MMEDIA 2012

The Fourth International Conferences on Advances in Multimedia

ISBN: 978-1-61208-195-3

April 29 - May 4, 2012

Chamonix / Mont Blanc, France

MMEDIA 2012 Editors

Philip Davies, Bournemouth and Poole College, UK

David Newell, Bournemouth University, UK

MMEDIA 2012

Foreword

The Fourth International Conferences on Advances in Multimedia [MMEDIA 2012], held between April 29th and May 4th, 2012 in Chamonix / Mont Blanc, France, was an international forum for researchers, students, and professionals where to present recent research results on advances in multimedia, and mobile and ubiquitous multimedia. MMEDIA 2012 brought together experts from both academia and industry for the exchange of ideas and discussion on future challenges in multimedia fundamentals, mobile and ubiquitous multimedia, multimedia ontology, multimedia user-centered perception, multimedia services and applications, and mobile multimedia.

The rapid growth of information on the Web, its ubiquity and pervasiveness, makes the www the biggest repository. While the volume of information may be useful, it creates new challenges for information retrieval, identification, understanding, selection, etc. Investigating new forms of platforms, tools, principles offered by Semantic Web opens another door to enable human programs, or agents, to understand what records are about, and allows integration between domain-dependent and media-dependent knowledge. Multimedia information has always been part of the Semantic Web paradigm, but it requires substantial effort to integrate both.

The new technological achievements in terms of speed and the quality expanded and created a variety of multimedia services such as voice, email, short messages, Internet access, m-commerce, mobile video conferencing, streaming video and audio.

Large and specialized databases together with these technological achievements have brought true mobile multimedia experiences to mobile customers. Multimedia implies adoption of new technologies and challenges to operators and infrastructure builders in terms of ensuring fast and reliable services for improving the quality of web information retrieval.

Huge amounts of multimedia data are increasingly available. The knowledge of spatial and/or temporal phenomena becomes critical for many applications, which requires techniques for the processing, analysis, search, mining, and management of multimedia data.

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

Also, this event could not have been a reality without the support of many individuals, organizations, and sponsors. We are grateful to the members of the MMEDIA 2012 organizing committee for their help in handling the logistics and for their work to make this professional meeting a success.

We hope that MMEDIA 2012 was a successful international forum for the exchange of ideas and results between academia and industry and for the promotion of progress in the field of multimedia.

We are convinced that the participants found the event useful and communications very open. We also hope the attendees enjoyed their stay in the French Alps.

MMEDIA Advisory Committee:

Dumitru Dan Burdescu, University of Craiova, Romania
Philip Davies, Bournemouth and Poole College, UK
Jean-Claude Moissinac, TELECOM ParisTech, France

MMEDIA 2012 PROGRAM COMMITTEE

MMEDIA Advisory Committee

Dumitru Dan Burdescu, University of Craiova, Romania
Philip Davies, Bournemouth and Poole College, UK
Jean-Claude Moissinac, TELECOM ParisTech, France
David Newell, Bournemouth University, UK
Francisco J. Garcia, Agilent Technologies - Edinburgh, UK
Noël Crespi, Institut Telecom, France
Jonathan Loo, Middlesex University - Hendon, UK
Patrice Rondao Alface, Alcatel-Lucent Bell Labs - Antwerp, Belgium
Trista Chen, Fotolugu Inc, USA
Alexander C. Loui, Kodak Research Labs / Eastman Kodak Company-Rochester, USA

MMEDIA 2012 Technical Program Committee

Max Agueh, LACSC - ECE Paris, France
Hakiri Akram, Université Paul Sabatier - Toulouse, France
Musab Al-Hadrusi, Wayne State University, USA
Nancy Alonistioti, N.K. University of Athens, Greece
Giuseppe Amato ISTI-CNR, Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo" - Pisa, Italy
Maria Teresa Andrade, University of Porto / INESC Porto, Portugal
Marios C. Angelides, Brunel University - Uxbridge, UK
Stylianos Asteriadis, National Technical University of Athens, Greece
Ramazan S. Aygun, University of Alabama in Huntsville, USA
Andrew D. Bagdanov, Universita Autonoma de Barcelona, Spain
Yannick Benezeth, Université de Bourgogne - Dijon, France
Jenny Benois-Pineau, LaBRI/University of Bordeaux 1, France
Sid-Ahmed Berrani, Orange Labs - France Telecom, France
Steven Boker, University of Virginia - Charlottesville, USA
Laszlo Böszörmenyi, University Klagenfurt, Austria
Marius Brezovan, University of Craiova, Romania
Dumitru Burdescu, University of Craiova, Romania
Helmar Burkhart, Universität Basel, Switzerland
Eduardo Cerqueira, Federal University of Para, Brazil
Damon Chandler, Oklahoma State University, USA
Vincent Charvillat, ENSEEIHT/IRIT - Toulouse, France
Bruno Checcucci, Perugia University, Italy
Shu-Ching Chen, Florida International University - Miami, USA
Trista Chen, Fotolugu Inc., USA
Wei-Ta Chu, National Chung Cheng University, Taiwan
Antonio d'Acierno, Italian National Council of Research - Avellino, Italy
Philip Davies, Bournemouth and Poole College, UK
Vincenzo De Florio, University of Antwerp & IBBT, Belgium
Manfred del Fabro, Institute for Information Technology, Klagenfurt University, Austria
Vlastislav Dohnal, Masaryk University, Brno, Czech Republic
Jean-Pierre Evain, EBU Technical - Grand Saconnex, Switzerland

Nick Evans, EURECOM - Sophia Antipolis, France
Fabrizio Falchi, ISTI-CNR, Pisa, Italy
Kaori Fujinami, Tokyo University of Agriculture and Technology, Japan
Eugen Ganea, University of Craiova, Romania
Francisco J. Garcia, Agilent Technologies - Edinburgh, UK
Valerie Gouet-Brunet, Conservatoire National des Arts et Métiers - Paris, France
William I. Grosky, University of Michigan-Dearborn, USA
Christos Grecos, University of the West of Scotland, UK
Stefanos Gritzalis, University of the Aegean - Karlovassi, Greece
Angela Guercio, Kent State University, USA
Victor M. Gulias, University of Corunna, Spain
Hermann Hellwagner, Klagenfurt University, Austria
Luigi Iannone, Deutsche Telekom Laboratories, Germany
Razib Iqbal, University of Ottawa, Canada
Dimitris Kanellopoulos, University of Patras, Greece
Eleni Kaplani, TEI of Patra, Greece
Yiannis Kompatsiaris, CERTH-ITI, Greece
Markus Koskela, Aalto University, Finland
Panos Kudumakis, Queen Mary University of London, UK
Mikołaj Leszczuk, AGH University of Science and Technology - Krakow, Poland
Hongyu Li, Tongji University - Shanghai, China
Anthony Y. H. Liao, Asia University, Taiwan
Antonio Liotta, Eindhoven University of Technology, The Netherlands
Suzanne Little, The Open University/Knowledge Media Institute, UK
Alexander C. Loui, Kodak Research Labs, USA
Eng Keong Lua, Carnegie Mellon University, USA / Cylab, Japan
Erik Mannens, Ghent University, Belgium
Sathiamoorthy Manoharan, University of Auckland, New Zealand
Michael Massoth, University of Applied Sciences - Darmstadt, Germany
Annett Mitschick, Technical University - Dresden, Germany
Ayman Moghnieh, Universitat Pompeu Fabra - Barcelona, Spain
Jean-Claude Moissinac, TELECOM ParisTech, France
Mireia Montañola, Université catholique de Louvain, Belgium
David Newell, Bournemouth University, UK
Petros Nicopolitidis, Aristotle University of Thessaloniki, Greece
Vincent Oria, New Jersey Institute of Technology, USA
Jordi Ortiz Murillo, University of Murcia, Spain
Marco Paleari, Italian Institute of Technology / Center for Space Human Robotics - Torino, Italy
Eleni Patouni, University of Athens, Greece
Tom Pfeifer, Waterford Institute of Technology, Ireland
Wei Qu, Graduate University of Chinese Academy of Sciences, China
Piotr Romaniak, AGH University of Science and Technology - Krakow, Poland
Patrice Rondao Alface, Alcatel-Lucent Bell Labs - Antwerp, Belgium
Reza Sahandi, Bournemouth University - Dorset, UK
Susana Sargento, University of Aveiro/Institute of Telecommunications, Portugal
Klaus Schöffmann, Klagenfurt University, Austria
Yuqing Song, Chinese Academy of Sciences, China
Alexei Sourin, NTU, Singapore

Peter L. Stanchev, Kettering University - Flint, USA
Liana Stanescu, University of Craiova, Romania
Cosmin Stoica, University of Craiova, Romania
Yu Sun, University of Central Arkansas, USA
Anel Tanovic, BH Telecom d.d. Sarajevo, Bosnia and Herzegovina
Georg Thallinger, Joanneum Research - Graz, Austria
Daniel Thalmann, EPFL, Switzerland
Christian Timmerer, Alpen-Adria-Universität Klagenfurt, Austria
Chien-Cheng Tseng, National Kaohsiung First University of Science and Technology, Taiwan
Kuniaki Uehara, Kobe University, Japan
Andreas Uhl, Salzburg University, Austria
Binod Vaidya, Instituto de Telecomunicações / University of Beira Interior, Portugal
Davy Van Deursen, Ghent University - IBBT, Belgium
Andreas Veglis, Aristotle University of Thessaloniki, Greece
Janne Vehkaperä, VTT Technical Research Centre of Finland - Oulu, Finland
Dimitrios D. Vergados, University of Piraeus, Greece
Anne Verroust-Blondet, INRIA Paris-Rocquencourt, France
Qin Xin, Université Catholique de Louvain - Louvain-la-Neuve, Belgium
Toshihiro Yamauchi, Okayama University, Japan
Lei Ye, University of Wollongong, Australia
Shigang Yue, University of Lincoln, UK
Sherali Zeadally, University of the District of Columbia, USA
Tong Zhang, Hewlett-Packard Labs, USA
Yang Zhenyu, Florida International University, USA

MMEDIA Additional Reviewers

Hamed Ketabdar, Telekom Innovation Lab, Germany
Semih Dinc, University of Alabama in Huntsville, USA

Copyright Information

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

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

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

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

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

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

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

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

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

Table of Contents

Synchronization Techniques in Distributed Multimedia Presentation 1
Shahab Ud Din and Dick Bulterman

Low Complexity Multiple Candidate Motion Estimation Based on Constrained One-bit Transforms 10
Changryoul Choi and Jechang Jeong

Popularity Based Distribution Schemes for P2P Assisted Streaming of VoD Contents 14
Sasho Gramatikov, Fernando Jaureguizar, Julian Cabrera, and Narciso Garcia

A Domain Pool Classification Method for Better Fractal Volume Compression 20
Mihai Popescu, Mihai Pancu, and Razvan Tudor Tanasie

Video Casting Application Oriented Key Exchange 24
Abdullah Rashed and Henrique Santos

Multi-Connected Ontologies 29
Philip Davies, David Newell, Abigail Davies, and Damla Karagozlu

Adaptive Virtualisation for Multi Modal Learning Objects 37
David Newell, Philip Davies, Suzy Atfield-Cutts, and Andrew Yearp

Towards Distributing Multimedia Applications on a Virtualized Cloud Infrastructure 44
Mak Sharma, David Newell, Philip Davies, and Benjamin Todd

Similarity Search by Earth Movers Distance using Nonmetric Ground Distances 50
Jakub Lokoc, Tomas Skopal, Christian Beecks, and Thomas Seidl

File Size Comparisons of Modeled and Pixel-Based Video in Five Scenarios 56
Juergen Wuenschmann, Christian Feller, and Albrecht Rothermel

A Q-Learning Approach to Decision Problems in Image Processing 60
Alexandru Gherega, Monica Radulescu, and Mihnea Udrea

A Video Semantic Annotation System Based on User Attention Analysis 67
Jin-Young Moon, Chang-Seok Bae, and Wan-Chul Yoon

Video Retrieval by Managing Uncertainty in Concept Detection using Dempster-Shafer Theory 71
Kimiaki Shirahama, Kenji Kumabuchi, and Kuniaki Uehara

A Database of Artificial Urdu Text in Video Images with Semi-Automatic Text Line Labeling Scheme 75

Imran Siddiqi and Ahsen Raza

Text Driven Recognition of Multiple Faces in Newspapers 82
Nicola Adami, Sergio Benini, and Riccardo Leonardi

Optimisation of JPEG XR Quantisation Settings in Iris Recognition Systems 88
Kurt Horvath, Herbert Stogner, and Andreas Uhl

The Design of an Adaptive Multimedia Presentation System 94
Nick Rowe

Image rotation rectification in steroscopic 3D on multi-core architectures 101
Ivan Velciov, Cormac Brick, Marius Predut, and Valentin Muresan

CrossTale: Shared Narratives as a New Interactive Medium 106
Joaquim Colas, Alan Tapscott, Ayman Moghnieh, and Josep Blat

Silent Voice Elements for Text Input 112
Peng Teng and Yunde Jia

Summarization of Real-Life Events Based on Community-Contributed Content 119
Manfred Del Fabro, Anita Sobe, and Laszlo Boszormenyi

Healthcare Multimedia Application for Multi-modal Mobile Device Interaction 127
Marek Penhaker and Jan Kijonka

Distribute the Video Frame Pixels over the Streaming Video Sequence as Sub-Frames 133
Hussein Muzahim Aziz, Markus Fiedler, Hakan Grahn, and Lars Lundberg

A Robust and Fast Gesture Recognition Method for Wearable Sensing Garments 142
Ali Boyali and Manolya Kavakli

Influence of culture in gesture behavior between Anglo-Celtic and Latin-Americans. 148
Karime Nasser Alvarez and Manolya Kavakli

Automatic Discovery and Composition of Multimedia Adaptation Services 155
Jean-Claude Moissinac

Synchronization Techniques in Distributed Multimedia Presentation

Shahab Ud Din
Department of Computer Science
Vrije University
Amsterdam, the Netherlands
s.uddin@student.vu.nl

Dick Bulterman
(SEN5) Distributed and interactive systems
Centrum Wiskunde & Informatica (CWI)
Amsterdam, the Netherlands
dick.bulterman@cwi.nl

***Abstract*— In the last two decades, the transmission of multimedia streams using best effort network has been an attractive research area in multimedia communication. Multimedia streams have well defined temporal relations within themselves, generated when captured at the sender. At receiver these temporal relations have to be reconstructed to ensure smooth and synchronized multimedia presentation. The characteristics of best effort network –delay and jitter- degrade the temporal relations present in multimedia streams. Many methods have been proposed in order to mitigate the effect of network delay and jitter on the media streams. This paper classifies the work in the field of distributed multimedia synchronization. We have illustrated the techniques used in the three different multimedia synchronization types, namely, intra-media synchronization, inter-media synchronization and inter-destination synchronization.**

Keywords-distributed; multimedia; jitter; temporal relations; synchronization.

I. INTRODUCTION

Due to the last decade's breakthrough in the communication technologies, new applications in the area of distributed multimedia communication emerged. Distributed multimedia applications like video conferencing, video on demand, distance learning and others, are made feasible due to developments in the communication network. In such applications, at sender's side, different media streams are captured and sent to the receiver via packet switching network. On the receiver side, streams are received for presentation. These media streams can be classified into continuous and non-continuous streams. The continuous media streams have well defined temporal relations between the subsequent *Media Units (MUs*), for example, audio and video streams. The non-continuous media streams like images, text and graphics do not have temporal relations among MUs.

