focuses on P2P broadcasting technology <www.p2p-next.org>.

The volume of popular unlicensed multimedia content on the Internet is also increasing. There is a variety of providers of open platforms for special purpose with content of various types. P2P networks again provide an alternative to make this information available worldwide in competition to other architectures, e.g., YouTube .

In this new role, peer-to-peer approaches must compete on even grounds with other methods of content distribution. Viability and usability of peer-to-peer networks is measured through technological and economical efficiency and competitiveness. Therefore we examine the opportunities and challenges of peer-to-peer content distribution schemes as compared to alternative overlays including content delivery networks (CDN), which are build in the infrastructure of network providers [17].

In section II we analyze the achievable delivery time of content over large scale P2P networks as a performance indicator for scalability. In Section III we compare the efficiency of content distribution via P2P and other approaches.

II. EFFICIENCY OF P2P CONTENT DISTRIBUTION

We investigate P2P content delivery schemes in more detail to get insight into the structures and features required to distribute data volumes over a large community in about the same time as for a single end-to-end data transfer. Content distribution has developed to include many multimedia applications with prevalent one way transmission from web browsing to video streaming. Initially, we make some simplifying assumptions focusing on a homogeneous overlay network with the same bandwidth B being available at each peer.

We presume unique end-to-end connectivity between peers without considering the underlying network structure in detail. The bandwidth B is bi-directional and can be independently used for up- and downloading in parallel. In fact, ADSL broadband access is prevailingly deployed with asymmetrical up- and downstream speed, but since current P2P protocols enforce a symmetrical traffic in both directions [5][8], we consider the upstream access capacity as the bottleneck for both directions.

At the beginning, the content of size C is located on and provided by a source as one of the peers. From the start of the content distribution, we include N peers who want to get the content and we presume that they build a cooperative P2P subnet to forward the content to each peer. As usual in P2P protocols [3][5][7], we assume that the content of size C is subdivided into K data chunks of equal size. The number K of such data units is an important parameter for the performance of delivery, since the subdivision into small data portions enables flexible parallel transfers.

Once a peer has completely received a chunk, it can start to forward it to other peers. Upload of received data is enforced as soon as possible by current file sharing protocols [5]. The transmission time for a data chunk is $(C / K + P) / B$ with regard to some overhead P for additional information e.g. a hash identifier is usually added for unique identification of each chunk. The overhead P may also cover connection setup times or other delays occurring between successive chunk transfers.

The following analysis is done for an unrestricted full mesh of connections between $N = 2^M$ peers. The source transmits the complete series of chunks one by one choosing different destinations. After the first chunk has been received by a peer, this peer join to distribute it in further transfer steps in parallel to transfer from the source and others. In a first phase, chunks are always transmitted to peers who still haven't got a chunk. In this way, the number of peers that have received a chunk is growing to 2^m after m chunk transmission steps.

Figure 1 illustrates the progress of the distribution after m chunk transfer times for an example with 2^4 peers. For $N = 2^M$, the first chunk is completely distributed after $M + 1$ steps. At the same time, the second chunk is already available at half of the peers, while the $(M+1)$-th chunk has just been delivered once from the source to a peer. In general, the delivery of chunk k is completed in step $M+k$, when again the $(M+k)$-th chunk is just transferred once from the source.

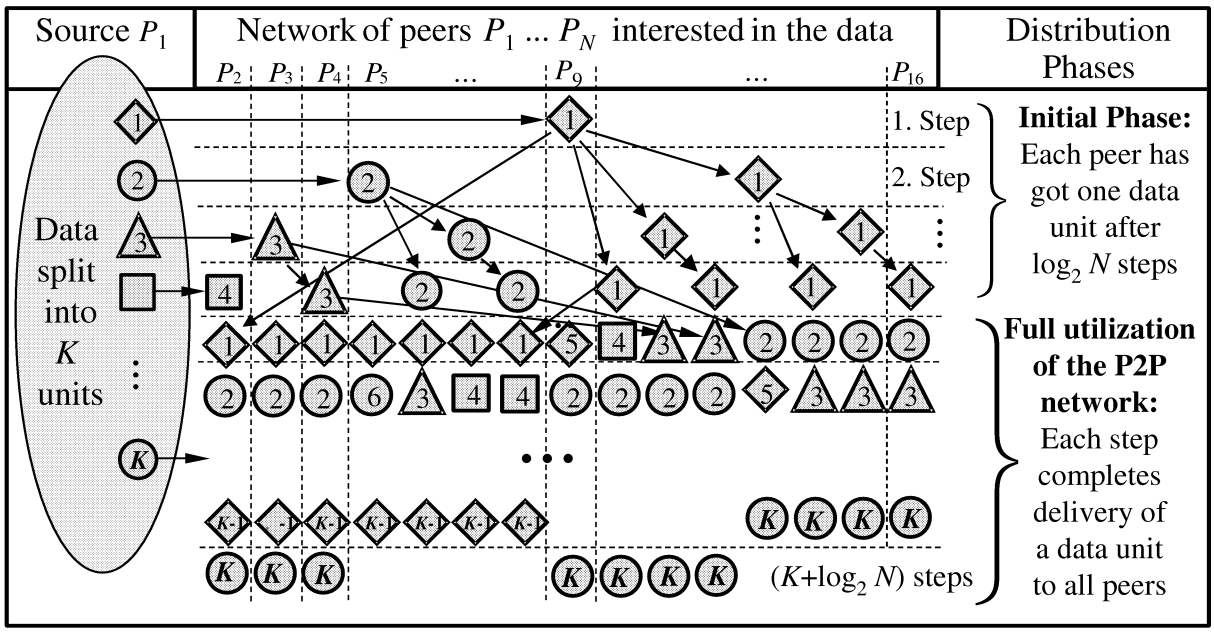


Figure 1: Scenario demonstrating the efficiency of peer-to-peer content distribution

The delivery process can be made periodical after the $2M$-th step, such that the $(j+M)$-th chunk is transmitted over the same peer-to-peer connection as the j-th chunk has been M steps ahead. There are many appropriate schemes for choosing the destinations for delivery in each step as suggested in Figure 1, which have to ensure that a chunk isn't transferred twice to the same destination.

Thus the complete content is distributed to all peers after $M+K$ chunk transfer times. As soon as a peer has received at least one chunk, it is assumed to upload a chunk in each step. Thus the content can be forwarded from a source within $M+K$ chunk transfer times to $N = 2^M$ other peers in a full mesh P2P network. It can be shown that the same delivery time T is sufficient for $2^{(M-1)} < N \le 2^M$ and we can conclude

$$T = (\lceil \log_2(N) \rceil + K)\,(C/K + P)/B$$
$$= [1 + M/K + (M+K)\,P/C]\,C/B$$

where $(C/K + P)/B$ is the time for one chunk transfer step and $\lceil \log_2(N) \rceil + K = M + K$ is the number of steps.

Usually, we can assume P/C to be small, since a few header or trailer bytes including a hash value should be sufficient as overhead P, whereas the content C is often much larger in the MB or GB range. In addition, M can be bounded in large P2P networks by $M \le 25$ even for $N = 2^{25} > 33$ million peers. From the preconditions $M \le 25$ and $P/C < 3\times10^{-4}$ we conclude that $K = 250$ chunks are always sufficient to obtain $M/K \le 0.1$ and $(M+K)\,P/C \le 0.1$.and finally we have $T < 1.2\,C/B$. Thus the delivery time T of content over a large P2P network can be kept close to the transfer time C/B to a single destination.

A. Comparison to analytic work including delivery via a tree or linear chain

A more detailed performance analysis is given by [3] including heterogeneous peers. In addition to transmission over a full mesh, cases of a delivery over a tree structure and in a linear chain are also investigated. Since no overhead per data chunk is considered in [3], their results are in accordance with the previous analysis for $P = 0$.

Comparing content distribution via chain, tree and a full mesh structures for homogeneous peers, the mesh achieves a close to optimum performance under fairly general conditions for scalable P2P network size. The restriction to a tree suffers from the fact, that the upload capacity of the leaf nodes do participate in uploading. In a regular tree structure with fan-degree h only a fraction less than $1/h$ of the network nodes can forward content. Therefore the speed of delivery is slowing down by about the same factor $1/h$.

In all cases, a subdivision of the content into small chunks essentially reduces the delay. The performance of the linear chain is essentially affected by additional overhead or delays per chunk and, in addition, depends on the number of chunks, which has to be optimized and chosen depending on the content size. For the performance of the tree and mesh distribution, the overhead per chunk is less important and a moderate number of chunks $K = 250$ is always sufficient.

With regard to fault tolerance, an unrestricted P2P mesh offers alternative paths to bypass failed links and nodes, whereas a linear or tree structure offers only a single source to destination path. In fact, only a part of the full mesh connections is required for content distribution. In each transfer step, N parallel connections are utilized. Since the delivery process can be made periodic after $\lceil \log_2(N) \rceil$ steps, only $N \lceil \log_2(N) \rceil$ out of $N\,(N-1)$ peer-to-peer connections are required, i.e. a partial mesh of no more than $N \lceil \log_2(N) \rceil$ connections already achieves optimum performance without failures.

B. Discussion of content delivery in generalized environment of realistic scenarios for current P2P protocols

In fact, the previous analysis makes some idealistic assumptions leading to a deterministic and synchronized distribution in a predefined schedule. Popular peer-to-peer networks have to adapt to different delays per data transfer depending on the length of transmission paths between peers in the underlying network. A source decides to upload data chunks based on incoming requests and a local prioritization scheme rather than on a network wide optimized schedule. Nevertheless, the self-organizing distribution of current P2P protocols is flexible enough to closely approach the optimum performance of the analyzed deterministic scheme.

After the initial phase, the delivery in a mesh structure exhausts almost the complete access bandwidth of a homogeneous peer-to-peer network. Therefore only the first M steps could be further improved. In the initial phase, it would be helpful to have a source or peers with larger upload speed involved. A powerful source peer with 2^m-fold bandwidth can distribute 2^m different chunks to 2^m peers in parallel in a single step and thus would skip $m-1$ steps, although only a limited fraction $m/(M+K)$ of time can be saved in this way.

Multiple uploads from the same source in parallel do not improve the performance for homogeneous peers. Simultaneous uploads with bandwidth B/n shared among n chunks of equal size C are all completed after a delay of nC/B. For sequential uploads with bandwidth B, chunks are delivered earlier after C/B, $2C/B$, $3C/B$, …, which is favourable in the starting phase to supply all peers with a chunk and to make their upload capacity available as fast as possible.

However, multiple up- and downloads in parallel are essential in communication between heterogeneous peers equipped with different access speeds. When many peers have a unique basic access speed and there are some peers with higher, e.g. n-fold speed, then it is easy to integrate them by subdividing their speed into n up- and downloads at unique speed in parallel. With a single transmission per peer, it would be difficult to fully utilize peers at different bandwidths due to the restrictions for transmitting between pairs of almost equal speed.

An extension of the performance bounds and estimations to heterogeneous P2P networks is more complex, but surely relevant, since different access speeds are offered for residential users and some very powerful hosts are often involved in P2P networks. Although multi source downloads are appropriate to adapt to the heterogeneity, the delivery time for content of size C starting from a source with bandwidth B_{Source}, is obviously limited by C/B_{Source} as a bottleneck at the source. When the data is distributed over many peers being available in the P2P network, then the access speed of most powerful hosts among the peers can be exploited.

More generally, we may distinguish classes of peers at different access speed but with the same speed per class. Then a class of high speed peers including the source can be viewed as a homogeneous subnet, whose delivery time is determined by the access speed of this class, even if the complete P2P network includes many peers of lower speed. In addition, the peers with lower access speed can profit from faster peers, especially in case of asymmetrical access (ADSL).

On the whole, bounds on the delivery time to all peers can be set up in terms of the mean upload speed per peer, since the upstream is assumed as the bottleneck in asymmetrical access, where classes of peers with higher speed may finish a download earlier than peers with smaller access bandwidth. From the view of each peer, the access speed of the peer can impose a bottleneck, as well as the upload speed of the source, if the content is only available from a source and not yet distributed over the network.

Involving more peers from outside of the subnet also would not improve the main distribution phase, since the bandwidth of the peers is already exhausted by the distribution scheme as suggested by Figure 1 within the subnet of all peers interested in the content. For asymmetrical ADSL access with a ratio R of download ÷ upload capacity, the situation is different. Then the download capacity can only be exploited by R peers uploading in parallel to the same destination.

Solutions for efficient content distribution apply in a similar way to non-real time as well as real-time applications, provided that real time data, e.g. a video stream is again subdivided into data chunks of suitable size.

III. TRANSPORT PATHS FOR CONTENT DELIVERY

A. Cross layer aspects

P2P protocols execute an application layer concept on the terminals of the users and are designed independent of the underlying network infrastructure. On the other hand, Figure 2 illustrates a usual topology of broadband access networks, where tree-shaped access areas are attached to edge routers of the backbone at points of presence (PoPs).

While overlay networks create new opportunities to launch services, an additional layer causes more overhead. P2P protocols establish their own routing schemes by selecting download sources via an implemented search scheme, including "hello" mechanisms to verify the availability of peers. When we compare the message exchange for downloads via the eDonkey and BitTorrent protocol, then we experience an overhead of 10-20% of the complete transferred data volume [14].

Another major performance issue is the length of paths for data transfers, i.e. the distance between source and destination. In P2P networks, frequently referenced data is replicated manifold according to the demands and becomes available from many sources. So, there is a potential to shorten the transfer paths by preferring the nearest sources. In this way, transmission delays and thus the user perceived QoS can be improved as well as the resource usage of network providers, avoiding unnecessary long paths through the backbone or on expensive peering and intercontinental links. Considering large provider networks serving millions of subscribers, it can be expected that a majority of the data of a global file sharing network can already be found to be replicated on the same ISP platform and often in the same access region.

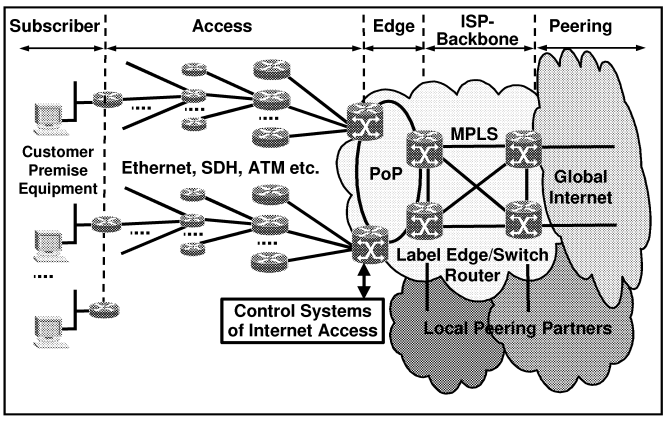


Figure 2: Structure of broadband Internet access platforms

This especially holds for the most popular and referenced data, since the major portion of downloads addresses a small set of the currently most popular files [14]. Consequently, an essential portion of data transfers could be handled at the edge of the platform without running through the backbone.

But peer-to-peer networks have no direct knowledge of the underlying IP network structure. The fact that the response time and availability of the peer nodes are related to the length of transmission paths on the network layer gives some preference for shorter transfer paths. But this effect is less relevant than the structure of communities separated by language and other social factors. The distribution of source locations offered by the eDonkey network for downloading German and English content has been investigated from a destination in Germany [14] with results shown in Figure 5. It is not surprising, that 78% of the sources for German content were also located in Germany, whereas most (83%) of the English content is transferred from peers from other countries, which puts load on backbone and peering links of the network providers. Similar experience for locality of sources due to language has been made for file-sharing in France [7]. In the BitTorrent network, source selection depends on the trackers, which control downloads of each file.

B. Transport path optimization in CDNs and caching

Content delivery networks (CDNs) provide an alternative overlay structure within the IP backbones of service providers as depicted in Figure 3. After a user contacts a web site, which is supported by a CDN, his request is redirected via a CDN overlay to a server in the near of the user.

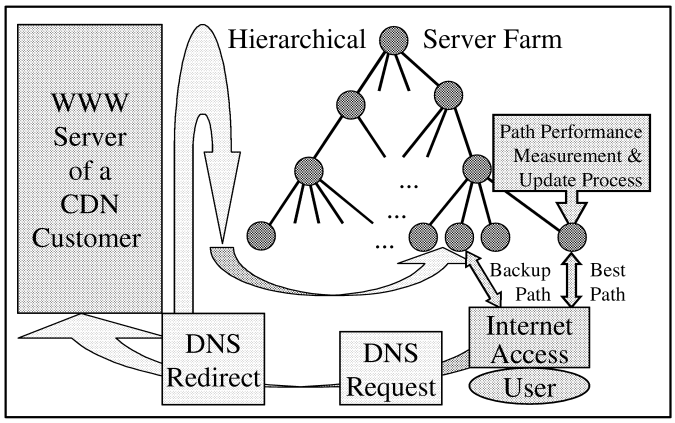


Figure 3: Transfer paths via content distribution networks

A performance study on Akamai's CDN [17] shows that a large hierarchical server farm is efficiently utilized to shorten the transmission paths. The solution includes frequent updating of the best path to a user based on measurement of the network and server performance. A backup path to another server is also provided for failure resilience. In this way, CDNs offer advantages as compared to P2P networks, concerning optimized reliability and low delay as main QoS measures.

Web caches provide another opportunity to optimize traffic paths. In principle, caches for P2P data are efficient since a high hit rate can be expected and problems of outdated data in classical web caches are avoided because the complete content and single data chunks are uniquely identified by hash values. Problems with illegal distribution of content through caches persist and are increased by current file sharing habits.

Since October 2004 the eDonkey network offers an option to use web caches of network providers by disguising P2P downloads as HTTP requests. In fact, it has been observed that data can be downloaded from caches, but only a small portion of less than 10% [14]. From the network provider's view, the caching option for support of P2P networks remains ambiguous. While transfers from the cache take load away from the backbone and reduce off-net traffic, other downloads may continue without caching and may generate a higher total P2P traffic load.

Last not least, network providers have opportunities to build content distribution architectures using their own infrastructure, e.g., via multicast for support of broadcast services.

SUMMARY AND CONCLUSIONS

Peer-to-peer networks provide an alternative for launching new services in a wide spectrum of mainstream applications. They are efficient for distributing large amounts of data in a scaleable and adaptable way. In meshed networks with sufficient connectivity degree the access bandwidth of the peers can be fully utilized after a short initial phase with scalable support for huge communities. Then the delivery time can be kept close to a single content transfer time for homogeneous peers. P2P protocols have a smoothing effect on the variability of the traffic in short time scales as well as in the daily profiles on broadband access platforms. These properties partly facilitate network planning, but the analysis of the P2P traffic and the prediction of its future development becomes difficult.

Transmission paths in P2P overlays depend on the protocol and on user communities. In comparison to content delivery networks, there is potential for more efficient download paths, which is essential for lower delay in real time applications and to reduce the traffic load in the backbone. The resource efficiency and suitability for content and network providers and last not least for the users will decide whether future Internet services will use P2P or other network infrastructure.

REFERENCES

[1] S. Baset and H. Schulzrinne, An analysis of the Skype peer-to-peer Internet telephony protocol, IEEE INFOCOM Conf. (2006) <www1.cs.columbia.edu/~salman/publications/skype1_4.pdf>

[2] BBC, Integrated media player (iMP) <www.bbc.co.uk/imp/> (2006)

[3] E. Biersack, D. Carra, R. Lo Cigno, P. Rodriguez and P. Felber, Overlay architectures for file distribution: Fundamental analysis for homogeneous & heterogeneous cases, Computer Networks 51 (2007) 901-917

[4] K. Cho, K. Fukuda, H. Esaki, A. Kato, The impact and implications of the growth in residential user-to-user traffic, ACM SIGCOMM Conf., Pisa, Italy (2006)

[5] B. Cohen, Incentives build robustness in BitTorrent, <http://bitconjurer.org/BitTorrent/bittorrentecon .pdf> (2003)

[6] C. Gkantsidis and P. Rodriguez, Network coding for large scale content distribution, IEEE INFOCOM Conf. (2005)

[7] G. Haßlinger, F. Guillemin, J. Ferreira and U. Halldorsson, The impact of peer-to-peer networking on network operators and Internet service providers, Eurescom study report P1553 (2005) <www.eurescom.de/public/projects/P1500-series/p1553/>

[8] G. Haßlinger, ISP Platforms under a heavy peer-to-peer workload, In: R. Steinmetz and K. Wehrle (eds.): Peer-to-Peer Systems and Applications, Springer LNCS 3485 (2005) 369-382

[9] M.M. Hefeeda, A framework for cost-effective peer-to-peer content distribution, Ph.D. Disertation, Purdue University (2004)

[10] T. Karagiannis, A. Broido, M. Faloutsos and K. Claffy, Transport layer identification of P2P traffic, Internet Measurement Conf., Taormina, Italy (2004) <www.imconf.net/imc-2004/papers/p121-karagiannis.pdf>

[11] A. Odlyzko, Internet traffic growth: Sources and implications, Proc. SPIE Vol. 5247 (2003) 1–15

[12] H. Sigurdsson, Peer-to-Peer aided streaming in a future multimedia framework, CICT Working Paper, no. 104, Lyngby (2005) <http://www.viskan.net/phd/publications.aspx>

[13] H. Sigurdsson U. Halldórsson and G. Haßlinger, Potentials and challenges of peer-to-peer based content distribution, Telematics and Informatics, Elsevier, Vol. 24 (2007) 348-365

[14] J. Schröder-Bernhardi, Analysis of the communication and traffic in P2P networks including web caches, Diploma/Masters Thesis, Report KOM-D-260, Darmstadt Univ. of Tech. (2006)

[15] Skype, 2006. Webpage: <http://www.skype.com/>

[16] K. Sripanidkulchai A. Ganjam, B. Maggs and H. Zhang, The feasibility of supporting large-scale live streaming applications with dynamic application end-points, ACM SIGCOMM'04, Portland, USA (2004)

[17] A.-J. Su, D.R. Choffnes, A. Kuzmanovic and F.E. Bustamante, Drafting behind Akamai, ACM SIGCOMM, Pisa, Italy (2006)

[18] K. Tutschku and P. Tran-Gia, Traffic characteristics and performance evaluation of P2P systems, Springer LNCS 3485 (2005) 383-398

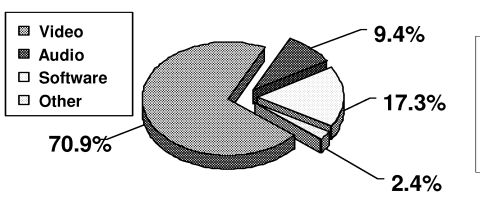


Figure 4: Types of content observed in eDonkey

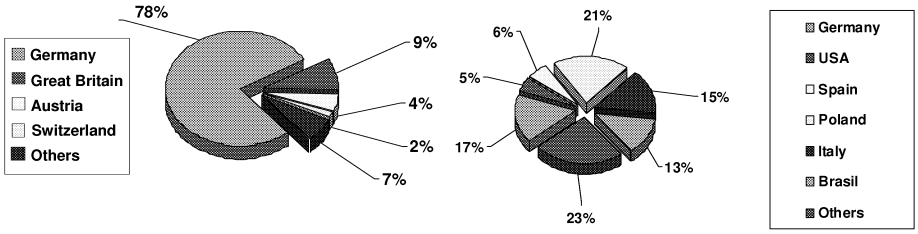


Figure 5: Distribution of sources for German and English eDonkey content

MODELLING GROUPING PRESSURES FOR EMERGENT AND SELF-ORGANIZING VISUAL PERCEPTION

J.C. Stevens(*)(**)
R. Dor(*)
Th.M. Hupkens(**)
L.J.M. Rothkrantz(*)

Man-Machine-Interaction Group, Delft University of Technology(*)
Mekelweg 4, 2628 CD Delft, The Netherlands
{j.c.stevens, r.dor, l.j.m.rothkrantz}@tudelft.nl

Netherlands Defence Academy (NLDA) (**)
P.O. Box 10000, 1780 CA Den Helder
Thm.hupkens@nlda.nl

KEYWORDS
Dynamic grouping, Emergent systems, Gestalt theory, Grouping pressures, Perceptual organization, Self-organization, Visual perception.

ABSTRACT

A lot of psychophysical and neurological evidence in primitive visual perception suggest that perception deploys emergent mechanisms for dynamic grouping with competition and cooperation of local grouping pressures, rather than `mathematical' bottom-up image-segmentation. We propose a computational model for primitive visual perception based on the general underlying theory of implementing perception by self-organization. By implementing several grouping pressures in an emergent architecture, we will demonstrate their importance and necessity for a computational model of visual perception.

INTRODUCTION

Traditionally, computational models of perception split human perception mechanisms into low and high level perception (Chalmers et al. 1992) regardless of the fact that all neuropsychological evidence points to mechanisms of deeply intertwined levels (Dor 2005). Such approaches tend to equate low-level perception with the primitive processing of the incoming data closer to the sense organs and high-level perception with mechanisms involving mental concepts such as object recognition and understanding. Most work to date in perception (mostly in the field of visual perception) has been targeted at either bottom-up processing (Viola and Jones 2001, Ullman 1996) or higher semantic levels (Mojsilovic and Gomes 2002). The main challenge for future models of perception is the integration of such top-down influences with bottom-up processing (Riesenhuber and Poggio 2000).

In the current paper we argue that an approach of modelling perception by means of emergence and self-organization is (neurologically) plausible, based on the fact that a lot of psychophysical and neurological evidence in primitive visual (and auditory) perception suggest that perception deploys emergent mechanisms for dynamic grouping with competition and cooperation of local grouping pressures (the actual processes that try to force a particular grouping onto incoming visual data), rather than `mathematical' bottom-up image-segmentation. Such interaction between pressures at the lowest level of perception give rise to emergent coherent structures (objects), which are novel with respect to the individual cues (pixels or tones). The reason why interaction among grouping pressures is such a key ingredient arises from the necessity for dealing with the contradictory and incomplete set of cues present at any real-world input caused, among other things, by occlusions, distortions, and reflections. By letting these pressures actively push each other with no centralized interference, structures may emerge, which amounts to a reconstruction of the shared fate of the constituent elements. In the school of Gestalt theory, Gestalt psychologists and their psychophysical and psychoacoustic experiments provide a theoretical framework with perceptual laws of organization, or in other words Gestalt grouping pressures (see Figure 1 for an illustration of Gestalt principles) with which they emphasize the interaction of the parts and the organizational process as a dynamic process (Wertheimer 1923).

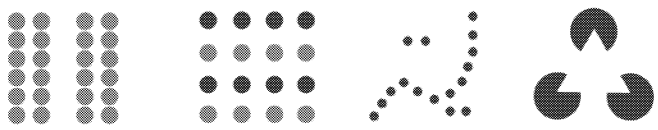


Figure 1: Gestalt principles from left to right, proximity, similarity, good continuation and closure (Kanizsa 1980)

Gestalt theorists often describe perception as a self-organizing system that spontaneously takes on the 'best' or simplest arrangement in given conditions. During the process of self-organizing perception Gestalts (organized wholes)

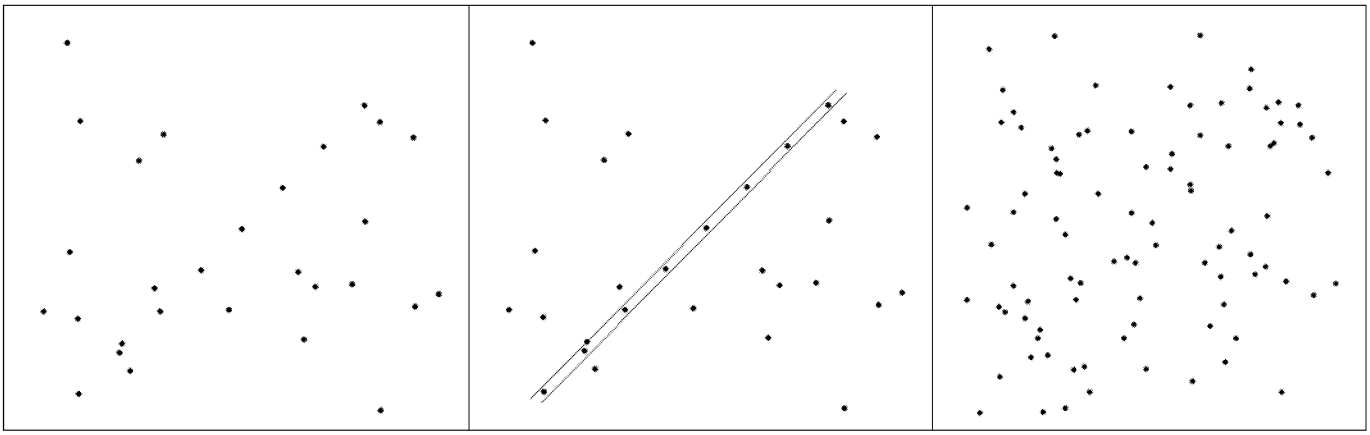


Figure 2*: Left: 20 uniformly randomly distributed dots, and 8 aligned added. Middle: this meaningful (and visible) alignment is detected as a large deviation. Right: same alignment added to 80 random dots. The alignment is no more meaningful (and no longer visible). In order to be meaningful, it would need to contain at least 11 points (Examples taken from Desolneux 2003)*

emerge from the data gathered by our sense organs. There is some evidence that the laws of gestalt have a physiological basis. In the area of Artificial Neural Networks and Brain computing special ANN-architectures and algorithms have been designed to recognize patterns (Hecht-Nielsen 1989, Cornet and Rothkrantz 2003).

Researchers have designed computational models for several Gestalt principles separately, e.g. good continuation, closure and organized contours (Desolneux 2003). However many have taken a strictly mathematical bottom-up approach, and computed absolute thresholds of meaningful groupings, where they neglected the crucial top-down (contextual) pressures and dynamic interaction among grouping pressures. Grouping pressures (like the Gestalt laws) on their own do not create strong coherent structures. Instead only those supported by an abundance of supportive evidence by other grouping pressures constitute coherent groupings. Take for example the left dot-pattern in Figure 2, which contains a visible (meaningful) alignment. Here seeing the alignment would not be the result of a single 'line-detection' grouping pressure, but is the result of a myriad of grouping pressures that interact and exploit redundant information. Different grouping pressures that propose similar groupings, provide more (redundant) evidence for a coherent strong structure. Examples of such grouping pressures are proximity, good-continuation, regular-orientation and regular-distance. Alternatively (bottom-up) mathematical line detection algorithms could quite easily find the same line in the left dot-pattern of Figure 2. However they would also still find the same line when more random points are added to the same example, even when we would no longer see the alignment (see the right example in Figure 2).

This paper continues on our ongoing work on audiovisual fusion (Stevens 2007), where in this paper we will go into more detail on the visual perception side of our framework and present our preliminary results. We propose a computational model for primitive visual perception based on the general underlying theory of implementing perception by self-organization, which is founded on "The Ear's Mind", an architecture that supports emergent processes, self-organization, and context sensitivity, for the primitive perception of sound, which represents also the auditory part in our fusion model (Dor, 2005). By implementing several visual grouping pressures that utilize the emergent architecture, we will demonstrate their importance and necessity for a computational model of visual perception. Our preliminary results agree with expected human visual grouping behaviour and support our ongoing work on audiovisual fusion.

The paper is organized as follows. First we start with 'Copycat and The Ear's Mind: emergent self-organization' – in which we describe the auditory perception architecture on which we based our visual perception architecture together with the Copycat model (Mitchell 1993). Thereafter we present our proposed vision model and implementation of visual grouping pressures and their grouping results on a visual example. We end the paper with our plans for the future and a conclusion.

COPYCAT AND THE EAR'S MIND: EMERGENT SELF-ORGANIZATION

The Ear's Mind theory, offers a general architecture for simulating emergent sensory perception and specifically for the primitive segregation of auditory scenes (Dor 2005). The model of The Ear's Mind was inspired by the 'Copycat' model (Mitchell 1993, Hofstadter and FARG 1995), which R. Dor abstracted from its original micro-domain and specific sort of analogy-making paradigm. This enabled R. Dor to support the implementation of emergent, context sensitive processes in general, and (human) perception mechanisms in particular.

Copycat

The Copycat computer program (Mitchell 1993) models the mechanisms of analogy-making in a letter-string micro-domain. Together with other models (e.g., Seek-Whence (Hofstadter and FARG 1995), which tackles linear patterns, Tabletop (French 1995), tailored for two-dimensional visual analogies, and Letter Spirit (Hofstadter and FARG 1995), which generates creative font variations), Copycat belongs to a lineage of stochastic sub-symbolic self-organizing cognitive models (Hofstadter 1995).

Copycat is based on the assertion that analogy-making at any given level (e.g., seeing two situations as 'the same' even when no one-to-one correspondences exist among their respective constituent elements) relies on emergent mechanisms at lower levels. Analogy is thus seen as an interpretation of a given arrangement or situation arrived at by lower level activities. Internal pressures at such levels stir the constituent parts (including both atomic features and groups thereof) to form higher-level coherent structures. Consequently, the Copycat implementation allows Sub cognitive pressures probabilistically influence the direction of processing. Both context-dependent and context-independent pressures make up a nondeterministic parallel architecture in both bottom-up and top-down directions. Moreover, since the resulting arrangements are not known in advance, such self-organizing systems could not be implemented by deterministic processes working at the same level of abstraction of the outcome itself. Instead, microscopic, nondeterministic, local processes interact with each other with no central control. The macroscopic outcome of such activities is emergent, rather than programmed.

The Ear's Mind

Based on Copycat, The Ear's Mind is designed to model the unconscious, automatic auditory grouping pressures in humans. Such pressures, it seems, steer the perception of sound by cooperative and competitive interactions, resulting in the grouping of sound elements into context-sensible entities. A software prototype, simulating the most basic functionality of the proposed model has already been implemented and presented with sound excerpts of standard psychoacoustic experiments (Bregman 1990, Bregman & Ahad 1990). Preliminary results agree with expected human auditory grouping behaviour. A fuller description of The Ear's Mind will be published in the Journal of Experimental and Theoretical Artificial Intelligence this year (Dor and Rothkrantz 2008). Though promising results have been reached so far, much effort still needs to be put into both theory and application aspects of emergent self-organising sensory perception. In a joint effort we contribute in this area by extending and tailoring the model for the visual domain.

EMERGENT VISUAL PERCEPTION MODEL

The emergent system works as a non-supervised collection of independent local primitive agents which represent and act as local grouping pressures (e.g., proximity and regularity) that will try to force a specific grouping onto the input. These agents compete and cooperate to build or destroy bridges in the data-landscape they work on, resulting in the creation of high-level structures out of low-level input. Different grouping pressures that propose similar groupings, provide more evidence for a coherent strong structure. We take the visual Gestalt laws of organization as our initial starting point for modelling several different grouping pressures, but we do not take only the Gestalt laws as an exhaustive list of possible pressures. Before delving into more detail on the grouping pressures we first turn to the overall architecture.

Architecture

The architecture, illustrated in Figure 3, is based on four major building blocks: the Pre-processor, Workspace, Agents and Slipnet.

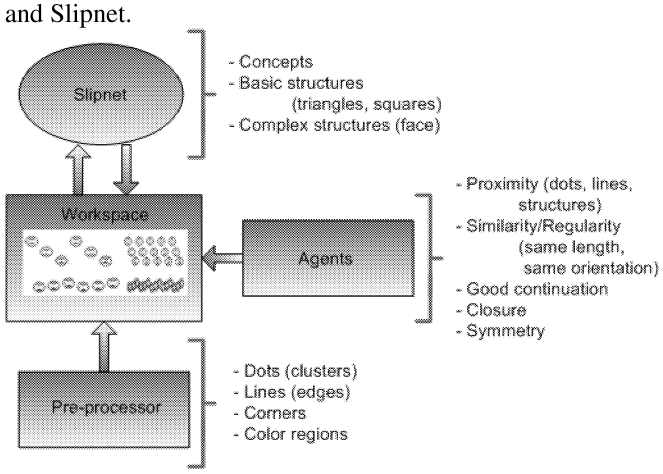


Figure 3: The architecture of the visual perception model

The *Pre-processor* analyses the input image and produces a list of salient cues, containing cue type, location and any other properties needed for a cue's definition. It is up to the pre-processor to fill the workspace with the most primitive cues and not with higher level interpretations or groupings of primitive cues. We do not propose an exhaustive list of primitive cues, but rather keep the option open to include more different cues as we go along. Currently we have only a single type of cue, namely dots, to study and model organizations of dot patterns. Later on for more complex input images we certainly need to resort to other primitive cues, e.g. density, gradients, colour and edges. One thing we have to keep in mind is that these artificially constructed dot patterns are in a way more difficult than real-life images, in the sense that in complex images we can find a lot more redundant information for building coherent structures.

The *Workspace* is where the actual construction is taking place with the building of perceptual structures on top of the primitive cues. When the workspace is filled with primitive cues, we are ready to start launching local agents. For now we implemented the following four different types of agents, which are described in more detail in the following sections: Proximity, Regular-orientation, Good-continuation and Regular-distance. Over time, through the actions of these agents, cues in the workspace gradually acquire various descriptions, and are linked (bonded) together by various perceptual structures. It is important to see that the strength of the architecture lies in the combination of multiple agents and their interaction, and not so much in any single agent. One imperfect agent that suggests a particular grouping that is in conflict with the grouping of a structure built by 20 other agents supporting each other, will by no means effect the overall good outcome of the system. Hence we do not try to build perfect agents, but we want to find and gather as much grouping evidence (pressures) as possible. Initially we randomly launch the agents on the workspace which means that the system works strictly in a bottom-up fashion. Later on the architecture provides the necessary top-down

influences to direct the launching of agents in an appropriate way and to focus on the most relevant cues and structures given the context of the input image. For example if we would not incorporate top-down pressures and in a particular case start and continue with a high portion of the agents being regular distance agents, one is bound to find regularity in the end, even though we might not perceive this regularity due to stronger structures in the context, which are not found because we only focused on finding regularity in the first place. Therefore we need to regulate the agents to be launched. If like in the previous case regularity is hard to be found, then less agents need to be launched to search for this type of grouping, especially when another structure based on non-regular evidence is being formed.

The *Slipnet* is responsible for the top-down influences, which is a network of interrelated concepts, where each concept is represented by a node and is surrounded by potential associations and slippages. Conceptual relationships represented as links have a numerical length, which resembles the `conceptual distance'. Conceptual links in the Slipnet adjust there lengths dynamically as the conceptual distances gradually change under the influence of the evolving structure in the workspace. In the Slipnet each of the concepts can become active when instances of them are noticed in the workspace. Also agents can provide feedback to the workspace by creating top-down pressure to look for further instances of themselves. Furthermore concepts can spread activation to their neighbours.

Building Bonds and Relations

On the Workspace we distinguish two types of bonds: Cue bonds and Relation bonds. The Cue bond is proposed between two cues, like in the middle example of Figure 4, where we have the three dot cues, from the left example of Figure 4, and a bond represented by an arrow that starts in the most right dot-cue and points to the middle one. What the bond represents is a local view from the right cue stating that it groups together with the middle cue, based on the grouping pressure that proposed and built the bond. In our model 'Proximity' is an example of a grouping pressure that constructs such cue to cue bonds. Some other grouping pressures, like regular distances, are not so much cue to cue bondings, but more bonds among the distance between two cues and the distance between two other cues. If they have equal distances, then we can speak of regular distances. To express these groupings between two sets of cues we use the Relation bonds. For example in the right dots-pattern of Figure 4 we displayed a bond between two relations, where the two dashed lines between the first and second dot, and the second and third dot represent two relations, which are bonded by the grey pointing arrow. Distance and Orientation are the two relations we used in our model. Relations are just like bonds to be build on the workspace.

Figure 4: Left: three dot cues. Middle: Cue to Cue bond. Right: Relation to Relation bond

The actual proposing and building of bonds is split into work for two different types of agents: scouts and builders. Initially we launch 'Propose Bond'-scouts, which follow two rules:

1. Land on two or three cues.
2. IF fitting pressure description THEN Propose bond and post Bond Builder scout ELSE terminate.

Next when the Bond Builder is launched it acts as follows:

1. Check for resistance to bond.
2. IF NO resistance THEN build bond (and relations) ELSE fight: Resulting in either building or deleting the proposed bond.
3. IF proposed bond is built THEN post Bond Extender scout.

The Bond Extender scout on his turn does the following:

1. Lands on the bond to be extended.
2. Checks for extensions to propose new bonds.
3. IF proposing a new bond THEN post Bond Builder scout. ELSE terminate.

Now that we have explained the general bond building mechanism, we can move on to the specific grouping pressures.

GROUPING PRESSURES

For now we have implemented the following four grouping pressures, which are modelled after the Gestalt laws of visual perception: *Proximity, Regular-orientation, Good-continuation and Regular-distance*. It is important to remember that the strength of our model lies in the combination of multiple grouping pressures and their mutual supporting evidence, and not so much in any of them in isolation. Different grouping pressures that propose similar groupings, provide more (redundant) evidence for a coherent strong structure. Next we will describe each implemented grouping pressure in more detail including their grouping results on our the alignment example from Figure 2. After which we present all the grouping pressures together.

Proximity

The proximity scout proposes and builds bonds between cues based on the relative distance between cues. The purpose of the proximity agent is to bond each cue from a local perspective to other cues which are relatively the closest. When we land with our proximity scout on a dot (cue) we take this cue as the centre point of a circular search zone for which we make a list of all the cues within this zone. The diameter of the search-zone should be sufficiently large, not to fool ourselves with a fixed threshold for proximity. For cue we find in our search zone we calculate the distance to the centre cue, and use these distances to set up a probability for being a candidate for a proximity bond. The shorter the (Euclidean) distance the higher the chance the cue will be chosen to build a proximity bond. Strong proximity relationships are between those cues that have built a two-way proximity bond, which shows that from the local perspectives of each of the two cues the other cue is relatively proximate.

The results on alignment examples are displayed in Figure 5 after 100 scouts had visited the workspace. In the top result we can see many two-way proximity bonds including between the 8 dots that form the visible line among 20 uniformly randomly distributed dots. Only the two bottom-left dots are not bonded together due to another (random) dot that lies really close by the alignment. This suggest that there is Proximity evidence for grouping some parts of the line together, but solely on proximity one would not perceive the line. It is interesting to see (although quite messy) that based on proximity the alignment in the bottom result of Figure 5 has no support whatsoever and is totally disturbed by interfering close by random dots. Just as one would expect to see based on proximity.

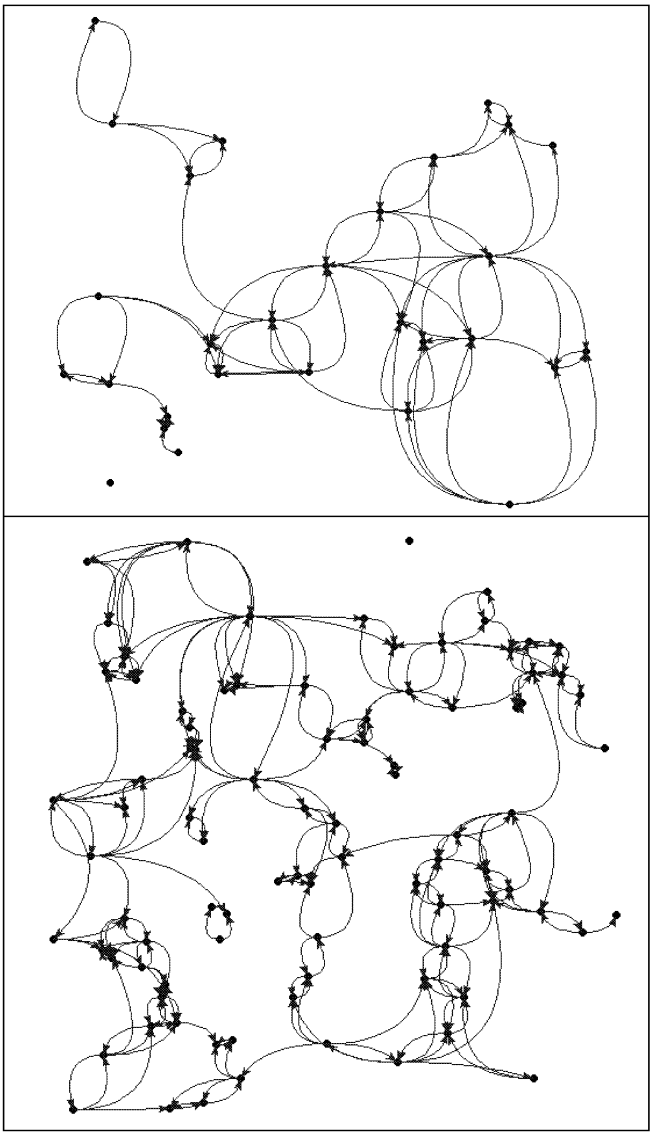


Figure 5: Top: Proximity bonds on alignment. Bottom: Proximity bonds on random dots

Regular-orientation

The Regular-orientation scout proposes and builds bonds between orientation-relations that have the same orientation and share one cue, which essentially means a straight line through three dots. We explain how the Regular-orientation scout operates by the use of the example given in Figure 6, where we initially start with three dots (leftmost dot-pattern).

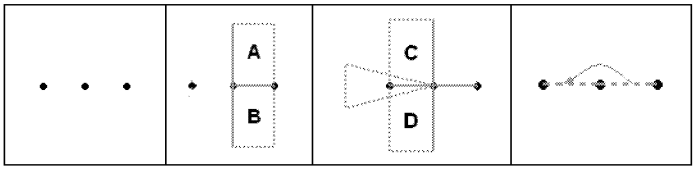


Figure 6: Leftmost: three initial dots. Middle left: Free zone check. Middle right: Free zone check and triangle search zone. Rightmost: Regular-orientation bond

First the agent lands on a random dot cue, which is in this example the dot (cue) to the far right and we take this cue as the centre point of a circular search zone for which we make a list of all the cues within this zone. The diameter of the search-zone should be sufficiently large, not to exclude the finding of lines over large distances. We find two cues and for both cues we calculate the distance to the rightmost cue, and use these distances to set up a probability for being a candidate for the second dot on the line. The shorter the distance the higher the chance the cue will be chosen (just like we did with the Proximity scout). Say we choose the middle point as the second dot. Now we check for free zones that need to be free of interfering dots, illustrated in the middle left example by two rectangular zones (A and B) Their size (height) depends on the distance between the first and second dot. If both A and B would contain any dots, then the scout will fizzle (terminate). On the other hand if only one of them includes a dot or if they are both free of them, then we continue the search for a third dot. The reason why we introduce the concept of free zones, is that it helps to home in on 'clear' lines by avoiding dense clusters of dots. In search for the third dot the scout constructs in the direction from the first to the second dot we have a triangle search zone starting from the second dot (see middle right example). The length and width of the triangle are relative to the distance between the first and second dot. From all the dots found in the triangle zone we calculate the distances to the middle cue and use these distances to set up a probability for being a candidate for the final third dot on the line. In our example we find only one dot, and also here we check for interfering dots between the second and third dot, just like we did between the first and second dot with rectangular zones (C and D). We have three conditions under which we abort proposing a regular orientation bond, because under these conditions both sides of the alignment would have interfering dots:

- If there are cues in rectangle A and D.
- If there are cues in rectangle B and C.
- If there are cues in rectangle C and D.

If none of these conditions apply then the scout proposes to bond the orientation relation between the first and second

cue, and the same relation between the second and third cue as shown in the rightmost example of Figure 6, where both relations have an orientation of 0 degrees.

The results of the regular orientation scout on alignment examples are displayed in Figure 7 after the actions of 100 scouts on the workspace. In the top result we can see that the scout for this grouping pressure easily finds and groups the alignment of dots together. Additionally it finds even more dots that form other alignments. These alignments are correct, yet when we look at the total dot pattern, we are not drawn to these other alignments and will not mark them as something interesting. However remember that this is just one view of one type of agent, which unless it is supported by any other grouping pressure remains a weak structure. In the bottom result of Figure 7 we see that the found alignments are all over the place and none seem to fit the 'hidden' 8-dot alignment, which matches expected human visual grouping in this particular example.

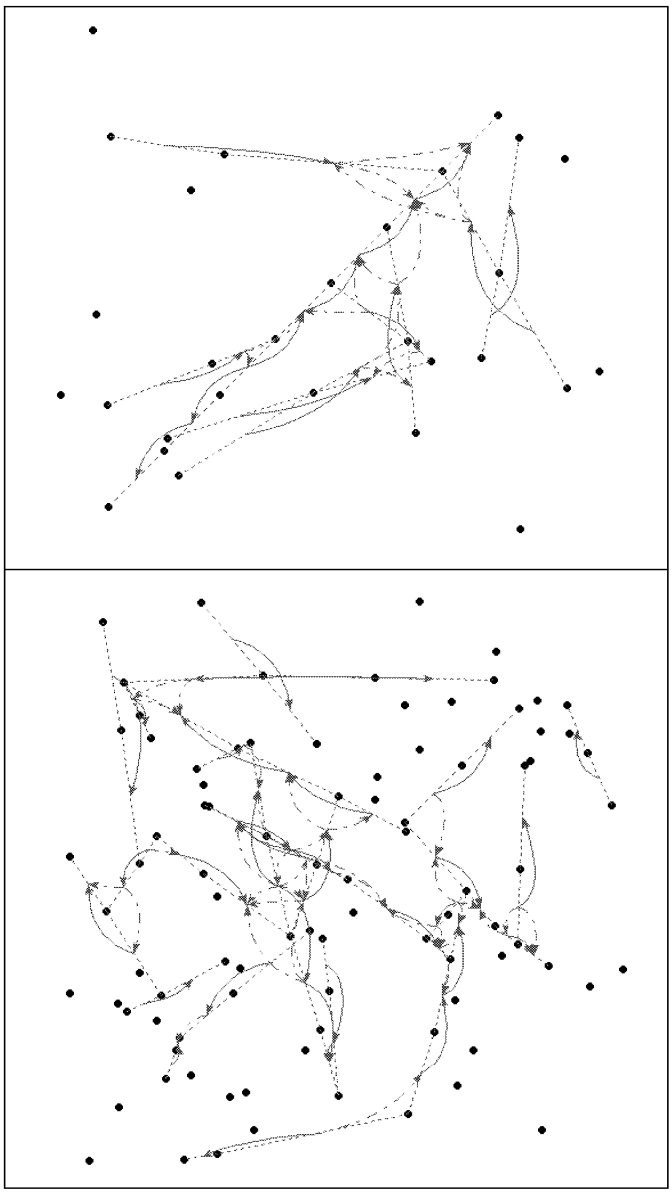


Figure 7: Top: Regular-orientation bonds on alignment. Bottom: Regular-orientation bonds on random dots

Good continuation

The Good-continuation scout works in an almost identical way as the Regular-orientation scout, working also with orientation relations. Only where the Regular-orientation scout spots straight lines (interdots with the same orientation), the good-continuation scout finds the best continuation of a line, which could be slightly curved (interdots with the best fitting adjacent orientation). For this behaviour the scout follows the same steps as the Regular-orientation scout, only allows the triangle search zone for the third dot to be much wider and has a different selection criteria for the 'best' candidate dot in the triangle zone. The selection criteria is no longer based on being nearer to the second dot (the point where the triangle cone begins), but based on best fitting of the orientation between the first and second dot. The results of the good-continuation scout are presented in Figure 8 resemble the results of the regular-orientation scout, with only minor differences.

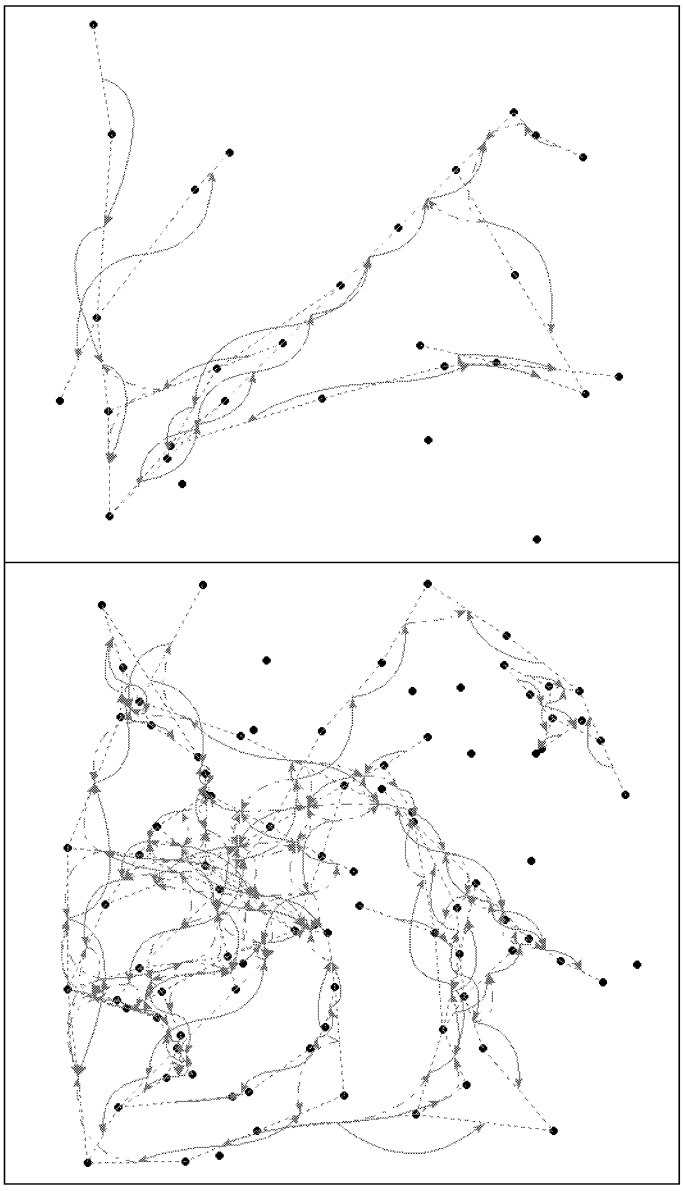


Figure 8: Top: Good-continuation bonds on alignment. Bottom: Good-continuation bonds on random dots

Regular-distance

Finally the fourth scout we have implemented, the regular-distance scout tries to bond cues together that have the same inter distance. With the help of Figure 9 we will demonstrate how this agents operates.

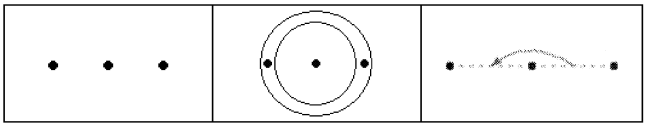


Figure 9: Left: three dot cues. Middle: margin space. Right: Regular-distance bond

First we land on a random cue in the workspace, which in this example is filled with three dot cues (leftmost dot-pattern). The agent lands on the middle cue and we use this dot as the centre point of a circular search zone for which we make a list of all the cues within this zone, and find two other cues (the leftmost and the rightmost).

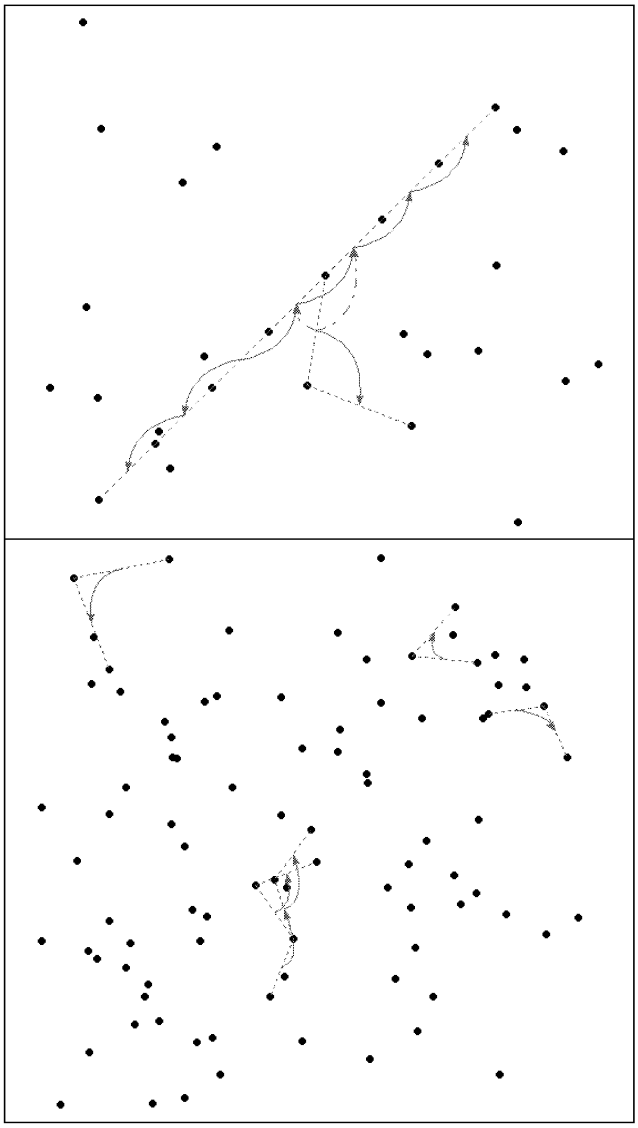


Figure 10: Top: Regular distance bonds on alignment. Bottom: Regular distance bonds on random dots

For both found cues we calculate the distance to the middle cue, and use these distances to set up a probability for being selected for the second step. The shorter the distance the higher the chance the cue will be chosen. Say we would have chosen the far right cue to perform the second step of the agent, which is finding other cues that have the same distance to the middle cue as the middle cue has to the selected far right cue. We make a list of all the cues with the same distance, given a small error margin, illustrated in the middle Figure by two circles. In our example we find the leftmost dot within the margins. The next step the scout proposes to bond the distance relation between the middle and rightmost cue, and the same relation between the middle and leftmost cue as shown in the rightmost example of Figure 9.

The results of the Regular-distance scout are presented in Figure 10, and just like with Regular-orientation and Good-continuation, this scout flawlessly discovers the alignment and this time it is almost the only thing it finds apart from one other bond. Furthermore as expected the scout finds none of the 'hidden' 7-dots in the Bottom example.

Joint Grouping pressures

Figure 11 depicts the combined results of the Regular-orientation, Good-continuation and Regular distance grouping pressures. The proximity grouping pressure was left out for clarity. The alignment of mutually supportive grouping pressures can be clearly seen in the top half of Figure 11. From this example, the advantage of using mutually supportive grouping pressures is contrasted with the interpretation power of each grouping pressure on its own. Consequently, only those bonds supported by multiple grouping pressures may lead to the formation of higher level structures.

FUTURE WORK AND CONCLUSIONS

The visual perception model was proposed as a novel emergent, self-organising model, which consists of an open architecture allowing the addition of new features, pressures and interaction methods, making it possible to define more agents and extend the model's capabilities. Following the design phase, the model was implemented as a software prototype, and was used for testing proximity, good-continuation, regular-orientation and regular-distance grouping pressures. Results so far suggest that the implemented model forms a promising foundation for further research and expansion for dealing with more complex images. Consequently, a plan for a second implementation phase has been devised for taking the computer program closer to the proposed theoretical model both by extending the cue and agent repertoire and by implementing higher-level inter-structure analogy capabilities. In addition, the visual perception model together with the Ear's Mind is used as a template for implementing a primitive visual segmentation model (see Stevens 2007). A fusion model consisting of both models is planned for multimodal perception. Such audiovisual model is expected to enhance the capabilities of real-world scene segmentation in comparison with single modality models.

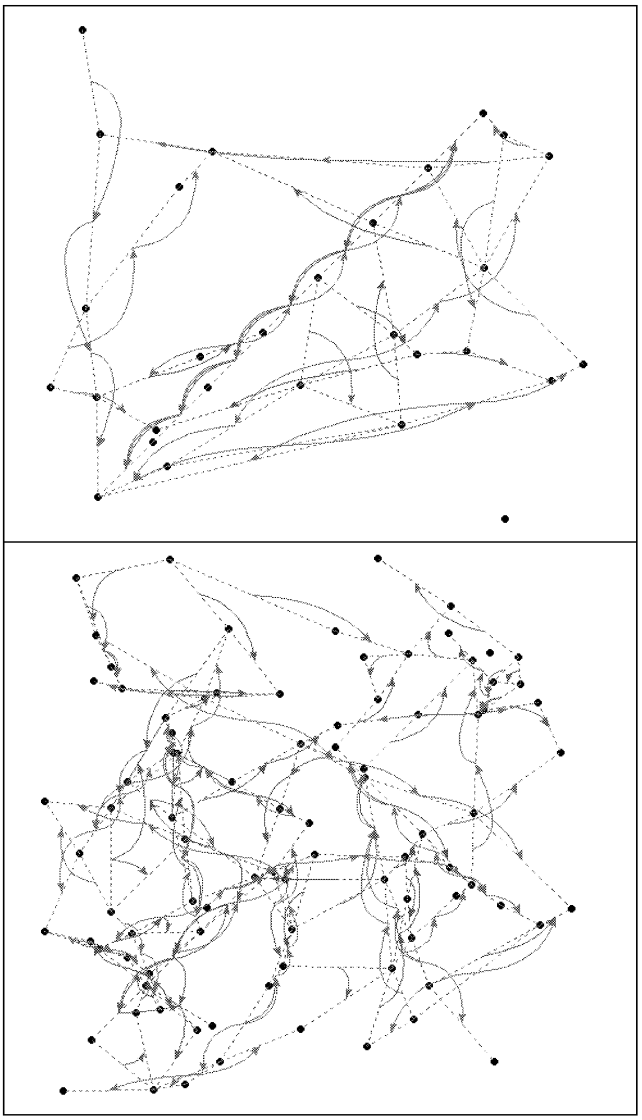


Figure 11: Top: Combined results on alignment. Bottom: combined results on random dots

REFERENCES

Bregman, A. S. 1990. *Auditory scene analysis, the perceptual organisation of sound*. 2nd paperback ed. 1999, MIT Press.

Bregman A. S. & Ahad 1990, P. A., *Demonstrations of Auditory Scene Analysis: The Perceptual Organization of Sound*, MIT Press.

Chalmers, D. J., French, R. M., and Hofstadter, D. R. 1992. High-level perception, representation, and analogy: A critique of artificial intelligence methodology. *Journal of Experimental and Theoretical Artificial Intelligence*, 4:185–211.

Cornet, B., Rothkrantz, L.J.M 2003, *Recognition of car license plates using a neocognitron type of artificial neural network*, Neural Network World, vol. 13, no. 2, pp. 115-132, Institute of Computer Science, Academy of Sciences, Prague.

Desolneux, A., Moisan, L., and Morel, J.-M. 2003. Computational Gestalts and perception thresholds. *Journal of Physiology* - Paris, 97:311–324.

Dor, R. 2005. *The ears mind: A computer model of the fundamental mechanisms of the perception of sound*. Technical report 05-16, Delft University of Technology.

Dor, R., Rothkrantz, L.J.M. 2008. *The Ear's Mind: An Emergent Self-Organizing Model of Auditory Perception.* Submitted to the Journal of Experimental and Theoretical Artificial intelligence. In press.

French, R. M. 1995. *The Subtlety of Sameness: A Theory and Computer Model of Analogy-Making*. The MIT Press.

Hecht-Nielsen, R., 1989 Neurocomputing, Addison –Wesley.

Hofstadter, D. R. and FARG 1995. *Fluid Concepts and Creative Analogies: Computer Models of the Fundamental Mechanisms of Thought.* Basic Books, New York.

Kanizsa, G. 1980 *Grammatica del vedere. Saggi su percezione e gestalt*. Il Mulino, Bologna.

Mitchell, M. 1993. *Analogy-Making as Perception: A Computer Model*. The MIT Press.

Mojsilovic, A. and Gomes, J. 2002. Semantic based categorization, browsing and retrieval in medical image databases. *In International Conference on Image Processing.*

Riesenhuber, M. and Poggio, T. 2000. Models of object recognition. *Nature Neuroscience*, 3:1199–1204.

Rothkrantz, L.J.M, Wojdel, J. C., and Wiggers, P. 2005. Fusing Data Streams in Continuous Audio-Visual Speech Recognition. In *Text, Speech and Dialogue: 8th International Conference*, Karlovy Vary, Czech Republic.

Stevens, J.C. et al. 2007 *Boiling Down Emergent Self-Organizing Soups to Solid Multimodal Perception*, Proceedings of Euromedia, 2007.

Ullman, S. 1996. *High-level Vision: Object Recognition and Visual Cognition*. Cambridge, MA: The MIT Press.

Viola, P. and Jones, M. 2001. *Robust real-time object detection*. In Second International Workshop on Statistical and Computational Theories of Vision-Modeling, Learning, Computing, and Sampling.

Wertheimer, M. 1923. *Laws of Organization in Perceptual Forms* Psycologische Forschung, 4, 301-350

APPLIED MEDIA TECHNOLOGY

DESIGN AND REALIZATION OF SMART AUDIO SYSTEMS

Zygmunt Ciota
Department of Microelectronics and Computer Science
Technical University of Lodz
90-924 Lodz, Al. Politechniki 11,
Poland
E-mail: ciota@dmcs.pl

KEYWORDS
Speech processing, Microprocessor, Integrted circuits.

ABSTRACT

The paper is focused on smart realization of small and mobile audio processors applied to voice signal analysis. Application of microprocessors and integrated circuits dedicated to analysis of audio frequency signals, demands very often modification of several algorithms, that are necessary for feature vector extraction and speech processing. Proposed methods permit to compare specific features of the current speaker voice with the base containing stored vectors of known speakers. The possibility of integrated CMOS realization of voice processing components has been also discussed.

INTRODUCTION

Signal processing concerning human voice seems to be very important task, especially in nowadays multimedia systems. Articulated sounds transmit some information, therefore it will be possible to find mutual relationship between acoustic structure of the signal and corresponding information. It is obvious, that speech processing cannot be confined to a time and a frequency domain analysis of electrical signal of microphone output. Researches concerning psychological and neurobiological analysis are also important and can improve the efficiency of intelligent human-machine interface. For example, psychological and linguistic researches show us, that emotions play significant role in decision-making process. Implementation of speech processing algorithms using hardware-software system, including mixed analog-digital approach, should improve the speed and possibility of on-line processing. Application of mixed digital-analog realization to the design process of sound processors may be better in comparison with purely digital solution and very often we can achieve better results, decreasing the chip surface and increasing the speed parameter of the system.
Therefore, the paper is focused on a smart realization of such small and mobile speech processors. We have taken into account two important tasks of speech analysis, which can be easily adopted in hardware-software implementation: speaker verification and emotion recognition. Development of multimedia systems demands very often specific procedures of people identification and verification. The most important methods base on iris and finger print analyses. However, human voice analysis becomes promising solution for people identification in modern telecommunication systems. Algorithms of speaker verification have potentially different arrays of application, like banking transactions, forensic purpose, shopping using telephone or Internet networks and other database services. The main idea of such verification is based on text-independent approach; it means that sentences applied in speaker verification system cannot be predicted.
One of the most important tasks is a proper definition of feature vectors. Each vector can contain several specific features of voice signals and finally, you have to calculate more than hundred features for each utterance. The following four vectors can be favored as very important and useful, covering the most important features of speech signal: the long-term spectra (LTS) vector, the speaking fundamental frequency vector, the time-energy distribution vector and vowel formant-tracking vector (Lee et al. 2003; Roberts and Ephraim 2005; Ciota 2005).

MICROPROCESSOR IMPLEMENTATION

Realization possibilities of autonomous speech processing using simple microprocessors are presented in this paper. The prototypes should be designed with real-time capabilities of the work. Therefore, the units of ARM7 family have been selected with respectively large SRAM memories. As an option, it is possible to add an external memory to increase the base of stored features for bigger number of speakers. In the case of speaker verification it was necessary, first to compute a spectrogram for given voice signal, afterwards corresponding LTS vector can be calculated. The time-frequency analysis gives spectrogram according to the following short-time Fourier transforms:

$$G(\nu,n)=\sum_{l=-\infty}^{\infty} g(l)h(n-l)\exp(-j2\pi\nu l) \qquad (1)$$

where $G(\nu,n)$ is a complex transform of discrete signal $g(l)$ in discrete time n for discrete frequency ν, and h is a window function.
Afterwards, it is possible to obtain LTS vector components according to the equation:

$$S(\nu)=\frac{1}{N}\sum_{n=0}^{N-1}10\log(G(\nu,n))^2 \qquad (2)$$

where N is a number of time windows.
It is possible to indicate the following procedures and algorithms: registration and playback of the speech signal; normalization of the recorded signal; voiceprint calculation using fast Fourier transform (FFT); calculation of feature vectors and normalization procedures; training block, which permits to create the base of referential templates for identification purpose; vector distance calculation and verification procedures. Using the above algorithms we can design embeded systems by using mixed hardware-software co-design.
In the case of emotion recognition fundamental frequency F_0 and its statistical behavior have been applied. It is possible to use cepstrum method or autocorrelation function for F_0 extraction. In our system simple and more precise method based on autocorrelation function has been implemented. The microprocessor has been applied for recognition of four basic emotions: joy, anger, sadness and neutral state of the voice.

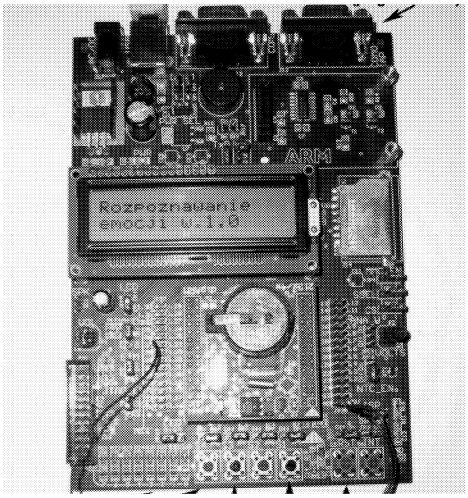


Figure 1: An Example of Speech Processing Prototype Using ARM Processor (Zloty 2007)

The observed efficiency of speaker verification is equal of about 65% for small utterances and 85% for longer speech (Dziubinski 2005). For emotion recognition the corresponding values are 56% and 75% respectively (Zloty 2007). In both cases the processing time was in the range of 1.5-5 seconds. In the case of speaker verification only 10 people participated in the test. A prototype of emotion recognition system based on ARM7 processor is shown in Fig. 1.

APPLICATION OF INTEGRATED CIRCUITS

More efficient systems can be obtained using mixed hardware-software approach. Current mode realizations of analog components permit to obtain an efficient tool for audio signal processing. As the most important analog realizations we can mention multiplier, analog-to-digital converter and current mode filters. Fig. 2 presents a prototype of CMOS integrated circuit containing amplifiers and analog-to-digital converter (Jankowski et al. 2000a). Computer simulations indicate that the proposed blocks have simple structures and have low sensitivity to variations in technological parameters. Additionally, such circuits can be easily manufactured in standard CMOS processes.
Low-power, high-speed and low-voltage comparators are more and more needed for market increasing of wireless devices, for instance cellular phones, global system for mobility and wireless local area networks, where verification of speaker emotions and speaker recognizing plays a significant rule. Comparators are often used in nowadays microelectronics, therefore different structures can be applied according to given requirements.

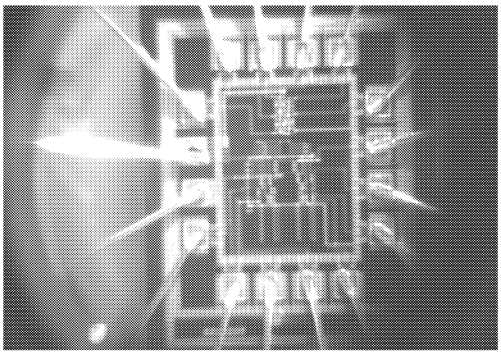


Figure 2: Microphotograph of Transconductance Amplifier and A/D Converter Suitable for Speech Processing

Comparison of the different design methods permits to choose the best one with regard to the kind of technology in which the system of voice processing should be integrated. Computer simulations indicate that the proposed blocks have simple structures and have low sensitivity to variations in technological parameters. Switched current (SI) circuits can be easily manufactured in standard CMOS processes because both the first- and the second generation. Additional advantage is the simplicity of implementing various mathematical operations, such as summation, comparison, inversion and multiplication. Therefore, current mode techniques are commonly used in discrete-time analog signal processing, neural networks, A/D and D/A conversion.
Taking into account the above remarks, a bank of switched current filters has been also designed and performed in CMOS technology (Jankowski et al. 2000a; Jankowski et al. 2000b). The prototypes contain high precision CMOS integrated filters with low sensitivity to technological mismatches. For prototype measurements, it was necessary to use special laboratory equipment with precise current-voltage converters. The proposed laboratory stand permits to obtain accurate current characteristics for such circuits. The prototype contains also digital memory to control all functions of audio signal processing, including software radio capabilities. The implementation of the proposed algorithm as hardware-software system, including mixed analog-digital approach, should improve the speed and also the quality of nowadays audio systems.

CONCLUSIONS AND FUTURE WORKS

Designing process of audio and also speech processing system is inherently a complex task involving human expertise as well as aids intended to accelerate the process. Furthermore, such efficient system has to have real-time capabilities, so the hardware-software co-design permits to achieve low cost and high speed performances. While microcontrollers and microprocessors are inherently digital components, some functions can be executed in analog or digital form.

The implementation of the proposed algorithms as hardware-software system, including mixed analog-digital approach, should improve the speed and also the quality and resolution of voice processing systems. The presented systems take into account only acoustic information of speech signal, but it will be possible to expand its including lexical, semantic and discourse information.

The results of emotion recognition and speaker verification system are rewarding, but sometimes we can observe mistakes in recognition process. However, emotion recognition makes difficulties also for human evaluation. The quality of the system depends also on the training processes, including duration of recorded voice signal. Unfortunately, it is difficult to obtain a proper base of voice examples for different emotions. Moreover, our prototypes have rather little memory capabilities, confined to 512 kB FLASH memory. The above problems give us direction of current and future researches, and we expect to solve such problems in a new prototype project.

REFERENCES

Ciota, Z. “Emotion Recognition on the Basis of Human Speech”, ICECom-2005, in Proc. *18th International Conference on Applied Electromagnetics and Communications*, 12-14 October 2005, Dubrovnik, Croatia, pp. 467-470.

Dziubinski, W. “Speaker recognition - hardware implementation” (in Polish), M.Sc. Thesis, Technical Univ. of Lodz, 2005

Jankowski M., Z. Ciota, A. Napieralski: "CMOS Realization of Switched Current Discrete-Time Filter", *Proc. IEEE Nordic Signal Processing Symposium NORSIG2000*, June 13-15, 2000, Kolmården, Sweden, pp. 169-172

Jankowski, M., Z. Ciota, A. Napieralski, "Methodology of CMOS VLSI design using current mode approach", *Proc. 7th International Conference Mixed Design of Integrated Circuit and Systems - MIXDES 2000*, Gdynia, Poland, pp. 457-461

Lee Ch., Donghoon Hyun, Euisun Choi, Jinwook Go, Chungyong Lee, “Optimizing Feature Extraction for Speech Recognition”, IEEE Trans. Speech and Audio Processing, no 1, January 2003, pp. 80-87

Lee, Ch., Donghoon Hyun, Euisun Choi, Jinwook Go, Chungyong Lee, “Optimizing Feature Extraction for Speech Recognition”, *IEEE Trans. Speech and Audio Processing*, no 1, January 2003, pp. 80-87.

Roberts, W.J.J., Yariv Ephraim, "Speaker Classification Using Composite Hypothesis Testing and List Decoding", *IEEE Trans. Speech and Audio Processing*, vol. 13, no 2, March 2005, pp.211-219.

Zloty, R. “Emotion recognition based on speech signals - hardware implementation” (in Polish), M.Sc. Thesis, Technical Univ. of Lodz, 2007

ENERGY SAVING IN INTELLIGENT BUILDINGS

Via Energy Harvesting In Wireless Sensor Networks

Lizzie Tang
Chris Guy
School of System Engineering
University of Reading
Reading, RG6 6AY, UK
Email: {z.tang|c.g.guy}@rdg.ac.uk

KEYWORDS

wireless sensor network (WSN); radio frequency (RF); energy harvesting; intelligent buildings; electromagnetic energy

ABSTRACT

Wireless sensor networks (WSNs) have been widely used in pervasive systems such as intelligent buildings. As a vital factor of product cost, energy consuming in WSN has been focused upon, but only via energy harvesting can the problem be overcome radically. This article presents a new approach to harvesting electromagnetic energy for WSN from useless radio frequency (RF) signals transmitted in WSN, with a quantitative analysis showing its feasibility.

I. Introduction

Wireless sensor networks (WSN), consisting of spatially distributed autonomous devices using sensors to cooperatively monitor physical or environmental conditions such as temperature and vibration at different locations[1] have been widely deployed in intelligent buildings. Energy consumption in WSNs has attracted the attention of researchers and various efforts on energy harvesting have been made. Here we present a brand-new approach to harvesting electromagnetic energy from the useless signals transmitted in WSN applied in intelligent buildings.

The rest of the paper is organized as follows: Section II explains why we propose to harvest electromagnetic energy from useless signals transmitted in WSNs. Section III describes a simplified model for the proposed energy harvesting mechanism and Section IV presents relevant quantitative analysis to discuss whether this approach is worthwhile. Section V gives the conclusions.

II. Source Of Energy

There are a number of reasons why it makes sense to try to capture electromagnetic energy from useless signals transmitted in WSNs in intelligent buildings, attempting to prolong battery life:

First, inside buildings where WSNs are applied, most ambient energy sources such as solar (light as well), wind and tides, are not suitable because they are not available. For instance, there will be hardly any energy harvested by solar harvesting system during the period when it continues raining. On the other hand, sources of piezoelectricity, vibration or motion are stable but none of them are ambient. That is why the utilization of electromagnetic energy in these applications is considered, as a general way to absorb energy from radio or microwave frequency and convert it into electricity via a rectifier-antenna circuit.

There are two different coupling techniques, near and far fields, in the utilization of electromagnetic energy, typical examples of which are research outcomes from MIT[2] and Powercast Co.[3] respectively. It is hard to tell from the published literature which is better. However, aiming at the aspect of application of WSN in intelligent buildings wherein long range (5-20 m) is more common, the latter is employed more often [4]. In contrast to near-field the two most common of which are 128 kHz (LF) and 13.56 MHz (HF), there is no restriction on the field boundary for far-field coupling. Only low carrier frequencies are used in near-field coupling applications, while one problem with the use of low frequencies is that a large antenna coil is required. An advantage of a far-field device operating at a high frequency is that the antenna can be small, leading to low fabrication and assembly costs. The attenuation of the EM field in the far-field region is proportional to $1/r^2$, which is smaller by orders of magnitude than in the near-field range (which is $1/r^6$), where r is the distance between the transmitter and receiver.

Far-field devices usually operate in the 860–960 MHz UHF band or in the 2.45 GHz Microwave band, representative of which are the frequency bands of ZigBee or Z-Wave (recent relevant promising technologies to be introduced in the following text) and that of Bluetooth respectively. RuBee[5], a relatively new active RFID technology, operates in the LF band and employs long-wave magnetic signaling. It can achieve a read range of 30 m. A great advantage that long-wave magnetic signaling has is that it is highly resistant to performance degradation near water and metal objects such as ironwork, a serious problem for UHF and Microwave far-field. But the problem of its antenna coil size can't be ignored, as

mentioned above. There is also another kind of RFID technique SAW[5] which relies on surface acoustic wave technology, whose tags are smaller, lighter and relatively cheaper, compared to silicon-based RFID ones. Their other advantage is a built-in ability to measure an object's temperature and to estimate real-time location, which is also claimed as to be able to achieve a longer read range and greater reliability in the presence of water and metal. However, consider the character of the environment in an intelligent building (not very distant nor with much water and ironwork around). Additionally, there is a conclusion from the Friis transmission equation that operating at 900 MHz exhibits a significantly longer range than is possible at 2.4 GHz[6]. So it seems a frequency around 900 MHz is optimal.

Two kinds of far-field coupling electromagnetic energy sources are usually considered currently: ambient radiation derived from radio and TV broadcasting, and energy deliberately broadcast by RF devices. The former is not reliable while the latter needs a deliberate energy supply. The useless signals transferred during communication are considerable. In a WSN, if a message doesn't need to be replied nor transmitted by the node which receives it, it is useless for this node. Thus a signal for this message can be treated as one designed just to charge the node in the wireless power supply system, without any power beacon to locate. So what we are concerned with is how much energy there is in the signal, rather than what details of this message are.

In addition, people have to face the problem of electromagnetic pollution, and it is considered that overall electromagnetic radiation should be reduced [7][8][9][10]. So it makes sense to make use of existing electromagnetic radiation distributed around rather than adopting additional electromagnetic power supply device.

III. Corresponding Model Of Energy Harvesting System

Suppose the antenna of the energy harvester is omni-directional. The amount of electromagnetic energy harvested from a message received, A, is related to its conversion efficiency η (up to 70% [11] for Powercast and above 84% for Buck-Boost Converter [12]), the received signal strength P and the time t that the signal lasts:

$$A = P \times \eta \times t \qquad (1)$$

When considering a large-scale attenuation model for indoor radio signal propagation, the simplest model only considering the radio propagation path loss is [13]:

$$P(d) = P(d_0) - 10 \times n \times \log_{10}(d/d_0) \qquad (2)$$

n is the attenuation exponent which is 2 for free space otherwise greater than 2; P(d) is the radio power at distance d from the transmitter, which equals to P in Equation 1; $P(d_0)$ is the power at a reference distance d_0, usually set to 1 meter.

Suppose d_0=1 meter and $P(d_0)$=1 Watt=$10\log_{10}$(1W/1mW) dBm=30 dBm, because most the power of indoor device using Powercast is around 1 Watt[14]. Suppose n=2, d=30 meters, then $P(d) \approx 0.458$ dBm, that is, about 1.1112 mW. But since real space is not free space and besides distance there are walls and floors, etc [15], to effect attenuation as well, the actual signal strength is usually within the range from -30 dBm to -100 dBm[16]. For ZigBee, nearly a quarter of the signals received from 30 meters away are -40 dBm[17] (that is 10^{-4} mW).

Let ρ be the mean proportion of useful ones among all the messages received, k be the total of sensor nodes that send messages to the energy harvester (usually with the range of 5 to 100 meters if indoors), and the frequency for a node to send a message be f. The total energy the harvester can capture during the period t is

$$At = \sum_{i=1}^{k} \rho \times fi \times Pi \times \eta \times t$$

Pi means the received power of the ith node.

If all the nodes are within the same distance from the harvester and have the same frequency to send messages,

$$At = \rho \times f \times P \times \eta \times t \times k.$$

Suppose there are 4 nodes in free space, all 30 meters away from the energy harvester, every second sending 10 messages, 9 in which are useless. $P(d_0)$=1 W, and η=50%. So in theory, the total energy captured in a second is $90\% \times 10 \times 1.1112 \times 10^{-3} \times 50\% \times 4$ J= 0.02 J. Even for the case of ZigBee ($P(d) = 10^{-4}$ mW, not 1.1112 mW), 2 μJ can be captured, which is acceptable compared to some pioneering work done in the field of electromagnetic energy harvesting [18].

It does not make much sense to consider the bit rate of the signal received, because the amount of energy converted depends on the speed that the flux of the coil in the receiver changes. If the signal is frequency modulated, its bit rate is independent of changing speed of the flux. If the signal is amplitude modulated, however, higher bit rates can result in a faster speed of flux changing, which will be very complex to calculate.

IV. Quantitative Analysis

Suppose an AA or AAA battery's capacity is x mAh (which means it can last at the rate of 1 mA for x hours, or 1 μA for 1000x hours), and its average operating current and voltage are y A and z V respectively. Therefore the original lifetime of the battery

$$t = 0.001x/y \text{ hours} \qquad (3)$$

Suppose the power of energy harvesting in the simplified model is m watt, and the lifetime of the battery after implementing energy harvesting to the simplified model be t' hours, then

$$z*y*t'=m*t'+ z*x*0.001 \qquad (4)$$

According to (3) and (4) we get

$$(t'-t)/t=m/(z*y-m) \qquad (5)$$

Let $m = 2*10^{-6}$ (the worst possible case of ZigBee mentioned before). The battery works at 1.5 v and 10 μA, thus $z = 1.5$ and $y = 10^{-5}$. In this case $(t'-t)/t \approx 15.385\%$, which means if the original lifetime of the battery is one year, it can be extended for more than 56 days after implementing energy harvesting described before.

Only when the values of z*y and m are at the same scale in quantity will the extension make sense: in the example above they are both at the scale of 10^{-6}, with their own international units. Otherwise the extension is too tiny: if the battery works at a current of 0.1 mA while all the other parameters remain the same, $(t'-t)/t \approx 1.35\%$, that is about 5 days based on one year original lifetime. Certainly if $z*y<m$ it would make sense since the battery's life would be extended endlessly. Fortunately, now almost all the wireless sensors consume the current when resting at the scale of μA, for example 8 μA [19], or even, say, 0.5 μA [20]. Although not many wireless sensors can achieve a consumption of transmission current around 1 μA [21][22][23][24], the resting period (especially if it's long) can be used to recharge the battery and prepare for other periods. The energy consumed during recharging mode on and off and battery creepage should be considered as factors of battery recharging efficiency.

V. Conclusion

A new approach to reduce energy waste and exploit a new source for an energy harvesting mechanism in WSNs applied in intelligent buildings has been proposed, following relevant quantitative analysis to demonstrate its feasibility. There are useless signals received by nodes in WSNs, which are electromagnetic energy worth utilizing. The battery's charge mode will be turned off when the sensor reaches the lower limit of energy that should be, or when fully charged. A protocol based on existing optimal routing to tell as fast as possible whether a message received is useful or not will be proposed later.