Multimedia presentation requires the integration of multiple media streams of both continuous and non-continuous streams. These streams have different temporal dependencies among the MUs of one or multiple streams. To ensure these relationships between the MUs of single and/or multiple media streams, a coordination process is required, which is called the *multimedia synchronization*. Typical synchronization solutions can be classified in to two basic types: (1) Intra-media synchronization deals with the reconstruction of the temporal relations between the MUs of the same media stream, at the presentation time. For example, maintaining the frame sequence and frame rate of the video stream to ensure a smooth presentation. (2) During presentation, reconstruction of the temporal relations between the MUs of the different but related media streams is referred as Inter-media synchronization. A typical example of the inter-media synchronization is lip synchronization [1] between the corresponding audio and video stream.

Developments in computer and communication technology led to the popularity of distributed multimedia applications. In these applications, a geographically separated sender and receiver are linked via a communication network. The sender is capturing the media stream with temporal relations and sending to receiver(s), which have to ensure these relationships during the presentation. The unreliability and unpredictability of best effort packet switching network make it hard for receiver to keep intact relations between the one or multiple streams. An accurate and explicit process of restructuring of the MUs at the receiver is required, which is called *distributed multimedia synchronization*. In distributed multimedia environment, apart from the two basic synchronization problems described earlier, another type of synchronization is required in case of multicast communication and is called inter-destination or group synchronization. This is required when geographically scattered group of receivers have to present the same stream(s) approximately at the same. With the emergence of Interactive Distributed Multimedia Applications (IDMA) a new type of interactive synchronization emerges and examples are [2-6]. In these types of applications, users can modify the presentation state of stream and this modification has to be communicated to all receivers to maintain the synchronized view of the presentation among them.

This survey is intended to study and classify research in the three types of synchronization solutions. The main objective is not to compare their techniques, but to classify them in such a way that is easy to comprehend for the multimedia research community. Although the classifications of the techniques presented in this paper are neither exhaustive, nor orthogonal, still they can act as a very good starting point for the researchers in field of distributed multimedia. The rest of the paper is structured as follows. In Section 2, we identify the causes of delay and present related work. Sections 3, 4 and 5 are dedicated to intra-media synchronization, inter-media synchronization and inter-destination synchronization techniques, respectively. The paper concludes in Section 6.

II. BACKGROUND

The current packet switching networks do not provide any guarantee on delay bounds of packet delivery. Rather they only promise best effort to deliver the data to the intended recipient. This characteristic of packet switching

 ISBN: 978-1-61208-195-3

networks make the success of the distributed applications challenging. It causes asynchrony (de-synchronization) in Distributed Multimedia Applications. In the following section, we will briefly discuss the factor of asynchrony.

A. *Causes of Asynchrony*

MUs of the media stream suffer different type of delays from the generation at source to presentation at receiver. These delays can be different for different MUs depending upon the load at sender, network and receiver. These delay variations for different MUs cause asynchrony in the media presentation at the receiver. We can divide delays into three types: the delay caused by sender, network and by the receiver. Figure1 gives a pictorial representation of all three components of end-to-end delay.

Delay at sender: Capturing, coding, packetizing, protocol layer processing and transmission-buffering delays depend on the sender load and clock speed. At the different time instances, the sender may have different loads variations, which can cause the variation in these delays for different MUs. Moreover, if the related sub-streams are captured or/and sent by different sources, then, these delays experienced by different sub-stream can be more variable.

Network delay: Network delay is the delay experienced by the MUs in the network to reach its receiver, which varies according to network load. This delay can include the propagation delay and queuing delay at the intermediate routers. Network jitters is delay variations of inter-arrival of MUs at the receiver due to varying network load. This is due to the fact that the queues of the intermediate routers between sender and receiver may have different loads at the different time instances. This delay can cause intra-media asynchrony. Network skew is the time difference in arrival of temporally related MUs of different but related streams, i.e. differential delay among the streams, which can cause inter-media asynchrony. Clock drift is the rate of change of clock skew because of temperature differences or imperfections in crystal clocks. Clock skew is the clock time difference between the sender and the receiver. This is possible if the sender and the receiver are using local clock information instead of global clock information. The sender and receiver are considering time synchronized with respect to clock only if they are using the Network Time Protocol (NTP) or Global Positioning System devices.

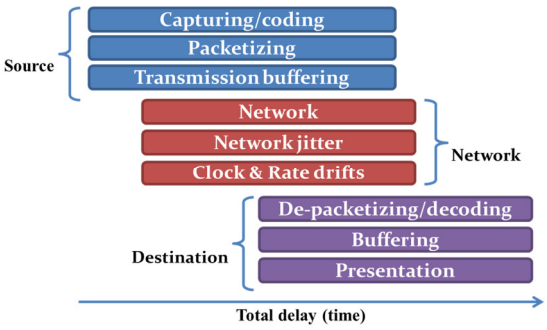


Figure 1. End-to-end causes of delay.

Delay at receiver: The presentation, decoding, de-packetizing, protocol layer processing, and buffering delay at the receivers can be different for different MUs. These delay variations are present at the receiver due to the fact that different receivers may have different processing capabilities and different loads at the different time instance.

Depending on the nature of the application some or all of these problems may be relevant to different applications. Different synchronization mechanisms are needed to cope with these problems to ensure the temporal ordering of streams and to maintain the presentation quality.

B. *Related Work*

Most synchronization mechanisms in the literature are either very abstract, independent of the application at hand or very application specific. There are some surveys of multimedia synchronization mechanisms [24, 28, 30, 31], which either are specific to type of synchronization, or partly cover synchronization mechanisms.

Perez-Luque et al. presented a survey of multimedia synchronization in term of temporal specifications [31]. They presented a theoretical reference framework to compare temporal specification schemes and their relationship with multimedia synchronizations. Ehley et al. classify synchronization schemes as distributed schemes and local schemes depending upon the location of the sender and receiver [30]. They further classify the distributed schemes as “distributed protocol based”, “distributed among nodes” and “distributed on servers or co-processors”. Similarly, they classify the local schemes in to two categories namely “local at different level at workstation” and “local on servers or co-processors”. Ishibashi et al. present very comprehensive survey of only intra-media and inter-media synchronization schemes [28]. They classify the techniques into common control, basic control, preemptive control and reactive control schemes. They also compare the different algorithms in terms of location, clock information, and type of media. They did not include inter-destination synchronization, as it was not very matured at that time. Similar to their pattern, Boronat et al. [24] present a recent survey, which includes the inter-destination synchronization, but exclude intra-media synchronization. To the best of our knowledge, there is no single survey, which covers all the three types of multimedia synchronization. Our effort is the first attempt in this regard.

III. INTRA- MEDIA SYNCHRONIZATION

The reconstruction of temporal relations between media units of the same continuous media stream is referred to as intra-media synchronization. For audio streams, the basic media unit is audio sample. The spacing between samples is determined by the sampling process. The goal of inter-media synchronization is to ensure the same spacing at the presentation time. For video streams, the basic media unit is the video frame and the temporal relation is the frequency of the video frames. The frame rate determines the spacing between the

 ISBN: 978-1-61208-195-3

frames. At presentation time, similar frame-rate has to be ensured by reconstructing the temporal relationship.

Many schemes have been proposed in literature to ensure the temporal relationship at presentation time. All the schemes use a receiver buffer for the temporary storage of incoming MUs. The audio/video samples/frames are then presented at appropriate time from buffer. The use of a MU buffer introduces delay in the application, which is directly proportional to the size of this buffer. The objective of the process is to provide a presentation that resembles as good as possible to the temporal relations that were created during the encoding process.

All Distributed Multimedia Applications (DMAs) have their own end-to-end delay tolerance requirement [33] that depends upon the nature of the application. Interactive bidirectional applications such as online quizzes have very strict end-to-end delay requirements and the applications like video conferencing have slightly less strict latency requirements. On the other hand applications like video on demand (VOD) can allow larger latency. All the proposed schemes provide for a compromise between the intra-media synchronization quality and the increase of end-to-end delay due to the buffering of MUs. On one extreme, there can be a buffer less scheme with minimum delay by presenting the frame as soon as they arrive and other can be assured synchronization that completely eliminate the effect of jitter on the cost of high end-to-end delay.

The perfect intra-media synchronization quality can be achieved by completely eliminating any kind of distortion in the temporal relationships of MUs and to completely restore the stream to its initial form. If the delay variability is unbounded, meaning that an infinitely long inter arrival period may appear, then no technique with a finite buffer can eliminate the distortion from the MUs. But, some assured services (QoS) guarantee the bounded network delay. In this case, one can achieve assured/perfect synchronization.

We divided the intra-media synchronization in to two basic categories: Time-oriented techniques and buffer-oriented techniques. In time-oriented techniques sender puts a time stamp on the MUs. The sender and the receiver use clock in order to measure the delay and jitter. Receiver on the basis of these measurements devises a technique to ensure synchronous presentation of streams. Buffer-oriented techniques do not use the clock rather they implicitly measure network delay and jitter by the occupancy of the receiver buffer. The summary is presented in Table 1.

A. *Time-oriented Techniques*

We divide time-oriented techniques into three subcategories, depending upon the timing information: techniques using global clock information, techniques using local clock information, and techniques using approximated clock information.

Techniques in which sender and receiver use some mechanism for the synchronization of their clock are said to use *global clock information*. The existence of having the globally synchronized clock allows the receiver to measure the exact network delay of MUs. Due to exact measurements of network delay, it can guarantee that MUs will be delivered and presented before or at the required time.

The techniques "using the global clock information" [7, 8] measure network delays of the first MU. They then add buffering delay in already measured network delay to compose it to total delay. They set the Maximum Delay equal to this total delay. The receiver keeps the first MU in the buffer for minimum of buffering delay time plus the extra interval before extracting from the head of the buffer for presentation. This extra buffering delay for the first MU protects the synchronization of the stream for the succeeding MUs. This way, it is guaranteed that no MU will experience a larger delay then the first MU, thus no loss of synchronization will occur. The amount of this extra buffering delay will decide the quality of synchronization. The larger extra buffering delay means assured synchronization and smaller means small end-to-end delay but no assurance of synchronization. The amount of this extra buffering delay can be adjusted according to the nature of the application. For more interactive application this amount can be set low.

TABLE 1: SUMMARY OF INTRA-MEDIA SYNCHRONIZATION

Type	*Sub type*	*Description*
Time-oriented	global clock information	Due to exact measurement of network transfer delay it can guarantee that MU will be deliver before a particular time.
	local clock information	Instead of delay duration it works on differential delay information. Due to absence of global clock it can guarantee bounded delivery.
	approximated clock information	Approximate clock information by RTT value between source and destination. Can give soft bound on MU delivery.
Buffer-oriented	Pause/drop MU	Measure the delay by buffer occupancy. Drop MU if the occupancy is high and pause when occupancy is low.
	dynamic regulation of MU duration	Instead of dropping/pausing the MU, it dynamically regulates the MU duration in accordance with buffer occupancy.

The global clock can provide the highest degree of precision in terms of clock synchronization. It is the technique which supports the strictest synchronization which requires all the MUs to be presented at a constant small delay. As a global clock is not always available, many of the techniques are based on the delay differences instead of absolute delays of MUs. In these techniques, the receivers decide presentation time for the frames using the timestamps, in varying network delay environment, in absence of global clock information on the sender and receiver end. The receiver estimates the one way network delay and its variability using local clock information. These techniques are suitable for applications that do not require a constant end-to-end delay. These techniques can be categorized as techniques *with local clock information* or without global clock information.

 ISBN: 978-1-61208-195-3

As these techniques [9-12] are based on the delay differences, the two clocks need not be synchronized because their offset will be canceled while calculating the timestamp differences. But the two clocks should not drift. In these techniques, the total delivery delay of MUs cannot be kept constant rather it will fluctuate due to changing network delay. In this way the requirement of the tradeoff between the synchronicity and delay will be relaxed. The network delay differences act as indication of the current level of the jitter between the source and destination and are used as the main parameter of these schemes.

Apart from the two above mentioned techniques there is another category of techniques using *approximated clock information*. These techniques do not require a global clock, so cannot guarantee constant end-to-end delay like the techniques based on global clock information. But, they are better than the techniques with local clock information, which only promise fluctuating end-to-end delay due to the variable network delay. In these techniques, the receiver establishes a total delivery delay by measuring round trip time (RTT) between the sender and receiver. The receiver ensures that no MU will be presented after maximum delay value calculated by some expression of the RTT between sender and receiver. As a result of this assurance, these techniques promise a soft delivery guarantee.

In [13, 14], by exchanging probe packets, a three way handshake protocol is established to measure the RTT value between the sender and the receiver. The receiver then synchronizes its clock with the sender' clock by adopting its local time as of the timestamps of the probe packets. The receiver uses RTT/2 as the estimate of the network delay and adds some delay component to achieve a fixed soft end-to-end delay. To update the RTT value according to the current network load, receiver sends the periodic probe packets. During all the communication period, the clock of the receiver is adjusted virtually with the sender's clock. Thus, the clock of the receiver is; RTT + additional delay time units behind the sender's clock. Due to this virtual clock synchronization of the sender and the receiver, these techniques are also considered as based on *virtual clocks*. Later the receiver decides about the action to take against the packet on the basis of the local clock. Packets arriving at the receiver with the timestamp larger than the local clock are buffered and the packets that arrive with timestamps smaller than the local clock are considered late. The packets are extracted from the buffer and played when the local clock is equal to their timestamps.

A part from time-oriented and buffer-oriented techniques, another classification of these techniques is possible on the basis of how the receivers deal with the late MUs. A technique is characterized as being *delay preserving,* if it does not present late MUs (MUs that have missed their scheduled time). In *none delay preserving* techniques, the receiver may accept and present a late MU, instead of discarding it, to protect the continuity of the stream from further degradation. These techniques are mostly applied with the time-oriented techniques, where the timing information is explicitly available.

B. Buffer-oriented Techniques

The class of buffer-oriented techniques deals with the fundamental synchronization/latency tradeoff but do not require timestamps of MUs or the use of any kind of clock information. Buffer-oriented techniques implicitly assess jitter by observing the occupancy of the receiver buffer instead of using timestamps. As these techniques do not rely on timing information, they cannot provide the absolute/constant end-to-end delay guarantee. The total end-to-end delay comprises of fluctuating network and buffering delay. The better stream synchronization quality can be obtained by increasing buffering delay, which will result in increased end-to-end delay. Using these techniques, delay performance can approach requirements of interactive applications but this cannot be guaranteed. Due to this lack of guarantee regarding the end-to-end delay, buffer-oriented techniques are usually employed in video applications, where the interactivity requirements are more relaxed than in audio applications. These techniques indirectly measure impact of the delay jitter on a receiver by observing the occupancy of the presentation buffer over time. The fundamental idea is to adjust the receiver's consumption rate of the frame according to the buffer occupancy. As a result of more frames in buffer, the receiver increases its consumption rate to avoid buffer overflow which will make the presentation smoother, while in case of less frames in the buffer the receiver will decrease its consumption rate to avoid buffer underflow, which will cause the increase in the presentation time of a frame and ultimately decrease the smoothness of the stream. Buffer-oriented techniques can be divided in to two broad categories: Pause/drop MUs techniques and Dynamic regulation of MUs duration.

In *Pause/drop MUs* techniques [15-18], the receiver pauses or drops the frame from the presentation buffer according to the occupancy of the buffer. If the buffer has the higher occupancy of the frame due to decrease in the network delay, it will discard the newest frames from the buffer considering them as late frames without using the timing information, which is the dropping of late MUs. Similarly, if the buffer is suffering from the underflow the receiver decrease its consumption rate by pausing the MUs in the buffer, which will increase the presentation duration of the MUs. In both cases, the objective is to bring the buffer occupancy between the underflow and overflow stage to present a continuous and synchronized presentation.

In [15, 16, 17], authors used a series of thresholds for every occupancy level and then associated these thresholds with counters for the derivation of the frame discard decisions. Biersack et al. [18] proposed a more complex technique for adopting the presentation schedule by associating it with the threshold for underflow: Low Water Mark (LWM), overflow: High Water mark (HWM) and also for Low Target Boundary (LTB) and Upper Target Boundary (UTB). For regulation of the buffering delay most buffer-oriented techniques used the

 ISBN: 978-1-61208-195-3

approach in which, they increase or decrease delay in constant amounts, that equal the duration of a MU, for example, discarding the late frame from an overpopulated buffer results in sharp delay reduction of constant duration of the video frame. Similarly, in case of under flow buffer, the presentation resumes after one or multiple MU periods. The benefit of these techniques is the simplicity of implementation. The drawback is that human visual system can detect this abrupt degradation of the perceptual quality. This will be more evident in case of low frame rate, where single video frame carries enough information. To solve this abrupt degradation problem, techniques [19, 20], which use *dynamic regulation of MUs* duration, were proposed.

These techniques demonstrated an improvement in the perceptual quality of the video, which was achieved through a fine grained regulation of presentation durations, based on the current occupancy of the presentation buffer. In [19], the receiver employs progressively reduced presentation rates, to avoid large underflow discontinuities as the buffer occupancy drops below a threshold value. The threshold value is selected prior to the stream initialization and it remains constant irrespective of the network delay jitter. It regulates the tradeoff between stream continuity and reduction of presentation rate. This work is enhanced in [20] by introducing a window based approach in which the sender optimizes the stream quality by changing network delay jitter values. This window acts as a dynamic threshold. By using a neural network approach, the sender estimates the network delay characteristics and then regulates the window accordingly. The regulation of the window will implicitly change the threshold values for the buffer occupancy. It results in dynamic selection of presentation durations for the buffered frames.

IV. INTER-MEDIA SYNCHRONIZATION

The inter-media synchronization is concerned with maintaining the temporal and/or logical dependencies among several streams in order to present the data in the same view as they were generated. At the receiver, MUs will not arrive in synchronized manner due to jitter in the network. The temporal relationship within sub-streams is destroyed and the time gaps between arriving MUs vary according to the occurred jitter. Thus, a synchronized presentation cannot be achieved at the receiver, if arriving MUs of sub-streams would be presented immediately. Hence, intra-media and inter-media synchronization is disturbed. To mitigate the effect of the jitter, MUs have to be delayed at the receiver so that, a continuous synchronized presentation can be guaranteed. Consequently, MUs have to be stored in buffer and the size of the buffer will correspond to the amount of jitter in the network.

For example, in video conferencing applications speech and video MUs must have the temporal relationships at the time the streams were captured at source. These speech and video MUs captured at the same time have to be presented together at receiver. Like any two different streams, the audio and video stream can be affected by the network delay differently. If these streams would be presented without any synchronization mechanism at the receiver, the audio and the corresponding lip movement in the video will not be synched. This temporal relation between the audio and the video stream is called inter-media synchronization or Lip synchronization. A pictorial representation of lip synchronization is presented in Figure 2.

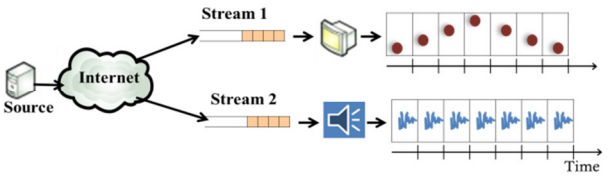


Figure 2. Inter-media synchronization.

The perfect inter-media synchronization quality is achieved by completely eliminating any kind of distortion in the temporal relationships of MUs among multiple streams and to completely restoring the stream to its initial form. This objective must be achieved on the fly as MUs arrive at the receiver, having crossed a network that alters the spacing between MUs, by imposing a variable network transfer delay.

There are many algorithms in literature that were applied in different applications to achieve the inter-media synchronization. Due to the different nature of the application, it is challenging to compare the performance of these algorithms. These algorithms used many synchronization techniques both at sender and receiver side. There is no benchmark found in literature to compare these techniques. Most of the algorithms evaluate their performance with the satisfaction of the users of the target application. So, instead of algorithm, we decided to survey these techniques that are the building blocks of algorithms. The study of inter-media synchronization technique is summarized in comprehensive manner in [24, 28, 29].

Several ways of classifying the technique are possible, we chose to categorize by location, purpose, type of content, and synchronization information. Before describing the categories of the technique, it is important to note that any algorithm can use multiple of these techniques to achieve the synchronization mechanism even from different categories. More over, these classifications are neither exhaustive, nor orthogonal, to each other, as one specific technique can be categorize according to the location, purpose, content and information used.

A. *Classification of Techniques*

Location of synchronization technique: The synchronization control can be performed either by source or receiver. The synchronization control on receiver is used more often as compare to source. If control is performed by the source, most of the time it will require some feedback information from the receiver. The receiver will tell the source about the degree of asynchrony at the current instance.

Purpose of synchronization technique: We divided the techniques into four sub categories with respect to its pur-

 ISBN: 978-1-61208-195-3

pose: The *basic Control techniques* are required in almost all the algorithms. These must be present in all algorithms to provide synchronization. Examples are adding synchronization information in MUs at source and buffering of MUs at receiver. *Preventive control techniques* are used to prevent the asynchrony in the streams. These are applied to synchronized streams to keep them in the same state. *Reactive control techniques* are used to recover from the asynchrony, once it occurred. The *common control techniques* are techniques which can be applied in both ways.

Type of media: Some of the techniques are used only for stored media and some for live media, while some can be used for both types of media. Both types of media may have different implications for a particular technique. Some techniques suit better to stored media and others to live media.

Information used for synchronization technique: The information included in the MU for the synchronization purpose can be different like timestamp, sequence number. Some techniques used either sequence number or timestamp, while the other may use both.

B. Introduction of Techniques

Here, we define the techniques shortly and then we categorize them according to the criteria said above. Most of the time these techniques are naïve and self-explanatory, so we decided to include only the short description of technique. The summary of all these techniques can be found in the previous surveys [24, 28, and 29].

Attachment of synchronization information to MU: In this technique the synchronization information is attached with MUs. Timestamps, sequence numbers are the example of the timing information.

Buffering MU: On reception, the receiver stores MUs, to compensate for network jitter. It then presents MUs according to synchronization information attached to MU.

Transmission of MUs according to synchronization information: The MUs are transmitted according to the synchronization information attached with them. This information can be a timestamp.

Decrease the number of media stream transmission: When it is difficult for receiver to achieve synchronization, the source can temporarily stop the transmission of one of the stream. It will restart the transmission of the paused stream when the receiver is synchronized.

Deadline based transmission: The source schedules the transmission of MUs to meet the associated deadline requirements. The output deadline and the delay bounds associated to each MU must be known for this technique.

Interleaving of MUs: Source interleaves the MUs of multiple streams to make a single stream. This can degrade the intra-media quality of the stream(s)

Preventive skipping/pausing: The destination skips/discards or pause/repeat the play out of MUs depending upon the state of the buffer. It can be discarding of one from every two MUs (when the buffer occupancy exceeds the threshold) or play out every MU twice (when the buffer occupancy decreases the threshold).

Change buffering time with network delay estimation: By estimating the network delay the destination alters the buffering time of the MU. If the delay is increased to avoid buffer underflow, the buffering time of the MU can be decreased and vice versa.

Adjustment of transmission time: Upon reception of MUs, the receiver sends feedback information to the source for changing the transmission timing. The source then change the transmission period.

Reactive skipping/pausing of MU: If the play out time of the current MU is late, the receiver can skip (drop) the already received succeeding MUs. Similarly receiver can pause (play out again) the play out of the previous MU until next MU is available for play out.

Shortening/extending of play out duration: To gradually recover from asynchrony, instead of abrupt change in play out, destination can shorten/extend the play out time of MU.

Virtual time contraction/expansion: If the receiver is using the virtual time for the play out of MU instead of actual time, the MU should be played out when virtual time equals the target play out time of MU. This technique of contraction/expansion of virtual time is similar to "shortening/extending of play out duration", but it gains same effect through indirect way.

Master Slave switching: The role of master and slave stream can be interchanged dynamically, when the slave stream asynchrony is increased to a certain threshold.

Source skipping/pausing: The source can skip or pause the MUs according to the received feedback information from receiver. The receiver can also insert the dummy data (or resend the previous MU) instead of pausing the MU.

Advancement of transmission timing with network delay estimation: The source advances the transmission timing of MUs according to network delay estimates. In this way the source can skip the MUs. It will dynamically schedule the transmission of MU. It is similar with the "deadline-based transmission", which also schedules the transmission time statically.

Adjustment of capturing rate: Source adjusts the clock speed of the capturing device according to synchronization quality.

Adjustment of play out rate: The receiver adjusts the presentation device frequency according to the synchronization quality.

V. INTER-DESTINATION MULTIMEDIA SYNCHRONIZATION (IDMS)

In multicast media communication, apart from intra-media and inter-media synchronization, we can find another type of synchronization called group or inter-destination media synchronization (IDMS). The objective is to present the same stream at all the receivers in a group, approximately at the same time. To add to the complexity of the problem, these different receivers may be located at different

geographical locations and may have different processing capabilities. These receivers may not only be of different type like smart phone and laptop computer but also may have the network connection of the different speeds. Network quizzes can be a good example of this scenario, where the objective will be to achieve the fairness among all the participants of the quiz. Solution will be required to display all the questions of the quiz to the entire participant at the same time.

The other example can be of the real time distance learning (tele-teaching), where the teacher multicasts a multimedia lesson to a number of students, who are located at different geographical areas. In this scenario, the teacher can also make comments about the live streaming of the lesson. Another similar example is of the interactive internet TV (Internet Social TV), where different groups of friends are watching a live online football match at different geographic locations. Consider the case, when these groups can chat (audio/video) to each other to comment on the game to experience of watching the match together from distinct location. It will be very important to synchronize the streams, so that they can watch the different events of the match at the same time to have the real experience of watching together. Figure 3 pictorially illustrates the scenario of inter-destination synchronization.

The level of required synchrony among the receivers depends on the application on hand. Considering the above three cited examples, to ensure the fairness among the participants of the online quiz, a hard synchronization will be required. In case of the other two examples, the required level of synchrony is softer as compared to the online quiz case.

The IDMS techniques cited in the literature fall under one of the three categories: *master/slave receiver scheme* (MSRC), *synchronization maestro scheme* (SMS) and *distributed control scheme* (DCS). The techniques presented in literature vary but the basic concept of the technique lies in one of the above. Here, we present the basic control scheme of each category. For better understanding of these three schemes, consider that M sources and N destinations are connected through a network. MUs of M different stream have been stored with timestamps in M source, and they are broadcasted to all receivers. The timestamp contained in an MU indicates its generation time. The streams fall into a master stream and slave streams. At each destination for inter-media synchronization, the slave streams are synchronized with the master stream by using inter-media synchronization mechanism.

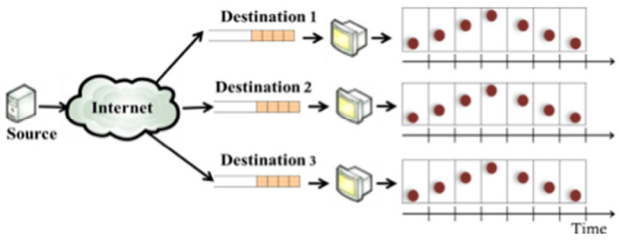


Figure 3. Inter-destination synchronization.

A. *Master/Slave Receiver Scheme (MSRS)*

In MSRS, the destinations are divided into a master destination and slave destinations. The master destination will be in control and will calculate the presentation timing of the MUs independently according to its own state of the received stream data. The slave destinations should present MUs at the same timing as the master destination. In practice multiple streams will be received at each destination and one of these streams will act as master stream for the purpose of inter-media synchronization at each destination. MSRS achieves group synchronization by adjusting the presentation time of the MUs of master stream at the slave destinations to that of the master destination.

In order to synchronize the slave destinations with the master destination, the master destination sends control packets to the slave destinations. In the beginning, the master destination multicasts a control packet including presentation time of its first MU of master stream to all slave destinations. This is called initial presentation adjustment. For the continuous synchronization among receivers the master periodically multicast control-packets whenever the target presentation time of the master destination is modified. The master notifies all the slaves about the modification by multicasting a control packet which contains the amount of time which is modified and the sequence number of the MU for which the target presentation time has been changed. Figure 4 presents the different type of message exchanges in the basic MSRS.

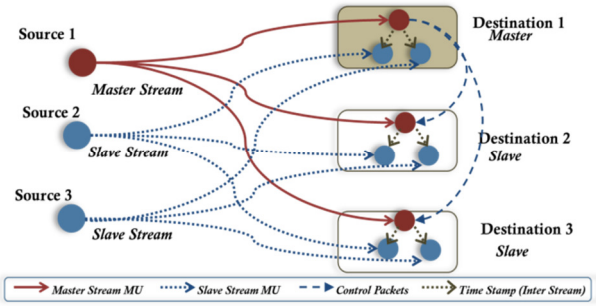


Figure 4. Master/Slave Receiver Scheme (MSRS).

This technique was initially presented in [21], and then presented in [22] by extending the RTP/RTCP messages for containing the synchronization information. The benefit of this technique is its simplicity and decreased amount of information exchange as control packet to support IDMS. Only the master destination will multicast the control packets occasionally when its target presentation time is modified or it will periodically multicast the control packets to accommodate the newly joined slave destination. Another factor which can influence the performance of the scheme is the selection of the master destination. If the slowest destination is selected as master, it can cause buffer overflow on fast slave destination, which will result as high packet drops at faster slave destination. On the other hand, if the faster destination is selected as the master destination, it can

 ISBN: 978-1-61208-195-3

cause the buffer underflow in the slower slave destination, which can result as the poor presentation quality at slow destinations. In [32], all the possible options with pros and cons are discussed for the master selection in this scheme. One issue with this technique is the associated degree of unfairness with the slave destinations. The other problem is that the master can act as bottleneck in the system.

B. *Synchronization Manager Scheme (SMS)*

SMS does not classify destinations into a master and slaves, therefore, all the destinations can be handled fairly. It involves a synchronization manager in order to synchronize the master stream among all destinations. The role of synchronization manager can be performed by one of the source or receiver. Each destination estimates the network delay and uses the estimates to determine the local presentation time of the MU. Each destination then sends this estimated presentation time of MU to the synchronization manager. The manager gathers the estimates from the destinations, and it adjusts the presentation timing among the destinations by multicasting control packets to destinations. SMS assumes that clock speed at the sources and destinations is the same, and that the current local times are also the same (i.e., globally synchronized clocks). The basic scheme is illustrated pictorially in Figure 5.