VI. Reference

[1] http://en.wikipedia.org/wiki/Sensor_network

[2] http://web.mit.edu/newsoffice/2006/wireless.html

[3] http://en.wikipedia.org/wiki/Wireless_energy_transfer

[4] Vipul Chawla and Dong Sam Ha, "An Overview of Passive RFID", IEEE Communications Magazine: Applications & Practice, pp. 11-17, Sep. 2007

[5] Vipul Chawla and Dong Sam Ha, "An Overview of Passive RFID", IEEE Communications Magazine: Applications & Practice, pp. 11-17, Sep. 2007

[6] http://agraja.wordpress.com/2007/03/14/friis-transmission-equation/

[7] http://en.wikipedia.org/wiki/Electromagnetic_radiation_and_health

[8] Chun De Liu, Cheng Bin Li, "Electromagnetic pollution and its control", 2000. 2nd International Conference on Microwave and Millimeter Wave Technology, pp. 461-464, 14-16 Sep. 2000

[9] Guanghou Jin, Gengyin Li, Ming Zhou; Yixin Ni, "Research on allocation of gross electromagnetic pollution emission right in power quality markets", 2005. IEEE Power Engineering Society General Meeting, pp. 320-325, 12-16 Jun. 2005

[10] Delhi, N., Behari, J, "Electromagnetic pollution-the causes and concerns", 2002. Proceedings of the International Conference on Electromagnetic Interference and Compatibility, pp. 316-320, 21-23 Feb. 2002

[11] http://www.powercastco.com/downloads/TECH_OVERVIEW.pdf, pp.1

[12] Elie Lefeuvre, David Audigier, Claude Richard, Daniel Guyomar, "Buck-Boost Converter for Sensorless Power Optimization of Piezoelectric Energy Harvester", IEEE Transactions on Power Electronics, Vol.22, No.5, pp.2022, Sep 2007

[13] Saˆad Biaz, Yiming Ji; Bing Qi, Shaoen Wu, "Dynamic Signal Strength Estimates For Indoor Wireless Communications", 2005 International Conference on Wireless Communications, Networking and Mobile Computing, 2005. Proceedings, pp. 607, 23-26 Sep 2005

[14] http://www.powercastco.com/downloads/HEALTH_SAFETY_QA.pdf, pp.3

[15] Yongguang Chen, Hisashi Kobayashi, "Signal Strength Based Indoor Geolocation", 2002. IEEE International Conference on Communication, pp. 437, 28 Apr-2 May 2002

[16] Stuart A. Golden, Steve S. Bateman, "Sensor Measurements for Wi-Fi Location with Emphasis on Time-of-Arrival Ranging", IEEE Transactions on Mobile Computing, Vol.6, No.10, pp. 1189, Oct 2007

[17] Wenping Chen; Xiaofeng Meng, "A Cooperative Localization Scheme for Zigbee-based Wireless Sensor Networks", 2006. 14th IEEE International Conference on Networks, pp. 2, Sep 2006

[18] Xinping Cao, Wen-Jen Chiang, Ya-Chin King, Yi-Kuen Lee, "Electromagnetic Energy Harvesting Circuit With Feedforward and Feedback DC-DC PWM Boost Converter for Vibration Power Generator System", IEEE Transactions on Power Electronics, Vol.22, No.2, pp. 679, Mar 2007

[19] http://www.dustnetworks.com/docs/M1030.pdf

[20] http://epubl.ltu.se/1402-1617/2007/118/LTU-EX-07118-SE.pdf, pp.30,32,35

[21] http://www.edn.com/article/CA601827.html

[22] http://www.opensourceinstruments.com/Electronics/A3013/M3013.html

[23] http://www.eeproductcenter.com/passives/brief/showArticle.jhtml?articleID=168600580

[24] http://young-sacl.stanford.edu/~jnrao/Stanford/stanford_presentation.ppt#283,10, Slide 10

A CONTEXT AWARE AND USER-TAILORED MULTIMODAL INFORMATION GENERATION IN A MULTIMODAL HCI FRAMEWORK[1]

Siska Fitrianie
Iulia Tatomir
Leon J.M. Rothkrantz

Man-Machine-Interaction Group, Delft University of Technology
Mekelweg 4 2628CD Delft,
the Netherlands,
E-mail: {s.fitrianie, l.j.m.rothkrantz}@tudelft.nl

KEYWORDS

Multimodal fission output, natural language generation, visual-language generation.

ABSTRACT

In recent years, we have developed a framework of human-computer interaction that offers recognition of various communication modalities including speech, lip movement, facial expression, handwriting and drawing, body gesture, text and visual symbols. The framework allows the rapid construction of a multimodal, multi-devices, and multi-user communication system within crisis management. This paper reports the multimodal information presentation module combining language, speech, visual-language and graphics, which can be used in isolation, but also as part of the framework. It provides a communication channel between the system and users with different communication devices. The module is able to specify and produce context-sensitive and user-tailored output. By the employment of ontology, it receives the system's view about the world and dialogue actions from a dialogue manager and generates appropriate multimodal responses.

INTRODUCTION

As information and media technologies develop and become widespread, people's ability to communicate and share information increases. Recent developments in electronic technology offer possibilities for more diverse computer types with ever-increasing capabilities: workstation, notebook computer, tablet computer, personal digital assistant (PDA), mobile telephone, etc. As mobility becomes ubiquitous, multimodality becomes the inherent basis of the interaction paradigm. Multimodal user interfaces designed for multi-devices scenarios can offer various ways for users to interact in a more natural fashion with provided services (Oviatt et al., 2004).

Multimedia services can be generated in a coordinated fashion on a combination of multiple modality channels (Maybury, 1999). Compared with existing traditional interfaces, which draw upon canned presentation using windows, menus, and dialogue boxes, this extension of interaction can provide users with more choices to use different modalities and devices and to find their own optimal experiences. The integration of interaction means where are available in the user side, into the human-computer interaction (HCI) provides challenges on several levels.

Firstly, the question arises regarding the actual physical capabilities of user communication devices and the type of input and output modality capabilities the devices have. An HCI system should be able to accommodate the flexible switching of communication modes and devices. This means a demand for such a system to deliver the same content of information and provide the same services using different sets of modalities for different devices.

Secondly, the system must be aware of context variables, such as user profile, user emotion and location, which usually are used to adapt applications to user's environment. User's contextual situation may change over time due to mobility and the change of environment. The user interface and presented information may change based on these variables.

Finally, the constraint and the change should not affect the capabilities of the system to be able to offer the user a unified view towards a wide scale of its applications. Conveying enough information and applying complex functionalities into different types of computers and putting them in the hand of different (mobile) users with different activities represent a serious interface and interaction design issue.

This work aims at the development of an automated interface and information presentation generation module that is able to select content to achieve given communicative goals, design the presentation, and allocate coordinate information across media (i.e. typed or spoken language, visual language, and graphics). The module can receive the communication plans (dialogue actions) from a dialogue manager (DAM) of a HCI system. Appropriate multimodal information is presented to the user by the employment of ontology. Here, we explore a method to specify and produce context-sensitive and user-tailored information presentation.

The structure of this paper is as follows. In the following section, we start with related work. Next, overview of our research domain is presented. We continue with the description of the world knowledge representation. Further, the information generation and presentation module is

[1] The research reported here is part of the Interactive Collaborative Information Systems (ICIS) project, supported by the Dutch Ministry of Economic Affairs, grant nr: BSIK03024

explained. Finally, we elaborate our experiments and conclude the paper with a discussion on future work.

RELATED WORK

To date, there are a number of multimodal systems in existence. SmartKom combines speech, hand gestures and facial expressions for both input and output and employs an embodied conversational agent, "Smartakus" (Wahlster et al., 2001). The system supports dialogue with mixed initiative with a restricted-domain question answering. All features of the modality generation are planned and encoded in Multimodal MarkUp Language (M3L). The planning includes user modality preferences, output devices constraints and user's native language. The planner specifies presentation goals for the text generator, the graphics generator and the animation generation. The language generation follows a template oriented approach based on Tree Adjoining Grammar (Joshi & Vijay-Shanker, 1999). The animation generator uses planned animations of Smartakus that are synchronized with the graphical display and speech output.

COMIC project's demonstrator system adds a dialogue interface to a CAD-like application used in bathroom sales situation to guide clients redesign their rooms (Foster, 2004). The input to the system includes speech, handwriting and pen drawing. The output combines synthesized speech, a talking head and control of the underlying application. The system's knowledge is stored in ontology represented in RDF-OWL (W3C). The presentation planner in COMIC translates specifications from its dialogue manager (which includes dialogue history and user model) into logical forms. The language generation uses combination of n-grams language model and Combinatory Categorial Grammar (Steedman & Baldridge, 2005) to select language configurations that satisfy the formulated forms. The output module controls the facial behavior of the talking head, the deictic gesture at objects and the graphical displays using the timing information returned by the speech synthesizer, which creates a full scheduled in advanced.

Match (Johnston et al., 2002) is an interface to restaurant and subway information of a city, which provides users with input using speech, handwriting, touch or composite multimodal commands. The system responds to the user by a talking head and dynamic graphical displays. The knowledge of the system is stored in the domain ontology that determines its next moves for different types of user actions. Depends on the dialogue history, topic space, user model and system beliefs, the language generation performs a simple template based generation for simple responses and text planner for more complex generation, e.g. to compare two restaurant based on user preferences. A stack of graphical actions is used to coordinate the facial behavior of the virtual agent, the graphical displays and speech synthesizer.

MACK uses a combination of speech, gesture and indication on a normal paper map that users place on a table between themselves and MACK (Cassell et al., 2002). It uses a template-based sentence generator. BEAT (Cassell et al., 2001) is responsible for the generation of appropriate speech with intonation and the facial and body behavior of MACK. Behind BEAT, a rule-based approach is used to synchronize the animation and the speech based on MACK characteristics with which estimated timings are scheduled prior to execution.

AdApt project (Gustafson et al., 2000) focuses user-system-dialogue dependent speech recognition and audiovisual synthesis in real-estate domain. It employs a static 3D wired-frame of a face and a visual map display. Template-based language generator selects the system responses based on planned output messages in XML format from a dialogue manager. The messages consist of the latest utterance, the visual map attributes and constraints, and a list of apartment using the current constraints. Using a frame-based approach, the GUI manager synchronizes the facial behavior of the face and graphical displays.

XISM (Sharma et al., 2003) employs input processing of natural gestures and speech commands for managing emergency scenarios on a large display to facilitate decision making in crisis control rooms. The implementation supports collaborative tasks among people present at remote sites with different computing platforms, communication devices and network connections. Using the knowledge of user, current task, and the current situation of the world, the output planner establishes the system's intention and belief and selects dialogue action. If further communication with the user is required, it will prepare agenda items. Using schema-based approach, the output module organizes the contents of the multimedia and multimodal responses.

COMPUTATIONAL HUMAN INTERACTION MODELING FRAMEWORK OVERVIEW

The introduction of novel information and communication technology (ICT) in the crisis management domain can help to provide more detailed and accurate situation overviews that are current and shared amongst all management levels (Moore, 2006). Towards this goal, we have developed a framework that allows the rapid construction and evaluation of multimodal human-computer interaction systems (Fitrianie et al., 2007). The development aims at module integration that is independent of the availability of modalities. A HCI system can be constructed using this framework to support communication between different actors via different devices to work collaboratively responding to crisis situation. Currently, we developed a project demonstrator system for reporting observations, which is able to collect many small but up to date observation reports, interpret them automatically and form a global view about the reported events.

Input Output User Interfaces

To support natural and multimodal interaction, the framework includes input recognition modules of different modalities, such as text, speech, visual language, gesture, pen writing and drawing, and facial expression. Figure 1 shows the architecture of a communication system for a single user system. The output combines text, synthesized speech, visual language and control of the underlying user interface. We also employ a presentation agent that is able to generate appropriate speech with intonation and facial behaviors based on annotated XML-based inputs (Bui, 2004).

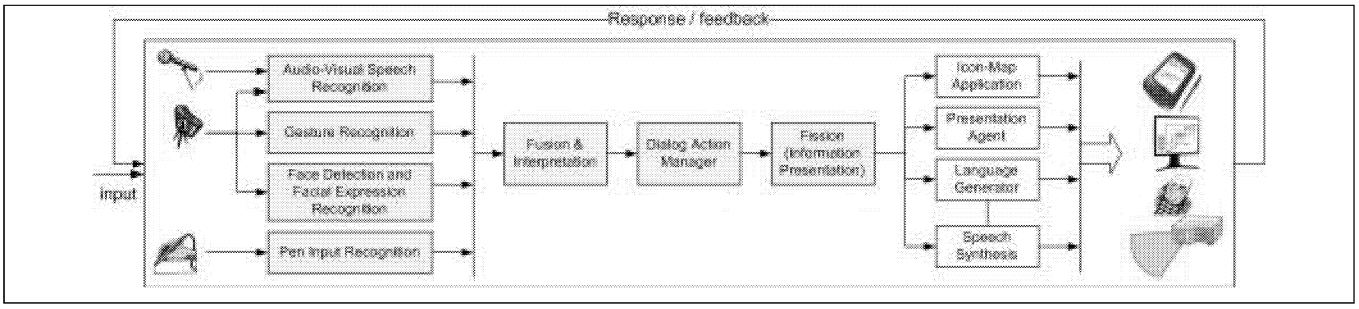


Figure 1 The developed framework architecture for a single user

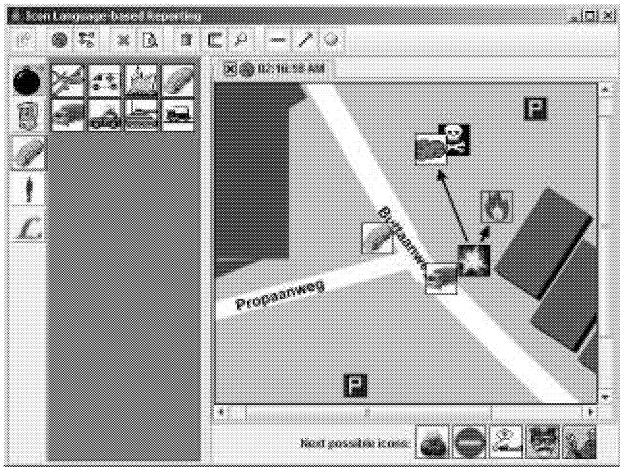


Figure 2 Map-based interface

We designed a map interface to support people with geospatial information (Figure 2) (Fitrianie et al., 2006). It provides drawing tools and predefined icons for a free and natural way of sketching and describing a crisis situation. In addition, the interface provides icon-strings, text and photos input. The icons represent objects and events in the world (for example: fire, car, smoke). This representation is chosen to support faster interaction (Kjeldskov & Kolbe, 2002), to reduce the ambiguity of the communicated information (Norman, 1993), and to provide a language independent message construction (Perlovsky, 1999). The drawing tools can be used to draw lines, arrows, and shapes to represent an area, for emphasizing icons or locations on the map, or for grouping icons. Icon strings (Fitrianie & Rothkrantz, 2005 & 2006), photos, and text are used to represent non-spatial information.

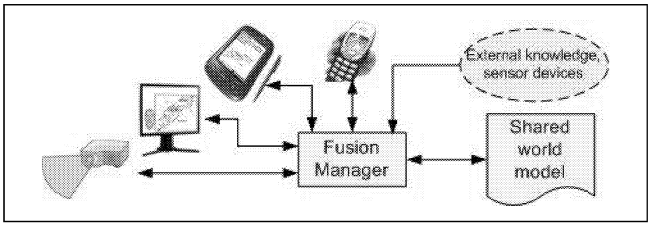


Figure 3 The developed communication system: multimodal, multi-devices and multi-users

The variety of modules allows us to apply the framework to support various roles within the crisis management, including rescue workers and civilians in the field and control room operators. A schematic overview of the system that supports multiple users is shown in Figure 3. A centralized Fusion Manager processes and integrates every newly reported situation from all users and adapts the world view accordingly then sent it back to the network and shared with the users.

Data Fusion and Dialogue Action Manager

Multimodal fusion module provides the mechanism of building coherent, context dependent interpretation of the communicated situation. It consists of a parser which builds concepts from various preprocessed input modalities on a user workspace. In tandem with the parser, an emergent self-organizing mechanism is designed based on the work of (Dor, 2007) to find coherent structures from activated world model concepts and parser input. Concepts from the world model may become activated by their instances on the workspace or by the activation of related concepts. Activation may in turn influence the building of structures among concepts on the workspace, trigger DAM for additional inputs, supply context for the parser, and finally, when enough structure coherence produce a representation of the current interpretation of the fused modalities. The interpretation results are handed over to the DAM and may assist it in forming feedback to the user. Using the knowledge of the world, the observed user's action and the observed user's affective state, the multimodal DAM maintains system's belief state about the user and selects appropriate actions to send to the user through output modules (Bui et al., 2007).

Communication Infrastructure

All modules are integrated in an *ad hoc* fashion using the iROS middleware system (Johanson, 2002). Modules can send messages to the middleware and other modules can subscribe to receive certain types of messages from it. In a multi-user system, the same module can exist in different instances for different users. Messages can be addressed to a specific user or be broadcasted to all users.

KNOWLEDGE REPRESENTATION

The architecture of HCI systems consists of a number of highly delicate components that were developed by different developers. Each component is designed around the task that its developer is targeting. To encounter the problem of different format of communication data inter-components, knowledge representation is proposed in this framework, which is referred to as the ontology and will be defined as a

body of knowledge about the world, the user, and the task. The ontology is stored in W3C-OWL.

World Model

Knowledge of the world is a critical component for a crisis management system, since it must be aware of situations in a crisis event to be able to interact with the user. The world model consists of two contexts: the dynamic and the static. The dynamic context is represented by a chain of temporal specific events and a group of dynamic objects in action at a certain location in the world. While the static context is the geographical information concerning the crisis location, such as buildings, streets and parcels. It has direct links to the visual symbols on the user interface. Figure 4 shows the class WorldObject, which refers to an entity involved in a crisis event. The icons are the instances of the subclasses of this class. For example, the icon "collision" is a Collision, a subclass of a CrisisEvent.

```
WorldObject
     StaticObjaect …
     DynamicObject
            Actor …
            CrisisEvent
                Collision
                Explosion
                …
            RescueEvent …
            Transportation
                Car
                Truck
                …
            Tool …
            Substrate …
            Preposition …
```

Figure 4 The WorldObject taxonomy

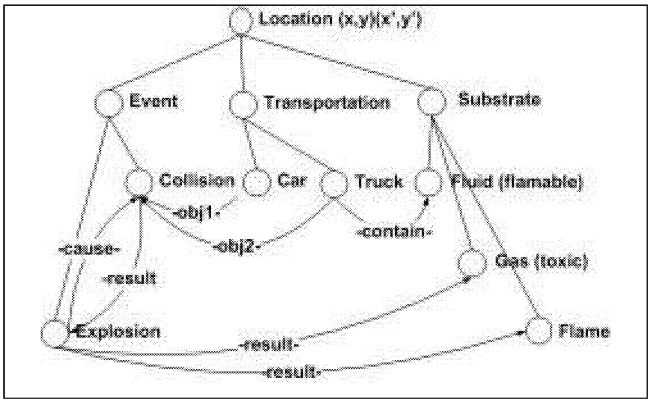


Figure 5 Graph-based symbolic representation of a crisis event on a certain location in the world

We represent geospatial knowledge of crisis situations using graphs for data modeling. The graph connects its nodes based on their approximated spatial coordinates in the world. The lower nodes represent objects, actions or events in the world. They do not only contain specific information of events and objects, but also their current status (for example living condition and dynamic spatial state), their temporal information (for example frequency and time point), and their spatial information (for example current location, origin, destination and path). The arcs represent the hierarchy of groups of individuals or show the relations among concepts (for example result, cause and contain). At the root, a node describes the perspective location of the crisis event. The illustration in Figure 5 shows some events: a collision of two transportation entities (car and truck), has resulted in an explosion, and the explosion has caused toxic gas and fire.

User Model

Although user model and world model are closely related, however, we define them separate to each other. Our definition about user model is a dynamic model of our users that are registered in our user database. The user information contains their identity, emotional states, interaction time, current communication media and modality(ies) and current location in the world. A user in our user model can be a part of actors in the world model, which is referred by its user id.

Figure 6 shows the class diagram of the user model. In a crisis situation, we define two types of users: civilians and professionals. We define a person as an independent entity of a user that stores most static information about the user. The dynamic information about users is stored in the class User. Every interaction with a user is archived in the class History. One or more communication channels are used by a user and a communication channel has one or more communication infrastructures. The communication infrastructure contains information about communication media and modalities.

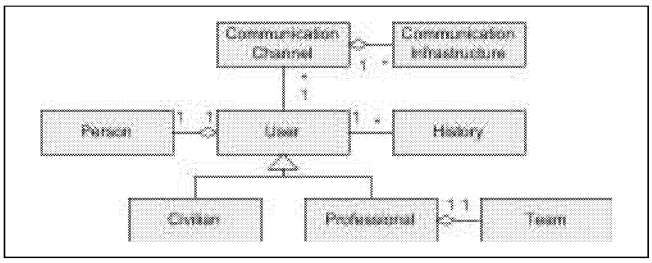


Figure 6 The class diagram of user model

Dialogue Model

In our multimodal HCI framework, the user system interaction flow is controlled by a DAM. The DAM defines dialogue acts and sends them to the multimodal output module. Together with the dialogue acts, the DAM gives information for the output module about the chosen output modalities to convey the message.

In the current implementation, we have 9 possible dialogue acts, such as: (1) statement: to notify the user about an event, (2) request: to command the user into a specific action, (3) ask: to request a specific information from the user, (4) confirmation: to ask the user to clarify a specific information, (5) acknowledge: to display incoming information on the user interface, (6) highlight: to highlight a certain informed information, (7) removal: to remove specific information from displays, and (8) affirmation or (9) negation: to notify the user whether the system agrees or not with his/her decision/action. The dialog acts are defined in the ontology as subclasses of a class "Process". A process contains information about its priority level, set of destination communication channels, timing, link to other processes, and reference to the world model. Figure 7 shows an example of a dialogue act message from the DAM to the output module.

```
…
<dialogueActType>statement</dialogueActType>
<priority><value>1</value></priority>
<communicationchannel>
  <communicationinfrastructure>
    <mode>short-message</mode>
    <mediatype>text</mediatype>
    …
  <communicationInfrastructure>
  …
</communicationchannel>
<concept>
  <name>Escort</name>
  <property>
    <name>agent</name>
    <value><instance id="WM01Police01"
           type="Police"/>
    </value>
  </property>
  <property>
    <name>patient</name>
    <value><instance id="WM01Paramedic01"
           type="Paramedic"/>
    </value>
  </property>
  <property><name>source</name>…</property>
  <property><name>destination</name>…</property>
</concept>
…
```

Figure 7 A dialogue act: "informing a user that a policeman is going to escort the paramedic from a source location to a destination"; the police and the paramedic are referred by an id in the world model

MULTIMODAL OUTPUT GENERATION

The output module receives the dialogue act from the DAM and the up-to-date world model for a specific user. Since the crisis is brought as the domain of the framework, the concept of the output module supports the rapid changes of a crisis situation. The generation of the output can be triggered by the change in the world model, not only by the selected actions from the DAM. The different components in our developed output module architecture (Figure 8) are explained below.

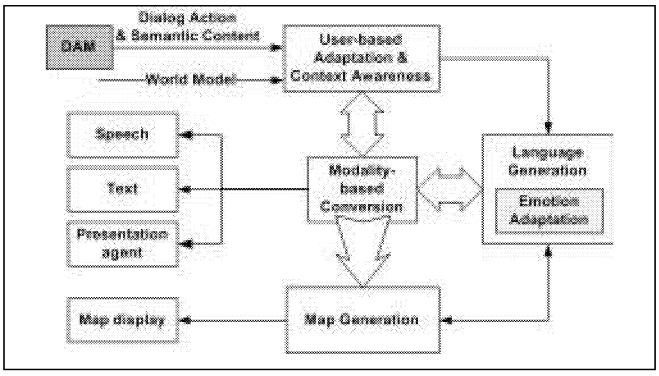


Figure 8 The architecture of the multimodal output module

User-based Adaptation

With direct access to three knowledge sources: task knowledge (dialogue acts - general ideas of how tasks should be done), user knowledge (who, where, what the emotion state of and what the communication device of the user), and world knowledge (world facts and events), the *user adaptation component* decomposes the dialog acts into presentation plans. This decomposition process includes contextual information (e.g. dialogue acts, user model, and user environment) and technical information (e.g. user preferences, output end-points constraint, and availability of modalities). Here, we took as an assumption the availability of real time context and technical information from a global representation while this component is actually planning the presentation tasks. The global representation includes characteristics of modality services of user communication devices based on the device capabilities. Using this knowledge, this component plans the presentation contents based on the following contextual information of the user.

User role. Some information may be private or confidential for a certain role, but public for others. For this purpose, the component uses predefined flags on each property of active concepts in the world model, such as public (i.e. available for all including general publics), protected (i.e. available for rescue workers and crisis room operators), and private (i.e. available for a certain rescue worker team or only crisis room operators), and filters the right messages for the right people.

User location. At any time, a user is associated with a location in the geographical space. At such time, the perspective view of a user may differ from the perspective area of the world model. Some information may become not important because it is far a way from the user. Therefore, the component uses a set of policies to locate the new perspective area of the world model based on the current user position. It is based on models of physical phenomena that include the dispersion of the crisis and its impact to other entities.

Available modalities. Knowing characteristics of modality services of user communication devices, the component uses a set of negotiation policies to negotiate and selects output modalities based on the modality choices from the DAM and input modalities used by the user. For example, if the input recognizer includes the body posture recognizer, the output modalities will include a large screen display, the map interface and speech; this policy is used for operators in a crisis room.

Interaction time. Using the user's dialogue history, the component selects the most up-to-date information to be presented to the user from the world model.

User emotion. The component passes this information to the language generation component.

Modality-based Conversion

The *modality-based conversion component* uses the results of the user adaptation component to design the presentation and allocate information across the language generation and the map display generation component. The DAM may include multiple concepts within a selected dialogue act. This component divides these concepts to be a set of sequence segments. All segments will be linked and synchronized with concepts that are processed in the language generation and map display generation. Other active concepts in the world model that are necessary to be displayed but are not linked to any language generation segment are passed to the map display generation component. These concepts will be displayed directly on the map interface. This component also

informs the map display generation component about those concepts that are necessary to be removed since they are not longer active.

A queue of presentation segments is processed by the language generation component. The modality-based conversion component then estimates the time needed for each processed segment. This depends on the chosen output modality. The component controls the system output using this timing information to create a full schedule for the turn of each segment to be generated. Each language segment will be generated one by one synchronized with the display (of the active or the highlighted concepts) on the map interface.

Language Generation

The *language generation component* works by simple dialogue act and concept-name recognition, and substitution of the property-names by their values controlled modified-AIML format. AIML (Wallace, 2003) is an extended-XML script that provides specifications for pattern matching and reply generation. The AIML also allows user-system dialogues focusing on a certain topic. In our modified version, the most important elements are:

- <aiml>, the tag that begins and ends an AIML document.
- <topic>, the tag that contains current dialogue topic pattern rule.
- <category>, the tag that marks a "unit of knowledge" in a dialogue.
- <dialogAct>, the tag that contains the dialogue act that matches to the DAM's selected action.
- <concept>, the tag that contains the concept name that matches the concept to be explained/reported.
- <properties>, the tag that contains the property names that are necessary for creating a complete text.
- <template>, the tag that contains the dialogue text. There are two types of templates: (a) short messages (e.g. for text generation in mobile phones) and (b) long messages.
- <concern>, the tag that marks user's emotion state.

Each category provides some possible templates that can be selected depending on the value of the <concern> tag. The same concept and dialogue act can have many categories, but different set of properties. The basic algorithm of text generation is the following:

- Identify the values of the dialogue act and the concept name, and some minimal context within a certain topic which the chosen values appears.
- Select a category by ensuring that the most specific known properties match first before any other categories or default (indicated by an asterisk "*").
- Choose an appropriate emotion state. If it matches, the system randomly selects one of the templates (for this emotion state).
- Use a selected template to construct the output text.

Inside the template, some <get> tags are used to be substituted by a value. The value can be generated (a) from the property of the chosen concept (e.g. `<get name = "source" type = "Location"/>`) and (b) by a function (e.g. `<get function = "getStaticObjectInfo (upperLeft, bottomUp" type = "String")>`). If the value type is an instance of a concept, the algorithm above will work recursively. Figure 9 shows an example of the text generation from some AIML units.

```
<aiml>
 <topic name="CAR ACCIDENT">
  <category>
   <dialogAct>statement</dialogAct>
   <concept name="Escort"/>
   <properties>
     <property name="patient"/>
     <property name="source"/>
     <property name="patient"/>
   </properties>
   <template type="short-message">
    <concern name="neutral" value=""/><random>
      <li>A <get name="agent" type="Class.name"/>
          will escort the
       <get name="patient" type="Class.name"/> from
       <get name="source" type="Location"/> to
       <get name="destination" type="Location"/>.
      </li>
      <li>A <get name="agent" type="Class.name"/>
       and a <get name="patient" type="Class.name"/>
       will come to
       <get name="destination" type="Location"/>
       from <get name="source" type="Location"/>
      </li>
     …
     </random></concern>
   …
   </template>
  </category>
  <category>
   <dialogAct>*</dialogAct>
   <concept name="Location"/>
   <properties>
     <property name="upperLeft"/>
     <property name="bottomUp"/>
   </properties>
   <template type="*">
      <get function="getStaticObjectInfo
       (upperLeft, bottomUp)" type="String"/>
   </template>
  </category>
 …
 </topic>
…
</aiml>
```

Example of resulted text:
"A police will escort a paramedic from A4 West BeneluxTunnel km 30 to Buttanweg no. 321 "

Figure 9 An example of the text generation from two AIML units; an asterisk ("*") means any

Map Display Generation

The *map display component* provides a representation of a map of the crisis event based on the selected perspective area (see *User-based Adaptation*). Since some concepts have a direct link to the icons on the user interface, the display of these concepts is considered as one-to-one mapping of the display of their correspondence icon on a certain location on a map. A function is used to map world coordinates to screen coordinates. This component works as follows:

Adding icons. The component displays all correspondence icons of active concepts by ensuring none of the instance of the concepts is displayed more than once.

Adding links between icons. An arrow is used to show a causality (i.e. a concept is activated because of another concept). The component uses the information from the

property "cause". A line is used to show a group of related icons. In the current implementation, a line is added for two conditions: (1) an icon is related to another icon if the former is the value of the latter's property and (2) two or more icons are related to each other if they are property values of an active concept that does not have a correspond icon, e.g. a concept "Passenger" has properties "vehicle" (whose type "Transportation") and "agent" (whose type "Actor"); this concept does not have any correspondence icon.

Remove icons and links. The component removes all correspondence icons (and their links), if the concepts are not longer active except those icons that are specified by the user. For such icons and links, only the user him/herself can remove them. This remove action can also be forced by the dialog act "remove".

Highlight concepts or certain locations on the map. This action is triggered by a dialog act "highlight". It is also used to support visually the information presented by the language component. The highlight is removed after some predefined time or forced by the dialogue act "remove".

To distinguish the user interaction and the system action on the graphical objects, three different colors are used: (1) black borders for the user's concepts, (2) green borders for the system's concepts and (3) thick yellow borders for highlighting concepts or locations. Figure 10 shows an example of a user's map display.

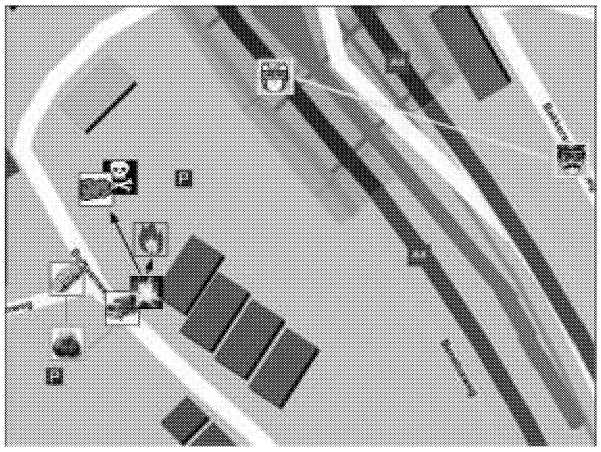


Figure 10 A display shown while informing the user that a policeman and a doctor will come to the crisis scene; the concept "Collision" (green bordered) is added to the user's workspace by the system; the information is retrieved from the world model

EXPERIMENTS

Unfortunately, the system cannot be tested in a real crisis, because (1) the AIML database is not yet sufficient, (2) it is difficult to create a controlled experiment of a disaster just to test the demonstrator system and (3) the current developed system can only handle one language, while in crises situations many languages can be used. Therefore, some preliminary experiments had to be done in simulated crises situations.

We have tested the first demonstrator in a crisis simulation scenario with participants playing certain roles and reporting back to the crisis center. The participants were divided into some groups of roles, such as firemen, paramedics, policemen and crisis center operators. They communicated to each other using different communication devices, such as a PDA, a tablet PC, a video phone and a large display. The scenario presented to respondents in the form of photographs and a map of the environment with explicitly information on what direction they are looking at. During the experiments, the respondents were asked to report what they might see, hear, or even feel and experience.

In the experiments we only focused on the integration of the HCI system using the developed framework. In particular to the output module, preliminary results show that the module is able to receive abstract representation of the world and dialog actions and generate appropriate multimodal responses based on available modalities and user contexts. However, some issues emerge during experiments, which can be classified in two categories:

- Users wanted to report about events, which were out of the system's domain.
- Not all functions of the module were tested. The combination of the unrealistic experiment setup and the respondents not being experienced reporters may have caused them to miss important events and showing only controlled emotions.

CONCLUSION

In this paper, we presented a module in a HCI framework that is able to produce and specify a context-sensitive and user-tailored multimodal information presentation. The output module can be integrated in a system built on the framework that is used as a supporting system for crisis management. With dialogue acts from a dialogue manager, the knowledge of user and the knowledge of the world, it can generate multimodal responses to the user-system interaction that are dynamic and dependent of the system's knowledge about the world and the observed emotion. The module is able to realize coherence and synchronized multimodal representation with constraints defined by the user's communication device and context information.

The developed language generation component is dialog act and concept-name matching and substitution of concept-property values approach. However, the order and timing in which a concept-property is activated cannot be determined beforehand, since the crisis situation may soon become chaotic and unmanageable. Because the module only generates language controlled by the AIML format, this becomes somewhat easier to manage. The format makes decisions and inferences easier to verify afterwards. Another advantage is that even though the concepts may be activated in a different order, the AIML-based dialogues may be designed and specified sequentially in order of the way users reporting about a crisis scenario. As a drawback, all AIML database have to be designed and specified in advance. This means that for the reporting to be fully expressive, all possible crisis scenarios have to be converted into the AIML database since reports about scenarios that have not been specified are not possible.

Using the geospatial graph for the modeling of the world, the output module is able to present the topographical information of the crisis event. By this way the developed framework can support crisis management teams who must collaboratively derive knowledge from geospatial information. Filtering presented information is still necessary to avoid redundancy displayed information and interfering user interaction.

A topic for further research will be the evaluation of the module performance in a system built based on the framework in real crisis exercise settings. This will allow us to determine how people use the systems developed, and in particular to assess whether they improve the efficiency of a given task using provided information.

REFERENCES

Bui T.D. (2004) *Creating Emotions and Facial Expressions for Embodied Agents*, Doctoral thesis, TU Twente, The Netherlands.

Bui T.H., Poel M., Nijholt A. and Zwiers J. (2007) A Tractable DDN-POMDP Approach to Affective Dialogue Modeling for General Probabilistic Frame-based Dialogue Systems, *Proc of IJCAI'07*, India.

Cassell J., Stocky T., Bickmore T., Gao Y., Nakano Y., Ryokai K., Tversky D., Vaucelle C. and Vilhjálmsson H. (2002) MACK: Media lab Autonomous Conversational Kiosk. *Proc. of Imagina '02*. Monte Carlo.

Cassell J., Vilhjálmsson H. and Bickmore T. (2001) BEAT: The Behavior Expression Animation Toolkit. *SIGGRAPH '01*, 477-486.

Dor R., *The Ear's Mind, a Computer Model of the Fundamental Mechanisms of the Perception of Sound*, tech. rep. 05-16, 2005.

Fitrianie S. and Rothkrantz L.J.M. (2005) Communication in Crisis Situations using Icon Language. *Proc. of IEEE ICME '05*, the Netherlands, 1370-1373.

Fitrianie S., Datcu D. and Rothkrantz L.J.M. (2006) Constructing Knowledge of the World in Crisis Situations using Visual Language, *Proc. of IEEE SMC'06*, 121-126.

Fitrianie S., Poppe R., Bui T.H., Chitu A.G., Datcu D., Dor R., Hofs D.H.W., Wiggers P., Willems D.J.M., Poel M., Rothkrantz L.J.M., Vuurpijl L.G. and Zwiers J. (2007) A Multimodal Human-Computer Interaction Framework for Research into Crisis Management, *ISCRAM 2007*, 149-158.

Foster M. E., White M., Setzer A. and Catizone R. (2005) Multimodal generation in the COMIC dialogue system. *Proc. of the ACL 2005*, 45-48.

Gustafson J., Bell L., Beskow J., Boye J., Carlson R., Edlund J., Granström B., House D. and Wirén M., (2000) Adapt - a Multimodal Conversational Dialogue System in an Apartment Domain, *ICSLP-2000*, (2)134-137.

Johnston M, Bangalore S., Vasireddy G., Stent A., Ehlen P., Walker M., Whittaker S. and Maloor P. (2002) MATCH: An Architecture for Multimodal Dialogue System, *Proc. of ACL*.

Johansson B., Fox A. and Winograd T. (2002) The Interactive Workspaces Project: Experiences with Ubiquitous Computing Rooms, *IEEE Pervasive Computing*. 1 (2): 67-74

Joshi A.K. and Vijay-Shanker K. (1999) Compositional Semantics for Lexicalized Tree-Adjoining Grammars, *Proc. of Computational Semantics*, The Netherlands.

Kjeldskov J. and Kolbe N. (2002) Interaction Design for Handheld Computers. *Proc. of* APCHI'02, Science Press, China.

Maybury M. (1999) Intelligent User Interfaces: An Introduction. *Proc. of* IUI '99. ACM, New York, 3-4.

Moore L.K. (2006) *CRS Report for Congress: Public Safety Communication Policy*. Confessional Research Service, the Library of Congress, USA.

Norman D. (1993) *Things That Make Us Smart*. Addison-Wesley Publishing Co.

Oviatt S., Coulston R. and Lunsford R. (2004) When Do We Interact Multimodally?: Cognitive Load and Multimodal Communication Patterns. *Proc. of ICMI '04*. ACM, 129-136.

Perlovsky L.I. (1999) Emotions, Learning and Control. Proc. of International Symposium: Intelligent Control, Intelligent Systems and Semiotics, 131-137.

Sharma R., Yeasin M., Krahnstoever N., Rauschert I., Cai G., Brewer I., MacEachren A.M. and Sengupta K. (2003) Speech-Gesture Driven Multimodal Interfaces for Crisis Management, Proc. of the IEEE, 91(9): 1327-1354.

Steedman M. and Baldridge J. (2005) Combinatory Categorial Grammar, in: R. Borsley and K. Borjars (eds.) *Non-Transformational Syntax*, Blackwell.

Wahlster W., Reithinger N. and Blocher A. (2001) SmartKom: Multimodal Communication with a Life-Like Character, *Proc. of Eurospeech'01*, Denmark.

Wallace, R. S. (2003). The elements of AIML style, A.L.I.C.E. Artificial Intelligence Foundation, inc.

W3C, *OWL: Ontology Web Language*, http://www.w3.org/TR/owlguide/.

BIOGRAPHY

SISKA FITRIANIE was born at Bandung, Indonesia and went to Delft University of Technology, the Netherlands, where she studied Technical Informatics and obtained her master degree in 2002. After doing her two years post-graduate programme at Eindhoven University of Technology, she involves in the Interactive Collaborative Information Systems (ICIS) project, supported by the Dutch Ministry of Economic Affairs, since 2004 as her PhD project at Delft University of Technology. Her project aims at designing and developing models of a computer-human interaction system.

E-mail: s.fitrianie@ewi.tudelft.nl
Webaddress: http://mmi.tudelft.nl/~siska

LEON J.M ROTHKRANTZ received the M.Sc. degree in mathematics from the University of Utrecht, Utrecht, The Netherlands, in 1971, the Ph.D. degree in mathematics from the University of Amsterdam, Amsterdam, The Netherlands, in 1980, and the M.Sc. degree in psychology from the University of Leiden, Leiden, The Netherlands, in 1990. He is currently an Associate Professor with the Man-Machine-Interaction Group, Mediamatics Department, Delft University of Technology, Delft, The Netherlands, since 1992. His current research focuses on a wide range of the related issues, including lip reading, speech recognition and synthesis, facial expression analysis and synthesis, multimodal information fusion, natural dialogue management, and human affective feedback recognition. The long-range goal of his research is the design and development of natural, context-aware, multimodal man–machine interfaces. Drs. Dr. Rothkrantz is a member of the Program Committee for EUROSIS.

E-mail: l.j.m.rothkrantz@ewi.tudelft.nl
Webaddress: http://mmi.tudelft.nl/~leon

MEDICAL IMAGING

BIOMECHANICAL ANALYSIS OF THE HUMAN MIDDLE EAR

Fernanda Gentil (fernanda.fgnanda@gmail.com)
IDMEC-Polo FEUP – R.Dr. Roberto Frias, Porto
ESTSP – Pr. Coronel Pacheco 15, Porto
Widex – Av. França 20, 3º, 307, Porto

Renato Natal Jorge* (rnatal@fe.up.pt)
Marco Parente (parente@fe.up.pt)
Pedro Martins (palsm@fe.up.pt)
Fátima Alexandre (mifa@fe.up.pt)
IDMEC-Polo FEUP – R.Dr.Roberto Frias, Porto

António Ferreira (ferreira@fe.up.pt)
INEGI – R.Barroco 174, Leça do Balio

Eurico Almeida (clinicaorlea@telepac.pt)
Clinica ORL – Av. Boavista 117, 6º, Porto

(*) author for correspondence

KEYWORDS
Biomechanical, 3D reconstruction, Finite Elements, Middle ear.

ABSTRACT

The human ear is a complex biomechanical system and is divided in three parts: external, middle and inner ear. The middle ear is formed by three ossicles (malleus, incus and stapes), ligaments and muscles that amplify the sound waves sending them to the inner ear.
In this work, a finite element modelling of the middle ear was made. For this purpose, we show an approximate 3D solid model of the ossicles and eardrum, based in imagiology. Ligaments and tendons with hyperelastic behaviour were include, using the Yeoh model.
With ABAQUS program, the discretization of these components was done. The connection between ossicles was made using contact formulation. With this model it is possible to do static and dynamic studies for better understand the ear behaviour and posteriorly to obtain improved solutions for hearing problems.

INTRODUCTION

When, for some reason, one or more parts of the hearing system do not work properly, it can result in hearing loss reflecting a serious health problem.
There are three different types of hearing loss: conductive, sensorineural and mixed (Paço 2003).
The conductive hearing loss happens when there is a problem in the external or in the middle ear. As examples of situations causing conductive hearing loss we can refer otitis, eardrum perforations, ear wax, tympanosclerosis or otosclerosis. At Otosclerosis, the hearing loss results from fixation of the stapes, donc this correction can be made by improving the application of mechanical prosthesis to replace partial or totality the ossicle, by surgery. The sensorineural hearing loss happens when inner ear is not working correctly. As examples we have presbyacusis, Ménière´s disease, ear tumors, ototoxic drugs, or exposure to loud noise. Finally, the mixed hearing loss combines conductive and sensorineural ones. Hearing losses can be corrected by medicine and/or surgery or hearing aids to guarantee wave sound travel.

In order to analyse the sound wave travel in the middle ear and its modulation nearly of real, it is very important to achieve a correct modelling of the vibro-acoustic behaviour of the middle ear.
Based in imagiology, a model of a normal human middle ear was obtained with an approximate 3D solid model of the ossicles and eardrum. The discretization of these components was carried out using the finite element method. The numerical simulation was achieved with the ABAQUS program, considering mechanical properties available in the literature (Prendergast et al. 1999; Sun et al. 2002). The connection between ossicles malleus/incus and incus/stapes was made using contact formulation (Gentil et al. 2007). The model includes ligaments and tendons with hyperelastic behaviour (Martins et al. 2006), using the Yeoh model (Holzapfel 2000). The boundary conditions on the ligaments, tendons, eardrum and stapes footplate were imposed.
Considering different acoustic pressure values applied on the eardrum, static and dynamic studies are possible and the results are obtained and compared with others, with the sense to improve the middle ear understanding. It is important to simulate different middle ear pathologies, as otitis media, eardrum perforations (Gentil et al. 2008), myringosclerosis, tympanosclerosis, or otosclerosis, and compare the obtained outcomes with the results for a normal ear.

ANATOMO-PHISIOLOGIE OF THE EAR

The hearing is the ability to perceive sounds by detecting vibrations from external ear, converting them into electrical signals in the auditory nerve, that are perceived by the brain.
The ear has three main parts: the external, middle, and inner ear. The external ear is formed by the auricle and external auditory canal. The eardrum, a thin membrane, divides the external from middle ear and transforms the acoustical energy into mechanical energy, which is moved along the ossicles to the inner ear. The middle ear is a cavity with air, which confines three small bones, (malleus, incus and stapes), ligaments (superior, lateral and anterior of malleus, superior and posterior of incus and a annular ligament of stapes) and muscles (tensor tympani and stapedius) that are involved in sound conduction (Paparella 1982). In the oval window, stapes footplate converts mechanical energy into electrical one and transfers it to the cochlea, in the inner ear,

which trigger the generation of nerve signals that are sent to the brain, by auditory nerve, called the 8th cranial nerve.