The SMS was initially presented in [23]. RTCP based schemes which follow the same basic principle were presented in [24]. The advantage of this scheme over MSRC is its fairness to the destinations, as the feedback information of all the destinations is accounted for determining the presentation time of the MU. But this fairness will cost more communication overhead among the destination and the synchronization manager. Like the MSRC, this scheme is also a centralized solution, so it can face the bottleneck problem.

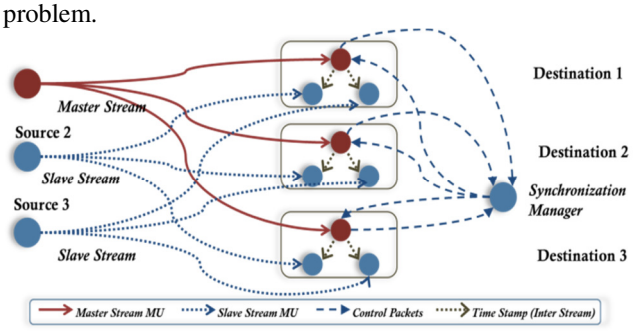


Figure 5. Synchronization Manager Scheme (SMS).

C. *Distributed Control Scheme (DCS)*

Unlike MSRC and SMS technique, DCS neither classifies the destination into master and slave, nor has a specific synchronization manager. In this technique, every destination estimates the network delay and then determines the presentation timing of the MU. It then sends (multicast) this presentation time to all destinations. Every destination will then have the entire view of the estimated time of MU. Each destination has the flexibility to decide the reference play out time among the timing of all the destinations. Figure 6 illustrates the pictorial representation of the scheme. This scheme was presented in [25-27].

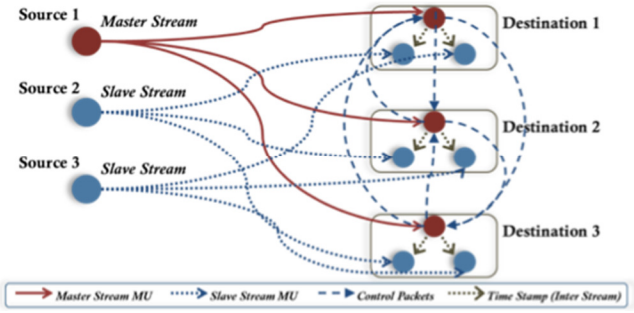


Figure 6. Distributed Control Scheme (DCS).

This scheme gives higher flexibility to each destination to decide the presentation time of MU. For example, it is possible that by selecting the presentation time of other destination, it can achieve higher IDMS quality but it may cause the inter-media or intra-media synchronization degradation. In this case, the destination has the flexibility to choose between the types of synchronization depending upon the nature of application on hand. If the application on hand demands the higher inter-media or intra-media synchronization and can sacrifice on the IDMS synchronization to certain limit, then destination can selects its own determined presentation time and vice versa. DCS is distributed scheme by nature and will not suffer from bottleneck problem. If one or more destinations leave the system, it will not disturb the overall scheme. This greater flexibility and the distributed nature of DCS make it complex in terms of processing, as before deciding presentation time of MU the destination have to do more calculations and comparisons. It has higher message complexity, as every destination will multicast the estimated presentation time.

VI. CONCLUSION AND DISCUSSION

The volume of research in distributed multimedia synchronization has increased significantly over the last decade. In this paper, we presented the three main types of synchronization, which are further categorized according to characteristics specific to each type. The issue of the intra-media synchronization is considered solved and no further research has been carried out for the last decade. The solutions of inter-media synchronization are challenging to compare qualitatively, since they are application specific and were evaluated subjectively. We included only initial research of group synchronization techniques, despite more solutions on these techniques have been developed lately. Although, there have been some research in inter-destination synchronization, more work is still needed to address its problems.

To the best of our knowledge, this survey is the first attempt that classifies the three main solutions at once. We hope that it will serve as a valuable asset for the research

 ISBN: 978-1-61208-195-3

community to comprehend the vast literature in the distributed multimedia synchronization.

VII. REFERENCES

[1] I. Kouvelas, V. Hardman, and A. Watson, Lip synchronisation for use over the Internet: analysis and implementation, Global Telecommunications Conference (GLOBECOM 96), Communications: The Key to Global Prosperity, 1996, vol. 2, pp. 893-898.

[2] E. Cronin, B. Filstrup, S. Jamin, and A.R. Kurc, An efficient synchronization mechanism for mirrored game architectures, Multimedia Tools Applications. vol. 23 (l) , 2004, pp. 7-30.

[3] C. Diot and L. Gautier, A distributed architecture for multiplayer interactive applications on the Internet, IEEE Network vol. 13 (4), 1999, pp. 6-15.

[4] C.M. Huang, C. Wang, and J.M. Hsu, Formal modeling and design of multimedia synchronization for interactive multimedia presentations in distributed environments, International Conference on Consumer Electronics, (ICCE 1998), Digest of Technical Papers, June 1998, pp. 458-459.

[5] Y. Ishibashi, S. Tasaka, and H. Miyamoto, Joint synchronization between stored media with interactive control and live media in multicast communications, IEICE Trans. On Commun. vol. E85-B (4), 2002, pp. 812–822.

[6] C.M. Huang, C. Wang, and C.H. Lin, Interactive multimedia synchronization in the distributed environment using the formal approach, IEEE Proc. Soft. 147 (4), 2000, pp. 131-146.

[7] N. Shivakumar, C.J. Sreenan, B. Narendran, and P. Agrawal, The concord algorithm for synchronization of networked multimedia streams, in international Conference on Multimedia Computing and Systems, May 1995, pp. 31-40.

[8] C.J. Sreenan, J.C. Chen, P. Agrawal, and B. Naendran, Delay reduction techniques for playout buffering, IEEE Transactions on Multimedia, vol. 2, no. 2, June 2000, pp. 88-100.

[9] R. Ramjee, J. Kurose, D. Towsley, and H. Schulzrinne, Adaptive playout mechanisms for packetized audio applications in wide-area networks, in Proceedings of the Conference on Computer Communications (IEEE Infocom), Toronto, Canada, June 1994, pp. 680–688.

[10] V. Jacobson, Congestion avoidance and control, in SIGCOMM Symposium on Communications Architectures and Protocols. ACM, Aug. 1988, pp. 314-329.

[11] J.C. Bolot, End-to-end packet delay and loss behavior in the Internet, in ACM Computer Communication Review, vol. 23 (4), 1993, pp. 289–298.

[12] S.B. Moon, J. Kurose, and D. Towsley, Packet audio playout delay adjustment: performance bounds and algorithms, ACM/ Springer Multimedia Systems, vol. 5 (1), pp. 17–28, 1998.

[13] M. Roccetti, V. Ghini, G. Pau, P. Salomoni, and M.E. Bonfigli, Design and experimental evaluation of an adaptive playout delay control mechanism for packetized audio for use over the internet, Multimedia Tools and Applications, vol. 14, (1), 2001, pp. 23-53.

[14] F.A. Cuevas, M. Bertran, F. Oller, and J.M. Selga, Voice synchronization in packet switching networks, IEEE Network, vol. 7 (5), Sept. 1993, pp. 20–25,.

[15] N. Laoutaris and I. Stavrakakis, An analytical design of optimal playout schedulers for packet video receivers, Computer Communications journal, vol. 26 (4), 2003, pp. 294-203.

[16] D.L. Stone and K. Jeffay, An empirical study of a jitter management scheme for video teleconferencing, Multimedia Systems, vol. 2 (2), 1995, pp. 267-279.

[17] K. Rothermel and T. Helbig, An adaptive stream synchronization protocol, in Proc. International Workshop on Network and Operating System Support for Digital Audio and Video (NOSSDAV), Durham, New Hampshire, Apr. 1995, Lecture Notes in Computer Science, pp. 189–202, Springer.

[18] E. Biersack, W. Geyer, and C. Bernhardt, Intra and interstream synchronisation for stored multimedia streams, in ICMCS (IEEE Multimedia Conference), Japan, 1996, pp. 372-381.

[19] M.C. Yuang, S.T. Liang, Y.G. Chen, and C.L. Shen, Dynamic video playout smoothing method for multimedia applications, in Proceedings of the IEEE International Conference on Communications (IEEE ICC), Texas, June 1996, pp. 1365-1369.

[20] M.C. Yuang, P.L. Tien, and S.T. Liang, Intelligent video smoother for multimedia communications, IEEE Journal on Selected Areas in Communications, vol. 15 (2), pp. 136–146, Feb. 1997.

[21] Y. Ishibashi, A. Tsuji, and S. Tasaka, A group synchronization mechanism for stored media in multicast communications, in: Proceed- ings of the Sixth Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM), vol. 2, Kobe, Japan, April 1997, pp. 692–700

[22] F. Boronat, J.C. Guerri, and J. Lloret, An RTP/RTCP based approach for multimedia group and inter-stream synchronization, Multimedia Tools and Applications Journal, June 2008, vol. 40 (2), 285-319.

[23] Y. Ishibashi and S. Tasaka, A group synchronization mechanism for live media in multicast communications, in: Global Telecommunications Conference, 1997 (IEEE GLOBECOM' 97), November 1997, pp. 746–752.

[24] F. Boronat, J. Lloret, and M. García, Multimedia group and interstream synchronization techniques: A comparative study. Inf. Systems, vol. 34 (1), March 2009, pp. 108-131.

[25] C. Diot and L. Gautier, A distributed architecture for multiplayer interactive applications on the Internet, IEEE Network vol. 13 (4), 1999, pp. 6-15.

[26] Y. Ishibashi and S. Tasaka, A distributed control scheme for causality and media synchronization in networked multimedia games, in: Proceedings of the 11th International Conference on Computer Communications and Networks, Miami, USA, Oct. 2002, pp. 144-149.

[27] M. Mauve, J. Vogel, V. Hilt, and W. Effelsberg, Local-lag and timewarp: providing consistency for replicated continuous app., IEEE Trans. Multimedia vol. 6 (l), 2004, pp. 47-57.

[28] Y. Ishibashi and S. Tasaka, A comparative survey of synchronization algorithms for continuous media in network environments, in: Proceedings of the 25th IEEE Conference on Local Computer Networks, Tampa, FL, USA, November 2000, pp. 337–348.

[29] H. Liu and M.E. Zarki, A synchronization control scheme for real-time streaming multimedia applications, in: Proceedings of the 13th Packet Video Workshop, Nantes, France, April 2003.

[30] L. Ehley, B. Furht, and M. Ilyas, Evaluation of multimedia synchronization techniques, in: Proceedings of the International Conference Multimedia Computing and Systems, (ICMCS 94), Boston, MA, USA, May 1994, pp. 514–519.

[31] M.J.P. Luque and T.D.C. Little, A temporal reference framework for multimedia synchronization, IEEE J. Sel. Areas Communications, vol. 14 (1), 1996, pp. 36-51.

[32] F. Boronat , M. Montagud, and V. Vidal, Master selection policies for inter-destination multimedia synchronization in distributed applications, in Modeling, Analysis & Simulation of Computer and Telecommunication Systems (MASCOTS), 2011 IEEE 19th International Symposium on. pp. 269 – 277

[33] D. Köhler and H. Müller, (Ed.) Multimedia playout synchronization using buffer level control, Multimedia: Advanced Teleservices and High-Speed Communication Architectures, Springer Berlin / Heidelberg, 1994, vol. 868, pp. 167-180.

 ISBN: 978-1-61208-195-3

Low Complexity Multiple Candidate Motion Estimation Based on Constrained One-bit Transforms

Changryoul Choi and Jechang Jeong
Dept. of Electronics and Communication Engineering Hanyang University
Seoul, Korea
e-mail : denebchoi@gmail.com & jjeong@ece.hanyang.ac.kr

***Abstract*— In this paper, we propose a low complexity multiple candidate motion estimation algorithm based on the constrained one-bit transform. We propose variations of constrained one-bit transform whose matching criteria are almost the same as the constrained one-bit transform. The motion estimation performances of the proposed variations are statistically similar to that of constrained one-bit transform in whole, but its local behaviors are very different. By adopting the multiple candidate search strategy into the typical constrained one-bit transform and its variation thereafter, we can efficiently determine two best motion vectors and enhance the overall motion estimation accuracy. Experimental results show that the proposed algorithm achieves peak-to-peak signal-to-noise ratio gains up to 0.66dB on average compared with the conventional constrained one-bit transform-based motion estimation with negligible complexity increase.**

Keywords-motion estimation; bit-wise matching; constrained one-bit transform

I. INTRODUCTION

Motion estimation (ME) is the key technique in video compression and has been widely used in many video applications such as video compression, video segmentation, and video tracking. ME is usually regarded as the computationally most intensive part, performing up to 50% computations of the encoding system [1]. The most popular technique for ME is block matching algorithm (BMA) which is deployed in many video compression standards [2][3] because of its simplicity and effectiveness. In BMA, a frame is partitioned into a number of rectangular blocks and a motion vector for that block is estimated within its search range in the reference frame by finding the closest block of pixels according to a certain matching criterion such as the sum of absolute differences (SAD) or the sum of squared differences (SSD). The full search block matching algorithm (FSBMA) can give optimal estimation of motion in terms of minimal matching error by checking all the candidates within the search range, but the prohibitively huge computational complexity makes it impractical for real-time video applications. Thus, many techniques have been proposed to reduce the high computational complexity of the FSBMA.

The techniques that exploit different matching criteria instead of the classical sum of absolute differences (SAD) such as one-bit transform (1BT), multiplication-free 1BT, two-bit transform (2BT), constrained one-bit transform (C1BT), and TGC-BPM were proposed to make the faster computation of the matching criteria using Boolean exclusive-OR (XOR) operations [5][6][7][8]. In [5], 1BT-based ME where the reference frames and the current frames are transformed into one-bit representations by comparing the original image frame against a bandpass filtered output was proposed. After this transform, the matching error criterion between two one-bit image frames, which is called the number of non-matching points of 1BT ($NNMP_{1BT}$) is given by

$$NNMP_{1BT}(m,n)=\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}\{B^{t}(i,j)\oplus B^{t-1}(i+m,j+n)\} \qquad (1)$$

where $B^{t}(i,j)$ and $B^{t-1}(i,j)$ are the 1BT representations of the current and the previous image frames, respectively, $\oplus$ denotes the Boolean XOR operation, the motion block size is $N\times N$, and $-s \le m, n \le s$ is the search range [5].

To reduce the computational complexity of calculating the 1BTs, the multiplication-free filter was also proposed in [6]. Although the 1BT-based motion estimation accomplishes a reduction in arithmetic and hardware complexity, the reconstructed image is degraded due to bad motion vectors resulting from the reduced bit-depth (particularly for small block sizes) [7]. A 2BT-based ME was proposed to enhance the ME accuracy of the 1BT-based ME algorithms [7]. In the 2BT-based ME, the values of local mean μ, variance σ^2, and the approximate standard deviation σ_a are used to convert frames into two-bit representations. The 2BT-based ME uses the number of non-matching points ($NNMP_{2BT}$) as a matching criterion given as :

$$NNMP_{2BT}(m,n)=\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}\{B_1^{t}(i,j)\oplus B_1^{t-1}(i+m,j+n)\} \,\|\, \{B_2^{t}(i,j)\oplus B_2^{t-1}(i+m,j+n)\} \qquad (2)$$

where $B_{1,2}^{t}(i,j)$ and $B_{1,2}^{t-1}(i,j)$ are the 2BT representations of the current and the previous image frames, respectively, $\|$ denotes the Boolean OR operation, the motion block size is $N\times N$, and $-s \le m, n \le s$ is the search range. The variations of

 ISBN: 978-1-61208-195-3

the 2BT-based matching criterion to increase the dynamic range of the matching criterion were proposed in [9]. These variations outperform the typical 2BT-based ME.