BUILDING THE FINITE ELEMENT MODELLING

Computed tomography (CT) is a medical imaging method where digital geometry processing can be used to generate a 3D image of the internals of an object from a large series of 2D X-ray images taken around a single axis of rotation. In this work, CT scan images of the middle ear, obtained from a 65 years old woman, with normal hearing, were used (Figure 1).

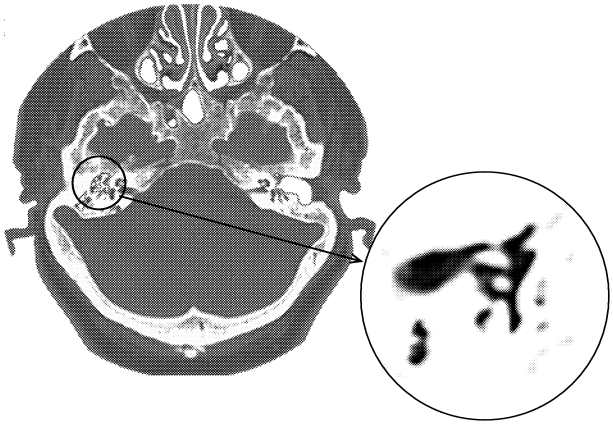


Figure 1: 2D visualization of TC image of middle ear

In order to eliminate the noise verified in slices of the images, the volume was filtered prior to further image analysis. The Analyze software was used to obtain noise reduction being utilized a tool image edit based on manual segmentation process and rendering.

Due to these images and structures were not sufficient distinct in the original data, manual options were used. This methodology was applied often to obtain more detail, but the resolution of the DICOM (Digital Imaging and Communications in Medicine) images was so poor that was quite impossible to have a clear shape of the desired structures (ossicles and eardrum).

Copy of many slices was performed, one by one, with Analyze software tool. Each slice was transformed to bitmap image format (BMP). Those images were opened in a CAD software in same order that it were saved. The reconstruction of 3D objects process that was adopted is based upon their transversal section (Alexandre et al. 2006):

1. Pre-processing of the 2D images. The steps were adopted within the methodology established. The scope of pre-processing is to extract, for each slice, a collection of representative object boundary points. A triangulation process is then made, based on that collection of points between consecutive slices (Folowosele et al. 2004).

2. Surface reconstruction between outlines and Stereo Lithography (STL) data generations files. This representation is obtained by connecting the object boundaries through triangular elements. The triangularization process generates a collection of triangular patches between consecutive pairs of contours, forming a precise approximation to the original object surface. This procedure is able to handle cases where several contours in each slice must exist, known as multiple branching problems. Basically, the reconstruction process consists of obtaining a 3D representation of the object under investigation (malleus, incus and the eardrum), allowing not only its visualization, but also a more detailed comprehension of its structure through the analysis of the geometric parameters of the object (Figure 2).

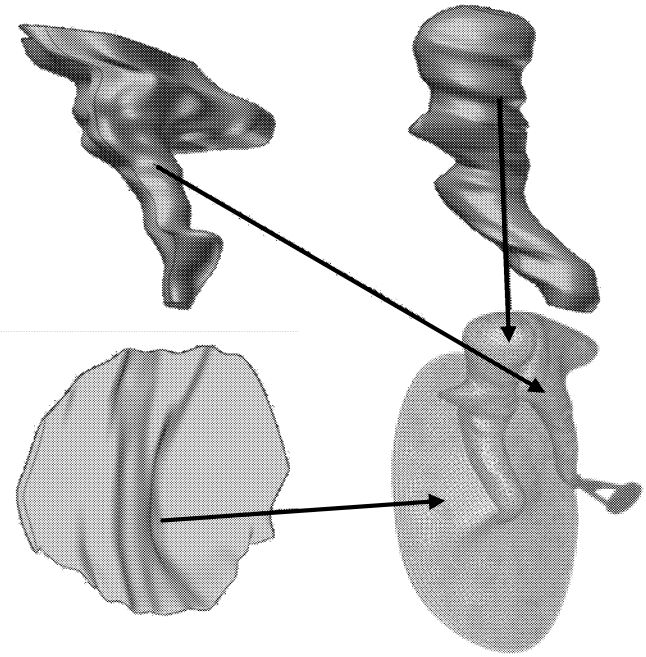


Figure 2: 3D representation of: incus (top left); malleus (top right); eardrum (down left); reconstruction of the middle ear (down right).

3. Visualization of the generated images. After the reconstruction, the new 3D models must be converted to a standard format that can be read by the finite element program. The STL format was used in this work because, first of all, it is an open-source format and an easy format to handle and manipulate.

Once generated, the STL model was read by FEMAP (commercial pre and post processing finite element program). The FEMAP software has the ability to generate different layers and also to rebuild the incomplete 3D elements (Folowosele et al. 2004). Afterward, that archive can be used directly into a fast prototype machine to obtain the prototype object, if that is the intention.

After obtained the solid model of the ossicles and eardrum, the discretization of these components was made using tetrahedral solid elements for the ossicles (C3D4) and hexahedral for the eardrum (C3D8) (Figure 3).

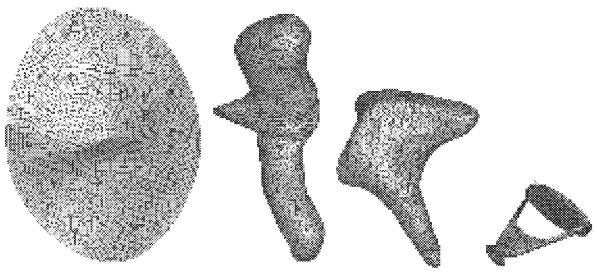


Figure 3: Finite element mesh of eardrum, malleus, incus and stapes, respectively.

The model includes three ligaments of malleus (superior, lateral and anterior), two of incus (superior and posterior), the annular ligament of the stapes, and two tendons (tensor tympanic and stapedius). These ligaments and tendons were modelled by linear elements (T3D2). The simulation of cochlear fluid was obtained by fluid elements (F3D3) (Figure 4).

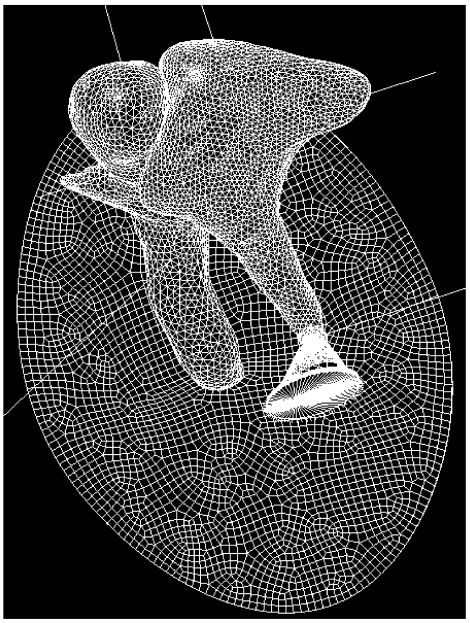


Figure 4: Finite element mesh of complete model of the middle ear.

Hyperelastic non-linear behaviour of these ligaments and tendons was taken into account (Martins et al. 2006), based on the Yeoh model (Holzapfel 2000).

The strain-energy function, ψ, for the Yeoh model can be written like a Equation (1) above.

$$\psi = c_1 (I_1 - 3) + c_2 (I_1 - 3)^2 + c_3 (I_1 - 3)^3 \qquad (1)$$

where I_1 is the first right Cauchy-Green tensor invariant and c_1, c_2 and c_3 are the material constants based in work of Martins et al. (2006).

The numerical simulation was achieved with the ABAQUS program, considering mechanical properties available in the literature (Prendergast et al. 1999; Sun et al. 2002). The connection between ossicles malleus/incus and incus/stapes, simulating incudomallear and incudostapedial joints, respectively, was made using contact formulation (Gentil et al. 2007). The boundary conditions include tympanic annulus (around the eardrum), stapedius annular ligament (around the stapes) and the suspensory ligaments and tendons.

CONCLUSIONS

This work presented a methodology of surface 3D reconstruction of the ossicles and the eardrum based on transversal sections of CT images. These 3D models can give a better comprehension of complex human anatomy, as the case for the middle ear. They also give us a kind of detail that is impossible to obtain with ordinary computer images.

After finite element modelling is done it is possible to achieve static and dynamic studies, about the human middle ear behaviour. The eigenvalues and the eigenvectors can be carried out. Considering different acoustic pressure values on the eardrum, the results can also be obtained. The present model allowed obtain harmonic responses of the ligaments tractions (Gentil et al. 2006), as well as the action and effects of the muscles. Different pathologies can be compared with a normal ear with the purpose of obtaining better clinical solutions.

Substantial improvements can be achieved in the prototype and prosthesis area, because the obtained models are now much more close to the real geometric dimensions of the bones of the middle ear.

ACKNOWLEDGEMENTS

The authors truly acknowledge the funding provided by Ministério da Ciência, Tecnologia e Ensino Superior – Fundação para a Ciência e a Tecnologia (Portugal) and by FEDER, under grant PTDC/EME-PME/81229/2006.

REFERENCES

ABAQUS Analyses User´s Manual. 2007. Version 6.5.

Alexandre F.; A.A. Fernandes; R.M. Natal Jorge; F. Gentil; P.A.L.S. Martins; T. Mascarenhas; C. Milheiro; A.J.M. Ferreira; and M.P.L. Parente. 2006. “3D Reconstruction of the Middle Ear for FEM Simulation”. *Simpósio Internacional CompIMAGE - Computational Modelling of Objects* Represented in Images: Fundamentals, Methods and Applications, JMRS Tavares, RM Natal Jorge (Eds.), 181-184, Coimbra.

Folowosele, F.O.; J.J. Camp; R.H. Brey; J.I. Lane; and R.A. Robb. 2004. “3D imaging and modeling of the middle and inner ear”. *Biomedical Imaging Resource*, *Audiology, Radiology*, Mayo Clinic, Rochester, MN.

Gentil F.; R.M. Natal Jorge; A.J.M. Ferreira; M.P.L. Parente; P.A.L.S. Martins; and E. Almeida. 2006. “Biomechanical simulation of middle ear using hyperelastic models”. *5th World Congress of Biomechanics*, Rik Huiskes, Farshid Guilak, (Eds.), CD-Rom (REF: 6035), Munique, Alemanha.

Gentil, F.; R.M. Natal Jorge; A.J.M. Ferreira; M.P.L. Parente; M. Moreira; and E. Almeida. 2007. “Estudo do efeito do atrito no contacto entre os ossículos do ouvido médio”. *Revista Internacional de Métodos Numéricos para Cálculo y Diseño en Ingeniería*. Vol. 23, 2, 177-187.

Gentil, F.; R.M. Natal Jorge; A.J.M. Ferreira; M.P.L. Parente; M. Moreira; and E. Almeida. 2008. “A static and dynamic study of the middle ear with different sizes of eardrum perforations”. In *8th International Symposium on Computer Methods in Biomechanics and Biomedical Engineering*, Porto.

Holzapfel, G.A. 2000. *Nonlinear solid mechanics*. John Wiley & sons, Ltd., New York.

Martins, P.A.L.S., R.M. Natal Jorge, and A.J.M. Ferreira. 2006. “A Comparative Study of Several Material Models for Prediction of Hyperelastic Properties: Application to Silicone-Rubber and Soft Tissues”. *Strain*, 42, 135-147.

Paço, J. 2003. *Doenças do tímpano.* Lidel, Lisboa.

Paparella, M.M. and D.A. Shumrick. 1982. *Otorrinolaringologia.* 2ª ed. Editora Médica Panamericana SA, Buenos Aires, 196-212.

Prendergast, P.J.; P. Ferris; H.J. Rice; and A.W. Blayney. 1999. “Vibro-acoustic modeling of the outer and middle ear using the finite element method”. *Audiol Neurootol*, 4, 185-191.

Sun, Q.; R.Z. Gan; K.H. Chang; and K.J. Dormer. 2002.

"Computer-integrated finite element modeling of human middle ear". *Biomechanics and Modeling in Mechanobiology*, 1, 109-122.

SEGMENTATION AND SIMULATION OF OBJECTS IN PEDOBAROGRAPHY IMAGES USING PHYSICAL PRINCIPLES

Patrícia C.T.Gonçalves, João Manuel R.S.Tavares and R.M.Natal Jorge
Faculty of Engineering of the University of Porto
Rua Dr. Roberto Frias
4200-465 Porto, Portugal
E-mail: {patricia.goncalves, tavares, rnatal}@fe.up.pt

KEYWORDS
Segmentation, simulation, deformable models, finite elements method, equilibrium equation, pedobarography.

ABSTRACT

The goals of the present work are to automatically extract the contour of an object and to simulate its deformation using a physical approach. Thus, to segment an object represented in an image, an initial contour is manually defined for it that will then automatically evolve until it equals the border of the desired object. The contour is modelled by a physical formulation, and its evolution to the desired final contour is driven by internal and external forces. To build the physical model of the contour used in the segmentation process, we adopted the isoparametric finite element proposed by Sclaroff, and to obtain its evolution towards the object border we used the methodology presented by Nastar that consists in solving the dynamic equilibrium equation between two consecutive instants.
As for the simulation of the deformation between two different instances of an object, or between two objects, after their contours have been properly modelled, modal analysis, complemented with global optimization techniques, is employed to establish the correspondence between their nodes (data points). After the matching phase, the displacements field between the two contours is simulated using the dynamic equilibrium equation.
The proposed approach will be here considered in dynamic pedobarography images.

INTRODUCTION

In the domain of Computational Vision, the identification of an object represented in an image – segmentation – is one of the most common and complex tasks. Usually, whenever it is intended to extract higher-level information from an image or even from image sequences, the used image analysis process starts by segmenting the input image(s). Thus, image segmentation is one of the working areas in Computational Vision with more research done and so it will probably continue to be throughout the times.
All real objects are deformable; that is why most of them cannot be accurately modelled if they are considered as rigid bodies. The simulation of non-rigid objects may be achieved using deformable models, which can be a challenging task because different application areas have different requirements; for instance, some require accuracy, like medical image analysis, and others require real time interactivity, like virtual environments. However, simulating the deformation of objects in a fast and accurate way is not an easy task.
The main goal of this work is the following: having two images of the same object in two distinct instants, or of two related objects, we want to simulate the intermediate shapes between the two. For that, we need to segment the object in each input image by extracting its contour after manually defining an initial contour for it. Each one of these coarse contours will then evolve throughout an iterative process until it reaches the border of the desired object. For that purpose, a deformable model is built for each of the contours using the finite elements method, namely the isoparametric finite element proposed by (Sclaroff 1995). Thereafter, each model will behave according to physical principles, as proposed by (Nastar 1994), using the dynamic equilibrium equation.
To simulate the deformation between two input contours, we use the approach proposed by (Terzopoulos et al. 1987) and (Terzopoulos et al. 1988) to do realistic deformation simulations considering an elastic model based on the resolution of the dynamic equilibrium equation. Thus, having the physical model for each contour, we consider: Shapiro and Brady's modal shape description to match their nodes (Shapiro and Brady 1992) complemented with an optimization search technique as proposed by (Bastos and Tavares 2006, Tavares and Bastos 2005); and Pentland and Horowitz' decomposition of object deformation into rigid and non-rigid vibration modes (components) (Pentland and Horowitz 1991).
In this paper, we propose a solution to apply this physical approach to contours that do not have all nodes successfully matched.
The first step in our methodology consists in drawing a rough contour on the two input images. Those shapes are considered as the initial segmentation contours for those objects. Next, the contours are modelled according to physical principles using the isoparametric finite element proposed by Sclaroff. To move the physical model towards the border of the object to segment, the dynamic equilibrium equation is solved, that describes the equilibrium between the internal and external forces involved. The internal forces are defined by the physical characteristics adopted for the model, determined by the adopted virtual material and the selected level of interaction between the nodes of the model. The external forces are computed by enhancing particular features of the object in the input images; namely, intensity, edges and the distance from each pixel to the nearest edge. After the extraction of the two contours, the nodes of their physical models are matched and the deformation of one into the other is simulated by solving the dynamic equilibrium equation.

PHYSICAL MODELLING

After defining the initial contour for the object to segment, it is time to computationally model it in physical terms; that is, to assign mass, stiffness and damp to each point of the contour, i.e., to each node of the used model.
To model the initial contour and simulate its elastic behaviour, (Nastar 1994) used affine interpolation functions together with finite differences. Instead, Gaussian interpolants

and the finite elements method are used in this work for the same purpose. To be precise, the isoparametric finite element proposed by (Sclaroff 1995) is considered to build the physical model. This finite element uses a set of radial base functions that allows an easy insertion of the data points in the model. Therefore, Gaussian interpolants are used and the nodes of the model do not need to be ordered. With this isoparametric finite element, when an object is modelled it is as if each of its feature points are covered by an elastic membrane (Sclaroff and Pentland 1995, Tavares et al. 2000).
Thus, starting with a collection of m sample points $X_i(x_i,y_i,z_i)$ of the object to be physically modelled, the interpolation matrix **H**, which relates the distances between object nodes and their interrelations, of Sclaroff's isoparametric finite element (Sclaroff 1995, Sclaroff and Pentland 1995) is built using:

$$g_i(\mathbf{X}) = e^{-\|\mathbf{X}-\mathbf{X}_i\|^2/2\sigma^2},$$

where σ is the standard deviation that controls the nodes interaction. Then, the interpolation functions, h_i, are given as:

$$h_i(\mathbf{X}) = \sum_{k=1}^{m} a_{ik} g_k(\mathbf{X}),$$

where a_{ik} are coefficients that satisfy $h_i = 1$ at node i and $h_i = 0$ at the other m-1 nodes. These interpolation coefficients compose matrix **A** and can be determined by inverting matrix **G** defined as:

$$\mathbf{G} = \begin{bmatrix} g_1(\mathbf{X}_1) & \cdots & g_1(\mathbf{X}_m) \\ \vdots & \ddots & \vdots \\ g_m(\mathbf{X}_1) & \cdots & g_m(\mathbf{X}_m) \end{bmatrix}.$$

Consequently, matrix **H** will be:

$$\mathbf{H} = \begin{bmatrix} h_1 & \cdots & h_m & 0 & \cdots & 0 \\ 0 & \cdots & 0 & h_1 & \cdots & h_m \end{bmatrix},$$

and the mass matrix of Sclaroff's isoparametric element is defined as (Sclaroff 1995, Sclaroff and Pentland 1995):

$$\mathbf{M} = \begin{bmatrix} \mathbf{M'} & 0 \\ 0 & \mathbf{M'} \end{bmatrix},$$

where $\mathbf{M'}$ is a sub-matrix $m \times m$ defined as $\mathbf{M'} = \rho\pi\sigma^2 \mathbf{A}^{\mathrm{T}} \mathbf{\Gamma} \mathbf{A} = \rho\pi\sigma^2 \mathbf{G}^{-1} \mathbf{\Gamma} \mathbf{G}^{-1}$, as matrix **A** is symmetric $\mathbf{A}^{\mathrm{T}} = \mathbf{A}$, ρ is the mass density, and the elements of matrix $\mathbf{\Gamma}$ are the square roots of the elements of matrix **G**.
On the other hand, the element stiffness matrix is given by:

$$\mathbf{K} = \begin{bmatrix} \mathbf{K}_{11} & \mathbf{K}_{12} \\ \mathbf{K}_{21} & \mathbf{K}_{22} \end{bmatrix},$$

where $\mathbf{K}_{ij}$ are symmetric $m \times m$ sub-matrices depending on the constants α, β and λ that are functions of the virtual material adopted for the object (Sclaroff 1995, Sclaroff and Pentland 1995) defined as:

$$K_{11_{ij}} = \pi\beta \sum_{k,l} a_{ik} a_{jl} \left[\frac{1+\lambda}{2} - \frac{\hat{x}_{kl}^2 + \lambda\, \hat{y}_{kl}^2}{4\sigma^2} \right] \sqrt{g_{kl}},$$

$$K_{22_{ij}} = \pi\beta \sum_{k,l} a_{ik} a_{jl} \left[\frac{1+\lambda}{2} - \frac{\hat{y}_{kl}^2 + \lambda\, \hat{x}_{kl}^2}{4\sigma^2} \right] \sqrt{g_{kl}},$$

$$K_{12_{ij}} = K_{21_{ij}} = -\frac{\pi\beta(\alpha+\lambda)}{4\sigma^2} \sum_{k,l} a_{ik} a_{jl} \hat{x}_{kl} \hat{y}_{kl} \sqrt{g_{kl}},$$

where $\hat{x}_{kl} = x_k - x_l$, $\hat{y}_{kl} = y_k - y_l$, $\hat{z}_{kl} = z_k - z_l$.
In this work, we use Rayleigh's damping matrix, **C**, which is a linear combination of the mass and stiffness matrices with constraints, μ and γ, based upon the chosen critical damping (Bathe 1996, Cook et al. 1989):

$$\mathbf{C} = \mu\mathbf{M} + \gamma\mathbf{K}.$$

MATCHING THE OBJECTS NODES

After having extracted the contours of the objects in the input images, called initial and target objects from here on, we need to find the correspondences between their nodes. For that, a generalized eigenvalue/eigenvector problem is solved for each object using:

$$\mathbf{K\Phi} = \mathbf{M\Phi\Omega},$$

where $\mathbf{\Phi}$ is the modal matrix of the shape vectors, the eigenvectors, which describe the modal displacement (u,v) of each node due to vibration mode i, and $\mathbf{\Omega}$ is the diagonal matrix whose entries are the squared eigenvalues increasingly ordered.
After building the modal matrix for each object, the nodes of the two objects can be matched comparing their displacements in the respective modal eigenspace (Shapiro and Brady 1992). The main idea of modal matching is that low order modes of two similar shapes will be very close even in the presence of affine transformation, non-rigid deformations, local shape variations or noise. Thus, to match the nodes of the initial object, I, with the ones of the target object, T, an affinity matrix, **Z**, is built with elements defined as:

$$Z_{ij} = \left\| u_{I,i} - u_{T,j} \right\|^2 + \left\| v_{I,i} - v_{T,j} \right\|^2, \qquad (1)$$

where the affinity between nodes i and j is 0 (zero) if the match is perfect, and increases as the match worsens.
In this work, two search methods are considered to find the best matches: a local method and a global one. The local search method was proposed by (Shapiro and Brady 1992), and consists in searching each row and each column of the affinity matrix for their lowest values. If the lowest value of row i is in column j, and that value is also the lowest of its column, then node i of the initial object matches node j of the target one. This procedure has the main disadvantage of disregarding the objects structure as it searches for the best match for each node.
On the other hand, the global search method proposed by (Bastos and Tavares 2006, Tavares and Bastos 2005) consists in describing the matching problem as an assignment problem, and solving it using an appropriate optimization algorithm. In this matching approach, cases in which the number of nodes in the initial and target objects is different can also be considered: initially the global matching algorithm adds fictitious nodes to the object with fewer ones, and then the nodes that are matched with the fictitious elements are

adequately matched with real nodes using neighbourhood and affinity criteria.

EQUILIBRIUM EQUATION

After having the initial contour transformed into an elastic physical model we need to estimate its evolution in the direction of the object edges to achieve the desired segmentation; and after having both contours extracted from the initial and target objects we need to simulate the deformation of one into the other. To achieve both of these goals, the second order ordinary differential equation, commonly known as Lagrange's dynamic equilibrium equation, is considered:

$$\mathbf{M}\ddot{\mathbf{U}}^t + \mathbf{C}\dot{\mathbf{U}}^t + \mathbf{K}\mathbf{U}^t = \mathbf{F}^t, \qquad (2)$$

for each time step t, where $\mathbf{U}$, $\dot{\mathbf{U}}$ and $\ddot{\mathbf{U}}$ are, respectively, the displacement, velocity and acceleration vectors, and $\mathbf{F}$ represents the external forces (Gonçalves et al. 2006, Pinho and Tavares 2004). This equation describes the equilibrium between the internal and external forces involved on the model nodes. The internal forces are defined by the physical characteristics adopted for the model, determined by the adopted virtual material and the level chosen for the interaction between its nodes, which is considered while building Sclaroff's isoparametric finite element. The external forces depend on whether we are dealing with the segmentation or the simulation of objects deformation.

External forces for the segmentation

To segment an object, the external forces, $\mathbf{F}$, are determined by the image features that best describe the object to segment. In this work we consider the intensity value of each pixel of the initial image, the value of the pixels of the edges image, and the distance from each pixel to the nearest edge. Thus, $\mathbf{F}$ is the sum of the forces due to the edges image, $\mathbf{F}_{edg}$, the intensity original image, $\mathbf{F}_{int}$, and the distance image, $\mathbf{F}_{dist}$:

$$\mathbf{F} = \mathbf{F}_{edg} + \mathbf{F}_{int} + \mathbf{F}_{dist}.$$

Here, the edges image is obtained by applying Shen & Castan's edge detection operator (Shen and Castan 1992) to the original image, and the distance image is obtained by calculating the distance of each pixel to its nearest edge using Chamfer's method.

After the physical modelling of the initial contour defined by the user, our algorithm calculates the line orthogonal to the line tangent to the contour at each node of the model. It is along each one of these lines that the external forces are calculated. Denoting as Q_i all the pixels belonging to the orthogonal line of node P, the edges force at point P is:

$$\mathbf{F}_{edg}(P) = k\frac{\sum_{i=1}^{N} Edg(Q_i)}{N}, \qquad (3)$$

where $Edg(Q_i)$ is the value of the pixel Q_i in the edges image, k is a stiffness constant and N is the number of pixels of the orthogonal line. The intensity and distance forces equations are similar to Equation (3).

If the mean of the edges values of the N pixels of the orthogonal line is lesser than a given threshold value, *val*, then the line will continually grow until the mean reaches *val*. Hence, each line has the length it needs to determine a sufficient force to move its associated node; that is, it has and adaptative lenght.

External forces for the simulation

To estimate the external force applied on each matched node i of the initial object, we consider that the force on each node is proportional to its associated displacement, as proposed by (Pinho and Tavares 2004):

$$\mathbf{F}(i) = q\left(\mathbf{X}_{T,i} - \mathbf{X}_{I,i}\right),$$

where $\mathbf{F}(i)$ is the force applied on node i, $X_{I,i}$ the coordinates of node i in the initial object, $X_{T,i}$ the coordinates of the node corresponding to node i in the target object and q is a global stiffness constant. Because this equation is updated after each iteration of the resolution of the dynamic equilibrium equation, its generalized form is:

$$\mathbf{F}(i) = q\left(\mathbf{X}_{T,i} - \mathbf{X}_{J,i}\right), \qquad (4)$$

where $X_{J,i}$ represents the coordinates of node i in the object shape obtained in the Jth iteration.

However, some nodes of the initial object may not be successfully matched with any of the nodes in the target object. To overcome this problem, suppose that b is an unmatched node between nodes a and c, matched with nodes a' and c' of the target object, respectively (Figure 1). Therefore, if b is the ith node in the Jth shape, then the ith component of the external force vector can be given by:

$$\mathbf{F}(i) = q\left(\sum_{p\ (\text{nodes between } a' \text{ and } c')}\left[W_p\left(\mathbf{X}_{F,p} - \mathbf{X}_{J,b}\right)\right]\right), \qquad (5)$$

where W_p is the weight of node p, according to its matching affinity with node b provided by Equation (1) – thus, the higher the matching affinity value, the lower the weight.

If there are no unmatched nodes between nodes a' and c', then a' and c' will be considered in the computation of the external force on node b.

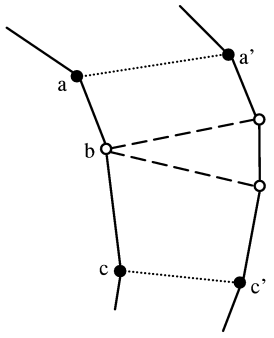


Figure 1: Estimation of the external forces applied on unmatched nodes.

EXPERIMENTAL RESULTS

Pedobarography is the measurement of dynamic variations in downward pressure on different areas of the foot sole, using a pedobarograph system (a device that records dynamic variations as a person stands upright or walks). The recording of pedobarographic data along the duration of a step in normal walking conditions permits the dynamic analysis of the foot behaviour, being an important tool for medical diagnosis and therapy planning (Pinho and Tavares 2004, Tavares et al. 2000).

To illustrate the results of the methodology here proposed to segment an object represented in an image by identifying its contour, consider the pedobarography images of the same foot taken in two different instants of a step, shown in Figure 2. In the first two images we can see, in red, the initial contour manually defined for the object. The next two images represent the segmentation obtained using a physical model made of rubber with 45 nodes, in the first case, and 39 nodes in the last one, and considering k=3,000N/m. The computational process took 5s and 4s, respectively, to achieve the final result. In this work we used a personal computer with an Intel Pentium D at 3GHz processor and 2GB of RAM.

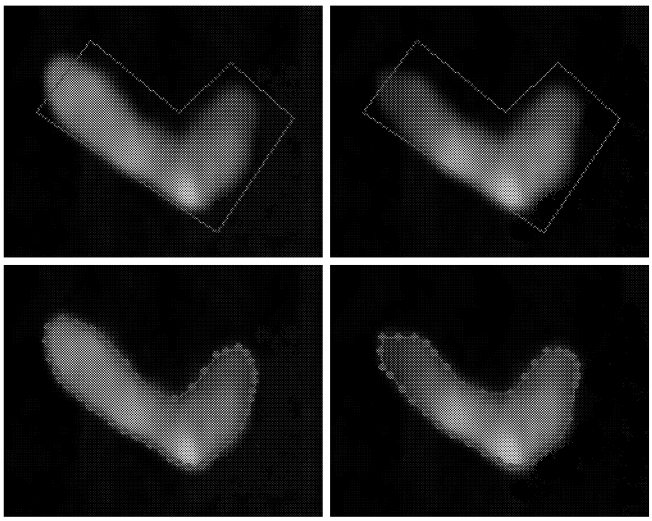


Figure 2: Two input images with the initial contours defined (top) and with the segmentation process results using k=3,000N/m and a model made of rubber (bottom).

Using the modal matching method based on global search to determine the correspondences between the nodes of the contours shown in Figure 2, all the 45 nodes of the initial object are successfully matched to the 39 of the target object, Figure 3. Adopting q=30,000N/m to calculate the external forces involved, the intermediate shapes in Figure 3 are estimated after 50s using Equation (2).

Figure 3: Matching found between the segmentation contours in Figure 2 using global search (left) and the obtained simulation (right – in black are the initial and target objects, and in grey four estimated shapes).

If we use the modal matching method based on local search instead, we only successfully match 34 of the 45 nodes of the model of the initial contour; however, with q=30,000N/m and employing Equation (5) instead of Equation (4) to determine the external force applied on each unmatched node, the simulation of the deformation between the two instances of the foot is identical to the one with all the nodes successfully matched. This behaviour shows that our approach to compute the external forces applied on the unmatched nodes seems to be adequate and valid.

CONCLUSIONS AND FUTURE WORK

The experimental results obtained using our physically driven segmentation methodology, applied here to pedobarography images, are quite satisfactory. However, in the near future some changes to fasten and improve the segmentation process will be introduced, such as trying different approaches for the definition of the external forces. The use of finite elements more suitable for large and nonlinear deformations is also a subject to be addressed in the following stages of this work.

In this paper we also described a physical approach to simulate the deformation of objects represented in images and proposed a solution that enables the application of the used approach to objects that do not have all their nodes successfully matched.

The experimental results obtained in the matching process and in the estimation of the involved deformation are coherent with the physically expected behaviour of the modelled objects, validating the used approach.

The tracking of objects along image sequences, using the methodology here proposed, complemented with stochastic methods to estimate the involved motion, is also a task that will be addressed in the near future.

ACKNOWLEDGEMENTS

The presented work was partially done in the scope of the research project “Segmentation, Tracking and Motion Analysis of Deformable (2D/3D) Objects Using Physical Principles”, with reference POSC/EEA-SRI/55386/2004, financially supported by *FCT - Fundação para a Ciência e a Tecnologia* in Portugal.

REFERENCES

Bastos, M.L. and J.M.R.S. Tavares. 2006. "Matching of Objects Nodal Points Improvement Using Optimization". *Inverse Problems in Science and Engineering*, No.14(5), 529-541.

Bathe, K.-J. 1996. *Finite Element Procedures*. Prentice-Hall, New Jersey, USA.

Cook, R.; D. Malkus and M. Plesha. 1989. *Concepts and Applications of Finite Element Analysis*. John Wiley and Sons, New York, USA.

Gonçalves, P.C.T.; R.R. Pinho and J.M.R.S. Tavares. 2006. "Physical Simulation Using FEM, Modal Analysis and

the Dynamic Equilibrium Equation", In *Proc. of the CompIMAGE - Computational Modelling of Objects Represented in Images: Fundamentals, Methods and Applications* (Coimbra, Portugal, 20-21 October). 197-204.

Nastar, C. 1994. *Modèles Physiques Déformables et Modes Vibratoires pour l'Analyse du Mouvement Non-Rigide dans les Images Multidimensionnelles*. Thèse de Doctorat, École Nationale des Ponts et Chaussées, Champs-sur-Marne, France.

Pentland, A. and B. Horowitz. 1991. "Recovery of Nonrigid Motion and Structure". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, No.13(7), 730-742.

Pinho, R.R. and J.M.R.S. Tavares. 2004. "Morphing of Image Represented Objects Using a Physical Methodology", In *Proc. of the 2004 ACM Symposium on Applied Computing* (New York, USA). 10-15.

Pinho, R.R. and J.M.R.S. Tavares. 2004. "Dynamic Pedobarography Transitional Objects by Lagrange's Equation with FEM, Modal Matching and Optimization Techniques". *Lecture Notes in Computer Science*, No.3212, 92-99.

Sclaroff, S. 1995. *Modal Matching: a Method for Describing, Comparing, and Manipulating Digital Signals*. PhD Thesis, Massachusetts Institute of Technology, Cambridge, USA.

Sclaroff, S. and A. Pentland. 1995. "Modal Matching for Correspondence and Recognition". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, No.17(6), 545-561.

Shapiro, L.S. and J.M. Brady. 1992. "Feature-based correspondence: an eigenvector approach". *Image and Vision Computing*, No.10(5), 283-288.

Shen, J. and S. Castan. 1992. "An Optimal Linear Operator for Step Edge Detection". *CVGIP: Graphical Models and Image Processing*, No.54(2), 112-133.

Tavares, J.M.R.S.; J. Barbosa and A.J. Padilha. 2000. "Matching Image Objects in Dynamic Pedobarography", In *Proc. of the RecPad 2000 - 11th Portuguese Conference on Pattern Recognition* (Porto, Portugal, 11-12 May).

Tavares, J.M.R.S. and M.L. Bastos. 2005. "Improvement of Modal Matching Image Objects in Dynamic Pedobarography using Optimization Techniques". *Electronic Letters on Computer Vision and Image Analysis*, No.5(3), 1-20.

Terzopoulos, D.; J. Platt; A. Barr and K. Fleischer. 1987. "Elastically deformable models", In *Proc. of the 14th Annual Conference on Computer Graphics and Interactive Techniques* (Anaheim, USA). 205-214.

Terzopoulos, D.; A. Witkin and M. Kass. 1988. "Constraints on Deformable Models: Recovering 3D Shape and Nonrigid Motion". *Artificial Intelligence*, No.36, 91-123.

BIOGRAPHY

PATRÍCIA C. T. GONÇALVES went to the Faculty of Sciences of the University of Porto, Portugal, where she obtained her degree in Physics and Applied Mathematics. In 2006, in the Faculty of Engineering of the same university, she obtained her master degree in Computational Methods in Science and Engineering, and is, since then, taking a PhD in Engineering Sciences studying the segmentation, simulation and tracking of objects in images using a physical approach.

JOÃO MANUEL R. S. TAVARES graduated in Mechanical Engineering from the University of Porto in 1992. In 1995 he obtained his master degree in Electrical and Computer Engineering, in the field of Industrial Computing, also at the University of Porto. In 2001 he obtained his PhD in Electrical and Computer Engineering from the same University. Since 2001, he is a senior researcher and project coordinator at LOME – Laboratory of Optics and Experimental Mechanics of INEGI – Institute of Mechanical Engineering and Industrial Management. He is an assistant professor of the Department of Mechanical Engineering and Industrial Management at the Faculty of Engineering, University of Porto, since 2001. His main research areas include Computational Vision, Computational Mechanics, Scientific Visualization, Human-Computer Interaction and New Product Development. Currently, he is co-editor of the International Journal for Computational Vision and Biomechanics (IJCV&B).

R. M. NATAL JORGE graduated in Mechanical Engineering in the University of Porto in 1987. In 1991 he obtained his master degree in Structural Mechanics, also at the University of Porto. In 1999 he obtained his PhD in Mechanical Engineering from the same University. Since then, he is an assistant professor of the Mechanical Engineering and Industrial Management Department of the Faculty of Engineering of the University of Porto. Since 2005 he is the Executive Director of Design and Experimental Validation Unit from the Mechanical Engineering Institute - *Pólo FEUP*. His main research areas include Numerical Methods in Engineering, including FEM and Meshless Methods, Computational Mechanics, Biomechanics and New Product Development and Computational Vision. Currently, he is co-editor of the International Journal for Computational Vision and Biomechanics (IJCV&B).

Tools for an Ultrasound Based 3D Bone Model Reconstruction, Registration and Visualization System

Paulo Jorge Sequeira Gonçalves[1,2] Joaquim Moisés Fernandes[3]
[1]Instituto Politécnico de Castelo Branco
Escola Superior de Tecnologia
Av. do Empresário, 6000-767 Castelo Branco, Portugal
[2]IDMEC/IST,Technical University of Lisbon (TU Lisbon)
Av. Rovisco Pais, 1049-001 Lisboa, Portugal
[3]Hospital Amato Lusitano
Av. Pedro Álvares Cabral, 6000-085 Castelo Branco, Portugal
email: pgoncalves@est.ipcb.pt

KEYWORDS
Ultrasound, 3D Imaging System, Orthopaedic

ABSTRACT

In this paper are presented the tools needed to develop a 3D Bone Model reconstruction, registration and visualization system. This system will assist orthopaedic surgeons in preoperative and in intraoperative environments. The preoperative tasks are the 3D bone model reconstruction, based on CT images, to assist the surgeon in planning the surgical procedures to be performed and, if needed, to help choosing the adequate prothesis. The main intraoperative task accomplished by the system is the identification of the bone position and orientation, based on ultrasound images of the bone. This task is called registration, because the ultrasound images are registered related to the 3D bone model. This approach is less expensive and invasive, when compared to the classical use of cameras and fiducial markers or images from x-ray c-arm, to obtain the position of the bone. In both tasks the system visualizes the bone and the procedures that are being performed by the surgeon or surgical robot.

INTRODUCTION

The use of visual information obtained from medical images is widely used in Computer Aided Orthopaedic Surgery (CAOS) (DiGioia and Nolte 2002). CAOS systems are increasingly available, with several commercial and research systems now well established. These systems assist surgeons in preoperative planning and simulation, from the obtained bone model, in intraoperative navigation, using tracking systems with fiducial markers attached to the patient, and in the robotic execution of the surgical procedure.

Knowledge of the surgical workspace is a major topic when the surgeon/robot must perform precisely a given task, and can be of greater importance if the workspace changes over time. Since several actors are present at the operating room: surgeon/robot, bone, tool, sensors; they all must be "kinematically connected" to each other in order to the overall system be as much precise and accurate as possible. For that, all the actors referred before must be referenced to a frame, which can be placed on the operating room. The exact location of the robot/surgeon, the bone, the tool and the sensors must be known, i.e. their calibration must be performed. The 3D bone model must also be known in order to plan the surgical procedure.

Some surgical orthopaedic procedures could be performed by a robot, for example the ROBODOC system (ROBODOC http://www.robodoc.com/home.html). If a perfect calibration is accomplished and the robot trajectory is correctly planned, it would be expected that the robot performs the task in a very precise and accurate way, under position and force control. This is true if the workspace is rigid and cannot move. In the operating room, the patient can move and this movement must be compensated using visual feedback. Several surgical centres use exclusively optical information based on two cameras and fiducial marks, or information from x-ray c-arm, to perform navigation. A less invasive and expensive system will use ultrasound images from the bone (Amin et al. 2003, Barratt et al. 2006), to close the visual feedback. This system is also smaller in the operating room, causes less collateral damages to the patient, no radiation, and no trauma due to the fiducial markers insertion on the patient.

The paper is organized as follows. The section 3D BONE MODEL RECONSTRUCTION, describes briefly the concept of 3D bone model reconstruction from CT images, and also presents the tools needed to perform the task. Section ULTRASOUND FOR REGISTRATION, describes the experimental setup needed to obtain the registration of the acquired ultrasound image to the 3D bone model, some initial experimental results are also presented. Finally, in the last section, conclusions and the future work are presented.

3D BONE MODEL RECONSTRUCTION

To obtain the 3D bone model, conventional computer tomography (CT) images are used in the system. The software developed is based on the well known visualization toolkit (VTK) libraries, developed by Kitware. This Open Source library allows 3D visualization, based on a large amount of modeling and visualization algorithms.
The comercial available software for 3D reconstruction are based in manual processing of medical images, this method is error prone, slow and requires medical expertise. An automatic solution is preferable, that will increase speed and precision. In the case of bones, where their gray intensity levels are clearly different from the rest of the image, algorithms for surface reconstruction of the Marching Cubes (Lorensen and Cline 1987) family can be applied with success.
Using VTK, several steps were implemented to perform the 3D bone model reconstruction:

- **Read CT images** - reads DICOM (Digital Imaging in COmmunications and Medicine) images. It is a standard for handling, storing, printing, and transmitting information in medical imaging.
- **vtkImageThreshold** - performs image adaptive binarization (Forsyth and Ponce 2003), where the threshold changes dynamically over the image. The method assumes that smaller image regions are more likely to have approximately uniform illumination, thus being more suitable for thresholding. The local threshold function is to statistically examine the intensity values of the local neighborhood of each pixel, e.g. using the mean or median values.
- **vtkImageIslandRemoval** - removes small sets of pixels not belonging to the bone, using the mathematical morphology operator, erosion (Forsyth and Ponce 2003).
- **vtkImageGaussianSmooth** - applies a gaussian filter to remove noise (Forsyth and Ponce 2003), that acts on both dimensions of the image, like a low pass frequency filter.
- **vtkMarchingCubes** - applies the marching cubes algorithm to the set of images. First, the space is subdivided into a series of small cubes, or voxels. Then, the algorithm "march" through each of the cubes, testing the corner points and replacing the cube with an appropriate set of polygons. The total sum of all polygons generated will be a surface that best approximates the one, that the data set describes.
- **vtkDecimatePro** - reduces the number of triangles in the mesh. The recursive procedure is based in (Schroeder et al. 1992).

Figure 1: The ultrasound equipment.

- **vtkSmoothPolyDataFilter** - smooths the surface. The effect is to "relax" the mesh, making the cells better shaped and the vertices more evenly distributed.
- **vtkPolyDataNormals** - computes point normal for a polygonal mesh, to insure consistent orientation across polygon neighbors. Sharp edges can be split and points duplicated with separate normals to give crisp surface definition.
- **vtkStripper** - exports the results and creates the mesh creating triangles strips from the surface, that render faster.

The software in development also uses the QT library, from Trolltech, that allows multi-platform development and is also Open Source. The VTK and Qt libraries are from different developers and the connection between them is accomplished using the QVTK tool provided from Kitware.

ULTRASOUND FOR REGISTRATION

In the previous section were presented the tools needed to obtain a 3D bone model from CT images. In this section the algorithms presented will be used to achieve the registration of ultrasound images to the bone model. First, is is obtained the model of an egg from a 3D ultrasound calibration phantom. This procedure is needed to calibrate the ultrasound probe used to register ultrasound images of the bone to the previously obtained 3D

Figure 2: 3D Ultrasound calibration phantom.

Figure 3: Image from the calibration egg.

bone model, in order to obtain an accurate position and orientation of the bone relative to a fixed frame.