In [8], a constraint mask bitplane was introduced to improve the performance of 1BT, which is called the C1BT. Although C1BT-based ME uses two bitplanes in matching criterion similar to 2BT, it is very simple to create the constraint mask bitplane in C1BT. Note that for 2BT, the computational complexity of transforming frames into two-bit representation is relatively high because it involves multiplication operations. And in general, C1BT-based ME provides slightly better ME performance compared to the 2BT based ME. In C1BT, image frames are filtered using the multiplication-free 1BT filter in [6]. Then, the filtered image frames are compared to the original pixel values as in 1BT and the corresponding constraint mask is calculated as follows:

$$CM(i,j) = \begin{cases} 1, & if \ \left| I(i,j) - I_F(i,j) \right| \geq D \\ 0, & otherwise \end{cases} \quad (3)$$

where I and I_F are original and filtered image frames, respectively and D is a threshold. The corresponding matching error criterion is as follows :

$$\begin{aligned} &CNNMP_{original}(m,n) \\ &= \sum_{i=0}^{N-1}\sum_{j=0}^{N-1}\{CM^t(i,j) \,\|\, CM^{t-1}(i+m,j+n)\} \\ &\quad \bullet \{B^t(i,j) \oplus B^{t-1}(i+m,j+n)\} \end{aligned} \quad (4)$$

where $B^t(i,j)$ and $B^{t-1}(i,j)$ are the 1BT representations of the current and the previous image frames, respectively. $CM^t(i,j)$ and $CM^{t-1}(i,j)$ are the constraint mask of the current and the previous image frames, respectively. $\|$, $\oplus$ and $\bullet$ denote the Boolean OR, XOR, and AND operation, respectively. And the motion block size is $N \times N$, and $-s \leq m, n \leq s$ is the search range [8].

In this paper, we propose a low complexity multiple candidate motion estimation algorithm based on the C1BT. By exploiting the almost identical operations in two different matching error criteria, we can efficiently determine two best motion vectors according to the respective matching criteria and can enhance the overall motion estimation accuracy. The rest of this paper is organized as follows. Section 2 presents our proposed multiple candidate ME algorithm. Experimental results and analyses are provided in Section 3. Finally, Section 4 provides conclusions.

II. PROPOSED ALGORITHM

To improve the overall ME performance of the C1BT-based ME, we adopt the strategy in [10] of multiple candidate ME exploiting the similar operations between two different matching criteria. However, the matching error criterion of C1BT cannot be effectively splitted as in [10] because of Boolean AND operation. Note that because of this AND operation the C1BT matching criterion does not satisfy the metric requirements. Therefore we tested several matching error criterion as in [9] to find some substitutes whose operations are very similar to the C1BT matching criterion and whose performance is somewhat different in sequence to sequence. Among the many variations of the matching error criteria, the following two matching criteria show the similar ME performance as the C1BT matching criterion, we call it as a C1BT-extension and a C1BT-hybrid.

$$\begin{aligned} &CNNMP_{extension}(m,n) \\ &= \sum_{i=0}^{N-1}\sum_{j=0}^{N-1}\{CM^t(i,j) \oplus CM^{t-1}(i+m,j+n)\} \\ &\quad + \{B^t(i,j) \oplus B^{t-1}(i+m,j+n)\} \end{aligned} \quad (5)$$

$$\begin{aligned} &CNNMP_{hybrid}(m,n) \\ &= CNNMP_{original}(m,n) + CNNMP_{extension}(m,n) \end{aligned} \quad (6)$$

Table 1 shows the average PSNR performance of the C1BT, C1BT-hybrid and C1BT-extension when the motion block size is 16×16 and the search range is ±16. Note that for C1BT, the best performance was achieved when D = 10, however for other two variations, the best performance was achieved when D = 30. Of the variations of C1BT, the average performance of the C1BT-hybrid is slightly better than that of the C1BT and C1BT-extension. And as we can see from the Table 1, the average performance varies from sequence to sequence. For example, for sequence of "hall", C1BT outperforms C1BT-hybrid by 0.74dB on average and for the other sequences C1BT-hybrid always outperforms C1BT.

TABLE I. AVERAGE PSNR RESULTS OF C1BT, C1BT-EXTENSION AND C1BT-HYBRID

	C1BT (D = 10)	**C1BT-extension (D = 30)**	**C1BT-hybrid (D = 30)**
stefan	25.23	25.31	25.39
akiyo	42.54	42.51	42.57
mobile	23.64	23.76	23.8
hall	33.98	33.17	33.24
coastguard	29.24	29.36	29.42
container	38.25	38.26	38.28
table	28.07	28.17	28.33
flower	25.78	25.83	25.88
average	**30.84**	**30.80**	**30.86**

To exploit those uneven performance differences, we propose to use multiple candidate motion search based on

 ISBN: 978-1-61208-195-3

TABLE II. AVERAGE PSNR RESULTS OF ALGORITHMS WHEN THE MOTION BLOCK SIZE IS 16×16 (SEARCH RANGE = ±16)

	FSBMA	C1BT	AM2BT	DCMCW2BT	Proposed
stefan	25.75	25.23	25.53 (113.42)	25.50 (0.24)	25.53 (0.34)
akiyo	42.84	42.54	42.60 (1.59)	42.59 (0.02)	42.79 (0.03)
mobile	23.92	23.64	23.72 (182.15)	23.8 (0.2)	23.86 (0.16)
hall	34.34	33.98	33.56 (17.26)	33.91 (0.22)	34.21 (0.36)
coastguard	29.62	29.24	29.43 (45.8)	29.46 (0.2)	29.51 (0.64)
container	38.33	38.25	38.13 (2.78)	38.22 (0.02)	38.33 (0.07)
table	28.87	28.07	28.37 (56.56)	28.45 (0.3)	28.54 (0.64)
flower	26.03	25.78	25.91 (218.39)	25.95 (0.1)	25.94 (0.44)
average	**31.21**	**30.84**	**30.91 (37.53)**	**30.99 (0.08)**	**31.09 (0.16)**

the C1BT and C1BT-hybrid (MCC1BT). Note that the matching error criteria of these two ME algorithms share many identical operations. The proposed algorithm is as follows:

1) Calculate the matching error criteria as in (4) and (6).
2) Find two best motion vectors according to the respective matching criteria.
3) If two best motion vectors are the same, declare it as the best motion vector for the current block and go to 5).
4) Calculate SADs of the two best motion vectors and declare the motion vector with less SAD as the best motion vector for the current block.
5) Go to the next current block.

Note that the calculations of SADs are needed only when two best motion vectors are different, which is very rare as will be seen in the experimental results.

III. EXPERIMENTAL RESULTS

The performance of the proposed algorithm (D = 30) was compared with the C1BT, the AM2BT [4], DCMCW2BT [10] and FSBMA using the metric of SAD. The first 100 frames of 8 CIF (352 × 288) sequences are used as test sequences. All the searching processes were in spiral order.

Table 2 and 3 show the average PSNR comparison results when the motion block size is 16×16 and the search range is ±16 and when the motion block size is 8×8 and the search range is ±8, respectively. The average numbers of SAD calculations per motion block for AM2BT, DCMCW2BT and the proposed algorithm are also shown in the Tables. Note that the maximum number of calculations of SADs in one motion block is two for comparison.

From the Tables, we can see that the performance of the proposed algorithm outperforms the other algorithms. To be specific, the average PSNR of the proposed algorithm is better than that of the C1BT by 0.25dB, that of the AM2BT by 0.18dB, and that of the DCMCW2BT by 0.10dB when the motion block size is 16×16 and the search range is ±16. The SAD calculations of the proposed algorithm are needed about 1 out of 12 motion blocks on average. Compared with the motion block size is 16×16 and the search range is ±16. The gap between the proposed algorithm and the FSBMA is within 0.12dB. And for the computational complexity increase, we can see that the calculations of SADs are needed about 1 out of 12 (≈2/0.16) motion blocks on average which is very small. Compared the AM2BT, the ratio between the proposed algorithm and the AM2BT is about 1 over 235 in terms of the number of SAD calculations when the motion block size is 16×16 and the search range is ±16. Also when the motion block size is 8×8 and the search range is ±8, the average PSNR of the proposed algorithm is better than that of the C1BT by 0.66dB, that of the AM2BT by 0.37dB, and that of the DCMCW2BT by 0.22dB.

 ISBN: 978-1-61208-195-3

TABLE III. AVERAGE PSNR RESULTS OF ALGORITHMS WHEN THE MOTION BLOCK SIZE IS 8×8 (SEARCH RANGE = ±8)

	FSBMA	**C1BT**	**AM2BT**	**DCMCW2BT**	**Proposed**
stefan	26.74	25.58	26.30 (44.83)	26.32 (0.3)	26.33 (0.5)
akiyo	43.48	42.71	42.36 (1.67)	42.61 (0.04)	43.22 (0.1)
mobile	24.83	23.89	24.25 (63.44)	24.45 (0.28)	24.56 (0.36)
hall	35.87	34.79	34.57 (4.58)	35.02 (0.24)	35.47 (0.55)
coastguard	30.68	29.43	30.05 (16.21)	30.14 (0.36)	30.19 (0.81)
container	38.42	37.87	37.99 (2.42)	38.02 (0.04)	38.37 (0.12)
table	30.55	28.94	29.71 (16.22)	29.79 (0.28)	29.85 (0.71)
flower	27.45	26.82	27.12 (58.43)	27.19 (0.12)	27.26 (0.44)
average	**32.25**	**31.25**	**31.54** **(25.98)**	**31.69** **(0.21)**	**31.91** **(0.45)**

IV. CONCLUSION AND FUTURE WORK

A low complexity C1BT-based multiple candidate motion estimation algorithm was proposed in this paper. By exploiting almost the identical operations in two different matching error criteria, we can efficiently determine two best motion vectors according to the respective matching criteria and can enhance the overall motion estimation accuracy. Experimental results show that the proposed algorithm achieves PSNR gains about 0.25dB and 0.66dB on average when the motion block size is 16×16 and 8×8, respectively compared with the conventional C1BT-based motion estimation without noticeable complexity increase. Note that the PSNR difference between the proposed algorithm and the FSBMA using the metric of SAD is only 0.12 dB on average, which is very small when the motion block size is 16×16. For future work, we plan to find an efficient local search algorithm to enhance the overall ME accuracy.

REFERENCES

[1] Z. He, C. Tsui, K. Chan, and M. Liou, "Low-power VLSI design for motion estimation using adaptive pixel truncation," *IEEE Trans. Circuits and Systems for Video Tech.*, vol. 10, no. 5, pp. 669-678, Aug. 2000.

[2] *Information Technology - Coding of Audio Visual Objects - Part 2: Visual*, JTC1/SC29/WG11, ISO/IEC 14496-2 (MPEG-4 Visual), 2002.

[3] Advanced Video Coding for Generic Audiovisual Services, *ITU-T Recommendation H.264*, May 2005.

[4] B. Demir and S. Erturk, "Block Motion Estimation Using Adaptive Modified Two-bit Transform," *IET Image Process.*, pp. 215-222, 2007.

[5] B. Natarajan, V. Bhaskaran, and K. Konstantinides, "Low-Complexity Block-based Motion Estimation via One-Bit Transforms," *IEEE Trans. Circuits and Systems for Video Tech.*, vol. 7, no. 5, pp. 702-706, Aug. 1997.

[6] Sarp Erturk, "Multiplication-Free One-Bit Transform for Low-Complexity Block-Based Motion Estimation," *IEEE Signal Processing Letters*, vol. 14, no. 2, 109-112, Feb. 2007.

[7] A. Erturk and S. Erturk, "Two-Bit Transform for Binary Block Motion Estimation," *IEEE Trans. Circuits and Systems for Video Tech.*, vol. 15, no. 7, pp. 938-946, July 2005.

[8] O. Urhan and S. Erturk, "Constrained One-Bit Transform for Low Complexity Block Motion Estimation," *IEEE Trans. Circuits and Systems for Video Tech.*, vol. 17, no. 4, pp. 478-482, Apr. 2007.

[9] Changryoul Choi and Jechang Jeong, "Enhanced Two-bit Transform Based Motion Estimation via Extension of Matching Criterion," *IEEE Trans. Consumer Electron.,* vol. 56, no. 3, pp. 1883-1889, Aug. 2010.

[10] Changryoul Choi and Jechang Jeong, "Low Complexity Weighted Two-bit Transforms based Multiple Candidate Motion Estimation," *IEEE Trans. Consumer Electron.,* Nov. 2011.

 ISBN: 978-1-61208-195-3

Popularity Based Distribution Schemes for P2P Assisted Streaming of VoD Contents

Sasho Gramatikov, Fernando Jaureguizar, Julián Cabrera, and Narciso García
Grupo de Tratamiento de Imágenes, ETSI Telecomunicación, Universidad Politécnica de Madrid, Spain
Email: {sgr, fjn, julian.cabrera, narciso}@gti.ssr.upm.es

***Abstract*—The Video on Demand (VoD) service is becoming a dominant service in the telecommunication market due to the great convenience regarding the choice of content items and their independent viewing time. However, it comes with the downsides of high server storage and capacity demands because of the large variety of content items and the high amount of traffic generated for serving all requests. Storing part of the popular contents on the peers brings certain advantages but, it still has issues regarding the overall traffic in the core of the network and the scalability. Therefore, we propose a P2P assisted model for streaming VoD contents that takes advantage of the clients unused uplink and storage capacity to serve requests of other clients and we present popularity based schemes for distribution of both the popular and unpopular contents on the peers. The proposed model and the schemes prove to reduce the streaming traffic in the core of the network, improve the responsiveness of the system and increase its scalability.**

Keywords-P2P; VoD; streaming; popularity.

I. INTRODUCTION

The great expansion of the IPTV [1] has made a good ground for the Video on Demand (VoD) to become one of the most popular services. Although VoD is a service that is also available on Internet, it has attracted special attention in the field of the Telecom-managed networks since they are already adapted to the implementation of a variety of TV services. Despite of its numerous advantages from client's point of view, the VoD service is a serious issue for the providers since it is very bandwidth demanding. Therefore, the design of systems and algorithms that aim at optimal distribution of the content items has become a challenge for many providers. Some of the solutions include a hierarchy of cache servers which contain replicas of the content items placed according to a variety of replica placement algorithms that depend on the users behaviour [2][3][4]. No matter how good these solutions might be, they all reach a point from where no further improvements can be done due to resource limitations. One possibility to overcome this problem is the implementation of the classical P2P principles for exchange of files over Internet for delivering video contents to a large community of users. Some systems designed for streaming VoD over Internet are presented in [5][6]. Despite of its numerous advantages, the P2P streaming over Internet lacks reliability. The environment where the implementation of P2P streaming perfectly fits are the telecom-managed IPTV networks. Some of the reasons for that are the considerable storage capacity of the set-top boxes (STBs) nowadays and the higher control of the operators over the devices on the clients premisses, which avoids the reliability issue of the classical P2P systems. The use of P2P in IPTV networks for live video contents and the contributions of various architectural designs are shown in [7]. In [8], a P2P assisted streaming system is proposed, where the peers are supported by one server to provide the missing parts or make up for any failures. Another IPTV network architecture that takes advantage of the P2P is presented in [9]. A solution that implements P2P streaming to reduce the load of hierarchically organized servers in busy hours is proposed in [10]. In this approach, only the most popular content items are stored in the peers.