The ultrasound equipment, presented in Fig. 1, used in the experiments is a Siemens Sonoline Versa Pro, with two probes: a convex probe with 3.5MHz and a linear probe with 7.5MHz. The 3D Ultrasound calibration phantom used is the model 55, from CIRS, presented in Fig. 2.

The first step to use the ultrasound equipment is to calibrate it. The calibration method used is based on the calibration phantom, whose 3D geometrical properties are known. The procedure is defined in (Rousseau et al. 2005) and is based on the Hough transform and robust estimators. Using the procedure defined in the previous section it is possible to obtain the 3D model. In Fig. 3 is presented a slice of the egg used to obtain the egg 3D model from the ultrasound images using the convex probe, depicted in Fig. 4.

At this stage all the prior procedures to perform registration of ultrasound images to the 3D bone model are presented. Our next step is to apply and further develop

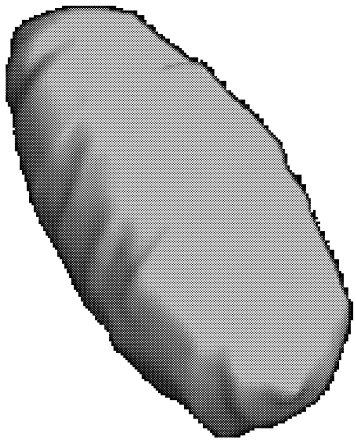


Figure 4: Image from the 3D reconstruction of calibration egg.

the algorithms for Ultrasound Registration of the Bone Surface for Surgical Navigation, proposed in (Amin et al. 2003, Barratt et al. 2006). One of the main milestones to accomplish is to detect the bone surface in the ultrasound image. Looking at Fig. 5, where is presented the bone surface of the radius on the a left arm, is is very difficult to detect the bone contour. This procedure will be performed using deformable contours, i.e. the classical Active Contour Models and Maximum Likelihood Parametric Deformable Models. The first approach was previously applied by the author to medical images in (Pinto et al. 1999) and the second was applied to fetal ultrasound images in (Jardim and Figueiredo 2005). These algorithms will be compared and the best solution used in the overall system.

CONCLUSIONS AND FUTURE WORK

In this paper were presented the tools needed to obtain a 3D Bone Model reconstruction, registration and visualization system. The software libraries and equipment needed were presented and also the algorithms to perform each task. Although this work is very useful for the orthopaedic surgeons, these first steps need further development in order to increase the precision of the approach regarding the clinical use of ultrasound based bone registration, and also regarding the interconnection of the several libraries used to increase efficiency and speed.

Future work is to get some insight about the possibility of using only the ultrasound based navigation with the advantage of being less invasive, when compared to fiducial markers and/or x-ray images. It will also be studied

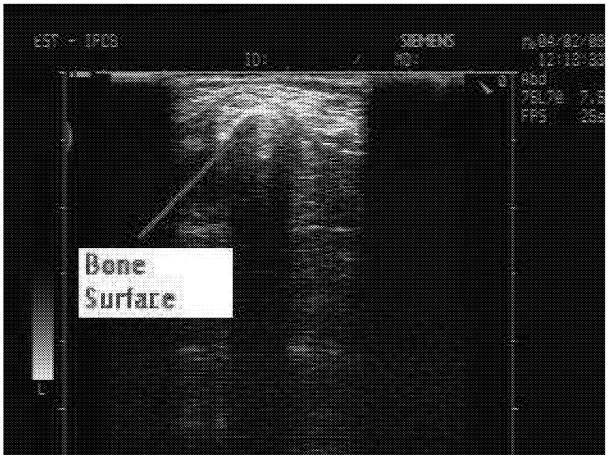


Figure 5: Ultrasound image from a bone (radius).

the advantages of fusing both techniques on the overall precision of the surgery.

ACKNOWLEDGEMENTS

This work was funded by FCT through "Programa POCI 2010, Unidade 10-46", subsidized by FEDER. The authors also want to thank the Virtual Reality and Robotics Lab from Miguel Hernandez University, Elche, Spain, for the help in the visualization system.

REFERENCES

Amin D.; Kanade T.; DiGioia A.; and Jaramaz B., 2003. *Ultrasound Registration of the Bone Surface for Surgical Navigation. Computer Aided Surgery*, 8, 1–16.

Barratt D.; Penney G.; Chan C.; Slomczykowski M.; Carter T.; Edwards P.; and Hawkes D., 2006. *Self-Calibrating 3D-Ultrasound-Based Bone Registration for Minimally Invasive Orthopedic Surgery. IEEE Transactions on Medical Imaging*, 25, 312–323.

DiGioia A.M. and Nolte L.P., 2002. *The challenges for CAOS: what is the role of CAOS in orthopaedics? Computer Aided Surgery*, 7, 127–128.

Forsyth D.A. and Ponce J., 2003. *Computer Vision: A Modern Approach.* Prentice Hall.

Jardim S. and Figueiredo M., 2005. *Segmentation of fetal ultrasound images. Ultrasound in Medicince and Biology*, 31, no. 2, 243–250.

Lorensen W. and Cline H., 1987. *Cubes: A high resolution 3D surface Construcction Algoritm. Computer Graphics*, 21, 163–169.

Pinto J.C.; Mendonça M.; Diniz D.; Gonçalves P.S.; and Ramalho M., 1999. *An On-Line Measuring System to Support Heart Surgery.* In *Proceedings of 7th International Conference on Image Processing and its Applications.* Manchester, UK, 377–381.

ROBODOC, http://www.robodoc.com/home.html. *visited in 16/03/2008.*

Rousseau F.; Hellier P.; and Barillot C., 2005. *Confhusius: a robust and fully automatic calibration method for 3D freehand ultrasound. Medical Image Analysis*, 9, 25–38.

Schroeder W.J.; Zarge J.A.; and Lorensen W.E., 1992. *Decimation of triangle meshes. Computer Graphics*, 26, no. 2, 65–70.

NUFFT-based Direct Fourier Methods and Regional Tomography

Silvia De Francesco
University of Aveiro – ESSUA
University of Aveiro – IEETA
Campus Universitário de S. Tiago
3810 – 193 Aveiro
silvia.francesco@ua.pt

Augusto Silva
University of Aveiro – DET
University of Aveiro – IEETA
Campus Universitário de S. Tiago
3810 – 193 Aveiro
augusto.silva@ua.pt

ABSTRACT

Direct Fourier methods, a class of reconstruction methods suitable for 2D reconstruction in x-ray Computed Tomography, have been known for the low computational complexity ($O(N^2 logN)$ compared to the $O(N^3)$ of the commonly used Filtered Backprojection method), but also for the lack of flexibility, difficult extension to projection geometries other than the parallel and poor image quality. The bad performance of these methods was essentially due to the need for interpolation in Fourier domain prior to inverse Fourier transformation on a certain step of the algorithm.

With the introduction of efficient computational methods for the calculation of Fourier transform in the case of non-equally spaced samples (Non-Uniform Fast Fourier Transform – NUFFT), a number of Direct Fourier algorithms have been proposed with no need for interpolation in Fourier domain. It has been shown that these new algorithms allow to improve image quality while keeping the computational complexity to $O(N^2 logN)$. The quality of the resulting images is equivalent to the quality of the images obtained by Filtered Backprojection.

In this paper we'll show that the introduction of NUFFT also allows to perform regional reconstruction, a feature that was considered exclusive of Filtered Bacjprojection method. A NUFFT-based Direct Fourier regional reconstruction algorithm will be described and the results will be compared with those obtained by Filtered Backprojection.

INTRODUCTION

In x-ray Computed Tomography (CT) the 2D reconstruction method of choice is the Filtered Backprojection (FBP) method (Kak and Slaney 1988). This method, due to its pixel oriented strategy has been claimed as the most versatile, suitable to any projection geometry, and capable to offer different image quality depending by the selected filter kernel. Moreover, this pixel oriented reconstruction strategy has been extended to 3D cone beam reconstruction in the so called FDK method (Feldkamp, Davis et al. 1984).

Another family of reconstruction methods, the direct Fourier (DF) methods, although characterized by lower computational complexity ($O(N^2 logN)$ compared to the $O(N^3)$ of FBP), has been left aside because of its lack of flexibility, difficult extension to projection geometries other than the parallel and poor image quality. The bad performance of DF methods was due to the need for interpolation in in Fourier domain prior to inverse Fourier transformation on a certain step of the algorithm.

With the introduction of efficient computational methods for the calculation of Fourier transform in the case of non-equally spaced samples (Non-Uniform Fast Fourier Transform – NUFFT), a number of DF algorithms have been proposed with no need for interpolation in Fourier domain. It has been shown that these new algorithms allow to improve image quality while keeping the computational complexity to $O(N^2 logN)$. The quality of the resulting images is equivalent to the quality of the images obtained by FBP method and the flexibility is also improved since these algorithms include a filtering step like FBP algorithm, allowing to adjust image quality by choosing different (harder or softer) filter kernels.

Another feature that is believed to be exclusive of FBP method due to its pixel oriented strategy, is the possibility to perform Regional Tomography, that means to reconstruct Regions of Interest (ROIs) in the tomographic field of view. In other words, Regional Tomography allows to obtain a zoomed version of any portion of the field of view.

In this paper we show that the introduction of NUFFT in DF algorithms also allows to perform Regional Tomography. A new NUFFT-based DF regional reconstruction algorithm will be described and evaluated by comparing its performance with the performance of the classical FBP regional reconstruction algorithm.

The described algorithms have been included in a comprehensive CT simulation environment, that is being developed in Matlab, based on the geometry of a real CT system (De Francesco and Silva 2002). The NUFFT calculation has been performed using the NFFT C-library developed by Kunis and Potts (Kunis and Potts 2001). The C-function calculating the NUFFT has been embedded in a Matlab *mex* function. More specifically, in the described algorithm we just used the algorithm for the calculation of NUFFT from non-equispaced data (NED).

This paper consists of four main sections. In the first and second we review some basic aspects of x-ray CT projection and reconstruction theory and of NUFFT-based DF methods. In the third section we introduce the theme of Regional Tomography and describe a new NUFFT-based DF algorithm for Regional Tomography. In the fourth

section the proposed algorithm is evaluated and its performance is compared with that of FBP.

BASICS OF X-RAY CT RECONSTRUCTION

The distribution of the attenuation coefficient on a transversal section of an object can be described as a 2D function *f* (object function) in the *(x,y)* plane of the section.

The two parameters θ and s univocally specify the line with equation

$$x\cos\theta + y\sin\theta = s \qquad (1)$$

in the *(x,y)* plane and the general formula for the line integral, known as the Radon transform of *f(x,y)*, is:

$$p(\theta,s) = \iint f(x,y)\delta(x\cos\theta + y\sin\theta - s)dxdy \quad (2).$$

The purpose of x-ray CT 2D reconstruction methods is to calculate *f(x,y)* given a proper set of measured line integrals, that means, from a mathematical point of view, to calculate the inverse Radon transform given a sufficient set of samples.

A projection consists of a collection of line integrals of *f(x,y)* taken along a set of straight lines in the plane and the projection data set is given by a number of projections taken with different orientations. Basically, two geometries have been defined for the sets of line integrals making a 2D projection: parallel and divergent (or fan-beam).

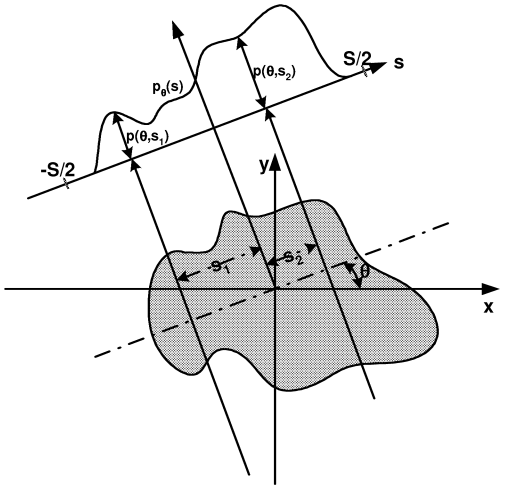


Fig. 1 Object function *f(x,y)* and its parallel projection in θ direction

In parallel geometry (the acquisition geometry of first generation systems, shown in figure 1), a projection $p_\theta(s)$ consists of a collection of line integrals taken along straight parallel lines in the plane, that means a collection of $p(\theta,s)$ with constant θ and $s \in [-S/2, S/2]$. A parallel projection data set is usually represented as a 2D matrix, called sinogram, each row of which corresponds to a value for the parameter θ (a parallel projection) and each column to a value for the parameter *s*.

A fundamental result in tomographic reconstruction is the Fourier slice theorem (details and demonstration can be found in (Kak and Slaney 1988)):

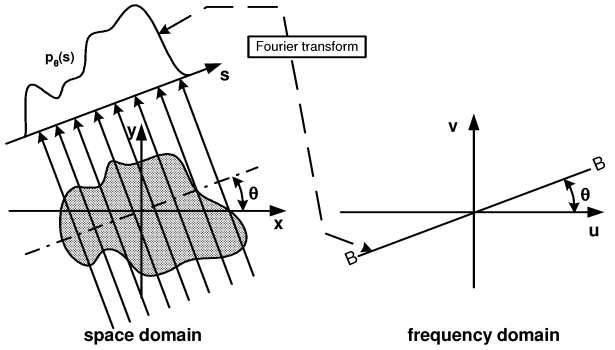


Fig. 2 Graphic representation of Fourier slice theorem statement

Theorem 1: The Fourier transform of a parallel projection of an object function *f(x,y)* taken at angle θ gives a slice of the two-dimensional Fourier transform of *f(x,y)*, *F(u,v)*, subtending an angle θ with the *u* axis.

In other words, the 1D Fourier transform $P_\theta(\sigma)$ of the parallel projection $p_\theta(s)$, gives the values of *F(u,v)* along line BB in figure 2.

FBP Reconstruction

For historical and practical reasons, the most successful 2D reconstruction method (chosen by all the manufacturers) is FBP. The reconstruction formula can be mathematically derived (Kak and Slaney 1988), but the method can also be introduced in a very intuitive fashion.

If we smear back (backproject) the measured samples along the direction with which they were taken, we obtain a blurred version of the image we were supposed to get. This problem is solved by filtering (with a ramp filter) the projections before backprojecting them. This method can be applied to divergent geometry by adding additional weighting to the projections and in the backprojection process.

DF Reconstruction

The Fourier slice theorem suggests a simple way to solve the reconstruction problem. Taking parallel projections of the object function *f* at angles θ_1, θ_2 .. θ_n and Fourier transforming each of them, we obtain the 2D Fourier transform of the object function *F(u,v)* on *n* radial lines. In ideal conditions (infinite number of projections and samples per projection) *F(u,v)* would be known at all points in the frequency domain and the object function *f(x,y)* could be recovered by 2D inverse Fourier transforming *F(u,v)*.

Fourier reconstruction methods (also known as direct Fourier or Fourier based methods) follow directly from this ideal procedure, adapted to the discrete case (Natterer

1985). Since only a finite number of projections and samples per projection are taken, *F(u,v)* is known just on a finite number of points along a finite number of radial lines (fig. 3) and, in order to obtain an approximation of *f(x,y)* by 2D inverse Fourier transform of *F(u,v)*, first we have to interpolate from the radial points to the points on a Cartesian grid.

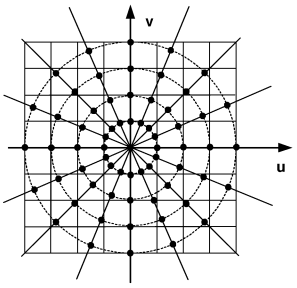


Fig. 3 In Fourier reconstruction, sample points of the 2D Fourier transform of the image are given on a conventional polar grid

Unfortunately, the results obtained with such a straight method suffer from artefacts due to interpolation in Fourier space and aliasing. Nevertheless, due to its low computational complexity ($O(N^2 logN)$) this method has been object of research and various techniques have been proposed in order to improve its performance. These techniques are based on peculiar sampling schemes – polar interleaved grid (Lewitt 1983), polar squared grid (Natterer 1985) and linogram (Magnusson 1993) – capable to reduce interpolation error and classical computational methods for aliasing reduction (zero padding).

NUFFT-BASED DF METHODS

With the development of efficient computational algorithms for the calculation of Fourier transform in the case of non-equally spaced samples (Non-Uniform Fast Fourier Transform – NUFFT), a number of DF reconstruction algorithms have been proposed in which 2D inverse Fourier transform of *F(u,v)* was calculated directly from its radial samples (Potts and Steidl 2000).

The NUFFT-based DF methods have shown a performance equivalent to the one of the more widely accepted FBP algorithm with a considerable saving in computational time. The fact that, in principle, these algorithms are not suitable for reconstruction from divergent projections doesn't seem to be an obstacle anymore since, in any case, 2D reconstruction is performed on interpolated data (after longitudinal interpolation), being possible to choose a parallel geometry resampling grid. Moreover, it has been demonstrated that, taking advantage of NUFFT, direct Fourier methods can be applied directly on divergent projections (De Francesco and Silva 2004).

The workflow of a NUFFT-based DF reconstruction algorithm for parallel projections (fig. 4) starts with zero padding (in order to avoid aliasing) and 1D Fourier transform of each parallel projection. At this point of the algorithm the calculated samples are samples of the 2D Fourier transform of the image uniformly distributed on a conventional polar grid, with radial (that means over the σ coordinate) oversampling. The image can be obtained by applying 2D inverse NUFFT to these non-equally spaced data (NED) being necessary, before this, a filtering step with a ramp-like filter (as those used in FBP) in order to compensate for the non-uniform density of the samples.

It is essential to notice that for the calculation of NUFFT, it is necessary to know the coordinates of each sample point.

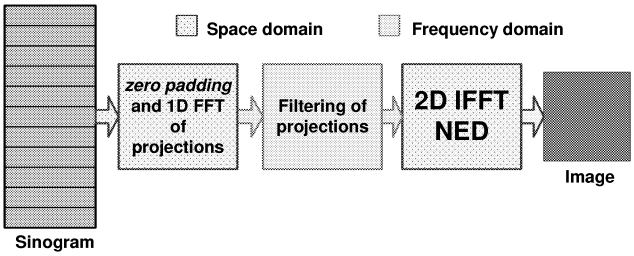


Fig. 4 workflow of NUFFT-based DF reconstruction for parallel projections

REGIONAL TOMOGRAPHY

The pixel oriented strategy of FBP algorithm allows to reconstruct any region of interest of the field of view over any number of sample points. So, FBP regional tomography is not an issue. On the contrary, the global reconstruction strategy of DF methods is hard to apply to the reconstruction of regions other than the all field of view. Nevertheless, we can demonstrate that, with the introduction of NUFFT, it is possible to perform DF regional reconstruction with considerable (possible) savings of computational complexity when compared with the corresponding FBP procedure.

NUFFT-Based Regional Tomography

From Fourier theory, it is known that if pixels' size is *a*, the highest possible spatial frequency present in the image is $1/2a$ and, as a consequence, 2D Fourier transform of the image is contained in a square of side $1/a$. To multiply pixels' size by a factor *k* corresponds to multiply by a factor $1/k$ the maximum spatial frequency in the image, being the support of 2D Fourier transform of the image a square of side $1/ka$. This means that, with a proper scaling of samples' coordinates in the Fourier space, we can change image support in order to reconstruct just a ROI, as shown in figure 5.

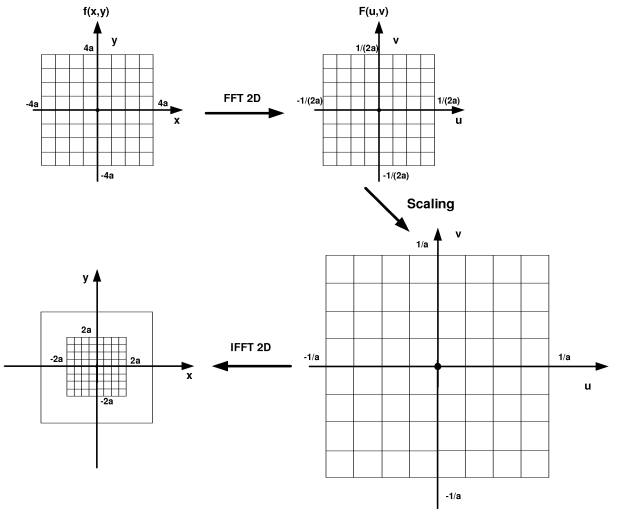


Fig. 5 Fourier reconstruction of a ROI

Since, in NUFFT-based DF methods, in order to apply the inverse NUFFT to the samples on the polar grid we have to know their coordinates, it is easy to introduce the necessary scaling of the coordinates at this point of the algorithm (fig. 6).

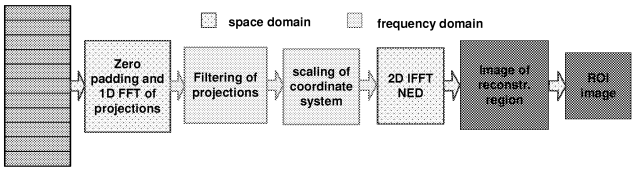


Fig. 6 Workflow of NUFFT-based DF regional reconstruction

The only limitation is that the reconstruction region should be squared and isocentrical with the field of view so, if the ROI doesn't satisfy this condition it has to de considered as reconstruction region the smaller squared region isocentrical with the field of view (fig. 7). Afterwords, the image of the ROI has to be extracted fom the reconstructed image.

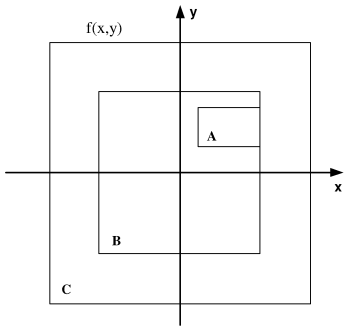


Fig. 7 NUFFT-based ROI reconstruction. A = ROI, B = reconstruction region, C = field of view

This means that, if the ROI is a square of side *l* and reconstruction region a square of side *L*, and if we want the ROI reconstructed on $N \times N$ pixels, the 2D inverse NUFFT should be calculated over $N\frac{L}{l} \times N\frac{L}{l}$ samples.

For this reason the computational cost of this method, which is $O((N\frac{L}{l})^2 \log(N\frac{L}{l}))$, depends not just upon the ROI's size and the required pixel density but also upon the ROI's localization. On the other side, the computational complexity of FBP regional tomography doesn't depend upon the localization of the ROI but it is always higher ($O(N^3)$).

RESULTS

The described method has been included in our CT simulation environment (De Francesco and Silva 2002) and tests have been performed with different 2D phantoms (Herman head, Shepp-Logan head, etc.).
As an example, we show here the results obtained using the modified Shepp-Logan phantom of size 256x256 pixels (fig. 8-a). The results obtained with other phantoms were similar. Acquisition of 256 parallel projections over 180º angular rotation (256 samples per projection) has been simulated both in ideal and noisy conditions. Noise has been described as a simplified Poisson model, which depends on the power of the generated beam, being N_{in} the number of incoming photons (De Francesco and Silva 2002).

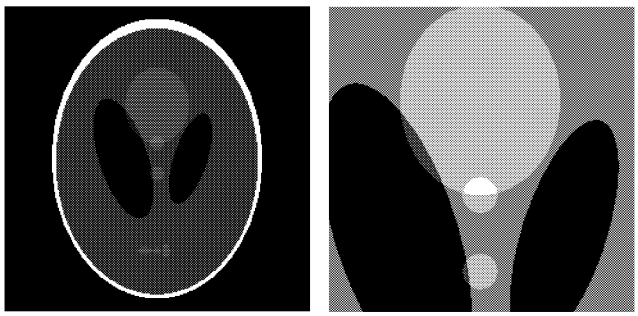


fig. 8 Modified Shepp-Logan phantom (a) and not isocentrical ROI of the phantom (b). Gray levels have been enhanced for better visualization

The images obtained with the described NUFFT-based DF regional reconstruction method have been evaluated and compared with those obtained with the corresponding FBP method for different ROIs (in size and localization) both for ideal and noisy projections. In all the cases the quality of the images obtained with the two methods were equivalent both qualitatively and quantitatively. As an example, in figure 9 we show the images obtained reconstructing the ROI of figure 8-b with the two methods from ideal projections and in figure 10 the corresponding results in the case of noisy projections ($N_{in} = 10^8$).

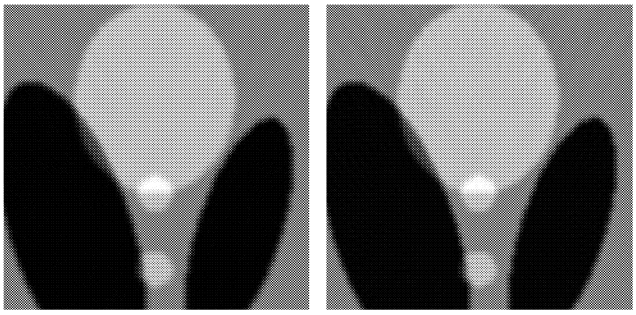


Fig. 9 Regional reconstruction of the ROI with FBP (a) and NUFFT based DF method (b) from ideal projections

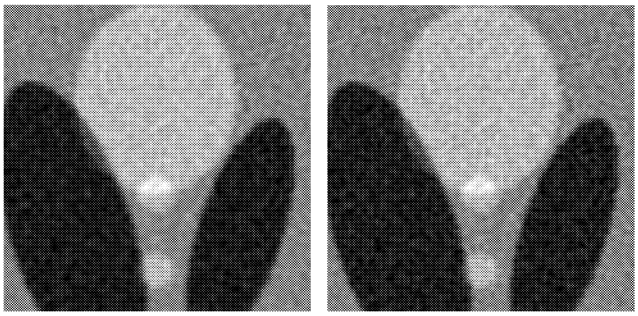


Fig. 10 Regional reconstruction of the ROI with FBP (a) and NUFFT based DF method (b) from noisy projections

As can be seen, from a qualitative point of view, there is almost no visible difference between the two methods, with the exception that the noise effect appears slightly increased in the image obtained with DF method (fig. 10).

Image quality has been quantitatively evaluated measuring the distance between the original (zoomed version of the phantom's ROI) and the reconstructed images. For this purpose we've used the distance metrics proposed by Herman (d – normalized root mean squared distance, r – normalized mean absolute distance, e – worst case distance measure (Herman 1980)). In table 1, the calculated distance values for the reconstructed ROIs of figure 9 and 10 are summarized.

Table 1:Quantitative evaluation of reconstructed images

Ideal proj.	d	r	e
FBP	0.1368	0.0444	0.0086
DF	0.1336	0.0449	0.0086
Noisy proj.	d	r	e
FBP	0.1794	0.1066	0.0088
DF	0.1882	0.1156	0.0091

From table 1 can be infer that, also from a quantitative point of view, the quality of the images obtained with the two methods is equivalent, with the exception of a slight increase of the error (confirming the perceivable noisy effect) in the case of DF method in noisy conditions.

It has to be pointed out that the worse performance of DF method in noisy conditions is not unexpected and is due to the need for a slightly softer filter than the used in FBP. In fact the choice of a ramp-like filter to compensate for the non-uniform density of the samples is due to common sense reasons (in order to have the same filters than in FBP) but it causes some "over-compensation". The two methods are perfectly equivalent if we use a softer filter in DF method than in FBP.

The great advantage of NUFFT-based DF method over FBP method is computational time. As an example, the reconstruction of the ROI of figure 8-b, were performed with DF method in just 34.8% of the time needed with FBP method.

CONCLUSIONS

A new NUFFT-based Direct Fourier regional reconstruction method has been described. The method has been included in a CT simulation environment and tests have been performed with different phantoms. It has been shown (selecting modified Shepp-Logan phantom as an example) that the performance of the proposed method is equivalent to the performance of the well known FBP regional reconstruction method with a significative reduction of computational complexity.

REFERENCES

De Francesco, S. and A. M. F. d. Silva (2002). "Multi-slice Spiral CT Simulator for dynamic cardio-pulmonary studies". *Medical Imaging 2002: Physiology and Function from Multidimensional Imaging*, S. Diego, SPIE.

De Francesco, S. and A. M. F. d. Silva (2004). "Efficient NUFFT-based direct Fourier algorithm for fan beam CT reconstruction". *Medical Imaging 2004: Image Processing*, S. Diego (CA), SPIE.

Feldkamp, L. A., L. C. Davis, et al. (1984). "Practical cone-beam algorithm." *J. Opt. Soc. Am.* 1(6): 612-619.

Herman, G. T. (1980). *Image Reconstruction from Projections*, Academic Press.

Kak, A. C. and M. Slaney (1988). *Principles of Computerized Tomographic Imaging*, IEEE Press.

Kunis, S. and D. Potts (2001). {NFFT C}-library. http://www.math.mu-luebeck.de/potts/nfft/, Mathematical Institute of the University of Luebeck.

Lewitt, R. M. (1983). "Reconstruction algorithms: transform methods." *Proceedings of the IEEE* 71(3): 390-408.

Magnusson, M. (1993). *Linogram ad other direct Fourier methods for tomographic reconstruction*. Department of Electrical Engineering. Linköping, Sweden, University of Linköping: 254.

Natterer, F. (1985). "Fourier reconstruction in tomography." *Numer. Math.* 47: 343-353.

Potts, D. and G. Steidl (2000). "New Fourier reconstruction algorithms for computerized tomography". *SPIE's International Symposium on Optical Science and Technology: Wavelet Applications in Signal and Image Processing VIII*, S. Diego (CA), SPIE.

MEDICAL IMAGING IN THE XXI CENTURY: THE PLACE OF FUNCTIONAL IMAGING AND NUCLEAR MEDICINE

Luís F. Metello and Lídia Cunha
Nuclear Medicine Department – ESTSP-IPP
Pr. Coronel Pacheco, 15
P-4050-453 Porto (Portugal)
E-mail: lfm@estsp.ipp.pt

The progressive and constant increment of the cerebral mass from the first *hominidae* forms until the *homo sapiens* is usually referred as one from the most critical factors on the evolution of humankind. This had supported a huge amount of diversity and any Evolution, but it must be recognized that only Science and Technology demonstrates real effective Progress.

Nowadays, Science and Technology are pretty much inseparable…even if easily distinguishable as Science includes the entire basis and fundaments, while Technology deals with practical applications. Technology mean always a scientific basis, no-one being able to develop anything technical without solid scientific capacities. A strong scientific education is nowadays consensually considered as essential at high technological level. On the other hand, it is also consensual that modern sciences can't subsist and/or further develop without "cut-of-the-edge" technology, so implying the recognition that it is mandatory to possess solid Technological Competences, concerning all that produces, or use, the Scientific Knowledge.

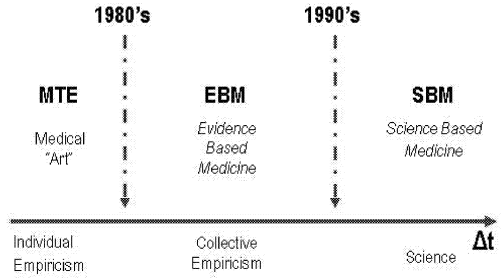


Fig.1 – Evolution on Medical Sciences

When considering the Medical Sciences evolution, it could be consider a first period, from the beginning till around the 80's, that has been characterized by an extreme individualization (since the patient is considered and assumed as unique, as well as the sickness is unique, the physician is unique, the approach is unique, …). Usually, the proposed therapeutic approaches within this period were essentially symptomatic.

With the second period, the Evidence Based Medicine period, it as been noticed a serious attempt of Systematization and Codification, essentially based on the production and application of specially designed algorithms, which were kind of "Standardized Procedures" for well defined and characterized situations. The therapeutic procedures were much more systematized and less episodic or symptomatic. Within this approach, the variability tends to decrease and it is more and more expected that a comparable situation receives comparable answers….the real problem remaining being that nothing really new were introduced: it has been a (simple) systematization effort.

After the 90's, it is usual to consider that Medical Sciences are more and more characterized by what is usually designed by a "Science Based Medicine approach", that it might be easy characterized by the "systematic application of Science progresses, at distinct levels and approaches, to Medicine", with that being the rational behind all the

enormous development of Medical Technology that it is observed actually.

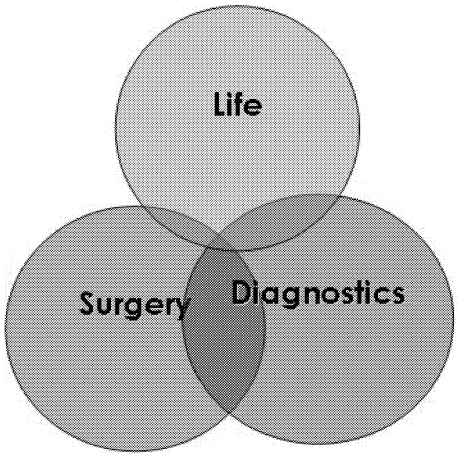


Fig.2 – The three main fields of actuation of Medical Technology

Medical Technology essentially deals with three main fields: Life's Support, Surgery Support and Diagnostic Support. This work pretends to deal only with the last one from those fields. It is a very important field: strictly as an example, e.g. it has been reported that 44.000 to 98.000 deaths/year are related with clinical errors (source: *Institute of Medicine – USA, 2005*), between them about 18.000 fatal myocardial infarcts due to lack of diagnostics and/or adequate therapeutic action (time reaction factors included).

Applying the old proverb "One picture is worth ten thousand words", based on the universally accepted fact that images are very powerful tools to communicate ideas without words - that is believed to applies on the medical field perhaps better than anywhere else - it is aimed to highlight one of the main fields of medical technology: Diagnostic Imaging Techniques, that supports medical diagnostic activities based on imaging, and between them the Functional Imaging and Nuclear Medicine.

Function Imaging and *Functional Imaging* are terms more and more widely applied to many medical imaging procedures. Nevertheless, it is consensually considered that the most appropriated sense is the one that mean "all imaging procedures that provide information beyond the mere morphological representation of a structure". On some situations, it is possible that the depiction of morphology already yields some information on organ function, as a morphologically altered organ may imply that its function is impaired. Hence, it is sometimes quite difficult to draw the line between what is functional imaging and what is morphologic imaging. Mostly for this reason, most of the authors currently consider that "functional imaging gives information that cannot be inferred directly from looking at the anatomic features in the image" as, while morphology sometimes may suggest function, it is often difficult or impossible to infer function from the anatomic features imaged.

Although some of us could probably trace the origin of functional imaging to the early days after the discovery of X Rays and more certainly to the time and moments when first radiologists started to use contrast agents, functional imaging, properly speaking, might be considered only a few decades old, having started to demonstrate some clinical relevance with Nuclear Medicine's advent. Specifically, the use of iodine-131 on the identification and adequate characterization of the regional distribution of thyroid hormone production, and thus allowing differentiating between "hot" and "cold" nodules as well as disseminated thyroid hyper-function could be considered as the beginning of clinical functional imaging.

Definitely the visualization of physiologic function is one of the essential points for further progress in the understanding of biomedical processes, in the design of new therapeutic procedures and in their clinical application. Fortunately for all of us – health professionals and patients on the first line - methods, techniques and equipment in biomedical imaging have considerably developed - including the enormous computational power increase and availability – making that, today, many techniques of visualization are available, providing a wealth of information just inconceivable twenty or thirty years ago…and their importance in medicine is consistently growing each day. It is easily observed, even for neophytes on the field, that the last twenty years has been responsible for a considerable growth on the number

(and the use) of Diagnostic Imaging Techniques (MRI, fMRI, US, Doppler, CT, SPECT, PET, TET,…).

Technological advances, including computational advances (for instance, concerning Image Processing and Quantization Techniques) radically transform the situation regarding Clinical Medicine and Surgery, as well as most of the other clinical fields. The professionals involved on the utilization of this modern imaging techniques need to be prepared to daily face challenges on informatics, physics, chemistry, mathematics, etc… obliging that, nowadays, in health sciences, a multidisciplinary high-skilled team is not only mandatory but also the only way to assure patients the provision of health cares at the desired level of Excellency.

Definitely, medical imaging techniques to assess function have quickly multiplied, having reached a level of sophistication that can no longer be mastered by a single individual, making difficult to keep track of the developments even in the field of one's expertise. Highly important information from related fields might therefore often go unnoticed.

In fact, one often observes some factionalism that more or less "actively" ignores developments in related fields. Furthermore, techniques long ago developed and explored in one field are being "rediscovered" and "sold" as new techniques in a second field. This factionalism is often a result of funding pressures, where researchers have to overstate that the methodology they propose to study a given problem is unique, and also a result of political pressures, as expensive imaging equipment – and this is what most of modern functional imaging is all about….- is often distributed among different departments in the same institution. Otherwise, not all equipment is available at a given institution, and the result might be "selective blindness" toward the equipment not available. The rather independent development of different disciplines working in different aspects of the same problem has led to very specific and idiosyncratic schools of thought, which makes communication difficult, even for those willing to "jump over the fences of their own scientific biotopes" (von Schulthess and Hennig, 1998).

Medical Imaging is a crucial part on most diagnostic processes, remaining (each day more and more) important in the adequate follow-up and post-therapeutics patient's precise condition assessment.

Medical Imaging might be divided on two main fields, according the approach followed:

A - Morphological Imaging (also referred as Anatomical Imaging, or Structural Imaging); it is imaging based on physical proprieties (attenuation coefficient, e.g.) and it might provides quite high resolution images (as it is the case with imaging modalities as the X-Rays, CT, US, MRI,…);

B - Functional Imaging, that is essentially imaging based on biochemical proprieties and it provides access most often to low spatial resolution images (as it is the case of images obtained by SPECT and PET techniques).

Ideally, Medical Imaging should deal with distinct tools and processes, regarding the distinct needs and situations: for instance, using functional MRI it is possible to depict complex brain processes such as thought and memory, using CT it is possible to display vivid 3D demonstrations of very tiny structures, such as colon polyps, small arteries and bronchi, etc…

Nevertheless, and despite all these advances, which indeed have radically - and positively - transformed patient care, clinicians have since long ago been unable to detect disease in either pre-clinical or very early stages. That is essentially because the mainstream clinical imaging modalities such as MRI, CT and Ultrasound can only identify abnormalities in tissue architecture at a point in time when around one billion abnormal cells are present, so for years it has been sought an imaging method with the high spatial and temporal resolution of these modalities

that also presents far better sensitivity; achieving this goal would allow to both

1) - identify disease earlier and

2) - anatomically localize disease to optimize therapy.

This sought as guided several lines of research and development, having conducted to one solution - hybrid PET/CT scanners – that is available from a short number of years (since around 2001), being potentially capable of detecting tumors or other abnormalities far earlier in their life cycles, when only a few millions of cells are present. Already from the initial evaluating results presented it has been conclusively demonstrated that the hybrid technique offers a great deal more than either technique used separately, even when software image fusions are available (Glazer, 2003).

Anyway, the Image Purpose should always depend on the specific Clinical Question; for instance, when the clinical case relies

1) on the detection of hairline bone fractures, most probably conventional X-rays techniques should be used, as it will allows access to High Spatial Resolution images;

2) on the detection of recurrent tumor tissue in the liver, most probably Computed Tomography (CT) Scanner should be used, as it will allows access to High Contrast Resolution images;

3) on the quantification and analysis of Left Ventricle Ejection Fraction, most probably planar Nuclear Medicine Images or gSPECT Images should be used, as it allows access do Low Noise (High S/N) images;

4) on the detection of metastatic tissue in soft tissue, most probably PET and/or SPECT should be used, as it allows access to High Specificity and High Signal/Noise ratio images;

Concerning the Image Quality – and, much more important, the information content and quality – it is mandatory to keep always in mind that the information supplied by most of medical image modalities concerns one and only one parameter, most often with indirect and small involvement with the pathology being studied.

There are also several limitations that always need to be considered:

A) Intrinsic Limitations (dues to the intrinsic technique itself), as Spatial Resolution, in the case of PET and Sensitivity, in the specific case of MRI;

B) Technological Limitations (dues to the technological approach used in each imaging modality), as Spatial Resolution, in the US case's, or Energy Resolution in the SPECT specific case;

C) Patient Protection Limitations (dues to the absolute need to respect Patient Protection related issues, namely regarding dosimetric aspects) that conduces to limitations concerning Pixel Dimension in CT's case, or Contrast levels limitations in Nuclear Medicine imaging modalities general case.

Theoretically, Nuclear Medicine is the Medical Imaging technique that better overtakes the limitation of informing about a sole parameter. For the same pathology different labeled molecules can be used, supplying different information, as if different techniques have been applied. That is indeed a major issue and it really distinguishes Nuclear Medicine studies from other imaging techniques, which map a single parameter (P. de Lima, 2002, personal communication).

Nuclear Medicine might be defined as "the best possible way to use radioactivity". Being an autonomous medical speciality, recognized by the UEMS since 1981 (the USA had already recognized it in 1971) it is essentially characterized by the fact that radioactive materials are used for medical purposes in many disease processes, using both imaging and non-imaging techniques.

The constant technical innovation in equipments, associated with the continuous development of new

molecules/ligands, for more and more specific aims and functions, together with the constantly increasing computational power that the medical community disposes nowadays, made possible to obtain essentially functional information, easily recognized as crucial in the adequate diagnosis and therapeutics definition of an each-day wider set of clinical situations.

Actually, detection capabilities are increasing more and more, so smaller lesions are currently being detected – so dramatically increasing the sensibility – at the same time that more and more specific labelled molecules are being produced/developed – so dramatically increasing the specificity.

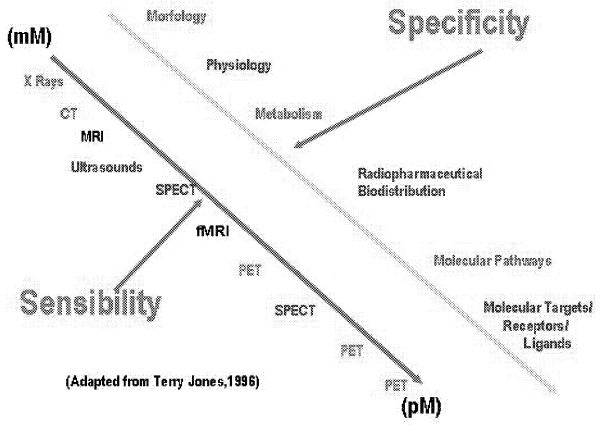


Fig.3 – Comparison Chart of Medical Imaging Techniques

Nuclear Medicine techniques are functional imaging techniques that involve the administration of radioactive substances, called radioisotopes (or radiopharmaceuticals), which demonstrate pathophysiology of target systems, organs and tissues. The physical decay inherent to radioactive atoms supplies the photons and/or the particles for the diagnostic or the therapeutic approaches, respectively.

Actually, the most important Nuclear Medicine technique is the combined modality PET/CT that is consistently growing in importance each day. Historians of science have traced how innovation and discovery is enhanced by, and often emerges from, convergences between technologies and disciplines. At this moment in time, the convergence of medicine, high technology and molecular biology is a burgeoning site for medical breakthroughs. The nascent field of *in vivo* imaging of gene expression is rapidly developing and it will not be long before it is a fundamental component of medical research as well as clinical applications. PET/CT is poised to become one of the platform technologies for exploring *in vivo* gene expression. This should result in both, enhanced diagnosis and identification of early response to therapy by PET/CT imaging in the near future (Glazer, 2003).

Nuclear Medicine means easy, cost-effective and non-invasive access to functional information and, as a matter of fact, it is a way to gather medical information that would otherwise be unavailable, require surgery or necessitate much more expensive diagnostic techniques.

Nuclear Medicine is a very powerful group of techniques that is able to study organs and systems – and the respective processes concerned - on a real-time approach, always without interference on the processes themselves and with the less possible invasiveness (usually just an intravenous injection) involving a minimal exposure to radiation. Most probably, the major advantage inherent to Nuclear Medicine imaging procedures is that it quite often identifies functional abnormalities very early in the disease process (long before many medical problems become apparent to other diagnostic tests) allowing the biggest improvements on prognosis.

Nuclear Medicine is indeed a very powerful medical tool that is helping to change the way of practicing Medicine with its ability to assess disease processes at the molecular level, most often giving the earliest answers to the clinical questions. Being clearly demonstrated as the most cost-effective approach, it is predictable a greater development and an increasingly wider application.

REFERENCES

Glazer, G.M.2003 „Foreword“. In *Clinical Molecular Anatomic Imaging*. Lippincott Williams & Wilkins, Philadelphia, PA

Von Schulthess, G.K. and J. Hennig. 1998 *Functional Imaging*. Lippincott – Raven Publishers, Philadelphia, PA

DETECTING ABNORMALITIES IN ENDOSCOPIC CAPSULE IMAGES USING COLOR WAVELET FEATURES AND FEED-FORWARD NEURAL NETWORKS

Carlos S. Lima, Daniel Barbosa, Jaime Ramos[(1)], Adriano Tavares, Luis Carvalho and Luis Monteiro

Department of Ind. Electronics of University of Minho, Campus de Azurém, Guimarães, Portugal
carlos.lima@dei.uminho.pt

[(1)] Capuchos Hospital, Alameda Santo António dos Capuchos, Lisboa, Portugal

ABSTRACT

This paper presents a system to support medical diagnosis and detection of abnormal lesions by processing endoscopic images. Endoscopic images possess rich information expressed by texture. Texture information can be efficiently extracted from medium scales of the wavelet transform. The set of features proposed in this paper to encode textural information is named color wavelet covariance (CWC). CWC coefficients are based on the covariances of second order textural measures, an optimum subset of them is proposed. The proposed approach is supported by a classifier based on multilayer perceptron network for the characterization of the image regions along the video frames. The whole methodology has been applied on real data containing 6 full endoscopic exams and reached 87% specificity and 97.4% sensitivity.

Index Terms— Color texture, computer aided diagnosis, image analysis, medical imaging, wavelet features

1. INTRODUCTION

Conventional endoscopy is limited to the upper gastrointestinal (GI) tract, at the duodenum, and to lower GI tract, at terminal ileum. So the vast majority of the small intestine, which has a medium length of six meter, isn't seen by these techniques. Therefore the capsule endoscopy allows the visualization of the GI tract, reaching places where conventional endoscopy is unable to. Images are captured, at the rate of two frames per second, by a short-focal-length lens as the capsule is propelled by peristalsis through the gastrointestinal tract. The result is a seven hours video with more than 50.000 frames per exam. Average small bowel transit time is about 90 minutes [2], then capsule reaches the cecum and visibility is severely decreased giving a total average of 15.000 useful images. Usually the physician is required to view 60.000 images and to select the ones that he considers important. This task is boring, time consuming and prone to subjective errors since most of the frames are normal, so it claims for computational assistance.

The automatic detection of lesions can be based in textural alterations of the small bowel mucosa surface. In the proposed approach the video frame sequences are transformed in scale by using the wavelet transform, since it has been observed that the textural information is localized in the middle frequencies and lower scales of the original signal [3]. The discrimination of normal and abnormal regions relies on a texture analysis scheme, supported by the statistical color wavelet features of each frame. The construction of the texture feature space follows the multiresolution approach on the wavelets extracted from the color domain. In this study, the features were obtained from the coocurrence matrices of the wavelet transform of different color spaces, at different scales, so that we have a coocurrence matrix for each band analyzed in the color space. Then second-order-statistics are computed between color channels, for the same orientation.

The feed-forward neural networks are, perhaps, the most commonly used networks for classification purposes. They were the first type of artificial neural network devised. In this network, the information moves in only one direction, forward, from the input nodes, through the hidden nodes (if any) and to the output nodes. There are no cycles or loops in the network. The Multi-Layer Perceptron (MLP) networks are commonly used in classification problems, because they have the ability to detect complex non-linear relationships in the data. There is an extensive range of applications of these neural networks, and so, a vast theoretical and practical background in this matter [4].

This paper is focused in the features extraction process from the wireless capsule video frames, with a method based in the correlation of statistical descriptors of cooccurrence matrix calculated for midband wavelet coefficients of each color channel of a given frame. These features are the input of a MLP network, in a classification scheme used to classify real data from Capucho's Hospital pacients.

The paper is organized as follows:

Section 2 describes the color wavelet covariance features, section 3 describes the multilayer perceptron architecture while section 4 describes the experimental results.

2. FEATURES EXTRACTION

This method relies on color textural features extraction process based in textural analysis. These features are estimated over the second order statistical representation of the cooccurrence matrix calculated from the wavelet transform of the colour image. The statistical descriptors calculated for each cooccurrence matrix give textural information about the properties of the decomposed subimages. These descriptors contain second order colour level information, which are mostly related to the human perception and discrimination of textures. For course textures these matrices tend to have higher values near the main diagonal whereas for a fine texture the values are scattered. The cooccurrence matrices encode the wavelet level (for each colour) spatial dependence based on the estimation of the second order joint-conditional probability density function $f(i,j,d,\theta)$, which is computed by counting all pairs of pixels at distance d having wavelet coefficients of colour levels i and j at a given direction θ. The angular displacement used is the set $\{0, \pi/4, \pi/2, 3\pi/4\}$.

It is considered only 4 statistical measures among the 14 originally proposed by Haralick [5]. They are angular second moment (F1), which gives a measure of homogeneity, correlation (F2), which is a measure of directional linearity, inverse difference moment (F3) and entropy (F4) defined respectively as

$$F1 = \sum_{i=1}^{N}\sum_{j=1}^{N} p(i,j)^2 \tag{1}$$

$$F2 = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} (i.j)p(i,j) - \mu_x\mu_y}{\sigma_x\sigma_y} \tag{2}$$

Where

$$\mu_x = \sum_{i=1}^{N} i \sum_{j=1}^{N} p(i,j) \quad \mu_y = \sum_{j=1}^{N} j \sum_{i=1}^{N} p(i,j) \tag{2a}$$

$$\sigma_x = \sum_{i=1}^{N} (i-\mu_x)^2 \sum_{j=1}^{N} p(i,j) \tag{2b}$$

$$\sigma_y = \sum_{j=1}^{N} (j-\mu_y)^2 \sum_{i=1}^{N} p(i,j) \tag{2c}$$

$$F3 = \sum_{i=1}^{N}\sum_{\substack{j=1 \\ i\neq j-1}}^{N} \frac{1}{1+(i-j)} p(i,j) \tag{3}$$

$$F4 = \sum_{i=1}^{N}\sum_{\substack{j=1 \\ p(i,j)\neq 0}}^{N} p(i,j)\log_2 p(i,j) \tag{4}$$

where $p(i,j)$ is the *ijth* entry of normalized cooccurrence matrix, N the number of levels of the wavelet and μ_x, μ_y, σ_x, σ_y are the means and standard deviations of the marginal probability $p_x(i)$ obtained by summing up the rows of the matrix p(i,j).

The proposed algorithm can be decomposed in the following categories:

A- Wavelet Domain Coefficients

Color transformations of the original image I result in three decomposed color channels, in the RGB color space:

$$I^i, \qquad i = 1,2,3. \tag{5}$$

where i stands for the color channel.

A four level discrete wavelet frame transformation is applied to each color channel (I^i). Daubechies filters of length 6 were used for multiresolution analysis purposes. These filters are very common since they are minimal phase filters that generate wavelets of minimal support for a given number of vanishing moments. This transformation results in a new representation of the original image by a low resolution image and the detail images. Therefore the new representation is defined as:

$$I^i = \{L_n^i, D_l^i\}, \quad i = 1,2,3 \quad l = 1,...,9 \tag{6}$$

where l stands for the wavelet band and n is the decomposition level.

Since the textural information is better presented in the middle wavelet detailed channels, the second level detailed coefficients were considered. Thus, the image representation consists of the detail images produced from (6) for the values l=4, 5, 6, as shown in figure 1. This results in a set of 9 subimages, where each color channel originates 3 subimages:

$$\{D_l^i\} \quad i = 1,2,3 \qquad l = 4,5,6 \tag{7}$$

B- Coocurrence matrix and statistical descriptors

For the extraction of the second order statistical textural information, cooccurrence matrices were calculated for the

nine different subimages. These matrices capture spatial interrelations among the intensities within the wavelet decomposition level, determining how often different combinations of pixel brightness values occur in an image. The cooccurrence matrices are estimated in four different directions resulting to 36 (3x3x4) matrices:

$$C_\alpha\left(D_l^i\right) \quad i=1,2,3 \quad l=4,5,6$$
$$\alpha=0,\frac{\pi}{4},\frac{\pi}{2},3\frac{\pi}{4}. \tag{8}$$

Where i stands for the color channel, l for the wavelet band and α for the direction in the coocurrence computation.

Four statistical measures given by equations (1), (2), (3) and (4) are estimated for each matrix C resulting in 144 wavelet features.

$$F_m\left(C_\alpha\left(D_l^i\right)\right) \quad i=1,2,3 \quad l=4,5,6$$
$$\alpha=0,\frac{\pi}{4},\frac{\pi}{2},3\frac{\pi}{4} \qquad m=1,2,3,4 \tag{9}$$

where m stands for statistical measure.

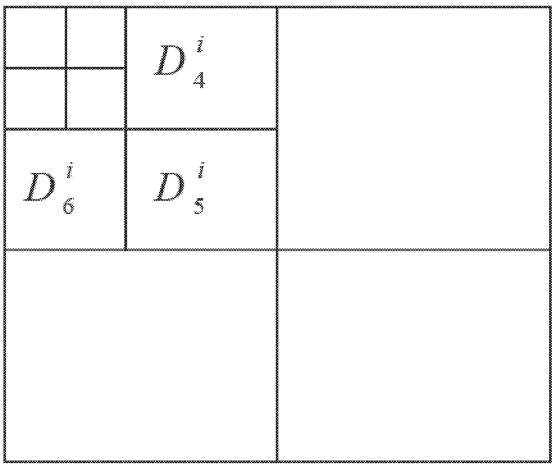


Figure 1. Three level wavelet decomposition scheme of the original image for color channel i.

C- Color Wavelet Covariance

Since each feature represents a different property of the examined region, the covariance among different statistical values between the color channels of the examined region, will statistically describe the textural behavior of the subimages, which will be very usefull information in our analysis. It is then expected that similar textures will have close statistical distributions and consequently they should have similar features. This similarity between features can be described by measuring the variance in pairs of them. Additionally the covariance between two features measures their tendency to vary together. The texture covariance has been proposed in the literature [6] as a measure used directly on image intensities or among the color intensities of the examined region. CWC coefficients can be computed based on the covariance of the same statistical descriptor, between different color channels, at different scales. This covariance can be computed as:

$$\gamma_{F_m,F_m}=\sum_\alpha\left[F_m\left(C_\alpha\left(D_{l+3}^i\right)\right)-E\left\{F_m\left(C_\alpha\left(D_{l+3}^i\right)\right)\right\}\right]X$$
$$\left[F_m\left(C_\alpha\left(D_l^j\right)\right)-E\left\{F_m\left(C_\alpha\left(D_l^j\right)\right)\right\}\right] \tag{10}$$

The color wavelet covariance features are then defined as

$$CWC_m^l(i,j)=\begin{cases}\gamma_{F_m,F_m}, & i<j\\ \sigma^2_{F_m,F_m} & i=j\end{cases} \tag{11}$$

Which results in a set of 72 components per frame. These components constitute the input of the feed-forward neural network.

3. MULTILAYER PERCEPTRON

The classification scheme described in this paper used a standard MLP network, with 72 input neurons, 2 output neurons (normal and tumour) and a variable number of neurons in the hidden layer. The performance of the network was tested for different configurations in the hidden layer, in the attempt of defining the most suitable number of neurons.

The training algorithm was the well known back propagation learning process, in which the values of each connection are adjusted in order to reduce the value of the error function. The two output neurons were used to classify the data into 2 classes, namely normal and tumour.

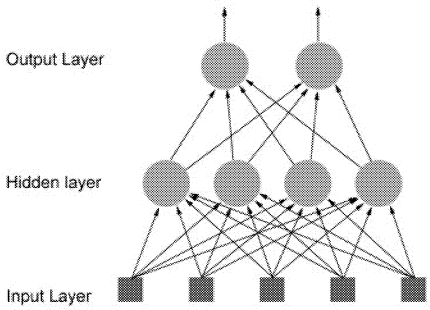


Figure 2. Example of a neural network with one hidden layer

4. EXPERIMENTAL RESULTS

The experimental set consisted of 6 full endoscopic exams taken at the Capucho's Hospital in Lisbon by Doctor Jaime Ramos. The system was trained in data that does not belong to the examined patients. The training set was

constructed with images from normal segments of capsule endoscopic videos, some of them taken from exams with pathological cases. The tumour images were taken from capsule endoscopy exams with this pathology. The final training dataset was composed by 100 normal images and 73 tumour images. Figures 3 and 4 show examples of normal tissue frame and a tumour tissue frame, respectively.

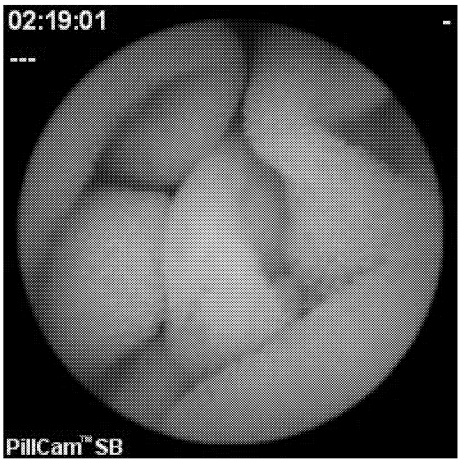


Figure 3. Example of a normal intestinal tissue frame

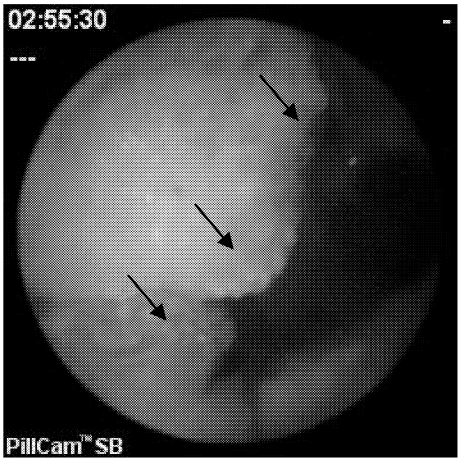


Figure 4. Example of an intestinal tumour tissue frame

Instead of measuring the rate of successful recognized patterns, more reliable measures for the evaluation of the classification performance can be achieved by using the sensitivity (true positive rate) and the specificity (100-false positive rate) measures. Therefore, sensitivity is the accuracy among positive patterns, while specificity is the accuracy among negative patterns [7]. These two measures can be calculated as:

$$Sensitivity = \frac{d}{c+d}.100\ (\%) \qquad (12)$$

$$Specificity = \left(100 - \frac{b}{a+b}.100\right)(\%) \qquad (13)$$

where a is the number of true negative patterns, b is the number of false positive patterns, c is the number of false negative patterns and d is the number of true positive patterns.

The classification performance is high when both Sensitivity and Specificity are high, in a way that their tradeoff favors true positive or false positive rate depending on the application.

A 3.2 GHz Pentium Dual Core processor-based with 256 MB of RAM was used with Matlab to run the developed algorithm. The average time processing per frame is about 2:15 minutes, which is fairly too much, but drops considerably without loss of performance if the size of the cooccurrence matrices is set to 64 X 64 instead of using 256 X 256 (full range). In any regular image, each pixel can assume 256 values (2^8 levels), that can be easily converted to 64 values (2^6 levels) as shown in equation (14). Note that with this operation, there is only a change in the number of the gradation levels for each position in the matrix. In this case the average time processing per frame is about 15 seconds.

$$V_{64}(i,j) = round\left[V_{256}(i,j)\right] \times \frac{63}{255} \qquad (14)$$

where $V_{64}(i,j)$ stands for the new value in 2^6 levels and $V_{64}(i,j)$ stands for the original value in 2^8 levels.

A mask was applied to the wavelet subimages in order to avoid computing cooccurrences in the image corners where no image information exists. The algorithm for computing cooccurrence matrices is implemented in such a way that only one passage on the matrix allows computing cooccurrences in the 4 required directions.

The first barrier in this work was the conversion of the wavelet coefficients to 256 or 64 gray levels, because the most of them are very close to zero, with some comparatively very large negative and positive values. The direct conversion into 256 levels, where the minimum wavelet coefficient is 0 and the maximum wavelet coefficient is 255, doesn't satisfy the performance criteria expected to this algorithm, because the most of the information is included in a few, very close, color levels, implying a very sparse coocurrence matrix, with only a few non-zero values. This can be solved with the proper dispersion of these very close wavelet coefficients to a more suitable interval. The extreme values are not appreciated, without loss of information. We can assume that the wavelet coefficients follow a normal distribution, with zero mean, so the data can be shifted and scaled, in order to be accommodated in the [0,1] range. As it is well known a random variable y with a given variance can be synthesized from a random variable x according to (15)

$$y = k \times x \rightarrow \sigma_y^2 = k^2 \sigma_x^2 \qquad (15)$$

The mean (μ_m) and variance (σ_m) of each wavelet coefficient matrix were calculated as the mean of the mean and variance of each row of the matrix. The new value for each wavelet coefficient was then calculated as:

$$W_f(i,j) = \sqrt{\frac{\sigma}{\sigma_m}}\left(W_f(i,j) - \mu_m\right) + 0.5 \qquad (16)$$

where W_f is the final value of the wavelet coefficient in the (i,j) position, W_i is the initial value of the wavelet coefficient in the (i,j) position and σ is the normalized variance of the wavelet coefficients after the dispersion process.

The effects of the variation of the variance in the wavelet coefficients and the number of neurons in the hidden layer of the multilayer perceptron network were tested, searching the optimal conditions for the performance of the algorithm. Table 1 shows the results for various variances of the wavelet coefficients assumed as normally distributed and also for different number of neurons in the intermediate layer. In table 1 *Nn* stands for number of neurons. The numbers between parentheses are the values for the normalized variance of the wavelet coefficients, according to (16).

Table 1. Experimental Results

Nn.	Se (0.3)	Sp	Se (0.5)	Sp	Se (1.0)	Sp
20	93%	80.5%	98.6%	79%	97.3%	71%
25	98.6%	81%	98.6%	80%	97.3%	86%
30	95.6%	83%	98.6%	81%	97.3%	87%
35	94.5%	86%	98.6%	82%	97.3%	84%

The results in Table1 show that the best sensitivity (98.6%) is achieved normalizing the variance of the wavelet coefficients to 0.5, for all of the network's sizes tested, while the best specificity (87%) is achieved by normalizing the variance of the wavelet coefficients to 1.0, in a MLP network with 30 neurons. The best overall results, considering the trade-off between sensitivity and specificity, are achieved in the MLP network with 30 neurons. There isn't any strong evidence that the further increase in the number of neurons in the network would lead to better results in the classification process. So, for this specific application and for a training set of 170 images, it can be concluded that the optimal number of neurons in the hidden layer is in the range [25,30], since the addition of more neurons consumes more computational resources, and simultaneously, doesn't improve the performance of the classification process. However the coocurrence matrix still had many zeros, with the most of the non-zero values clustered in the center of the matrix, so we can still increase the value of the normalized variance.

5. DISCUSSION AND FUTURE WORK

The results of this paper show that colour textural information can be adequate to classify images from endoscopic capsule. This colour textural information can be obtained from the covariances of the second-order statistical measures calculated over the wavelet frame transformation of different colour bands. The information present in the covariance of the selected features was successfully used in the classification of the images by a multilayer perceptron network.

However the performance of this method can be improved with some minor modifications. For instance, the colour space used is RGB, so it will be tested soon the proposed algorithm in other colour spaces as HSV, CIE-Lab or K-L. The enlargement of the dataset for training purposes is another important task to improve the classification process, since there is a wide range for normal tissue frames. The use of different wavelet bases will also be considered.

At the same time, they will be considered different approaches to the features extraction, namely a multiband algorithm based in the method proposed in this paper. The utilization of different classification systems, as Radial Basis Functions neural networks and Support Vector Machine classifiers, will also be subject of investigation.

In the future, our main goal is extend our work to other pathologies and develop a tool for automatic abnormalities detection to support the medical diagnosis.

6. REFERENCES

[1] Karkanis, S. A., Iakovidis, D. K., Maroulis, D. E., Karras, and Tzivras, M. (2003). Computer-Aided Tumor Detection in Endoscopic Video Using Color Wavelet Features. IEEE- Transactions on Information Technology in Biomedicine, vol. 7, Nº3, pp. 141-151.

[2] Iddan G., Meron, G., Glukhovsky, A., and Swain,P. (2000). Wireless capsule endoscopy. Nature, pp. 415-417.

[3] Abyoto, R. W., Wirdjosoedirdjo, S. J., and Watanable R. G. (1998). Unsupervised texture segmentation using multiresolution analysis for feature extraction. J Tokyo Univ. Inform. Sci., vol. 2, no. 9, pp 49-61.

[4] Haykin, S. (1994). Neural Networks. A comprehensive foundation. Mcmillan College Publishing Company New York.

[5] Haralick, R. M., (1979). Statistical and structural approaches to texture. Proc. IEEE, vol. 67 pp. 786-804.

[6] Chen, C. H., Pau, L. F., and Wang, P. S. P. (1998).The handbook of Pattern Recognition and Computer Vision, 2nd ed., Eds., World Scientific, Singapore, pp. 207-248.

[7] Swets, J. A., Dawes, R. M., and Monahan, J. (2000). Physiological science can improve diagnostic decisions. Psych. Sci Public Interest, vol. 1 pp.1-26.

MAPPING PELVIC FLOOR CLOSURE FORCES USING NOVEL MULTIDIRECTIONAL VAGINAL PROBE

Qiyu Peng,
Christos E. Constantinou CE
PAVA Medical Center and Stanford University
(Urology), CA, 94305, USA

Sadao Omata
NEWCAT Institute
College of Engineering, Nihon University, Koriyama,
Fukushima, 963, JAPAN

Email: ceconst@stanford.edu

ABSTRACT

Activation of the Pelvic Floor Muscles PFM to contract generates zonal compression of closure pressures on the urethra and vagina thereby contributing to continence. Localization of the direction and distribution of the muscular forces applied to these structures can potentially be of diagnostic value. Using a multidirectional vaginal probe, sampling of the muscle strength produced by voluntary PFM contraction was measured in asymptomatic female subjects. Individual contact pressure measurements were obtained, interpolated and used to generate a map of the forces acting along the length and circumference of the vagina. Visualization of the spatial organization of closure forces was computed to identify their topographical distribution.

INTRODUCTION

Contractions of muscles of the pelvic floor (PFM) contribute to closure forces acting on the urethra and are implicated in the maintenance of urinary continence. For this reason quantitative measurement of the forces generated are of clinical value in understanding the mechanism of stress urinary incontinence (SUI) and in developing approaches requiring conservative intervention (Bo 1992). Characterization of the magnitude and direction of the forces involved are not only of diagnostic importance but also can act as feedback in treatment if appropriately localized measurements can be made. Previously we developed a direction sensitive vaginal probe to measure the magnitude and orientation of generated contact pressures (Constantinou and Omata 2007, Constantinou et al 2007). In the present paper we report on a 3D visualization approach of force distribution along the vaginal wall of normal healthy volunteers.

METHODS

Human Subjects: Approval of this study was secured through the Institutional Review Board of Stanford University, who approved the experimental protocol. Informed consent was obtained from all subjects prior to the commencement of the investigation. Data presented in this report were obtained from 25 asymptomatic volunteers, age:51.5±5.3y, parity:one.

The Vaginal Probe: The mechanical details of the vaginal probe and details of its use have already been reported [ ]. Briefly the probe consisted of 4 force transducers mounted circumferentially at 90 degrees on a 23 mm shaft. Transducers were covered with a lubricated female condom. Upon insertion into the vagina, with patient supine, sensors made contact with the vaginal wall and contact pressures in the different four directions were measured. Data were acquired in real time, stored and also presented graphically on the computer screen as a feedback to the patient. The sample rate of the force and displacement signals was 25 Hz. Contact pressure *P,* derived from the contact area (*S*) and the force along the vaginal wall (*F*) is given by P=F/S where the diameter of the force sensor was 0.54 cm. Contact pressure, expressed in N/cm^2 is given by $P = 0.0428 \cdot F$. Calibration of force sensors were done using an electronic balance, with the data interpolated using a 4th degree polynomial equation.

Measurement procedure: Prior to any measurements, subjects were instructed to avoid straining or contracting their pelvic floor before generating a single cough at a time. The pressure obtained from each location, represents the peak value of contact pressure elevation for each cough and the average value of three coughs was computed and defined as: C_p . To increase the sampling of contact pressure measurements at any given depth, the probe was rotated by approximately 45º in some experiments and the procedure was repeated. The probe was moved from the superficial, to the middle and deep vagina for each set of measurements.

Mapping of contact pressure distribution The schema illustrated by Figure 1 was used to map the individual measurements of vaginal closure pressures. As indicated by Figure 1, the pressures at multiple points were measured under the assumption the pressure distribution on the vaginal wall is continuous.

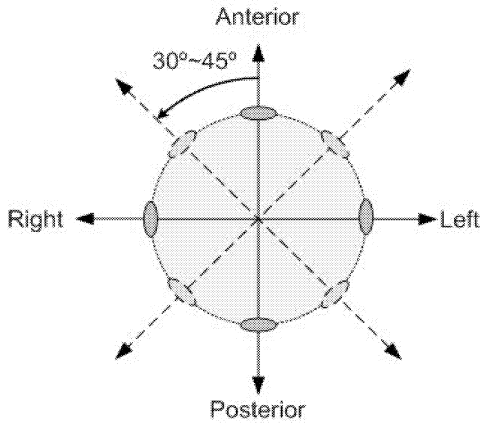


Figure 1 Radial orientation of positions measured by the contact pressure sensors of the probe. The four regions, anterior, left, posterior, and right are shown. Rotating the probe 30º~45º provides another set of data points. Therefore, the pressure on the whole vaginal wall can be derived by bi-linear interpolation. The rectilinear representation of all points is shown below by Figure 2.

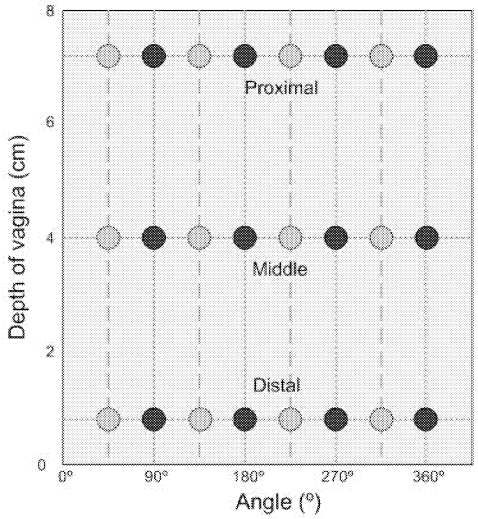


Figure 2 Rectilinear representations of contact pressures illustrating the 2D mapping of the vaginal wall from proximal to distal regions

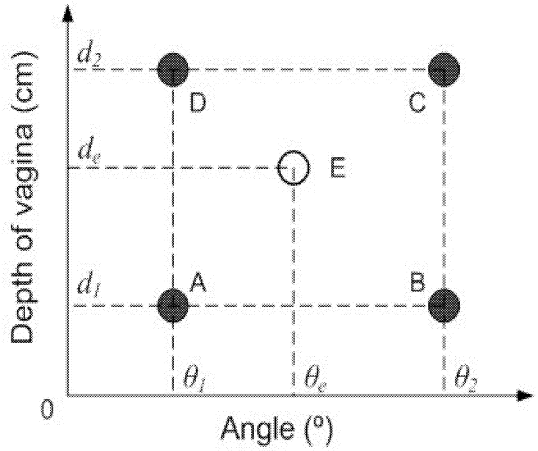


Figure 3. Bi-linear interpolations to calculate pressures at any given position.

For example, the pressure $P_e(t)$ on position "E" in Figure 3 can be derived from the pressures $P_a(t)$, $P_b(t)$, $P_c(t)$ and $P_d(t)$ on positions "A", "B", "C" and "D":

$$P_e(t) = k_a \cdot P_a(t) + k_b \cdot P_b(t) + k_c \cdot P_c(t) + k_d \cdot P_d(t) \quad (1)$$

Where:

$$\begin{cases} k_a = \dfrac{\theta_2 - \theta_e}{\theta_2 - \theta_1} \cdot \dfrac{d_2 - d_e}{d_2 - d_1} \\ k_b = \dfrac{\theta_e - \theta_1}{\theta_2 - \theta_1} \cdot \dfrac{d_2 - d_e}{d_2 - d_1} \\ k_c = \dfrac{\theta_e - \theta_1}{\theta_2 - \theta_1} \cdot \dfrac{d_e - d_1}{d_2 - d_1} \\ k_d = \dfrac{\theta_2 - \theta_e}{\theta_2 - \theta_1} \cdot \dfrac{d_e - d_1}{d_2 - d_1} \end{cases}$$

angles θ_1, θ_2 and θ_e, distances d_1, d_2 and d_e are shown in Figure 2.

During a cough, the pressure $P(t)$ on arbitrary position of the vaginal wall is a function of time *t*. Therefore, we can derive the pressure $P(\theta_e, d_e, t)$ on arbitrary position of the vaginal wall at any given time during the recording.

Statistical analysis: One-way ANOVA with Tukey's post-hoc test was performed to compare the peak pressures on three positions (Superficial, middle and Deep) of the vaginal wall. Results are presented as mean (standard error).

RESULTS

Figure 4 illustrates the amplitude and time course of a cough induced change detected at the posterior vaginal wall. As indicated the duration of the pressure elevation is approximately a second and its intensity 2.5 N/cm^2.

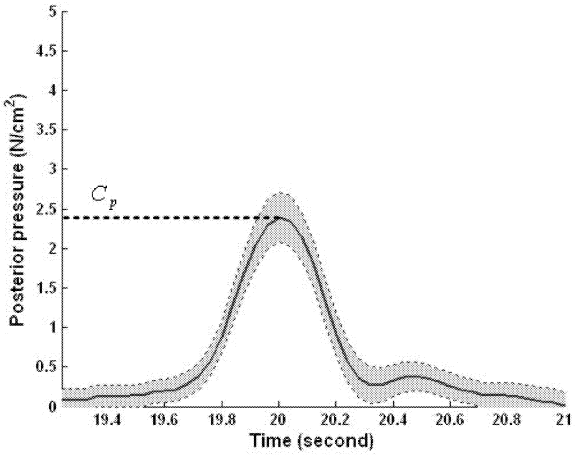


Figure 4: Characteristic pattern of a cough induced contact pressure rise measured at the posterior orientation in the middle region of the vagina. The maximum amplitude C_p is indicated and solid line denotes the mean value while the width of the dotted lines represents the SE from the mean.

A summary of the numerical values of C_p measured from all subjects evaluate is given by Table 1, showing that peak contact pressures at the Superficial region of the vagina wall are larger than those on the middle and deep positions of the vagina.

Table 1: Distribution of C_p relative to vaginal depth.

Orientation		Vaginal Depth
Anterior (0^0)	Superficial	4.34±0.71
	Middle	1.44±0.19
	Deep	1.61±0.54
Posterior (180^0)	Superficial	2.95±0.37
	Middle	2.40±0.32
	Deep	1.22±0.13
Left (90^0)	Superficial	1.63±0.40
	Middle	0.92±0.14
	Deep	1.07±0.19
Right (270^0)	Superficial	1.56±0.17
	Middle	1.38±0.14
	Deep	1.10±0.05

Furthermore at the Superficial position of the vagina wall, C_p in the anterior direction is significantly larger than that in the posterior direction. Whereas in the posterior direction is higher than those in the left and right directions and in the left and right directions are similar. Considering the middle position of the vaginal wall, C_p in the posterior direction is significantly larger ($p<0.001$) than those in the anterior, left and right directions. C_p in the anterior direction, of the deep position of the vagina wall, is also larger in the posterior, left and right directions.

The results of accumulating the individual cough responses obtained at the different vaginal orientations and depths are shown by Figure 5. The 2D mapping illustrated is captured at the peak of the cough reflex where C_p is at its maximum value.

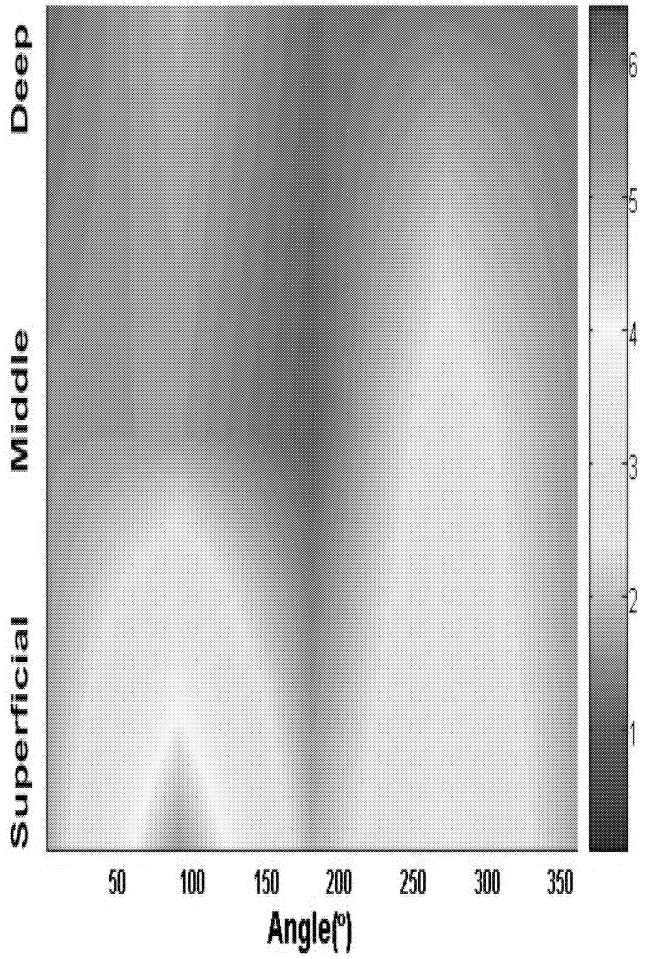


Figure 5 Mapping of the 2-D distribution the vaginal wall during cough captured at the C_p instant.

DISCUSSION

Pelvic Floor Muscles are important in maintaining urinary continence at any age but particularly in aging women. Clinical evaluation undertaken by vaginal examination detects the strength of these muscles by palpation which is qualitative at best and lacks spatial resolution. In the present study we generated a map of forces acting on the vaginal wall circumferentially and longitudinally, using a novel direction sensitive probe. The results presented identify the dynamic changes occurring during stress. These changes are suoerimposed on resting levels of closure forces (Peng et al 2007). We anticipate that the data, collected from healthy asymptomatic volunteers, recruited specifically for the study, can be used as a basis to compare the differences between normal and incontinent patients. In this sense using this protocol we expect to identify the precise location within the vagina where stresses such as coughing are likely to induce incontinence. The need to localize the zone at which PFM produce the maximum contact pressure generated a variety of instrumentation applied to the urethra (Constantinou and Govan 1982), Vagina (Domulin et al 2003, Whyte et al 1993) and (Guaderrama et al 2005). While these methods provided quatitative PFM measures of contractile force, ultrasound imaging (Peng et al 2007) and MRI (Constantinou et al 2002 and Dumulin et al 2007) topologically confirmed the anatomical distribution of closure forces. Ultimately it would be constructive to resolve the relative importance between the strength of the contraction and the anatomical region of its application given that imaging can be done non invasively.

In interpreting strength measurements its important to keep in mind that it is appropriate to incorporate the influence of insertion of pressure/force measuring device to the vagina. Given that currently the evaluation of the pelvic floor is qualitative, various quantitative measures of pelvic floor strength were developed using various trans-vaginal PFM dynamometers (Dumoulin et al 2002) enabling us to measure the overall force generated along the length of the vagina. The disadvantage of the dynamometers currently in use is that there is no spatial resolution in localizing the zones upon which forces are imposed by the contraction.

In summary, because multiple neuromuscular inputs converge to the pelvic floor muscles that contribute to urinary continence by activating the guarding reflexes. In this study we demonstrated that anatomically correct mapping of the role of the pelviv floor can be accomplished using a novel mechanical probe (Constantinou and Omata 2007) It is essential that the mapping of the closure forces be spatially identified and localized by mapping the different zones and their region of influence. Clearly, the mapping in 3D represent an approximate account of the dynamic events associated with refex contractions. A more complete visualization can be provided by mapping in 4D space to completely define the mechanisms involved.

ACKNOWLEDGEMENTS

This work was funded in part by NIH, grant 1R21 EB001654. We would like to acknowledge the contributions of R. Jones and V. Wolfe for clinical assistance in data collection and also Dr. I. Perkash in availing the urodynamic resources and facilities of the SCI unit of the Palo Alto VA medical system to undertake these studies.

REFERENCES

Bo K: "Pressure measurements during pelvic floor muscle contractions: 1992. The effect of different positions of the vaginal measuring device." Neurourol Urodyn 11: 107-113.

Constantinou CE & Omata S. 2007 "Direction Sensitive Sensor Probe for the Evaluation of Voluntary and Reflex Pelvic Floor Contractions. Neurourol & Urodyn 26(3):386-91".

Constantinou CE, Omata S, Yoshimura Y, Peng Q. 2007 "Evaluation of the Dynamic Responses of Female Pelvic Floor Using a Novel Vaginal Probe" Ann NY Acad Sci.1101:297-315.

Constantinou C.E., and D.E Govan. 1982 “Spatial distribution and timing of transmitted and reflexly generated urethral pressure in asymptomatic women.” J Urol;127:964-969.

Constantinou CE, A. Ryhammer, L.L. Nagel, J.C.Djurhuus. 2002 “Determining the displacement of the pelvic floor and pelvic organs during voluntary contractions using magnetic resonance imaging in younger and older women” BJU International (90) 408-414

Dumoulin, C., D. Bourbonnais, and Lemieux MC: 2003 “Development of a dynamometer for measuring the isometric force of the pelvic floor musculature.” Neurourol Urodyn 22: 648-653.

Dumoulin. C., Q. Peng, H. Stodkilde-Jorgensen, K. Shishido, and C. E. Constantinou. 2007 “Changes in Levator Ani Anatomical Configuration Following Physiotherapy in Women With Stress Urinary Incontinence”. J Urol 178 970-977.

Guaderrama N.M., C.W. Nager, J. Liu, D.H Pretorius, and R. K.Mittal: “The vaginal pressure profile.” Neurourol anUrodyn 24:243-247, 2005

Peng Q, R. Jones, K. Shishido, S. Omata, and C.E. Constantinou. 2007 “Spatial Distribution Of Vaginal Closure Pressures Of Continent And Stress Urinary Incontinent Women Physiological Measurements” 28(11): 1429-1450.

Peng Q, R. Jones R, K. Shishido, C.E. Constantinou CE. 2007 Ultrasound evaluation of dynamic responses of female pelvic floor muscles. Ultrasound Med Biol.

Whyte TD, D.S. McNally, E.E. James, 1993 “Six-element sensor for measuring vaginal pressure profiles.” Med & Biology Engineering & Computing 31, 184-186.

D-TV

OVERCOMING SOME LIMITS OF OUR INFORMATION BEHAVIOUR - CHOOSING RATIONALLY BY INTERACTIVE DIGITAL BROADCASTING

Larry Steindler
GTD - Gesellschaft für Technisches Dienstleistungswesen mbH, Düsseldorf
Visiting Lecturer at the Kunstakademie Münster
Germany
E-mail: artesliberales@web.de

KEYWORDS: *User behaviour, interactive broadcasting, interactive TV, reflective vs. determining judgement, knowledge gap, socio-technological barrier*

ABSTRACT

The growth of information and its possibility to be evaluated has a reciprocal relation to reflective judgement. Our respect for systematic knowledge and our general confidence in science grew enormously, but by knowledge explosion our capacity to find specific information fitting to any situation and any question has declined. The user staggers between excessive demands of new technology and his curiosity, and the particularisation of target groups seems to build up information barriers before the innovation process even reaches the quantum leap necessary. The analogue switch-off scenario has got to be the only reasonable way to master the situation. Yet for the sake of better arguments it is important to be aware of some limits of information behaviour.

Arguing that interactive digital broadcasting corresponds well with out ability of determining judgement this paper discusses limits of educational concepts, the knowledge and ability gap, socio-technological barriers, technological innovation as status-symbol, steady promotion and cross media reference, autonomy in media consumption, fragmentation of audience and uses, and some general information problems.

INTRODUCTION

Because of popular search engines in the internet more people are able to retrieve information. But this doesn't happen as adequate as by the accurate yield of information through data base systems. Their design has the purpose to target specific information dissemination. Since the internet is better known to a broader public information demand of the individual has increased enormously, but very often it really isn't aware of it.

However, facing the gap between accurate search results and practical use search engines more and more developed the character of giant but still quite inaccurate data base systems. The situation in watching TV is quite different: There are a lot of programs but you can't fail in watching a particular category. A movie play is a movie play and an animal documentation is an animal documentation.

THE USER AND THE INTERNET

At the beginning of the decade the statistical amount of estimated miss-retrieval was striking as the common accuracy of search engines indicated didn't exceed 20%–30%.[1] With the success of Google this percentage improved, but still remains behind from particular databases with specific input, specific queries, and exact output results.

According to a study of search engines' results by the endeavour of www.dogpile.com there was less than one percent average difference in the first page results of four major search engines (Google, Yahoo, MSN and Ask). This study from April 2007 measured about 19.000 user queries and proves by the dissimilarities in overlapping results, that searching the internet does not cope with scientifically reliable and demanded search methods. Search engines neither function similarly, nor index all available content on the web or deliver the same results.[2]

According to the study the percentage of total results unique to one search engine, not overlapping with the results of another, was established to be 88.3%, the percentage of total results shared by any two search engines was 8.9%, and the percentage goes down to little more than 2 % for results overlapping by the use of three different search engines. The endeavour to find a total amount of information on a particular topic by means of the internet turns out to be an illusion.

As people get used to the arbitrary sides of the web, hundreds of millions of users developed their information behaviour similarly – at least to a certain extent. With more than 100 million searches per day and nearly 20 million hours per month spent with Google (already a couple of years ago) this search engine proves not only that search engines are different in quality and quantity measures but puts a light on rational user behaviour. It demonstrates the

[1] cf. Carlo Tasso: Intelligent Digital Platforms for the semantic Web: New Technologies for Accessing and Filtering Information and Knowledge, Artificial Intelligence Laboratory, University of Udine/ infoFACTORY Group, Udine 2000 – euindia.dimi.uniud.it/wo6presentation/Tasso.ppt

[2] cf. Arnold Zafra: Search Engine Advocates Metasearch for Search Result Accuracy, June 13th 2007 – www.searchenginejournal.com/search-engine-advocates-metasearch-for-search-result-accuracy/5103/

frequent reuse of media as long as results are satisfactory.[3] This user behaviour also shows that it is more important to have a platform which presents sufficient quality results than to fulfil the demands of an objectively or scientifically omniscient database.

So by searching the web users seem to be very faithful: Jay McCarthy, vice president of web server log analysis company Websidestory, pointed out at the Search Engine Strategies Conference in Toronto 2005, that the number of referrals to pages deriving from search engines has surpassed those from direct links on particular pages. This means that people navigate the Web by searching more than by browsing[4] and that implies reflective judgement in order to be able for doing this.

REFLECTING AND DETERMINING JUDGEMENT

According to Immanuel Kant (1724-1804) reflective judgement is used to find facts and examples matching common concepts and notions, whereas we apply our capability of determining judgement to subsume facts and examples under broader terms and concepts.[5] Searching the web, we need both of these abilities, and we have to make use of the first and more difficult type – sometimes also called inductive method – even more often than we apply the second type. This determining judgement, our ability to subsume facts and examples under what we already know – sometimes called deductive way of reasoning – is less difficult but still requires some motivation. Though intuitive in manner and appearance the activity of browsing the web – or what is still also called "surfing" – requires determining judgement. In the whole a very complex rational behaviour is the basis to do the right choice relating questions, tasks and interests weather by the web or by any media.