Assuming that the content items in the IPTV networks are distributed in a way that the most popular content items are stored in servers that are closer to the clients, the idea of storing copies of the popular contents in the STBs is quite a reasonable solution that could significantly reduce the traffic in the edge of the network, particularly in the busy hours. However, there is a large number of contents that are not in the high popularity range, but still take significant part of the overall traffic. Since they are stored in more distant servers in the core of the network, the traffic generated for their streaming is a burden for the backbone of the network. The opposite case of distributing the unpopular contents in the STBs contributes to reducing the traffic in the core of the network because it concentrates most of the traffic in the periphery of the network: the popular contents are streamed by the servers on the edge, and a great part of the unpopular contents are streamed by the STBs. This is important when one of the objectives is reducing the transport cost in the network. Although both of the distributions bring improvements by reducing the overall traffic, they do not provide improved service for the entire set of contents in the cases of busy hours. When the popular contents are stored in the STBs, the response time for service of unpopular contents is increased because the servers cannot serve all the incoming requests. The same happens when the unpopular contents are stored in the STBs with the difference that now, not all the requests for popular contents can be immediately served.

Therefore, we propose a solution for a network with popularity based distribution of contents, both on the streaming servers and STBs, that aims to reduce the traffic in the core of the network and, at the same time, tends to provide immediate

 ISBN: 978-1-61208-195-3

service in the cases of high demand scenarios. One of the objectives targeted with the reduction of the traffic in the core is offloading the backbones from video traffic so that it can be used for other type of traffic and enabling growth of the number of clients subscribed to VoD service without additional changes and costs in the core of the network. Although the schemes that we propose consider all the contents, we put an accent on the low popularity contents by reserving more storage space in the STBs than the popular contents, thus providing locally close availability of most of the videos.

In our model, we take advantage of the unused upload and storage capacity of the STBs to assist in the streaming of the VoD contents. The streaming is done by parallel streaming sessions of multiple STBs in order to compensate for their limited streaming capacity. Unlike many P2P solutions where the peers self-organize themselves, in our model, the peers have a role of passive contributors to the streaming process, having no knowledge of the existence of other peers. They are only capable of serving the videos that they have already stored. All the decisions regarding redirection of the clients are taken by the servers on the edge of the network.

The rest of the paper is organized as follows. In Section II, we describe the architecture of the proposed model for peer assisted VoD streaming, the division of the contents for better utilization of the storage of the STBs and the request process for VoD contents. In Section III, we define the sizes of the streaming and storage capacity of the servers for their optimal utilization. In Section IV, we define the popularity based distributions and in Section V, we present the simulation environment and analyze the obtained results. Eventually, we give our conclusions in Section VI.

II. Proposed model

The model that we propose for optimal distribution of VoD contents is a hybrid solution that unites the advantages of both the IPTV and P2P architectures: the high reliability and scalability of the IPTV architecture and the storage space and unused up-link bandwidth of the P2P architecture.

A. Model arhcitecture

The proposed model's architecture consists of hierarchically organized streaming servers, management servers and STBs. The management servers are responsible for monitoring the system and taking decisions about redirection of the requests and the placement of the contents. We consider a company owned network which can be managed and configured according to the intensity of the requested traffic. The main streaming functionality is provided by the streaming servers, while the peers have the role to reduce the overall traffic in the network. Unlike the classical P2P solutions, where the clients decide whether to share content or not, in an IPTV managed network, the STBs are owned by the service provider and, therefore, part of their unused storage and streaming capacity can be reserved for the needs of the peer assisted streaming.

The streaming servers are organized in a hierarchical tree structure according to the distance from the clients (Figure 1).

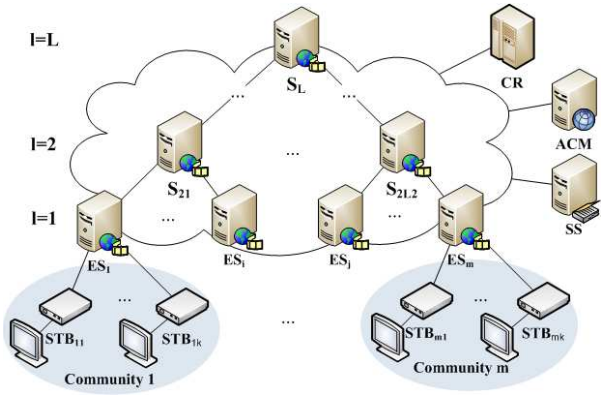


Fig. 1. Model architecture

These servers have limited storage and streaming capacity, so they can host a limited number of contents and can serve a limited number of clients. The servers that are in the edge of the network, called Edge Servers (ES), serve only one group of locally connected clients. All the clients assigned to one ES form a local community. Clients assigned to one ES cannot be served by an other ES because the tree structure of the servers architecture implies longer distances between them. Each peer can serve only clients within the same local community. The clients from a community cannot be served by other communities because that would cause additional traffic burden in the core of the network. Each ES keeps track of the popularity of the entries it currently hosts and sends it to the Automatic Content Movement server (ACM) server for redistribution purposes. The ES also maintains availability data of the portions of the content items stored in its assigned peers. It uses these data to redirect the clients whenever there is request for contents that are already stored in the peers.

Another part of the system is the Central Repository (CR) which is a server with capacity to store all the contents. It is highest in the hierarchy and it is entry point for new items. It does not directly serve the clients, but it supplies the streaming servers with the missing contents when it is necessary. The management servers are represented by the ACM and the Service Selection server (SS). The ACM server has the role to monitor the state of the network and to take decisions for a new replica distribution on the servers. When necessary, the ACM server runs a redistribution algorithm which, using popularity and state data, decides the number and the position of the replicas for each content item within the network. The objective of the redistribution algorithm is to place the contents in a way that the most popular contents will be stored in the edge servers and the less popular contents in the servers higher in the hierarchy [4]. The ACM server periodically gathers information for the current state of the streaming servers. Upon the execution of the redistribution algorithm, the ACM server issues commands, which may include insertion or deletion of contents on particular servers.

The SS server is responsible for redirection of the requests to the right servers in a way that the transport cost is minimized and the load between the servers is equally distributed. In

 ISBN: 978-1-61208-195-3

order to take the best redirection decisions, the SS server is frequently updated by the ACM server with the state of the system and the new position of the replicas.

The clients make requests to their assigned ES. If the ES is not able to serve the client, it addresses it to the SS server, which then redirects it to the most appropriate server. Clients can be served only by servers that are parents of their assigned ES. In the case when there are peers within the same community that contain parts of the requested content item, the ES takes the role of an index server. In the proposed model, we consider that this functionality does not require additional hardware upgrades and delays of service. Additionally, the server redirects the client to the SS server for completing the streaming of the rest of the content. In case of failure of any peer, the missing parts are compensated from other peers or from the streaming servers.

The contents are distributed in the STBs in off-peak hours, but we also use the volatile nature of popularity of the content items as an advantage for reduction of the distribution traffic. This property comes as result of the behaviour of the users for not repeating a request for the same content. Soon after a video is introduced in the system, it reaches high popularity, but as the time passes, the popularity decays because the clients who already saw the video are unlikely to request it again. Therefore, a content item that is already viewed and stored in the STB of many clients is very likely to be later removed from the ESs as not popular. In such a way, most of the contents with reduced popularity will be already stored in the STBs and available for peer assisted streaming. This saves a lot of additional traffic for distribution of contents from the streaming servers to the STBs. The decisions about the content placement in the peers are taken by the ES depending on the distribution determined by the ACM server.

B. Content division

The division of the contents into smaller strips is inevitable in the implementation of the P2P assisted streaming. The main reason for that is the limited up-link capacity of the STBs, which is several times smaller than the necessary playback rate. For immediate and uninterrupted playing, a content item has to be streamed in parallel by as many peers as it is necessary for reaching its playback rate. Each peer streams a portion of the content item. When all the portions reach the peer, they are assembled and the content is played. The size of the streamed portion Δ is determined as a product of the minimum STB's streaming capacity u and the maximum acceptable initial viewing delay, defined as the time necessary for the entire length of a portion to be received. Each strip consists of consequent streaming portions that are on distance $k\Delta$ between each other, where k is the ratio between the play rate r_s and the minimum up-link capacity u.

The division of the contents also contributes for increasing the storage efficiency of the peers and the contents availability. Considering that each peer is capable of streaming only a portion of the content makes it reasonable to store only those portions that it is capable to stream. Since the strips are k times smaller in size than the original content, each peer can store k times more different content items, assuming that all the contents have, on average, the same size. All the contents that are stored in the STBs are entirely stored in the servers so that they can be delivered whenever the STBs are not able to provide any of the strips.

C. Requesting process

The requesting process is initiated by the client which sends a request for a content item to its designated ES server. According to the content availability, there are the following cases: the ES already has the content; the server does not have the content, nor any of the peers; the ES does not have the content, but it knows which peers partially contain it; and the server is overloaded. In the first case, the ES sends acknowledgement to the client which is followed by a direct streaming session. In the second case, the ES redirects the client to the SS server which then chooses the best server to serve it and sends it the address of the chosen server. Once the client has the address, the process is the same as in the first case. In the case when some strips are stored in the peers, the ES looks up in its availability table and sends a list of the available strips and their location. If there is not sufficient number of strips available on the peers, the ES redirects the client to the SS server. Just like in the previous case, the SS redirects the client to the best streaming server for the delivery of the missing strips. When the client receives the availability data of all the strips, it initiates streaming sessions with each peer of the obtained list and at the same time initiates streaming session for the missing strips with the server assigned by the SS. The streaming sessions on the peers occupy the uplink capacity of the STBs and therefore, once an ES sends the availability of the strips, it marks all the peers that contain those strips as unavailable until the end of the peer streaming session. When the streaming is over, the client updates the ES, and the strips become available again. In the case when the server is overloaded, the request is rejected, and the client retries requesting the content after determined time.

III. System dimensioning

The system we are considering consists of S streaming servers which belong to one of the L levels of a tree structure. Each server s has a streaming capacity $U(s)$ and a storage capacity $S(s)$ for storing a limited number of C content items. Each content item c has a size $s(c)$ and a playback rate $r_s(c)$. There are N clients in the system which are connected to one of the m edge servers.

One of the important issues for estimating the contributions of the proposed model is planning the streaming capacity $U(s)$ and storage capacity $S(s)$ of the servers so that they can comply to the requests of the N clients. Because the storage capacity of a server is more easily upgradeable than the streaming capacity and the capacities of the links that interconnect the servers, we will consider adjusting the storage space for a fixed streaming capacity. We assume that the servers at the edge of the network serve approximately the

 ISBN: 978-1-61208-195-3

same number of clients and therefore, have the same streaming and storage capacities. We also assume that all the content items have the same streaming rate r_s and average size $\overline{s}$.

We model the system size according to the popularity distribution of the content items and according to the way the servers are organized within the hierarchy. We consider that the popularity of the content items obeys the Zipf-Mandelbrot distribution and that they are previously ranked according to the past request data and estimation of the recently inserted items. According to this distribution, the relative frequency (popularity) of the content item with *i*-th rank in is defined as:

$$f(i) = \frac{(i+q)^{-\alpha}}{\sum_{c=1}^{C}(c+q)^{-\alpha}} \quad (1)$$

where α is a real number that typically takes values between 0.8 and 1.2 and q is a shifting constant. This distribution is a generalized form of the Zipf distribution, which includes the shifting constant q in order to characterize the behaviour of the clients for not repeating requests of already seen content items [11]. We consider that the distribution algorithm always places the most popular videos in the servers that are closest to the clients. The higher level a server has, less popular contents it will contain. Having this in mind, the condition that the streaming capacity $U(s)$ has to fulfil so that all the requests directed to server s can be served is

$$n(s) \cdot \sum_{c=a(s)}^{b(s)} f(c) r_s(c) \leq U(s) \quad (2)$$

where $n(s)$ is the maximum number of simultaneously served clients by server s. For the first level of the tree, $n(s)$ is the number of active peers in the local community of server s and in the rest of the levels it is the sum of all active clients in the communities that can be served by that server. The indexes $a(s)$ and $b(s)$ note the ranks of the first and the last most popular content items stored in server s. Considering the assumption that the edge servers serve the same number of clients, $n(s)$ can be expressed as

$$n(s) = \mu \frac{N}{m} T(s) \quad (3)$$

where $T(s)$ is number of served local communities and μ is the percent of active clients. The same assumptions let us define the initial rank of the contents on server s as one value above the rank of the least popular content stored in the servers in the level below. Thus, the problem is reduced to finding the rank $b(s)$ of the contents that will be placed in server s. If we substitute (1) and (3) in (2), we get

$$\sum_{c=a(s)}^{b(s)} (c+q)^{-\alpha} \leq \frac{U(s)m}{\mu r_s T(s) N} \sum_{c=1}^{C} (c+q)^{-\alpha} \quad (4)$$

Once the indexes $a(s)$ and $b(s)$ are determined, the optimal storage capacity of the server is determined from the following condition

$$(b(s) - a(s) + 1)\overline{s} \leq S(s) \quad (5)$$

Since $b(s)$ cannot be expressed in closed form, it is determined by using numerical methods.

IV. Distribution schemes

In this paper, we propose mixed schemes for distribution of the contents on the STBs which include both the popular and unpopular content items. By combining these simple distributions, we take advantage of the contributions of each one of them: the distribution of popular contents makes the network more responsive in highly congested conditions, and the distribution of the unpopular contents makes the streaming process locally closer to the clients for all the available contents and thus reduces the traffic in the core of the network. One of the key factors in the definition of the distributions is the percentage h of dedicated storage space for popular and unpopular contents. We should keep in mind that the STBs store only strips of the contents and, therefore, increasing the storage reserved for popular contents would keep more of the STBs busy and the strips of the unpopular contents could be rarely used.

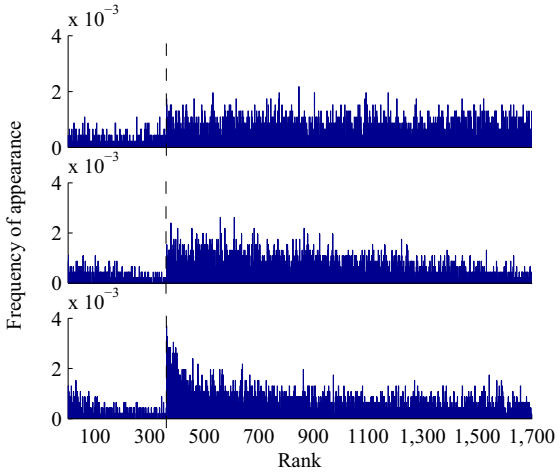


Fig. 2. P2P content distributions: U-Uniform, L-Linear and Z-Zipf

Since our main objective is to concentrate the traffic in the periphery of the network, we dedicate most of the storage capacity to the unpopular contents. The reservation of a small portion of the STBs storage space for the popular content items will provide sufficient alleviation of the edge servers in the busy hours and the rest of the storage will enable reduction of the backbone traffic. By means of simulation we obtained that our objectives are best fitting for values of h that belong to the interval between 10 and 15%. The distributions are based on the contents popularity and determine the number of strips of each content that will be distributed in the peers. Each distribution consists of two equal distributions applied to the popular and unpopular contents. The single distributions applied to both the popular and unpopular contents are Uniform, Linear and Zipf distribution. Another important issue is determining the border between popular and unpopular contents. Following the 80-20 rule of the Pareto distribution which states that 80% of the total number of requests is addressed for the first 20% most popular contents, we will consider 20% to

 ISBN: 978-1-61208-195-3

be the border which will distinguish the contents as popular or unpopular, although we consider different distribution of contents' popularity. Figure 2 shows some of the considered content distributions.

V. Simulations and results

We developed a simulation environment for testing the behaviour of the proposed model with various distributions of the content items on the STBs. In our experiments, we consider a network of $S = 13$ streaming servers organized in a tree structure with $L = 3$ levels (Figure 1) where each level $l = 1, 2$ and 3, contains 10, 2 and 1 servers, respectively. The streaming capacities of the servers in the same order of levels are $U(s) = 500, 1000$ and 1500 Mbps. The links that interconnect the servers have enough capacity to support the maximum streaming load of all the servers. The streaming servers host $C = 1700$ Standard Definition (SD) quality contents with playback rate $r_s = 2$ Mbps and average duration of 60 min.

The servers are serving $N = 5000$ clients divided into $m = 10$ communities, each community directly served by one ES. The maximum percentage of active clients in the system in the peak hours is $\mu = 85\%$. The clients posses STBs with capacity to store the entire length of 3 content items. The portion of this storage reserved for strips of the popular contents is $h = 0.12$. The STBs are connected to the network with links that have download capacity much higher than the playback rate of the SD video quality and uplink capacity $u = 200$ kbps, which is 1/10 of the SD playback rate ($k = 10$).

The popularity of the content items obeys the Zipf-Mandelbrot distribution with shifting coefficient $q = 10$ and $\alpha = 0.8$. The process of generating requests is modelled as a Poisson process. Taking into consideration these data, the storage and streaming capacities of the servers are dimensioned according to (5) and (4) in a way that they are optimally used. The contents are previously distributed on the servers.

In the simulations we considered several different scenarios. The first scenario is the reference for comparison and represents the simple case when the streaming process is completely done by the streaming servers (no P2P). The number of clients that simultaneously request a content item is set to such a value that would keep the streaming servers constantly overloaded and the same request rate will be later used in all the simulation scenarios. In order to compare the contributions of the proposed distributions, we also consider the two simple cases when only the high popularity contents, proposed in [10], are uniformly distributed on the STBs and when the low popularity contents are distributed on the STBs.

Because in our simulations, the servers are kept in a state of high utilization, some of the requests directed to the overloaded servers are rejected and the clients are demanded to request the content latter. The percentage of requests that are rejected for immediate service due to overloaded state of the servers and the time they have to wait until they are served are shown in Figures 3 and 4. The high miss rate in 3, in the scenario with no P2P assisted streaming, is

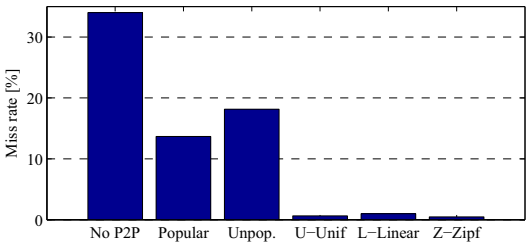


Fig. 3. Miss rate

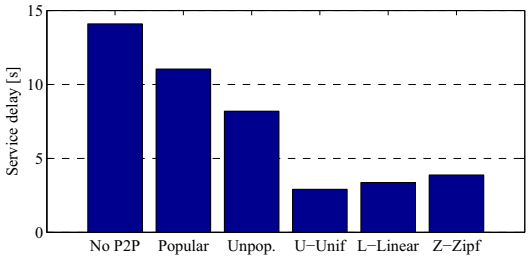


Fig. 4. Service Delay

quite expected result since the request rate of the clients is higher than the available resources. The figure shows that the implementation of P2P assisted streaming for any distribution of the contents on the STBs introduces reduction of the miss rate. The simple distribution of unpopular contents reduce the number of rejected clients to half. This effect is even more emphasized in the distribution of popular contents [10]. The proposed mixed distributions, however, introduce significantly lower miss rates with values around 1%. Although there is only slight difference, the lowest miss rate is obtained for the Z-Zipf distribution, followed by the U-Uniform and L-Linear distribution.

The advantages of the mixed distribution schemes are also visible in the reduction of the service delay (Figure 4). Whenever a client is denied, it has to wait much shorter time when the contents are distributed according to the proposed mixed distributions compared to the other cases.

Another measure that we analyse in order to estimate the contribution of the considered distributions is the transport cost for delivering the streaming traffic from the streaming servers to the clients. This measure is mainly based on the distance of the servers from the clients and their current load and it is expressed as

$$Cost = \sum_{s=1}^{S} d(s)u(s) \tag{6}$$

where $d(s)$ is the distance of server s from the local communities it is serving, counted as number of links, and $u(s)$ is its current streaming rate. Since the P2P streaming is done over the unused uplink rate of the clients, we omit it in the calculation of the transport cost.

Figure 5 shows the average transport cost reduction obtained as a result of the implementation of the various distribution schemes for P2P assisted streaming relative to the case of pure server streaming. The P2P streaming of the most popular contents introduces lowest reduction because it only contributes

 ISBN: 978-1-61208-195-3

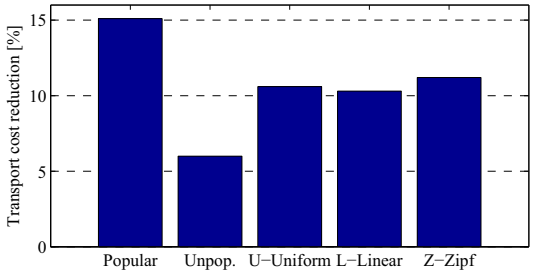


Fig. 5. Transport cost reduction

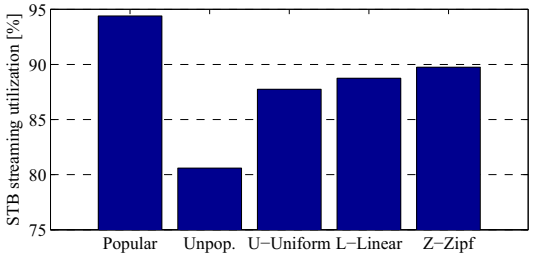


Fig. 6. STB capacity utilization

to reducing the load on the edge servers. On the contrary, the distribution of the unpopular contents on the STBs reduces the traffic in the higher layers and therefore it reaches the maximum reduction of the transport cost. Although the difference is almost insignificant, the Z-Zipf distribution contributes the most for reduction of the transport cost, followed by the L-Linear and U-Uniform distribution.

The various distribution schemes also contribute to a different streaming capacity utilization of the STBs. This dependence is shown in Figure 6. The utilization of the proposed schemes lays between the maximum value obtained for the popular contents distribution and the minimum value obtained for the unpopular content distribution, which is a good compromise considering the improvements that the schemes introduce in the transport cost and the quality of service. The results show that although the mixed distribution schemes do not reach the maximum cost saving and peer utilization of the simple distributions, they are a good compensation for the weak points of both of them. In addition, they significantly improve the number of immediately served clients and the average service delay, which under no condition can be reached by the simple distributions.

One important contribution of the reduction of the traffic in the network core is the possibility to serve more clients with the same streaming capacity of the servers in the core of the network. The advantage of the higher number of clients in the system is that it also implies higher storage and streaming capacities for serving more requests. The only price that has to be paid for the higher number of clients is the installation of new ES on the periphery of the network that would satisfy the demand of the most popular contents. In the case when the popular contents are stored in the STBs, a higher number of clients would require both installation of additional ES and increasing the capacity of the links and the streaming servers in the core of the network. Therefore, the proposed distribution schemes not only reduce the transport cost, miss rate and service delay, but also reduce the installation costs in case of increasing the number of clients in the system.

VI. Conclusions

In this work, we proposed a P2P assisted VoD streaming model that uses the unused storage and uplink capacities of the STBs. We also proposed popularity based distribution schemes of the contents on the STBs determined by assigning different portions of the available storage capacity for the popular and unpopular contents. These schemes prove to reduce the transport cost in the core of the network and to well utilize the uplink capacity of the STBs. In addition, the proposed schemes improve the quality of service that receive the clients by reducing the percentage of rejected request for immediate service as well as the time they have to wait to be served. The reduced traffic in the core of the network and the improved responsiveness give the possibility to increase the number of clients in the system without high installation costs and additional changes in the core of the network, making the system highly scalable.

Acknowledgement

This work has been partially supported by the Ministerio de Ciencia e Innovación of the Spanish Government under project TEC2010-20412 (Enhanced 3DTV).

References

[1] W. Simpson and H. Greenfield, *IPTV and Internet Video: Expanding the Reach of Television Broadcasting*. Elsevier Science & Technology, 2009.

[2] D. De Vleeschauwer and K. Laevens, "Performance of caching algorithms for iptv on-demand services," *IEEE Transactions on Broadcasting*, vol. 55, no. 2, pp. 491 –501, June 2009.

[3] M. Verhoeyen, D. De Vleeschauwer, and D. Robinson, "Content storage architectures for boosted IPTV service," *Bell Labs Technical Journal*, vol. 13, no. 3, pp. 29–43, 2008.

[4] S. Gramatikov, F. Jaureguizar, J. Cabrera, and N. Garcia, "Content delivery system for optimal vod streaming," in *Proceedings of the 11th International Conference on Telecommunications*, june 2011, pp. 487 – 494.

[5] N. Carlsson, D. L. Eager, and A. Mahanti, "Peer-assisted on-demand video streaming with selfish peers." in *Networking*, ser. Lecture Notes in Computer Science, vol. 5550. Springer, 2009, pp. 586–599.

[6] C. Huang, J. Li, and K. W. Ross, "Peer-Assisted VoD: Making Internet Video Distribution Cheap," in *6th International Workshop on Peer-to-Peer Systems*, 2007.

[7] M. Cha, P. Rodriguez, S. Moon, and J. Crowcroft, "On next-generation telco-managed p2p tv architectures," in *Proceedings of the 7th international conference on Peer-to-peer systems*, 2008, pp. 5–5.

[8] K. Suh, C. Diot, J. Kurose, L. Massoulie, C. Neumann, D. F. Towsley, and M. Varvello, "Push-to-peer video-on-demand system: Design and evaluation." *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 9, pp. 1706–1716, 2007.

[9] Y.-F. Chen, Y. Huang, R. Jana, H. Jiang, M. Rabinovich, J. Rahe, B. Wei, and Z. Xiao, "Towards capacity and profit optimization of video-on-demand services in a peer-assisted iptv platform." *Multimedia Syst.*, vol. 15, no. 1, pp. 19–32, 2009.

[10] Y.-F. Chen, R. Jana, D. Stern, B. Wei, M. Yang, H. Sun, and J. Dyaberi, "Zebroid: using IPTV data to support STB-assisted VoD content delivery," *Multimedia Systems*, vol. 16, no. 3, pp. 199–214, 2010.

[11] B. Krogfoss, L. Sofman, and A. Agrawal, "Caching architectures and optimization strategies for IPTV networks," *Bell Labs Technical Journal*, vol. 13, no. 3, pp. 13–28, 2008.

 ISBN: 978-1-61208-195-3

A Domain Pool Classification Method for Better Fractal Volume Compression

Mihai Popescu, Mihai Pancu, and Razvan Tudor Tanasie
Department of Software Engineering
Faculty of Automatics, Computers and Electronics
University of Craiova, Romania
Email: {mpopescu, mpancu, razvan.tanasie}@software.ucv.ro

***Abstract*—In this paper we solve the problem of domain pool classification in fractal volume coding using all symmetries of a cube. Using the algorithms described in the article we are able to perform useful Domain-Range comparisons and achieve better compression rates while not loosing fidelity. In this paper we take advantage of the symmetric permutation group of cube and decrease the number of classes (from 10080 to 840). The same transformations are used in the compression and so we will have a better approximation of the final compressed volume. This paper represents a work in progress so no experimental results can be provided at this time.**

Keywords- domain pool, classification, volume compression, fractals.

I. INTRODUCTION

Fractal compression is a coding technique originally proposed by Barnsley [1] and it's based on the fact that data entropy is self-similar. For this, fractal compression is considered a special form of vector quantization method having the codebook vector self-contained rather than external.

Due to a huge compression time the method was considered impractical at first achieving almost the same rate-distortion curves as the DCT methods. The only attractive properties of the fractal coding that also attracted a lot of research was resolution independence (meaning that data can be decoded at any resolution) and fast decoder convergence.

These properties made fractal compression more suited for off-line applications rather than for real-time ones, like video compression, even if the first attempts [2] to extend the fractal image compression method to the 3D realm was to treat video signal as a volume.

II. STATE OF THE ART

The majority of fractal compression techniques are based on the method developed by Jacquin [3] and later by Fisher et al. [4] by which data is partitioned into blocks named *ranges* and *domains*. The bottleneck of the method resides in the search for the optimal pairing between a range and a domain. Even if the domain size is restricted to be twice the range size, the overlapping lattice of the domain can generate huge domain pools.

Moreover, to improve quality, a domain block is transformed using isometric symmetry operations such that the encoder will find nearly optimal domain-range pairing. If maximum speed is required the symmetry operations can be ignored but the overall quality will suffer dramatically [5].

The surest way is to use a brute force approach and to consider the entire domain pool but the time complexity will be $O(n^2)$ and it becomes impractical as the volume size increases.

Opposed to faster searching, less searching is a promising approach. If one can determine a-priori if a domain is not likely to be used in the final fractal code, eliminating it from the domain pool can improve performances while not loosing fidelity.

One way to reduce the domain pool is by considering only domains at even lattice locations. Another method is by searching in a restricted spatial partition defined by the parent quadrant/octant of the range block or in the near vicinity of it. Using these methods, the spatial information from the fractal code can be efficiently packed with a minimum amount of bits.

Local 2D spiral search from [6] can easily be extended to 3D using a Hilbert scan curve. Other methods can imply just the elimination of a specified fraction of the domains having small variance [7].

For images, Jacquin sorts the domains into three classes (shade, edge and mid-range blocks) and restricted the search only within the same class. On the other hand, Fisher [4] used 72 classes taking into account not only intensities but also the intensity variance across the domains.

The first complete research in fractal volume compression was elaborated by Cochran [8]. He proposed a new classification method by using Principal Component Analysis (PCA) where domains are classified by their variance. PCA is a well known technique for finding orthogonal basis vector that express the direction of progressively smaller variance in a given data set.

In our approach, the volumetric fractal coding algorithm [9] worked by segmenting the volume into domains using an octree and classified the domains using only one rotation. In this paper we take advantage of the complete symmetric permutation group of cube rotations and reflections in order to find the best candidates for Domain-Range comparisons. We are targeting to achieve a good rate-distortion curve disregarding the speed although experimental results cannot be provided at this time.