INFORMATION BEHAVIOUR AND INTERACTIVE BROADCASTING

Mass media as vehicles for information, education and communication procedures are becoming linked within each other more and more. For this future TV functions and interactive programs have to be exploited as effectively as internet information. User habit is already adapting to new technologies as a study on young TV-watchers recently proved, that everyday TV-consumption is declining in comparison with the use of internet (including internet games). Ordinary TV-consumption requires "only" the ability of determining judgement because by switching to a certain program willingly you already made your choice to watch a geographical feature, a comedy, a western movie or a sequence of a criminal series.

Therefore we may say that to a certain extent complexity of the technology is not a big barrier for the use of any media but rather the content and its accessibility. Young users learned how to access to internet games though it is more complicated than to push a button at their TV set.

Limits of educational concepts: The demand for snappy imparting of key abilities and of practice-oriented knowledge is unignorable. Critics on school knowledge and the lack of it have became powerful in Germany since the first PISA-shock and also since the first signs of well suited endeavours which brought some relief according latest better educational results. The answer of the TV program consists very much in knowledge - or quiz shows, in which thrilling entertainment is produced by query of data and multiple choice questions instead of education which means coherent knowledge. Coherent knowledge is long lasting, communicable and fits to practical life. Quiz shows don't have to do anything with this type of practical knowledge: with key abilities which life and experts are asking for as well. Though the chance to attain what we call "education" is bigger by specified TV channels than by traditional TV broadcasting because of the conscious activities requested.

Knowledge and ability gap: One of the main borders of information acceptance lies in a lack of knowledge how to use tools and the absence of the wish to learn it. Interactive broadcasting gives an opportunity to simplify access to certain services. A recent survey stated that among a big majority of viewers EPG-services (electronic program guide) turned out to be as easy-to-use like remote control.

Socio-technological barrier: Trust in technological innovation is an important force to try new services. Interest to learn about them is highly limited if there is not enough confidence. As innovation doesn't occur always in a skilful step-by-step manner users become often demotivated and remain extraordinarily patient to wait for later improvements to gain better results relating to cost-benefit ratio. Only strengthening self-esteem related to the ability to make a good choice and technological curiosity overcomes information barriers like ignorance, demotivation, and the lack of time or money.[6]

Non-participating in status symbols: Or, on the contrary, innovative technical approaches of linked media appear to be activating and make users curious. They even become accustomed with malfunctions and rare content in beta-phases. To belong to an in-group already using highly developed technology sometimes is more important than a substantially proper utilization of the equipment. This process is a question of the right timing.

Steadiness and cross media reference: Continuity in building up information awareness and in innovation

[3] cf. 1 cog Webdesign, Bristol UK – www.1cog.com/search-engine-statistics.html

[4] cf. Websidestory, May 2005. www.seroundtable.com/archives/001896.html

[5] cf. Critique of Judgement, B XXXVI, (Germ. Edit.)

[6] „Kompetenz im Umgang mit Medien wird in der digitalen Informationsgesellschaft immer wichtiger. … Bürger und Bürgerinnen [müssen] mehr denn je beurteilen können, welchen Quellen sie vertrauen können und wo sie in der Flut der digitalen Informationen zuverlässige und seriöse Inhalte finden. Dies wird zu einem wichtigen Bildungs- und Erziehungsziel ... Noch fehlen aber konsequente Digitalstrategien in sehr vielen Bildungsbereichen." Deutsche Digitalcharta, Berlin 2007, hg. von Jo Groebel und Bernd Schiphorst, Deutsches Digitalinstitut, Berlin 2007, Leitsatz III, p. 16 – www.deutsches-digital-institut.de/downloads/IFA_CHARTA.pdf

promotion will fulfill their purpose. Repeated cross media reference by conventional TV-program on VoD or NVoD (Video-on-Demand or Near to Video-on-Demand) in the internet or in digital broadcast strengthen users' awareness. For the viewer it is important that he can choose content at a certain time, individually from a particular type of news or from an amount of trailers and features.[7]

Autonomy in media consumption: Consumers' expectations of digital television have to face their current viewing habits and their daily life. According to an IBM-survey on consumer anticipations people are aware of the advantages of individual choice of programs.[8] A survey in the Netherlands proves that young families "tend to regard the active, selective and individualistic viewing promised by the industry more as a threat than as an improvement to their current television use". Thats why "in the hectic and – by necessity – rigidly organised daily lives of these families passive, random and shared television viewing seems to make the best suited leisure activity".[9]

Fragmentation of the audience and uses: There is a gap between the forecasts of television operators and the outcomes of independent consumer studies as a result of fragmentation of the audience and use. Still moving towards individual accessibility of information and entertainment like through the internet and by interactive digital broadcasting is the only way to face consequences of this fragmentation gap. Until now PPV (Pay per View) and betting are the only media activities where the consumer is willing not only to use it but also to pay for it. Or, to put it more general, users usually take interest in digital advantages but they very often have either an aversion to the technical equipment or to the subscription. Nevertheless, promoters and producers of digital broadcasting have to conceive fragmentation of purposes and needs as a great chance for the switch-over to the new technology.[10]

Avoiding general information problems: Information retrieval or knowledge research through the internet faces four major problems: 1. oversupply, overload, 2. miss-retrieval, 3. untimeliness, and 4. information waste.[11] In its core digital broadcasting avoids these factors so that the future of information supply may be quite free from them.

[7] cf. Kabel Deutschland (ed.): Stellungnahme zur Revision der RL 89/552/EWG („Fernsehen ohne Grenzen") Themenpapier für die Liverpooler Konferenz zur audiovisuellen Politik. Regeln für Audiovisuelle Inhaltsdienste, München 2005, p. 3f.

[8] cf. IBM (2006): Konvergenz und Divergenz? June 2006; cf. DocuWatch Digitales Fernsehen. Eine Sichtung ausgewählter Dokumente und wissenschaftlicher Studien, (ed. Hans-Bredow-Inst. für Medienforschung an der Univ. Hamburg), 2/2006, p. 4

[9] DokuWatch. Digitales Fernsehen, 1/2006, p. 30 – short review on Aalberts, Chris/ van Zoonen, Liesbet: Televisiekijken in het digitale tijdperk, in: Tijdschrift voor Communicatiewetenschap, 33rd year, 2005 No. 4, p. 347-364

[10] cf. Digital Switchover in Broadcasting. A BIPE Consulting Study for the European Commission, April 12, 2002, p. 48-54, Cf. for the fragmentation of the audience and uses also Eric Karstens: „Fernsehen digital: Eine Einführung, Wiesbaden (VSVerlag für Sozialwissenschaften) 2006

[11] cf. Carlo Tasso, ibid. (note 1)

CONCLUSION

At least the first three problems don't occur by using digital broadcasting, whereas we should speak of information waste only if people don't switch on their TV-set.

Information waste in digital broadcasting would mean a waste of program capacity and economic resources as well. For the sake of the TV-watcher or interactive program user producers will avoid information waste as efficiently as they can. Relating to technological education in the long run digital broadcasting doesn't only face our predominant capability of determining judgement but serves information needs more accurately than ordinary TV program or the internet did before the time of web 2.0 approach.

REFERENCES

Ardissono, L., C. Gena, P. Torasso, F. Bellifemine, a.o.: Personalized Recommendation of TV Programs, in: AI*IA 2003 – Advances in Artificial Intelligence, Lecture Notes in Computer Science, Vol. 2829, Berlin, Heidelberg (Springer) 2003, p. 474-486.

Broszeit, Jörg: IPTV und interaktives Fernsehen: Grundlagen, Marktu bersicht, Nutzerakzeptanz, Saarbrücken (VDM-Verlag) 2007

Chorianopoulos, Konstantinos: Virtual Television Channels. Conceptual Model, User Interface and Affective Usability Evaluation, (Diss./ Department of Management Science and Technology at Athens University of Economics and Business) Athen 2004 –http://uitv.info/about/editors/chorianopoulos/thesis/phd.pdf

Cover, R: Changing channels: Scheduling, Temporality, New Technologies (and the Future of "Television" in Media Studies). In: Australian Journal of Communication, 32nd year, 2006, No. 2, p. 9-24.

Iosifidis, Petros: Digital Switchover in Europe, in: The International Communication Gazette, Vol. 68 (No. 3), London, Thousand Oaks, New Delhi (SAGE PUBLICATIONS) 2006, p. 249-268, here p. 250f. - www.global.asc.upenn.edu/docs/anox06/secure/july21/starks_tambini/21_starkstambini_reading4.pdf

Kunert, Tibor: User-centred interaction design patterns for interactive digital television applications, Diss./ Technische Universität Ilmenau 2007

Schröfel, Ariane: Interaktives Fernsehen. Grundlagen, Anwendungen, Perspektiven. Saarbrücken (VDM Verlag) 2006

Tegge, Svenja: Die Auswirkungen der Digitalisierung auf den Markt für Fernsehprogramme. Arbeitspapiere des Instituts für Rundfunkökonomie, Heft 220. Köln, 2006 www.rundfunkinstitut.uni-koeln.de/institut/pdfs/22006.pdf

DESIGN OF NON-COLLISION BROADBAND WIRELESS CHANNEL FOR DELIVERING OF MULTIMEDIA INFORMATION

V.M. Vishnevsky

Institute for information transmission
problems RAS
Bolshoy Karetny per. 19, Moscow, 127994
Russia
Email: vishn@iitp.ru

Tatiana Atanasova

Institute of Information Technologies -BAS
Acad. G. Bonchev 2, Sofia
Bulgaria
Email: atanasova@iinf.bas.bg

H. Joachim Nern

Global IT&TV GmbH
200710, Duesseldorf
Germany
Email: nern@global-ittv.com

KEYWORDS: *Wireless access, non-collision channel, multimedia information, radio relay.*

ABSTRACT

The main objective of the paper is to introduce an idea of constructing a new technology for wireless broadband to deliver multimedia information with high efficiency. The needed fundamental and applied investigation for the development of the new technology and the design of the software and hardware tools on the base of this new technology are outlined. New technology to deliver multimedia information and services with data speeds of up to 100 Mbps within broadband multimedia wireless last-mile solution is proposed.

INTRODUCTION

Digital TV gives a possibility to provide a wide spectrum of new services. This means more programming choices for viewers and interactive video and data services.

Digital programs broadcasting, customization of the TV content, increasing demand in interactivity and mobility – all of these factors insist on working-out of new technologies for multimedia information delivery.

Recently the development of the wireless information networks are recognized as one of the major directions in the progressing of the telecommunication industry. On one hand, this is determined by the intensive growing up of Internet and on the other hand – it is due to adopting new progressive methods for coding, modulation and transferring of wireless information. Now it is obvious that the wireless broadband networks are beyond competition concerning the operativeness and efficiency of settling, mobility, price, and variety of possible applications.

Wireless broadband offers for improving multimedia information and service delivery in digital TV (Atakishchev et al 2004).

One of the main advantages of the dynamically developing wireless technologies is in their potentialities for organization of “point-to-point” broadband radio channels that are profitably distinguished by their cost and productivity from the microwave radio relay lines.

The strong demand for broadband information delivery and access on the base of the popular standard *IEEE802.11x* is now satisfied by devices, that realize half-duplex regime and CSMA/CA (Carrier Sense Multiply Access/Collision Avoidance) method for multiply access, but this does not correspond to the modern requirements for the efficient transferring of fast expanding volumes of multimedia information. Sudden drop in the productivity evolves from the collisions that arise between packets, which are transferred in the opposite directions in the half-duplex channel.

Investigations on the increasing of the broadband wireless channels efficiency are conducted in many research and development world centers because of the current demand for wireless broadband information transfer.

For example, OFDM (as specified by the IEEE 802.11a) is a multicarrier modulation (MCM) scheme in which many parallel data streams are transmitted at the same time over a channel, with each transmitting only a small part of the total data rate. OFDM is used for wireless digital radio and TV transmissions, particularly in Europe.

OFDM is competing with CDMA (Code Division Multiple Access). While it has more robust transmission capabilities, it is currently more expensive to implement. Other techniques, such as UWB (Ultra WideBand) are also available. Cisco has developed a scheme called VOFDM (Vector OFDM).

In wireless networks different transmission techniques are used to overcome wireless transmission problems and to improve bandwidth. In (Glinos et al 2004) the algorithmic issues related to the delayed multicast technique for video-on-demand delivery are examined.

Now the wireless broadband access is preferred by the users that change their connections from analogue lines while the companies increase the speed of proposed channels. With the increasing number of various mobile devices now it is more convenient to connect to the IP networks by wireless

radio channel and not by cables which limit the mobility. At that, the IEEE 802.11 protocol ensures several megabits per second as transfer speed in all of its modifications.

It is observed that the market demands the multifunctionality of services. There is a need of ability to transfer the multimedia information during videoconferencing, to provide of VoIP and IMS services and the wide spectrum of services in the digital TV (VOD, games, media-adapters, pear-to-pear connections, etc.).

So the wireless networks devices with well developed communications capabilities and high speed of transfer of the multimedia information have to be developed.

BUSINESS MODELS RELATED TO DIGITAL TV

Digital TV combines two specific types of communication channels: channels for TV broadcasting and interactive feedback channels. The technical platform can ensure new business possibilities providing an infrastructure for creating new value-added services and facilitating e-business (Daskalova and Atanasova 2007).

Interactive feedback channel allows the user to be an active participant in:

- Receiving additional information
- Playing an online game
- Video on Demand

The new functionality requires different degree of feedback. In the (Atanasova et al 2007) framework for interaction with the semantic content describing multimedia information in digital TV is considered.

New segments of the business market, such as small business and home office customers, while utilizing its existing infrastructure may be covered by wireless technology. Wireless is the cheapest way to provide multimedia information access, personalized content services and quality of service to the customers.

Different user interactivity defines different services and these services can be provided by mobile tools that give wireless access for users with Wi-Fi devices such as laptop or personal digital assistant (PDA). Mobility is defining as the future of the Digital TV and multimedia data transfer.

The mobile technology may make the digital TV to be ubiquitous. The wireless technology can be implemented in remote areas where traditional networks are not feasible. And specifically wireless can be used in video broadcasting. New wireless broadband systems put the focus not just for voice but also for multimedia traffic at comparable highbit rate of several Mbits/s.

The most beneficial direction is to develop and deploy wireless systems where they are unique: mobility and access over inaccessible areas.

Multi-service (All-in-one service, broadband Internet, VoIP, video conferencing, IPTV, e-commerce and surveillance) for broadband data transmission in the highly constrained networking environment requires an emphasis on networking and data protocol design as much as on wireless transmitter and receiver technology.

NON-COLLISION BROADBAND CHANNEL

With the improvements in the bandwidth and associated speed it is now possible to transfer video and other forms of streaming data communications over packet-based wireless networks.

In packet-based networks such as asynchronous transfer mode (ATM) network or frame relay networks the communication is realized by packets delivering. However, during heavy traffic conditions, packets may be delayed and lost. This may cause poor performance of communications and multimedia data transferring. Collisions in the multimedia data (packets) transfer that is lost or delayed due to inadequate or unavailable capacity may result in gaps, silence, and clipping of video and audio at the receiving end (Vishnevskiy et al 2005).

Data packets are manipulating by queuing that is the commonly accepted tool for data communications flows.

Queuing is constructed to examine packet headers and to make decisions for routing data flows. But this techniques results in traffic delay or jitter. In wireless environment the queuing construction is used only to enable packet and radio-frame processing (Agamanolis and Bove 2003). For multimedia data delivery the overall added delay in real-time traffic should be held less than 20 milliseconds because of wireless systems are usually more bandwidth constrained and therefore more sensitive to delay than their wire line counterparts.

Also, there can be losses in speed in the packet-based networks because of the time needed for communication establishment.

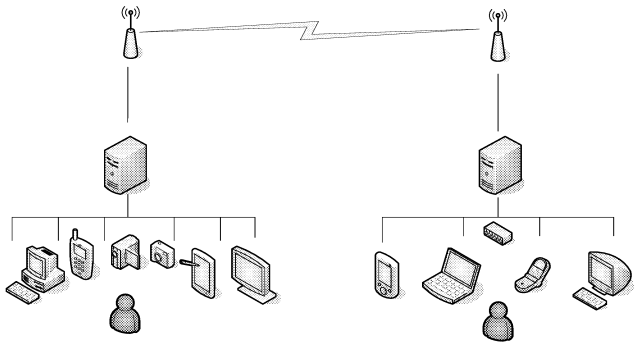


Figure 1: Wireless Network Distribution of the Multimedia Information

The bandwidth limitation of a wireless communication channel also has to be taken into the consideration. A bandwidth of a wireless channel (narrow or broadband) is always limited due to the limitation of available frequency spectrum. With this limitation, the only way to accommodate more users is to devise techniques that enable more efficient use of the given spectrum. That is exactly the reason why the latest the third generation (3G) wireless systems adopt Code Division Multiple Access (CDMA) for

channel sharing since it is more efficient than other sharing techniques.

Traffic congestion, out-of-sequence data packets, latency and jitter may influence on the multimedia information delivery (Vvedenskaya and Suhov 2007). In addition, wireless access introduces high inherent bit error rates, limited bandwidth, user contention, and radio interference.

Thus, there is a need for a method and apparatus that can ensure and increase the quality level (Sharma G., Ganesh A.., and Key) of the multimedia communications system.

This paper is aimed to introduce a concept and algorithm for the optimal dividing of the radio bands and receiving/transmittion channels into two channels working in opposite directions that excludes a possibility of collision and ensures significant simplification of the access channel mechanism. Transferring only in one direction will avoid collisions between packets. Every transition direction will utilize a full duplex regime.

Thus the reduction of needed resources with sudden increase of the operating speed of the wireless channel can be achieved (Figure 1).

For the design of the non-collision wireless channel, it is also necessary to develop:

- a concept and new algorithms on the channel level that will use in maximal range the non-collision nature of the channel by excluding random impediments obstacles during transfer which are foresighted by CSMA/CA method for collision avoiding;
- control algorithms to govern the queue in the transmission with multiplexing of different kinds of traffic. The algorithms have to render limitations on the efficiency indices for every traffic category.

Beside that, new methods of adaptive control for reliability of the transfer are needed that regulate the bounds for the repeated transfer and duration of packets lifetime. The methods depend on the type of packets transferring from the multimedia traffic and take into account the bounds of efficiency criteria as minimal value of the traffic capacity and maximal admissible variation for the time of packets transfer (Vishnevskiy and Semenova 2008).

The new algorithms for the channel level have to prevent collision by avoiding random delays during transfer. These random hold-ups are used by the CSMA/CA method for collision handling. The proposed non-collision channel has to rule the queue of the transfmission during multiplexing of the traffic with different categories. Furthermore, limitations on the performance measure for every traffic category have to be taken into account.

Computer simulations have shown that the proposed approach (Figure 2) can prevent the collisions in the broadband wireless and to increase the speed approximately by 30-40%.

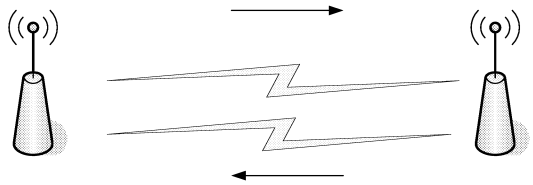


Figure 2: Non-collision Wireless Network

EXPECTED RESULTS

To realize the idea to its industrial implementation, it is required to conduct research on the creation of new technology for high speed transfer of multimedia information. On the base of this research, the experimental sample model of the duplex channel apparatuses can be constructed.

The main results are expected in the field of investigations on:

- principles of high frequency tract that can provide possibility of full duplex work with minimal interference:
- schematic design solutions for realization of non-collision broadband wireless channel for multimedia data transfer;
- new methods and algorithms for minimization of fluctuation of the time for packet transfer and selection of optimal parameters for block confirmation and packet aggregation that are adaptive to changed conditions of wireless transfer of information;
- a set of mathematical and simulation models for estimation of productivity of non-collision wireless channel where the input traffic has a highest priority;
- methods for differential handling of packets during multiplexing of different types of the traffic.

The industrial model of the constructed gears of the duplex channel can prove the research ideas and simulation results.

The software and hardware realization of the new technology for the wireless transfer of the multimedia information will provide construction of the reliable non-collision channel with the nominal speed of the transfer up to 100 Mbps and the distance up to 30 km that will be better than existing solution. It has to be mentioned that the cost of the existing microwave radio relayed channels with similar characteristics are approximately tenfold comparing to the proposed solution.

CONCLUSION

The Digital TV has its visible benefits. But the demanded interactivity and mobility insist on the development of new technology for wireless multimedia information transfer.

The goal of high-quality data transmission over a shared wireless broadband access system requires new and creative approaches to system hardware and software design.

An additional challenge is the problem of contention among users for limited wireless bandwidth. The system must handle service requests from multiple users in a medium of radio which is subject to interference and noise. This makes efficient bandwidth allocation difficult in existing schemas.

The proposed approach for the wireless network infrastructure may enhance the quality of the service provided to users by various communications services (including data, voice and multimedia services).

The simplicity of the proposed schematic design and software realization of the devices constructed on the base of the proposed technology will provide low cost and high competitiveness on extensive market of the wireless appliances.

The research work will bring as a result the industrial production of the apparatus, instrument tools and appliances for the wireless non-collision channels. The need for such instruments is very large in the industry, sciences and culture when information networks are designed and the "last mile" problem is resolved.

REFERENCES

Vishnevskiy V., Lyahov A., Portnoy S., Shahnovich I., "Broadband wireless information networks", Moscow, 2005

Vishnevskiy V., O. Semenova, "Adaptive dynamical polling in wireless networks", *Cybernetics and Information Technologies*. 2008. Vol. 8, No. 1

Atanasova T., H. Joachim Nern, A. Dziech, N. M. Sgouros, **"**Framework Approach for Search and Meta-Data Handling of AV Objects in Digital TV Cycles"**,** Proc. of Scientific Conference *EUROMEDIA 2007*, April 25-27, 2007, Delft, The Netherlands, pp.145-147.

Daskalova Hr., T. Atanasova, "Web Services and Tools for their Composition Considering Aspects of Digital TV Workflow", In: Proc. of Scientific Conference on Web Technology, New Media, Communications and Telematics Theory, Methods, Tools and Applications *EUROMEDIA 2007*, April 25-27, 2007, Delft, The Netherlands, pp.139-144.

Agamanolis S., M. Bove, "Viper: A Framework for Responsive Television" *IEEE MultiMedia*, 2003, vol.10, no.3, pp.88-98.

Atakishchev O. I., Emelianov S. G., Zakharov Ivan S., Klujkov B. V. "Features of Multimedia Information Transfer in Telecommunication Networks and Distributed Corporate Safety Networks", *Telecommunications and Radio Engineering*, Vol. 62, 2004 Issue 1-6, pp. 421-436

Vvedenskaya N.D., and Suhov Yu.M., Multiuser Multiple-Access System: Stability and Metastability, *Information Transmission Problems*, 2007, Vol. 43, No.3, pp.105-111

Sharma G., Ganesh A., Key P., "Performance Analysis of Contention Based Medium Access Control Protocols", http://research.microsoft.com/~ peterkey/papers/IEEE_ MAC_Diffusion.pdf

Glinos N., D. B. Hoang, C. Nguyen and A. Symvonis Video-on-Demand Based on Delayed-Multicast: Algorithmic Support. *The Computer Journal,* 2004 47(5):545-559

.

CRM 2.0 – Service Delivery and Value Chain of an Interactive Media Broadcast Platform

Wolfgang Rothe
Travel TV OOD, BG
CIO - Travel Television International Germany
E-mail: rothe@travel-television.de

KEYWORDS

Customer Binding, Semantic Web, Innovative Advertising, Value Chain , Customer Profiling, WebTV, Click-through Rate , Semantic Zoom

ABSTRACT

Interactive and collaborative elements are well established in the Internet community. For the Internet trading industry the impact of marketing which reflects an interactive customer approach is not yet established with mayor business relevance.
Although the media industry started multiple activities to launch IP-TV services a commercial success is still outstanding.
The proposed Interactive Media Platform integrates proven interactive communication aspects with semantic search technologies and an effective transaction behavior. This platform can easily be integrated in existing ecommerce platforms to attract social networkers, to enable a higher customer binding.
The customer is provided with an effective search algorithm which focuses supplier and demand aspects. The supplier can link his product and service world via a semantic web to the customer demand context and serve these demands with optimized One-to-One marketing activities.

MOTIVATION

Customer binding is a central challenge of Internet trading platforms. To attract a new customer traditional advertising via print, Internet or TV has limited success and is very expensive. New channels to the customer are established via online-communities and affiliate marketing. To add value to the customer the supplier has

- to deliver effective and pragmatically ergonomics and the transaction behaviour.
- to relay on personalised content to be able to present only products of interest
- to offer appealing product and service information to fulfil the demand of the customer for advisory service
- to incorporate recommendations of other customers and instances

Interactive communities are established on a broad scale. Nearly one third of al Internet users joined one of the existing platforms. From user research it is known, that the **social networker** has a 20% higher income than average /Zattoo 2007/ and has a higher online spending than average (**Figure 1**). These communities can be attracted by various marketing strategies like customer review and ratings, online customer forums, peer-to-peer transactions, product focused blogs and community related products.

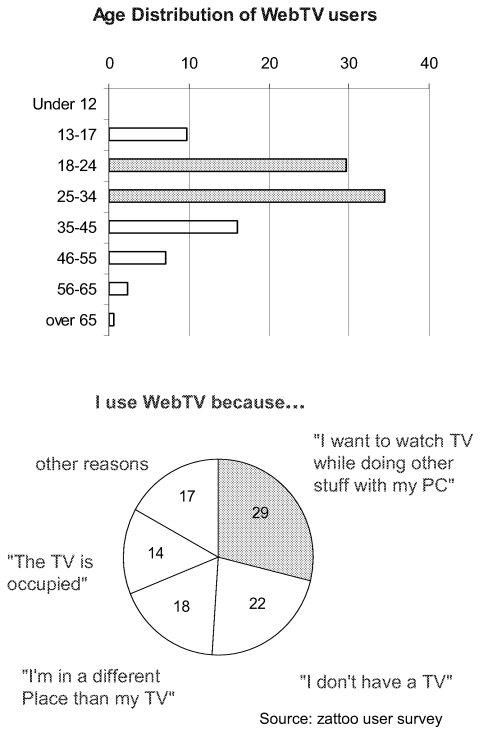


Figure 1: Online spending – Social Networker

In the TV-industry plenty of special interest channels have been established. The business model of these channels mainly relay on marketing revenues. Due to the limited range of coverage these revenues are limited. To leverage the technical infrastructure new hubs are created and offered as a package by various service providers. These offerings are available on the Internet with reasonable bandwidth. As a result for the German market 75% of all DSL customers are able to receive **WebTV** offerings on their local PCs.

Additionally the accessibility of advertising is depending on the advertising-dose, thus the pressure of advertising. Thereby is the contact-dose of a defined target group responsible for the recall performance, further the after deduction coverage express calculative fundamental contact-prospects.

STIMULATED CUSTOMER BINDING

Stimulation of customer loyalty is a challenge. To support attractive customers and to integrate their multiplication power into a effective marketing strategy proactive emails for new products, customer self service, follow up activities and bonus programs are well established. The incorporation of social networkers is the most effective way to support existing loyal customers and to generate new loyal customers to the supplier. The introduction of a managed community platform into an existing eCommerce portal attracts the social networker and their multiplication power can be used to increase customer loyalty.

Advertising with moving images

Due to information overflow in most of the Internet portals online advertising click-through Rates (CTR) are falling. Click-through Rates of existing banners like Popup layers, skyscraper, wide skyscraper, medium-, full size rectangles and leader boards vary between 0.6% and 0,11% . Only video ads have a higher rate of 4,6 % (**Figure 2**). Effective product advertising has to reflect this customer behavior and to use this type of product placement. The introduction of moving images to boost attraction combined with high quality TV-content related to the product is the only way to fulfill the requirements of an appealing product presentation in an Internet portal.

Reflecting product and customer context

The product catalogue of an Internet portal is normally structured via a fixed keyword based product hierarchy. The navigation in this product catalogue is often ineffective from the customers perspective. To solve this problem, a search string is offered in most the Internet portals. Because the customer does not know the internal product hierarchy, the results of his individual product search are ineffective, unless the correct semantic phrase is used, or the portal owner has introduced a full text search. After 3-5 tries the customers leaves the portal and the customer is lost. The introduction of a semantic web on both sides, the customers context derived from former visits or current navigation behavior and product context derived from an semantic import process which covers all relevant product descriptions solves this information gap. Thus the ‘ puma’ can be identified clearly as an animal or as a brand for sportive clothes. To enable semantic search capabilities in a portal, a new platform has to be introduced to existing Internet

Click-through rates by IAB format and country:

	Average	DE	UK	F	Italy	DK	FI
Pop-ups/ Layer	0,58	0,59	0,69	2,39	0,16	0,13	0,04
Video Ads	4,64	4,79	5,31	3,92	5,12	3,51	3,72
Button 2	0,05	0,42	0,13	0,38	0,07	0,02	0,08
Skyscraper	0,11	0,10	0,16	0,07	1,00	0,12	0,02
Wide Skyscraper	0,15	0,11	0,16	0,11	0,19	0,10	0,20
Medium Rectangle	0,20	0,20	0,25	0,09	0,19	0,12	0,14
Fullsize	0,20	0,14	0,08	0,19	0,38	0,17	0,08
Leaderboard	0,12	0,14	0,32	0,13	0,24	0,11	1,00
Ø	0,18	0,17	0,20	0,24	0,23	0,11	0,09

The table below shows the click-through trends of the last three years:

Nov 2004	March 2005	June 2005	Nov 2005	June 2006	Sept 2006	Nov 2006	Dec 2006
0,33	0,24	0,27	0,23	0,35	0,20	0,19	0,22

Source : www.e-finch.info/en/pr_07_10.html

Figure 2: Click-Through Rates

VALUE CHAIN

The value chain of the eCommerce portal consists of various parts (**Figure 3**).

- **Access:** This is the first layer which grants the technical access of a customer to a portal solution. The access via Internet is to be realized by existing service providers. This layer is not focus of the described platform. To enhance reaches via additional channels the TV content can be delivered via broadcast services via cable or SAT and HTV later on.

- **Solution:** This layer represents the suppliers portal in the web. This solution is to be enhanced with community services as described above and standard capabilities to serve affiliates and video on demand services. The deployment of these services will be supported by the platform.

- **Products/Content**: This layer represent all items to be sold via the portal. The existing product base is enhanced with video content from third party suppliers and user generated from the community. These items are linked via semantic

annotation to the a semantic web and made available to the customer context using fuzzy set technologies. This is the core of the Interactive media platform.

- Fulfillment: The media platform can integrate existing fulfillment services. CRM, billing, payment and logistic services will be incorporated from the supplier.

To support the market entry the technical platform will be offered as a managed service and accompanied with a consulting and marketing service to optimize the effort of product indexing and support the marketing strategies of the supplier.

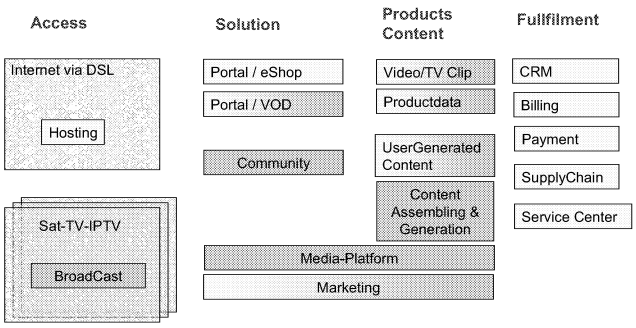


Figure 3: Value Chain and Fullfillment

SYSTEM ARCHITECTUE

The proposed Interactive Media Platform /Nern 2008-1, Nern 2008-2,/ consists of three layers:

- A backend layer which builds the interface to the content owners (B2B). Via standardized interfaces using web services, the moving images and the TV-content, which is associated with semantic values of their domain are imported to a **semantic annotation** module. This module generates a fuzzy set based organizational memory of the product domain. A semantic indexing is performed for every introduced product. The backend layer has an administrator console to maintain and control the semantic indexing.

- A frontend layer for the end user (B2C) which covers a **semantic zoom** functionality to guide the customer from a TV spot to the related products of his personal context, community services like chat, forums and an upload portal to collect the user generated content which is passed through a semantic filter for quality assurance.

- A technical infrastructure layer which consists of web-servers, streaming servers, load balancers etc.

MARKET OFFERING

As a consequence the proposed Interactive Media Platform enables a complete new service offering. Beside the existing product placements this platform can combine multiple service offerings to a combined service world. Within this service world products of various partners can be arranged around a predefined context (e.g. wellness, leisure, adventure, health) and enriched with high value video content. This service offering can be enriched with user generated content to increase new revenue and enhance customer binding. The context based match of products of various partners to the domain of interest of the customers via semantic fit gives a new service focus. In this sense it is a contemporary CRM 2.0 service offering.

AUTHOR BIOGRAPHY

Wolfgang Rothe was born in 1954 in Bochum., Germany. He studied Nuclear Physics at the University of Applied Physics in Bonn and conferred to doctorate in Metal Science at RWTH Aachen, Germany. After his research activities at the Max-Planck Institute for Iron Research in Düsseldorf, Germany he worked as an IT-Consultant for Industrial Automation and was later on responsible for Marketing and Consulting in the Telecommunication, Media and Internet Industry. He is currently active as a Principal Consultant in the TIME industry.

REFERENCES

Zattoo 2007, Zattoo News;
" User Survey in 2007";
http://www.comdays.ch/de/prog.html; Ref "Zattoo- Der Laptop wird zum Fernseher"

Adtech 2008, Adtech;
" Adtech Statistics and News";
http://www.adtech.info/news/pr-070510.htm

Nern 2008 -1, H Joachim Nern, Tatiana Atanasova;
"Semantic Television – a New Vision or a New Business Case Approach ? - Interactive Media Based Edutainment Realized as WEB 3.0 Environment"; 2nd Workshop on Digital Television, Euromedia 2008, University of Porto, Porto, Portugal

Nern 2008 -2, H Joachim Nern, V.M. Vishnevsky;
"Digital Broadcast Environment using Web Service Technology – New Approaches for Fuzzy Content Detection and Service distribution"; 2nd Workshop on Digital Television, Euromedia 2008, University of Porto, Porto, Portugal

SURVEY ABOUT RUNNING INTERNATIONAL IPTV PROJECTS – STANDARDS, SPECIFICATIONS AND FUTURE DIRECTIONS

Jan Elsner

Mastertape.TV
Langerstr.32 , Cologne, Germany
Email: elsner@mastertape.tv

(Trad. by H. Joachim Nern and Larry Steindler)

KEYWORDS: *IPTV, TV broadcasting, internet TV, economic development, internet portals, business policy, P2P technology (Peer-to-Peer), P2P-NEXT-project, open source, development study, conventional private TV*

ABSTRACT

Along the development and requirements of moving picture as part of file transmission technology in the Internet, interdependencies between radio, TV and internet broadcasting (IPTV) the author makes an outline of running business models. The article explains how different these projects are, which motivations stand behind them and which limits they have to face. It names evolving standards, specifications and organizational methods. The other main focus besides the technological one is on the economic impact of IPTV, its relation to Google's powerful advertising market initiatives, to international projects and to the private sponsoring program of the European Union and the EBU (European Union Broadcasting). Further emphasize is put on the P2P-Next-project, its technological principle, its community impact and its accelerative effects in the arising IPTV market. Regarding to this the author shows the dimensions and limits of capacity growth (Cisco Study), the role of open source solutions and licensing within the vivid scenario of IPTV as well as some long term consequences to private TV broadcast market.

INTRODUCTION

At the end of the 90's around the turn of the century, the first affords were made to spread moving picture through the internet. The companies Apple with its famous program Quicktime and Realnetworks were responsible for the development of streaming technologies. After internet technology by alternative transmission means became ever more interesting for a growing amount of business models, also software giant Microsoft engaged itself on the slow-growing market. Techniques bet on server/client systems. Most of the internet users at that time were still attached with ISDN or dialup modems to the internet. In common, basis of technology at that time consisted of two factors:

1. The development of file formats, which did not have to be downloaded completely from the internet, but were playable already during the loading procedure;

2. The media were stored by software (buffering) on the computer of the consumers, to guarantee a trouble free reproduction (practically as a virtual data dam).

BUSINESS MODELS

The offers at that time were not competitive because of missing range and rendering quality compared with similar means of transmission like classical television or DVD. However within the field of radio broadcast this looked already different: by the introduction of DSL-technology and new account models, with ISPs (flat rates) it was no more problem to receive radio through internet. Today you hardly find any radio broadcast which does not offer live stream. For quite a time (from 1998 to 2004) it was to be good form at classical television stations to offer their formats over internet too. The magic word "video on demand" went through the whole industry. For pure news broadcasting services like CNN, N-TV, or BBC immediately it became a practical obligation to make their program accessible live by the internet.

The business model was simple: the initiators expected more spectators since television reached the consumer also at their workplace. That business model did not work. The internet television became a victim of its own success. After the tragedy of 09/11 many news stations were forced to switch off their streaming offers from their websites. One the one hand the capacity of lines and servers were overloaded in such a manner that there was no transmission possible anymore. The costs on the other hand took their toll: while a pure audio stream of an internet radio requires in an hour per user only about 100 MB, video files in TV quality require nearly one Gigabyte (GB) capacity. If you even wish to offer HD-TV it requires considerably more. In this case 1.8 GB per user and per hour traffic capacity is required and this costs a lot of money.

Beside classical senders again and again there were also efforts of content providers to use the streaming technologies for the spreading of special interest formats. Until the end of the 90's a strong growth market was expected here too. The key word for this was "Business-TV", where another trend became visible: Many start-up companies purged into an euphoric gold rush attitude, as it sounded perfect to disseminate television fast and simply over the common internet. The technology became ever less expensive. But extremely high traffic costs should be avoided as nobody expected an audience of millions and airtime should be sold excellently, since the intention was to address a selected target group directly – in conventional TV business, a fantastic precondition. However you mostly didn't see anything from live-TV by the internet at these dotcoms – most of the offers resembled a

conglomeration for self-advertisement. Spectators were missing completely. The most eminent problem of the "internet TV" broadcasters consisted in the quality of their content. Television is a mass media and money can be gained in this business only with a measurable ratio earn. In order to achieve a good rating, you must offer good content – and this costs money. The situation of internet TV broadcast is the case as in ordinary television broadcast: it is the most expensive what exists in this business, the production of contents, which is accepted by the consumer. Of course for this you need expensive professionals.

YOUTUBE AND OTHERS

After 2004 the internet became ever more a general information medium. Everything could be found on the internet in digital form, on what also conventional mass media reported. Live content was not anymore the guarantee for the success of a portal. The user behaviour of the internet developed more and more in such a way that the user arranges his information for his own purposes. 2005 the portal YouTube started. In November 2005 YouTube received 3.5 million US dollar from the silicone Valley venture capital provider **Sequoia Capital**, which had helped also Google during the initial financing. In April 2006 the young company got additional 8 million US dollar from Sequoia. Soon about 65.000 new videos were high-loaded day by day and 100 million clips were watched daily (status: October 2006).

The popularity of **YouTube** can be explained with the large community, which can up-load video files and is able as well to evaluate as to put its comments on them. Since its establishment YouTube rapidly ascended to be the most prominent video portal in the internet. At the present we presume a market share of approximately 45 per cent. The evaluation of YouTube rose from US $ 600 million in the spring 2006 to 1.5 billion US dollar in the autumn of the year. According to a New York Post report there were companies such as Viacom, Disney, AOL, eBay and Rupert Murdoch's News Corporation – the parent company of the New York Post – interested in a purchase of YouTube. By the purchase of the portal MySpace for 580 million US dollar in the year 2005 Murdoch's media empire sounded the bell for new internet purchase intoxication. On October the 9th 2006 YouTube was bought from the search engine operator **Google** for 1.31 billion Euro (in shares).

For conventional mass media meanwhile YouTube & Co. has become a genuine problem. The advertising economy recognized since long that such on demand content portals represent an outstanding platform. Not only that you can advertise purposefully, it is also easily possible to receive by evaluation of the "user behaviour" direct conclusion of his habits and preferences.

CURRENT SITUATION

IPTV is the keyword of the media industry in the present – but what is behind it?

An exact definition does not exist. IPTV is a mixture of technical achievements deriving from transmission of moved picture contents by the internet. IPTV is not even defined whether it concerns itself with the transmission of Live pictures or video on demand services.

In the video on demand business (**VoD**) the fight is in full activity. YouTube, according to the statement of a study of a prominent American IT enterprise, was responsible for scarcely 10 % of the world-wide data traffic in the internet in the year 2007. This conclusion is a result of own computations of Ellacoya network, a provider of telecommunications solutions. This US company at the same time determined a decrease of Peer to Peer Traffic (P2P). After four years, in which the **P2P**-data transfer was absolutely dominating, it is now for the first time HTTP traffic prevails again. 10 % of data traffic is a lot; until now each company went bankrupt which tried to put at stake one per mill of it. Server costs must be about the range of a million.

However, **Google** seem to speculate on a long-term basis on the success of the market. Nearly since two years YouTube is already in the possession of Google and for not more than a couple of weeks Google began to experiment with advertisement. Also in this matter Google shows up to be extremely generous: if somebody places his content at YouTube he can select whether he wishes additional advertisement or not – if the upload procedure permits it. Thus he gets a good piece from the cake. The advertisement runs directly over the Google cash cow service Adsense, the only division of the company, which makes profits. Since a long time Google is no search engine anymore. It is the **largest** advertising marketing company in the world. The expensive investment in YouTube will be worthwhile for Google; it is a credit on rates. It is not all about for YouTube to earn money compellingly by advertisement by videos being switched, but rather to bring a market promptly under control.

Google already took in the internet advertising market for scarcely 2 years. Because of the purchase of DoubleClick Google nearly can't be caught up any longer. In the future Google intends to be not only the number one provider in on line advertisement, but also in newspaper advertising, in radio spots **and TV** ad and to make thus media agencies and even advertising agencies redundant. Google has built up its own department with not less than 1000 employees, whose task consists of adapting and transferring AdWords to radio broadcast purposes – project name: AdSpots. This shows to be a thorn in the side of the advertising industry since Google can offer more favourable Spot seconds with the systematic of AdWords by practically selling an entire AdWords package: beside advertisement in the Web also that in print, radio and TV.

Concerning the dissemination of video content YouTube is meanwhile a important factor. TV producers start to bring television sets on the market which access the portal by network interface. The classical media companies recognized the trend too: Meanwhile every media giant has its own **Myvideo.com**, Clipfish, Sevenload or LiveLeak.com portal. After the fight for the VOD market seems to be

nearly decided, a new fight for a long-contested market begins – the spreading of conventional TV broadcasters on the internet. The difference to the well known dot.com blister is small but fine and is more detailed than before. Formerly the promoters tried to spread internet television or to bring the internet into the television – today thoughts tend to go rather to spread the television with the support of the internet.

In the last six months systems appeared on the market, which makes conventional television program by the internet accessible. In **China** this branch is mostly widespread by the CCTV service at the moment. There are two main reasons that conventional television is moving along this way just now and as slowly as it does:

1. While VOD services like YouTube only offer many different clips of an average play length below 3 minutes, and above that very often in bad quality, you have to admit that the users of **live-TV contents** by live stream watch the program for several hours and in a substantially high quality. This needs a multiple from traffic than YouTube.

2. There are problems with the **rights** and their clarification. While YouTube bet on user generated content, of which rights are free usually, TV broadcasters have to clarify the rights for every film they show. German TV market e.g. is one the most expensive ones, since a broadcaster who sends in Germany, and offers its service to German speaking audience also in Austria, Liechtenstein, Luxembourg, in parts of Switzerland, Italy, and Belgium, has to buy the rights for any of these countries. Above this, in such contracts at present the spreading in the internet is excluded since years and would cause additional costs.

EUROPE AND RELEVANT PROJECTS

In the last years TV broadcasters negotiated well with their partners, and film lenders slowly accommodate the media corporate groups with their needs. Also on Peer2Peer there are already existing alternatives for quite some time. **Octoshape**, a Denmark company, developed a plug-in which makes it possible to receive conventional live streams as P2P stream. The list of the customers is long, especially international, national foreign broadcasters like the Deutsche Welle, TVE – Televisió n Españ ola as well as the EBU European Broadcasting Union, the union of the broadcasting stations of the European Union ruled under public law. Also the parliamentary television of the European Union offers a live stream through Octoshape. But the European Union as well as the EBU slowly withdraw themselves from the Octoshape program and invest into its own open source solution.

At the beginning of March a consortium and institutes of European universities were established in order to launch the P2PNext project. Altogether 21 partners from the field of private economics and research are involved in the European Union promotion project P2P-Next. One goal is the development of a European-wide "NEXT generation" distribution system for internet television on the basis of P2P-technology. The European Union provides 14 million Euro for the development of the related programs (in its **"7th Framework"**-program Financing program for the promotion of the research and the capacity to compete of the technology industry in the European Union). Private investors and the EBU invest further 5 million Euro. A substantial technical aspect in the context of the project is the efficient distribution of new content by P2P-technology solutions. The dissemination of content which already many users received by P2P-technology solutions is simple: Everybody can offer contents once purchased for his part as „Seed". The distribution of new content in a P2P-network can prove itself to be difficult because of small Seed numbers. This applies particularly with contents, which very many users wish to see in real time if possible, as for instance sports broadcasting.

The **P2P-Next-Project** is going beyond the pure technical aspects and dimension. Also legal and regulation aspects are important and covered, because P2P-protocols are known for the use in illegal file-sharing and therefore have some kind of "bad reputation". To strike against the provisos against P2P-technology activities are going on to show up to network providers in which way "legal" P2P technologies can be applied for serious distribution of services. A further juristic aspect is already clarified: the outcome of the main Software technology will be **open source**. The project duration will be up to 4 years and first system tests will run in May 2008 within the online broadcast of the Eurovision Song Contest. The leading hand is with **VTT**, the Technical Research Centre of the University of Helsinki. The technological basis of the project has also been provided by the VTT – the P2P exchange bourse **Tribler**.

The P2P-Next project is showing up new dimensions – it is the first one, where different partners from different disciplines (research / industry) and countries are engaged to set up adequate specifications and standards. Furthermore it is the first project in this area which not only related to PC / Internet reception but also related to **set top based television** (TV sets).

The American company Zattoo has developed a system for cost free distribution of a broadcast. The software is working based on the P2P principle – like the exchange bourses. Accordingly no traffic costs occur on the provider side. Every body, who is receiving the broadcast is also distributing it. The network load is distributed among the users. For Zattoo just the provision of manageable network capacities is necessary. An interesting feature of the **business model** is that the advertisements are faded in just in case of a channel change – and also user specific data are collected, reflecting the user behavior. At present **Zattoo** achieved to gather up to 25 German-speaking broadcasts and in the following weeks the complete broadcast offer of ARD and ZDF (again approx 25 programs) would be applied. In this way Zattoo has established a service based on the broadcast of 50 German-speaking broadcasts within just one year.

AGAIN GOOGLE AND YAHOO AND OTHERS

As well some days ago **Google** has pronounced to provide life stream services via YouTube for this year. What

Google is planning in detail is not clear at the present and published. – it is just quite mysterious.

Yahoo is already present with an online live streaming offer. Within live.yahoo.com everybody can switch on and connect his webcam. It is to be assumed that Yahoo is more oriented on somekind of „live video blog“ than on broadcast of high quality content.

A further interesting and serious project is Joost, which is driven by developers and P2P pioneers of the first exchange bourse **Kazaa**. After several struggles in copyright and IPR issues Kazaa was closed. But the technological development was followed up – and out of Kazaa the VOIP messenger **Skype** has derived – also capable for video conferencing via P2P.

Joost is just oriented on English speaking clients and VOD end users. Since end of March also Joost is experimenting with the broadcast of live content, e.g. sport broadcast of NCAA (National Collegiate Athletic Association). With respect to VoD services Joost has access to content of considerable and namable content providers – mostly from the US market. For the beta phase contracts are closed with **Paramount**, Warner Bros., Viacom, Endemol and Turner Broadcasting System. Joost defines itself not as a pure IPTV service, but is grounding in the community aspect and issue.

FUTURE DIRECTIONS

Meanwhile it is not the question if “IPTV will come up”, but rather “which kind of technology will be prosperous”. For the international network providers the broadband capacity is the main problem. A prognostic study of Cisco says that already next year the consumer IT traffic will reach and overtake the business traffic.

Accordingly within the next four years the consumer traffic will increase by 58% a year, whereas the business traffic will increase just by 21% a year. So in the year 2011 monthly **28 Exabytes** (28 Billion Million Bytes) will flow via the networks as IP Traffic – hereby 3/5 as consumer traffic – as predicted by Cisco.

The biggest contribution to this fast growth will be the IPTV traffic. So in 2011 only 40% of the IP traffic will be conventional internet traffic. The rest of 60% will be due to commercial video services, which are distributed via IP (in 2005 the part of pure internet traffic was proportionally 80%). The major part of the IPTV signals will be the « Internet Video-to-TV traffic» . That is due the fact that at present the predominant « Internet Video-to-PC traffic» is dominated by short formats of low quality however the « Video-to-TV traffic» consistsof long term formats with higher quality. The so far biggest part of IP traffic – the P2P traffic – will be quadrupled and will reach in 2011 a volume of **two Exabytes** per month as the Cisco prediction says.

On the long run it is expected that open standards will come up in the market. The providers of IPTV offers and the network providers are called for being oriented on **common unique standards** to get the traffic under control. This will be handled by mirror- and Proxy-server distributing the net capacity on the network.

The media industry has to be focused on the development of new business models. So far used licensing rights for content are going out of time. General and international wide new rules for distribution licenses for program supplier have to be established and/or reworked. At present every broadcast station needs a state-approved license – vendors like Joost are broadcasting already without any license. The **governmental regulation** could not been fulfilled. – It seems that the regulation will be achieved by the market itself. In future sanctions against program vendors will be ruled by an international adaptation of the civil and criminal law. The entertainment industry has to agree on open standards like already made by DVB in the field of digital broadcast. In the past the industry could impose the user with special types of systems – but nowadays the industry has to learn to be oriented on the demands given by the user.

The future will be **OPEN SOURCE** – individual offers – open standards. On the long run software packages like WindowsMedia or RealNetworks will loose their market influence due missing compatibility. Who in future still will prefer old license models, DRM and proprietary codecs will be “punished” by the user and audience via disinterest. At the present companies like Google, LiveLleak, Zattoo and Joost already show up, that it is possible to make winnings just based on “new” business models. Already at the present in the US most of 20% of the current ABC TV series hits, like CSI, Dr. House and Co are not consumed via classical TV but via the streaming platform http://go.com (a joint venture of ABC, Disney, ESPN).

CONCLUSION

What will be the television in future? What is the influence of IPTV on today’s markets?

The future economical situation of conventional private broadcast stations, whose business model is solely based on advertisement and whose market value is still determined by audience ratings will be not the best – no to say: it will be worst and black.

By entering of IPTV the advertisement industry has the main advantage. Based on IPTV there will exist no uncertain representative quotas – there will exist certain and precise accounts and statistics. On the long run the influence of IPTV will result in the decrease of the conventional private television. In Europe public broadcaster will only exist furthermore, if a common European rule framework could be established based on a European common broadcast governmental contract. In the future the costs for the technical installation of a broadcast television station via Net will be just a fraction of the conventional ones. This yields self-evidently in an increase of the broadcast vendor spectrum. Accordingly this will influence the advertisement market – the incomes via advertisement will decrease. The distribution of News and information will solely be done by international cooperating broadcast joint ventures.

SEMANTIC TELEVISION – A NEW VISION OR A NEW BUSINESS CASE APPROACH ? - INTERACTIVE MEDIA BASED EDUTAINMENT REALIZED AS WEB 3.0 ENVIRONMENT

H. Joachim Nern

Global IT&TV GmbH
40710, Duesseldorf
Germany
Email: nern@global-ittv.com

Tatiana Atanasova

Institute of Information Technologies -BAS
Acad. G. Bonchev 2, Sofia
Bulgaria
Email: atanasova@iinf.bas.bg

Georg Jesdinsky

Big7.Net GmbH
Liebigstr. 14, Duesseldorf
Germany
Email: gj@big7.net

KEYWORDS: *semantic television, semantically based content formalization, Web 2.0, Web 3.0, semantically based knowledge distribution*

ABSTRACT

The main objective of this paper is to introduce an idea and vision of new kind of interactive television and broadcast – the semantic television, as a merging of Web 2.0 characteristics and Semantic Web technologies and methods. In this sense semantic television is defined as accessing and processing existing knowledge as well as creating new TV related content and information pools formalized using semantic web techniques and methods. Distributing of adequate formalized TV content is realized as a Web 3.0 interactive media platform, whereas the acquisition of content and production of TV content is oriented on the Web service paradigma.

INTRODUCTION

New challenges and possibilities to provide a wide range of new services are given by Digital TV – might realized as a pure Internet application in case of IPTV and/or WebTV /Elsner 2008/ or as interactive set top based digital television. This implies also increased programming choices for users as well as the access for interactive video and further data services.

Digital programs broadcast, customization of the TV content as well as the increased demand for interactivity and mobility insist on developing new technologies for multimedia resp. television specific information delivery.

Meanwhile several international vendors provide digital broadcast streaming. Also the German providers Deutsche Telekom, Arcor and Alice (Hansenet) distribute via DSL broadband live streaming products. However the base taxes and fees are quite high and are comparable with cable TV fees. The streamings consist of MPEG2 streams, which are provided via set top boxes. Using the VLC player this content is also technologically accessible via PC – but officially the usage is not supported. The Telco providers deploy this reception within a closed network (IPTV).

THE DESIGN OF THE MEDIA-PLATFORM

As depicted in Figure 1 the broadcast media screen is realized as a split screen window consisting of mainly 7 sub-screens. The main characteristic and main realization condition has been that each sub-screen is changing undependable from each other. This type of visualization requires either the application of a pure frame technique or the provision of portlet techniques /Doussier-2008/. The sub-screens are given as follows:

- Screen-Window 1 (SC 1): The main navigation window, for selection of Program, Semantic-Zoom, Communication (user specific actions), Marketplace, and Services
- Screen-Window 2 (SC 2): The main broadcast window, content depends on the chosen selection in Navigation in Screen 3
- Screen-Window 3 (SC 3): Navigation window for the main broadcast window
- Screen-Window 4 (SC 4): Gives the main title of the current broadcast resp. communication action
- Screen-Window 5 (SC 5): Gives the running broadcast as small window, the Zoom button shifts the stream to the main broadcast window, below the stream keywords are given, representing the current program
- Screen-Window 6 (SC 6): Window for information and communication activities selected by the current user (choice selecting the current program scheme, video selection, photo selection etc)

In the present state the screens are realized in conventional frame technique for demo purposes.

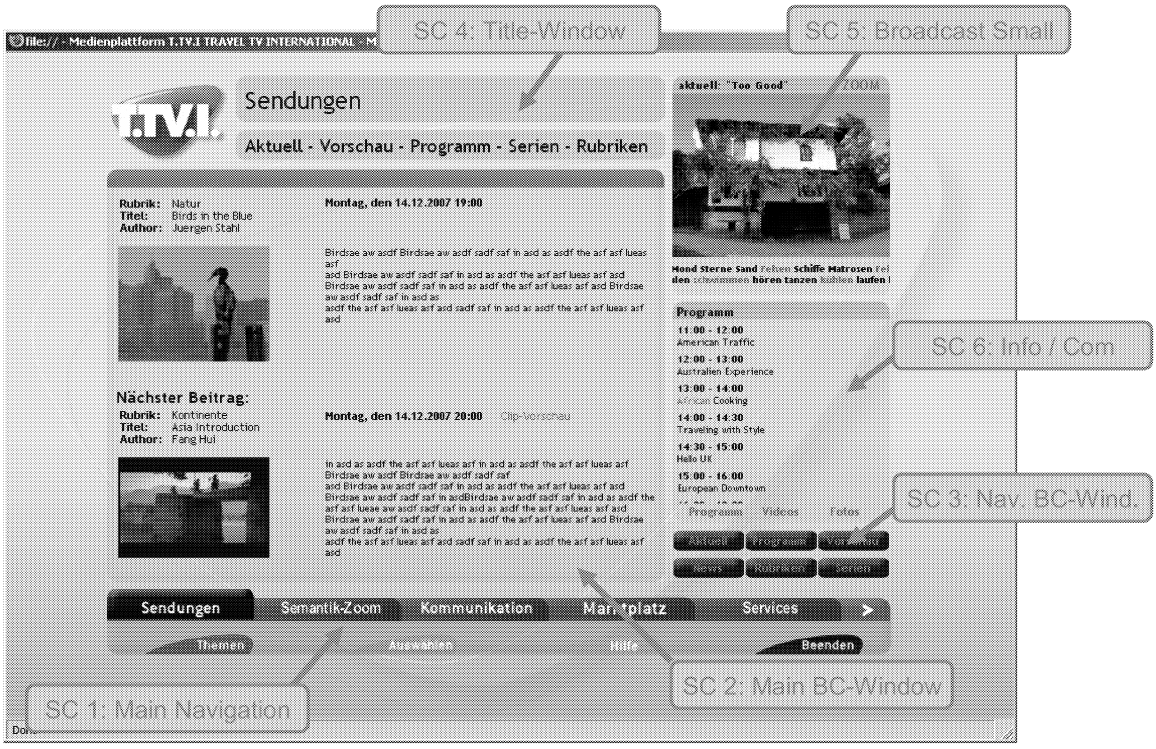


Figure 1: Start Screen of the Media-Platform

The click on the button ZOOM in SC 5 shifts the running stream to the main broadcast window as illustrated in Figure 2. In the screen window SC 5 the former stream is replaced by a news and information window, which gives access to news section, the as well as the forum and the chat section. Further screens are given in the appendix.

Figure 2: The main broadcast window – the change from small to full broadcast view

SEMANTICALLY REPRESENTATION AND ANNOTATION OF THE BROADCAST CONTENT

The semantically processing of multimedia content is a quite young activity and research topic – one of the first papers related to the use of Semantic Web technologies with respect to the handling of multimedia content is given by /Dowman 2005/. The paper describes the Rich News annotation system, which semantically annotates television news broadcasts using websites as a resource to aid in the annotation process. The decision to extract the semantically annotation data from web sites was due the poor quality of ASR (automatic speech recognition) of the audio track of the broadcasted stream. The authors furthermore pointed out that Rich News is essentially an application-independent annotation system applied in the first step to BBC news. In this work the first time the term of semantic television was used.

In the current media platform the annotation process is fulfilled by processing the redaction workflow data assigned to each AV object (reportage, documentation, clips etc.). The broadcast workflow is supported by a set of software tools, e.g. the program generation and handling, the processing of AV-Objects as well as as the semantically part. In Figure 3 the semantically parser is illustrated.

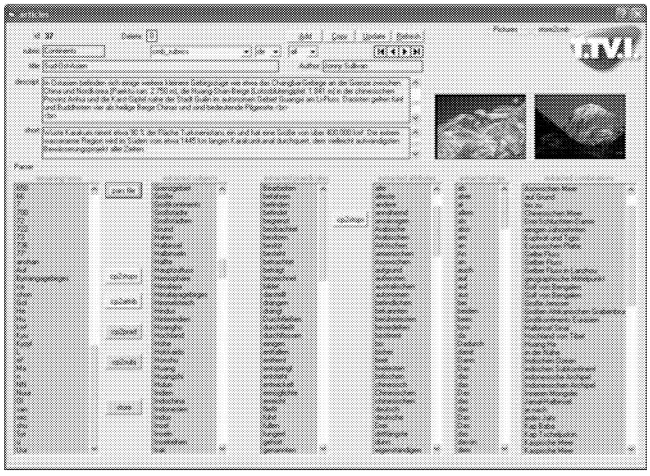


Figure 3: Comprehensive parsing and RDF oriented structuring of concepts of AV object content

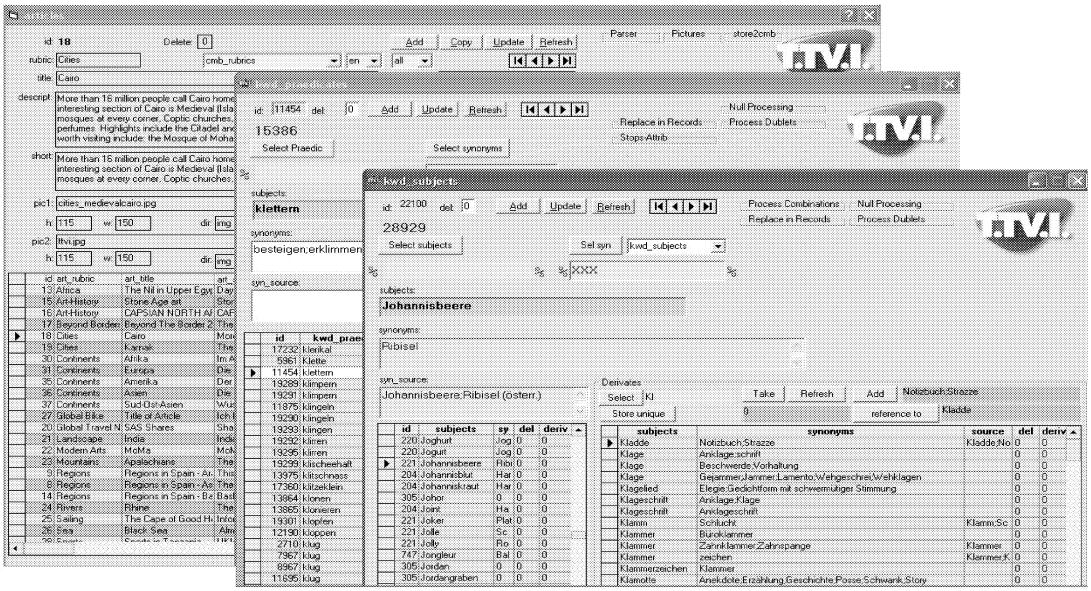


Figure 4: Tools for processing TV broadcast content and concepts

As depicted in Figure 4 several software tools are realized for processing the TV broadcast content and the semantically enrichment of the AV objects. The semantically processing is executed in two main steps:

1) the broadcast content as well as the edutainment products (given as textual descriptions) are RDF oriented processed using several flexible dictionaries; in the first step as classification and/ structure a three layered or categorization tree is used

2) after the annotation process the objects (AV-objects; edutainment products) are semantically assigned by a Fuzzy assignment procedure /Andonova-2006/

This assignment is one of the main features of the discussed media platform: as depicted in Figure 5 a semantic annotation and assignment procedure determines the relation between broadcasted content and edutainment products. In real-time – during the broadcast stream - the corresponding edutainment products are nested by links and recommended to the user.

The main effect here is an edutainment supporting one: the running broadcast is supported and enriched by recommended objects reflecting background knowledge and initiate the user and consumer for further interaction related to the given broadcast content. However this kind of conceptional assignment is also useful for arbitrary objects – beyond the edutainment section and area.

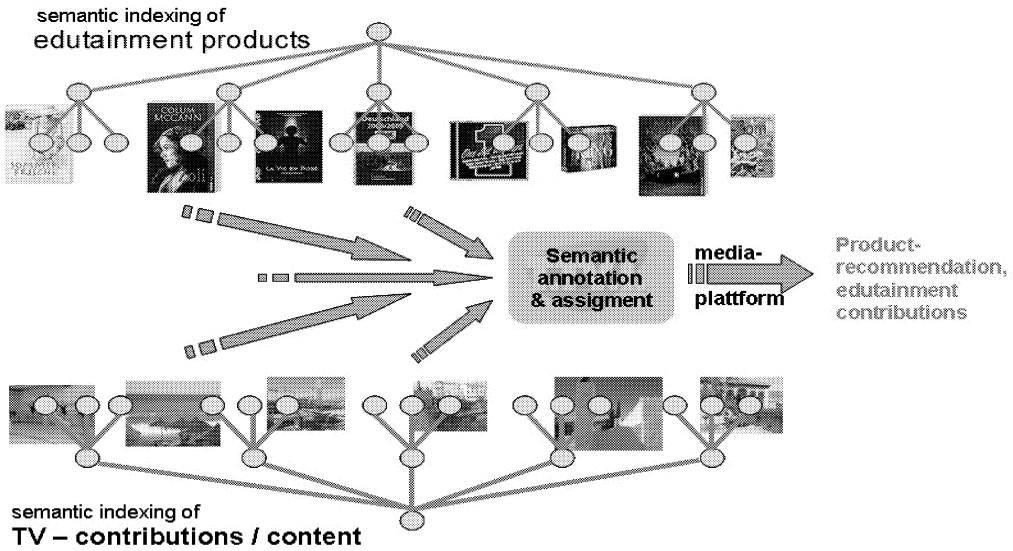


Figure 5: Semantically indexing of edutainment products and broadcast content and adequate assignment and recommendation

In this meaning the provided annotation and assignment algorithms and software tools are application independent in a double sense: the broadcast content as well as the edu-objects can be given in arbitrary manner.

EXPECTED RESULTS

The realization of the media platform has already been started. It is planned to apply the platform in the first run in the edutainment sector.

REFERENCES

Elsner-2008, Jan Elsner, “Survey about running international IPTV Projects – Standards, Specifications and Future Directions”; Euromedia Workshop DTV, 2008, Porto, Portugal

Doussier-2007; “Short Overview about Portlet Technology for Realization of Interactive Broadcast Platforms”; Euromedia Workshop DTV, 2008, Porto, Portugal

Dowman-2008; Dowman, Tablan, Popov; “Semantically Enhanced Television News through web and video integration”; Workshop Multimedia and the Semantic Web, 29th May 2005, ESWC05, 2005 www.ontotext.com/publications/semantically-enhanced-television-news.pdf, accessed March 2008

Buerger-2007; T. Buerger; “Towards a Semantic Turn in Rich-Media Analysis”; ELPUP2007, Conf. on Electronic Publishing, Vienna, Austria, June 2007

Andonova-2006 G. Andonova, G. Agre,.H.-J. Nern, A. Boyanov "Fuzzy Concept Set Based Organizational Memory as a Quasi Non-Semantic Component within the INFRAWEBS Framework." IPMU2006 IPMU 2006 proceedings. (2006): pp. 2268-2275. V., Semenova, “Adaptive dynamical polling in wireless networks”, *Cybernetics and Information Technologies*. 2008. Vol. 8, No. 1

APPENDIX (SCREENSHOTS)

SHORT OVERVIEW ABOUT PORTLET TECHNOLOGY FOR REALIZATION OF INTERACTIVE BROADCAST PLATFORMS

Axel Doussier

IT- Departrment
LTU Lufttransport-Unternehmen GmbH,
Duesseldorf, Germany
Email: axel.doussier@arcor.de

KEYWORDS: Open Source, Liferay, Hibernate, Model View Controller, Apache Struts

ABSTRACT

In this paper a rough overview about the main realization aspects of an interactive WEB 2.0 media platform is given. The main issues related to the technical objectives, the user demands as well as the chosen open source environment are shortly overviewed.

INTRODUCTION

Establishing new media platforms needs the requirements definition to finally design the targeted platform. In the first step the services to be applied and their functionality have to be defined on a conceptional level.

The most effective way to create media platforms is to set up on standard WEB 2.0 products /O'Reillynet-2008/. In this case the software product needs to cover most of the technical and functional requirements.

Media Broadcast platforms are not just streaming platforms. Effective media broadcast platforms need to integrate personal functionality improving the customer loyalty as well. Accordingly to this the customer binding has to be realized by offering comprehensive services and support functionalities. To fulfill this demand the broadcast platform is splitted into different areas with different user rights:

- **Global functionality**:
  any user is able to use services like viewing online TV, gathering general Information , wikis, blogs or buying products advertised on the pages

- **Community functionality**:
  registered users will be able to use some social services like dashboards, polls, personalized image galleries and/or public video pools

- **Personal functionality**:
  Special features (fee required) like access to personal video pools, video on demand or special ratings in case of purchasing acts

Most of these Portal services will be offered as global functionality.

PLATFORM INFRASTRUCTURE

Based on the conceptional definition of the requirements (e.g. given in /O'Reillynet-2008/) the decision is made about the infrastructure used. The technical requirements for realizing a media broadcast characterized by the objectives given above should focus the view to the following main topics:

- Data sources
- Services (SOA)
- Platform scalability
- Numerousness services ready to use

In fact the Liferay portal /Liferay-2008/ gives inherently a comprehensive fulfillment of these topics. In the Liferay Portal several services are implemented, like:

- Blogs
- Blogs Aggregator
- Chat
- Message Boards
- Polls
- Wiki
- Page Comments
- Page Ratings
- Bookmarks

For realization purposes the Liferay portal software package has been chosen due to the following reasons:

Liferay Portal /Liferay-2008/ supports all major application servers, databases and operating systems.

- „Liferay Portal complies with key industry standards
- „A highly granular permissioning system allows the user customization considering the user experience at the organizational and personal level.

Several aspects concerning the Enterprise Architecture are fulfilled /Liferay-2008-2/:

- Service Oriented Architecture (SOA) - Liferay uses SOA design principles throughout and provides the tools and framework to extend SOA to other enterprise applications
- The ServiceMix enterprise service bus (ESB) allows applications and 3services to be added quickly to an enterprise's infrastructure.

- Support for Web Services makes it easy for different applications to communicate with each other. Java, .NET, and proprietary applications can work together easily because Web Services use XML standards2 .
- Support for REST-style JSON Web Services for lightweight, maintainable code and to support AJAX-based user interfaces
- Single Sign On – Liferay offers customizable single sign-on with Yale CAS, JAAS, LDAP, Netegrity, Microsoft Exchange, and more. Yale CAS integration is offered out of the box2 .
- High Availability – Maintain zero down time for business critical applications with Hardware/Software Load Balancing, HTTP Failover, Session Replication, and Distributed Cache (using Lightweight Multicast Protocol)2 .
- Dynamic Virtual Hosting – Granting individual community members their own page with a user-defined friendly URL2 .

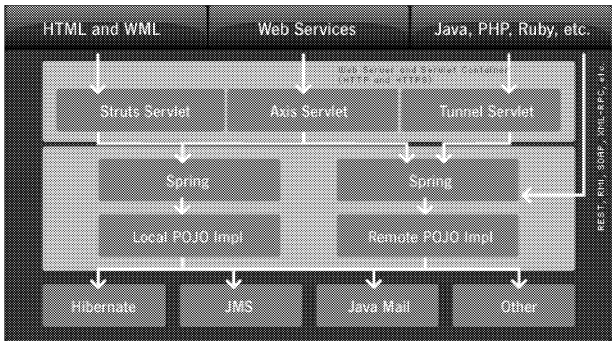


Figure 1: Liferay architecture

The Liferay Portal also includes a lot of technologies like „Hibernate“, „Struts“. In fact this an enormous advantage, because no proprietary technology is used.

Hibernate is an object-relational mapping (ORM) library for the Java language, providing a framework for mapping an object-oriented domain model to a traditional relational database. Hibernate solves Object-Relational impedance mismatch problems by replacing direct persistence-related database accesses with high-level object handling functions.

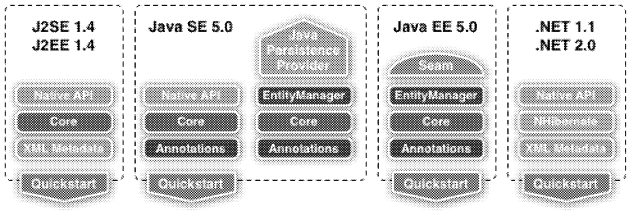


Figure 2: Hibernate modules

Apache Struts is an open-source web application framework for developing Java EE web applications. It uses and extends the Java Servlet API to encourage developers to adopt a model-view-controller (MVC) architecture. It was originally created by Craig McClanahan and donated to the Apache Foundation in May, 2000. Formerly located under the Apache Jakarta Project and known as Jakarta Struts, it became a top level Apache project in 2005.

It is common to split an application into separate layers: presentation (UI), domain logic, and data access. In MVC the presentation layer is further separated into view and controller. MVC encompasses more of the architecture of an application than is typical for a design pattern.

Within the ongoing developments the paradigma of a Model-view-controller is used, devided into a Model and a View:

Model: The domain-specific representation of the information on which the application operates. Domain logic adds meaning to raw data (e.g., shipping charges for shopping cart items).Many applications use a persistent storage mechanism (such as a database) to store data. MVC does not specifically mention the data access layer because it is understood to be underneath or encapsulated by the Model.

View: Renders the model into a form suitable for interaction, typically a user interface element. Multiple views can exist for a single model for different purposes.

Controller: Processes and responds to events, typically user actions, and may invoke changes on the model.

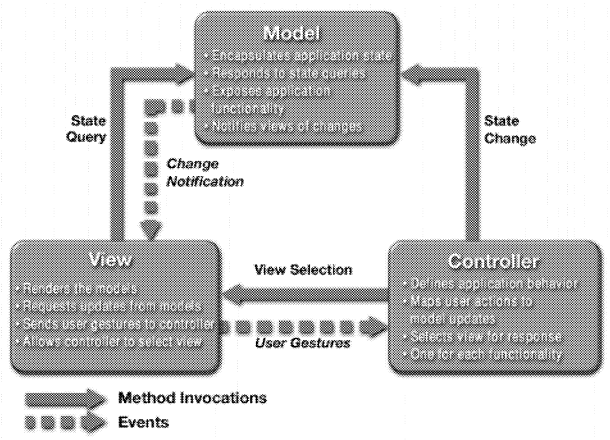


Figure 3: The model view controller paradigm used within this development /Rebag-2008/

REFERENCES

/O'Reillynet-2008/ http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html, accessed March 2008

/Liferay-2008-1/ http://www.liferay.com, accessed March 2008,

/Liferay-2008-2/ http://www.liferay.com/web/guest/community/tech_specs , accessed March 2008

/Rebag-2008/ http://www.rebag.it/rebag/index.php/Rebag_Ware#J2EE accessed March 2008

DIGITAL BROADCAST ENVIRONMENT USING WEB SERVICE TECHNOLOGY – NEW APPROACHES FOR FUZZY CONTENT DETECTION AND SERVICE DISTRIBUTION

H. Joachim Nern

Global IT&TV GmbH
200710, Duesseldorf
Germany
Email: nern@global-ittv.com

V.M. Vishnevsky

Institute for Information Transmission
Bolshoy Karetny per. 19, Moscow, 127994
Russia
Email: vishn@iitp.ru

Tatiana Atanasova

Institute of Information Technologies –BAS
Acad. G. Bonchev 2, Sofia
Bulgaria
Email: atanasova@iinf.bas.bg

KEYWORDS: *Digital Broadcast, DVB-C, DVB-S, DVB-H, IPTV, WEBTV, fuzzy classification, service net, convergent media platform*

ABSTRACT

The main objective of this paper is to introduce the software environmental approach for semantic television. The main idea is to combine WEB 2.0 characteristics with Semantic Web as well as AI features to achieve a flexible application independent media platform. In this paper the main realization aspects concerning the used web service technology, the AI methods as well as the environmental structure is described.

INTRODUCTION

Digital TV (IPTV, DVB, DVB-C, DVB-S) is one of the biggest challenges of the broadcasting market. IPTV denotes delivering of TV over the IP protocol. Internet as a platform for distributing TV services implies the possibility of customized transmission and facilitates new forms of interactivity and personalization of services. The concept WEB TV is used both when transmitting TV over the WEB and WEB services over TV networks. Because of interactivity on the Internet, it is possible to add other values to these services. The most successful Internet TV business models are likely to involve syndication to or from other media. On the user side (Video on demand, for example), all services are provided through the network.

Most of the features and functionalities of digital Television are depending on external market conditions. In spite of that several governments including Europe have determined an analog terrestrial broadcast cutoff date. Accordingly in many geographical regions, analog reception, or free TV, is the most prevalent television access technology. The analog broadcast cut-off date will yield in several activities on the provider side as well as the industrial vendors. The end of the analog television area implies also the development of new broadcast structures as well as the creation of new tools for digital object (especially moving pictures) handling.

DIGITAL BROADCAST REALIST AS CONVERGENT MEDIA-PLATFORM

Within the last few years the area of television is in a phase of change – several concepts related to Digital Television are discussed and meanwhile brought to realization. One of the main approaches in this context is the concept of "**convergent media-platforms**". Although a clear and distinguished definition is missing – and the term convergent media-platform is used within different contexts, the authors try to give a definition related to the introduced media platform approach. As illustrated in Figure 1 the term convergent is used in the sense that several digital media streams are converged to a pool of digital objects. Furthermore different distribution streams are bundled within one platform – also somekind of converging – bundling = converging. Although the distribution process itself is not focused on one distribution channel but several: within convergent media platforms the digital content (AV-objects, "moving pictures") is distributed and broadcasted via several streams (channels) like IPTV, DVB-S, DVB-C, DVB-H etc). It depends on the structure of the media platform itself, which broadcast and distribution channels are used – possible are a lot due to the digital feature of the objects.

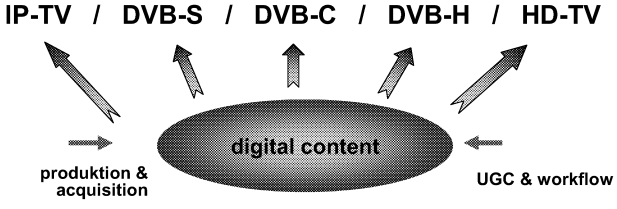


Figure 1: Convergent Media-Platform with different distribution channels

CLASSICAL VIEW: BACKEND – MIDDLEWARE - FRONTEND

The proposed type of media-platform is designed as a classical three-fold structure (illustrated in Figure 2): Backend, Middleware, and Frontend. The input for the backend is delivered by

1) edutainment partners, who post their edutainment objects and products and
2) by the TV editing team, processing the AV-objects (mainly moving pictures) and
3) further partner and clients

The backend module itself is related to several processes, like

1) semantic indexing of the AV-objects and edutainment products and
2) the adequate assignment of annotated AV- as well as edutainment objects based on the given annotation
3) a CMS system for organizing the objects
4) as well as special tools for handling user generated content (UGC)

The middleware is mainly devoted to streaming, encoding and web server applications. In a fully expanded media platform the server applications also include the satellite cable TV uplink resp. provision of the broadcast signals to the uplinking provider.

The user gets access to the platform via the frontend module and is provided with streaming products (WebTV / IPTV streams), Video on Demand (VoD) and further communication services as well as - in the expanded version - with DVB-C and/or DVB-S program broadcast.

As described in /Nern-2008/ the described media platform is realized in the first step in the edutainment sector.

The threefold architecture structure is already mentioned in /Rothe-2008/.

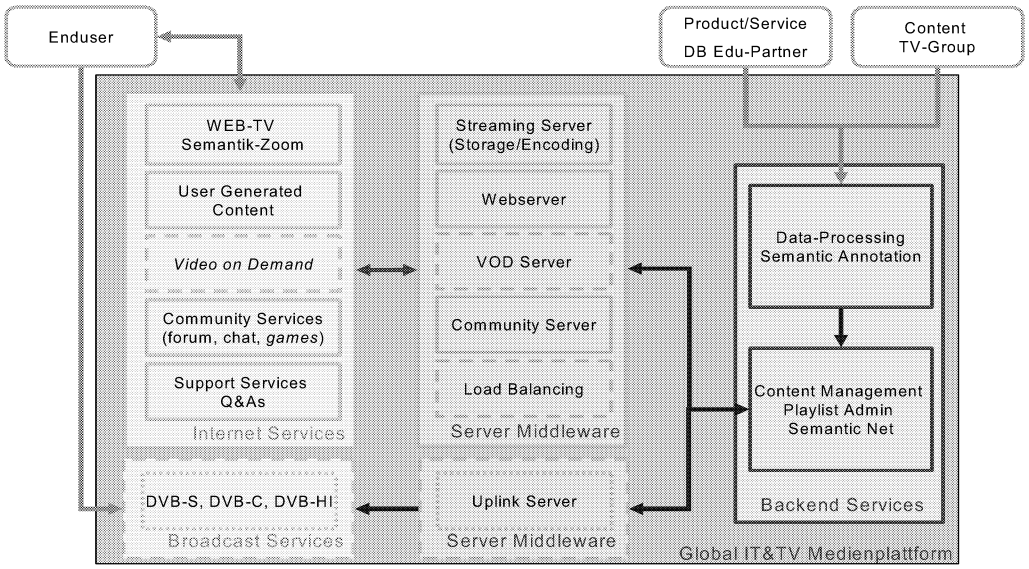


Figure 2: Detailed view of the three fold realization structure

FRONTEND – FUZZY CONTENT DETECTION AND PROCESSING

Data Processing & Semantic Annotation

The data handling procedure benefits from the fact that the AV-objects are given in known digital formats (mpg2, mpg4, avi, swf, rm,etc). This allows a (quasi-) automated preprocessing and detection resp. classification of known as well as unknown objects. The main idea here is to treat the object – especially the UGS objects - as services /Atanasova-2007/ and to build up net structures consisting of coupled objects and object fragments (illustrated in Figure 3). For enabling an optimized creation process the AV content objects are described as semantic web services resulting in a net of coupled objects and subnets of coupled objects. As illustrated in Figure 3 the advantage of the net structure is the feasibility to extract sub-nets and to create out of the nested objects and fragments new merged objects /Nern-2007/.

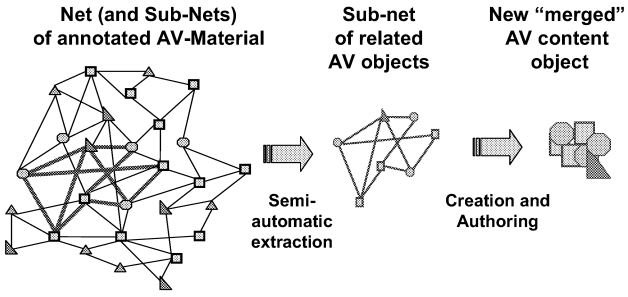


Figure 3: Net and sub-nets of objects and object fragments

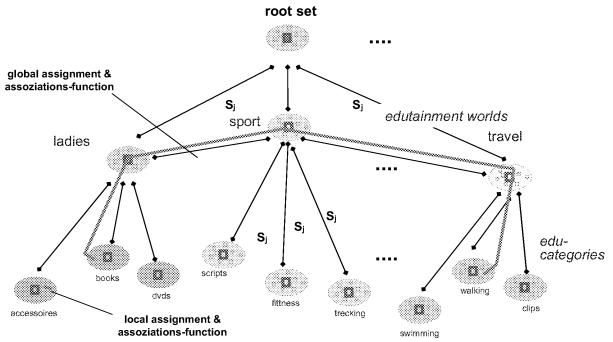


Figure 4: Flexible three layered classification structure

The algorithms for the fuzzy set based classification of the AV-objects is given in /Andonova 2006/. The result of the extraction and classification process is illustrated in Figure 4: a flexible three layered classification tree. The assignment functions for a special layer are illustrated in Figure 6.

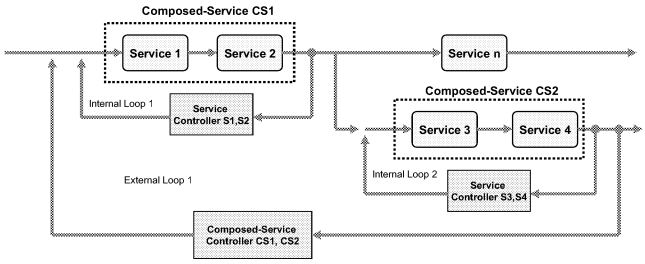


Figure 5: The service (AV-object) compsing process realized as a closed loop cycle

Application of the SWS paradigm to AV-objects

For representing the AV-objects as SWS the WSMO /WSMO-2008/ specification is applied, especially the choreography and orchestration feature. To optimize the orchestration procedure of AV-objects to be merged a controller structure in case of serialization and composition of services (representing the AV-objects) is used. As depicted in Figure 5 the composing process is designed as a closed loop process ensuring quite stable self regulating features. The details about the composing process is given in /Atanasova-2005/.

REFERENCES

Nern-2008 V., H.J.Nern, T.Atanasova, G.Jesdinsky; "Semantic Television – a New Vision or a New Business Case Approach ? - Interactive Media Based Edutainment", Workshop DTV at the Euromedia 2008, Porto, Portugal

Rothe-2008., W. Rothe; "CRM 2.0 - Service Delivery and Value Chain of an Interactive Media Platform". Workshop DTV at the Euromedia 2008, Porto, Portugal

Atanasova-2007, T.Atanasova, H.-J. Nern et al; "Modules for an Integrated System Approach for Advanced Processing of AV-objects in Digital TV Workflow"; Euromedia 2007, 25-27, April 2007, Delft University, The Netherlands, EUROSIS, pp 150-154

Nern-2007 H.J. Nern, T. Atanasova, M. Sgouros; "Framework Approach for Search and Meta-Data Handling of AV Objects in Digital TV Cycles"; Euromedia 2007, 25-27, April 2007, Delft University, The Netherlands, EUROSIS, pp 145-147

Andonova-2006 G. Andonova, G. Agre,.H.-J. Nern, A. Boyanov "Fuzzy Concept Set Based Organizational Memory as a Quasi Non-Semantic Component within the INFRAWEBS Framework." IPMU2006 IPMU 2006 proceedings. (2006): pp. 2268-2275. V., Semenova, "Adaptive dynamical polling in wireless networks", Cybernetics and Information Technologies. 2008. Vol. 8, No. 1

WSMO-2008 http://www.wsmo.org; accessed March 2008

Atanasova-2005 T. Atanasova, G. Agre, H Joachim Nern,, INFRAWEBS Semantic Web Unit for Design and Composition of Semantic Web Services"; EUROMEDIA 2005, Workshop for "Semantic Web Applications", April 11-13, 2005, IRIT, Université Paul Sabatier, Toulouse, France

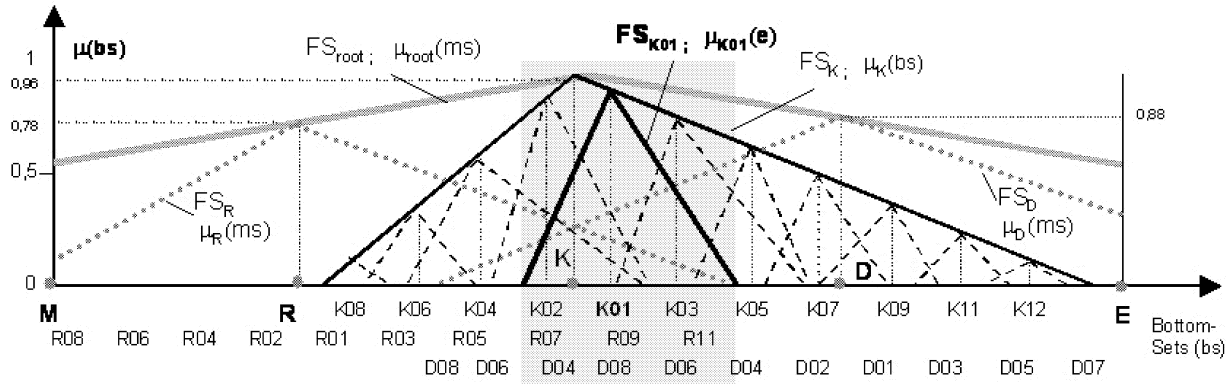


Figure 6: Example of assignment functions after a fuzzy set classification for a pseudo three layered classification tree

AUTHOR LISTING

AUTHOR LISTING