 ISBN: 978-1-61208-195-3

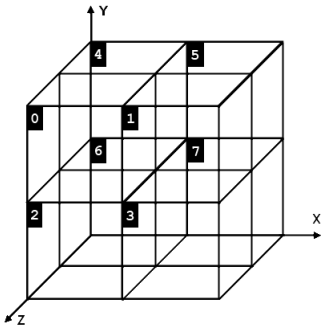


Figure 1. Cube

III. Cube Symmetries

The cube has octahedral symmetry like all the other octahedron solids (e.g., regular octahedron, truncated octahedron, truncated cube) but the cube is the only Platonic solid (among triangular prism, hexagonal prism, truncated octahedron) and also a hierarchical space-filling polyhedron. The octahedral symmetry group O_h has symmetry order 48 including transformations that combine a reflection with a rotation. It is isomorphic to $S_4 \times C_2$ and has 24 direction-preserving symmetries. The elements are:

1) 1 identity rotation that leaves the cube unchanged
2) 3 rotations (by $\pm\pi/2$ or π) around the centres of the 3 pairs of opposite faces
3) 1 rotation (by π) around the centres of the 6 pairs of opposite edges (that pass through the centre)
4) 2 rotations (by $\pm 2\pi/3$) around the 4 pairs of the opposite vertices (on diagonals)

To sum up, we have $1+3*3+1*6+2*4 = 24$ rotations.

The other 24 symmetries do not preserve directions because the transformation include a reflection and this implies changing the face normals along the symmetry plane. There are two types of symmetry planes for the cube. One is perpendicular to the coordinate system unit vectors $\vec{i}, \vec{j}, \vec{k}$ (one for each axis) and the other type is across diagonals (two for each pair of opposite vertices). So there are $3+2*3 = 9$ planes of symmetry for a cube.

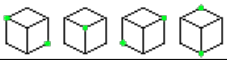
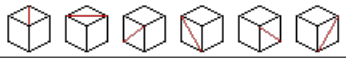
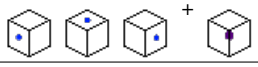
The other $24-9 = 15$ symmetries are turn-reflections and they combine a rotation and the antipodal reflection plus the antipodal reflection itself. All 48 cube symmetries are summarized in the Table I.

IV. Cube Symmetric Permutation Group

As we said in previous sections, cube's symmetric permutation group has 48 elements. For example, the right-hand rule rotations around the system axes (see Figure 1) can be encoded in permutations using the cycle notation as:

$$\begin{cases} \sigma_x = (0\ 2\ 6\ 4)(1\ 3\ 7\ 5) \\ \sigma_y = (0\ 1\ 5\ 4)(2\ 3\ 7\ 6) \\ \sigma_z = (0\ 2\ 3\ 1)(4\ 6\ 7\ 5) \end{cases} \quad (1)$$

Table I
Symmetries of the Cube

	identity
	$\pm\pi/2$ face rotation
	π face rotation
	π edge rotation
	$\pm 2\pi/3$ diagonal rotation
	axis plane reflection
	diagonal plane reflection
	$\pm\pi/2$ face rotation + antipodal reflection
	$\pm 2\pi/3$ edge rotation + antipodal reflection
	antipodal reflection

We know that σ_z can be expressed as a combination of σ_x and σ_y permutations because when rotating around the unit vector $\vec{k} = \vec{i} \times \vec{j}$, both $\vec{i}$ and $\vec{j}$ are rotated as well. We can generate the rotation permutation group using just the canonic generators σ_x and σ_y using the following algorithm.

The $GetPermutationNumber$ procedure just returns a unique number identifying the permutation (for example, the associated radix integer $r = \sum p[i] * 10^i$) where $GetPermutationName$ returns the generator name (e.g., e, x, y).

If we supply to *GeneratePermutationGroup* algorithm the $PermutationGenerators = \{\sigma_e, \sigma_x, \sigma_y\}$ it will generate the rotation symmetric permutation group, equivalent with the finite 3D rotation permutation group $SO(3)$. Note that $\sigma_e = (0)(1)(2)(3)(4)(5)(6)(7)$ is the identity permutation that leaves the cube unchanged. Its associated Cayley graph can be depicted in Figure 2. We can observe that the algorithm has found all 24 orientation-preserving permutations without providing σ_z as a generator because $z = xxyx$.

To find the other 24 permutations we just have to add antipodal reflection $\sigma_r = (0\ 7)(1\ 6)(2\ 5)(3\ 4)$ as a generator, $G = \{e, x, y, r\}$.

V. Classification of Permutations

We are making an isomorphism between the symmetric permutation of the 8 vertices of a cube and the arrangement of the density of its 8 partitioned octants. With this we can use the 48 symmetries at the octant level.

 ISBN: 978-1-61208-195-3

Algorithm 1 GeneratePermutationGroup

Require: Generators $G[?]$
Ensure: PermutationGroup $PG[48]$
  {Initialize permutation group map with generators}
  $Q \leftarrow \emptyset$
  $count \leftarrow Length(G)$
  for $i = 1 \rightarrow count$ **do**
    $n \leftarrow GetPermutationNumber(G[i])$
    $map[n] \leftarrow GetPermutationName(i)$
    $PG[i] \leftarrow G[i]$
    $Push(Q, G[i])$
  end for
  {Generate permutations in Breadth-First order}
  while $Length(Q) > 0$ **do**
    $p \leftarrow Front(Q)$
    for $i = 1 \rightarrow Length(G)$ **do**
      if $G[i] \neq \sigma_e$ **then**
        $q \leftarrow GeneratePermutation(G[i], i, p)$
        if $q \neq nil$ **then**
          $count \leftarrow count + 1$
          $PG[count] \leftarrow q$
          $Push(Q, q)$
        end if
      end if
    end for
  end while

Algorithm 2 GeneratePermutation

Require: Generator g, Generator Index i, Permutation p
Ensure: Permutation
  $n \leftarrow GetPermutationNumber(p)$
  $pit \leftarrow Find(map, n)$
  if $pit = nil$ **then**
    return nil
  end if
  $name \leftarrow Second(pit)$
  $q \leftarrow p \circ g$
  $m \leftarrow GetPermutationNumber(q)$
  $qit \leftarrow Find(map, m)$
  if $qit = nil$ **then**
    return nil
  end if
  $name \leftarrow Concat(name, GetPermutationName(i))$
  $map[m] \leftarrow name$
  return q

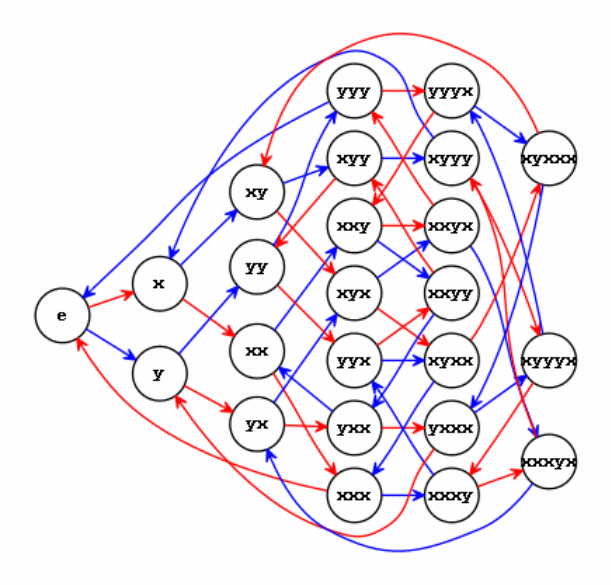


Figure 2. Cayley graph

There are $N = 8! = 40320$ possible configurations of the permutations of the order of density values for each octet $A_i, i = \overline{0..7}$. All 48 permutations form a conjugacy class so there are $N/48 = 840$ different classes.

For example the identity permutation σ_e is a permutation that corresponds to the first conjugacy class and to the ordering $A_0 \leq A_1 \leq A_2 \leq A_3 \leq A_4 \leq A_5 \leq A_6 \leq A_7$ being a 1-to-1 correspondence to the permutation canonic representation [10]: $\{0, 1, 2, 3, 4, 5, 6, 7\}$.

If we are rotating σ_e permutation along the x axis, we will get: $\sigma_x \circ \sigma_e = \{2, 3, 6, 7, 0, 1, 4, 5\}$ that corresponds to $A_2 \leq A_3 \leq A_6 \leq A_7 \leq A_0 \leq A_1 \leq A_4 \leq A_5$, which will be also in the same class.

If we do this for all 40320 permutations we will find the class for each permutation.

The class map is symmetric with itself, having the property that: $\forall i = \overline{1..40320}$, $C[i] = C[40320 - i - 1]$. This happens because the algorithm $GetNextPermutation$ (Knuth's L-Algorithm [11]) generates permutations in lexicographic order starting from $\{0, 1, 2, 3, 4, 5, 6, 7\}$ and with the antipodal reflection $\{7, 6, 5, 4, 3, 2, 1, 0\}$ the permutations will be in the same class.

The volume is partitioned using an octree generating for each level 8 octants. We must sort the octants by their average density to find the equivalent permutation but we are more interested for their initial indices than for the final sorted array. For this we are going to use a simple modified selection sort algorithm.

In Algorithm 4 we use the $GetPermutationIndex$ function (Knuth's P-Algorithm [12]) to find the index of the permutation and we use it with the pre-computed array class such that each octant will be classified.

 ISBN: 978-1-61208-195-3

Algorithm 3 GeneratePermutationClassMap

Require: PermutationGroup $PG[48]$
Ensure: PermutationClassMap $C[40320]$
  $k \leftarrow 1$
  $p \leftarrow \{0, 1, 2, 3, 4, 5, 6, 7\}$
  repeat
    $i_p \leftarrow GetPermutationIndex(p)$
    if $C[i_p] \neq nil$ **then**
      for $i = 1 \rightarrow 48$ **do**
        $q \leftarrow p \circ PG[i]$
        $i_q \leftarrow GetPermutationIndex(q)$
        if $C[i_q] \neq nil$ **then**
          $C[i_q] \leftarrow k$
        end if
      end for
      $k \leftarrow k + 1$
    end if
  until $GetNextPermutation(p) = nil$

Algorithm 4 FindPermutationClass

Require: Octants $O[8]$, PermutationClassMap $C[40320]$
Ensure: PermutationClass
  {Compute average density}
  for $i = 1 \rightarrow 8$ **do**
    $count \leftarrow VoxelCount(O[i])$
    $A[i] = \frac{1}{count} \sum_{j=1}^{count} VoxelDensity[j]$
  end for
  {Selection Sort}
  for $p = 1 \rightarrow 8$ **do**
    $min = p$
    for $i = p + 1 \rightarrow 8$ **do**
      if $A[i] < A[min]$ **then**
        $min \leftarrow i$
      end if
    end for
    $Swap(A, p, min)$
    $P[p] \leftarrow (min - 1)$
  end for
  return $C[GetPermutationIndex(P)]$

VI. Conclusion and Future Work

This paper uniquely presents how to use symmetric permutation groups to classify octree nodes into similar blocks. This classification is used to feed the Fractal Volume Coding algorithm such that the Domain-Range search will provide better distortion-curve results.

Future work will focus on speeding the domain pool search. We can extend the work of Kominek [5] to 3D by implementing a multi-dimensional r-tree indexing of the domain pool or by using other spatial data structures like the M-Tree [13]. Another interesting approach is to exploit SIMD architecture of the graphics commodity hardware available [14].

In this paper we complete our preliminary work done in Fractal Volume Coding [9] by finding a method for classifying the Domain Pool so that to have a good compression ratio. Having this done we can advance into the full implementation and have some useful experimental results. Unfortunately we do not have any experimental results to provide at this moment but a detailed description of the classification method will provide a better understanding of the method in matter.

References

[1] M.F. Barnsley, and L.P. Hurd. *Fractal Image Compression*, AK Peters, Ltd., Wellesley, Ma., 1992.

[2] M. S. Lazar and L. T. Bruton. *Fractal block coding of digital video*, IEEE Transactions on Circuits and Systems for Video Technology, vol. 4, no. 3, pp. 297308, 1994.

[3] A. Jacquin. *A Fractal Theory of Iterated Markov Operators with Applications to Digital Image Coding*, PhD thesis, Georgia Institute of Technology, 1989.

[4] Y. Fisher, E.W. Jacobs, and R.D. Boss, *Fractal image compression using iterated transforms*, Technical Report 1408, Naval Ocean Systems Center, San Diego, CA, 1991.

[5] J. Kominek, *Advances in fractal compression for multimedia applications*, Multimedia Systems, 5:255-270, Springer-Verlag, 1997.

[6] J. M. Beaumont, *Advances in block based fractal coding of still pictures.*, Proc IEEE Colloquium: The Application of Fractal Techniques in Image Processing, pp 3.13.6, 1990.

[7] D. Saupe, *Lean domain pools for fractal image compression*, Proceedings of PSIE Electronic Imaging'96, Science and Technology, Still Image Compression II, Volume 2669, San Jose, 1996.

[8] W. O. Cochran, J.C. Hart, and P.J. Flynn, "Fractal volume compression", IEEE Transactions on Visualization and Computer Graphics, Page(s):313 - 322, 1996.

[9] M. Popescu, M. Tudorache, and R. Tanasie, *Volume Content Indexing using a Fractal Coding Algorithm*, IEEE Computer Society, 2010.

[10] Donald E. Knuth, *The Art of Computer Programming*, Section 1.3.3, p.178-179, Volume 1, 3rd Ed, Addison-Wesley, 1997.

[11] D. E. Knuth, *The Art of Computer Programming*, Section 7.2.1.6, p.3, Volume 4, Fascicle 4a, Addison-Wesley, 2005.

[12] D. E. Knuth, *The Art of Computer Programming*, Section 3.3.2, p.65-66, Volume 2, 3rd Ed, Addison-Wesley, 1998.

[13] M. C. Mihaescu, D. D. Burdescu, "Using M Tree Data Structure as Unsupervised Classification Method", Informatica Journal, Ljubljana, 2011.

[14] U. Erra, *Toward Real Time Fractal Image Compression Using Graphics Hardware* Advances in Visual Computing, 2005 - Springer

 ISBN: 978-1-61208-195-3

Video Casting Application Oriented Key Exchange

Abdullah Rashed and Henrique Santos
Algortimi Centre
University of Minho
Guimaraes, Portugal
rashed@dsi.uminho.pt / hsantos@dsi.uminho.pt

***Abstract*— Within video stateless receivers, a central server should deliver information securely to the authorized users, over a public channel, even if receivers do not update their state from session to session. This is the case of a multimedia conditional access systems based on one way broadcasting. This paper suggests a new approach to assure a secure communication in such environments. The proposed approach is an efficient key exchange scheme for stateless receivers. It reduces the number of private keys used in traditional conditional access systems and the number of encryptions operations as it does not need to encrypt the ciphering keys. Furthermore, the presented approach eliminates the required key refreshment presented in other approaches. We tested the proposed system using AES algorithm. A numerical example is used to demonstrate the effectiveness of the presented approach. This technique can be very useful for small devices, with limited resources and strict power consumption requirements, which are becoming prevalent in multimedia Conditional Access Systems (CAS) one way broadcasting.**

Keywords-Key exchange; Broadcast Encryption; Conditional Access Systems.

I. INTRODUCTION

Security of digital multimedia transmission is very important due to the communication explosion [1]. It is widely used over PDAs (Personal Digital Assistants), mobile phones and other network devices, over public channels (cable, satellite, wireless networks, Internet, etc.) [6]. Naturally one main concern is copyright protection and access control.

To protect copyright and access control, broadcaster should use an encryption system [1] [6]. However, the limited batteries life of these devices obliges to reduce encryption computational complexity. Furthermore, access control mechanisms can enforce security protecting mainly confidentiality and integrity – availability is not addressed here, despite its importance, since it is normally enforced by different controls. Standard cryptographic techniques can guarantee high level of security but at the cost of expensive implementation and important transmission delays [11]. Selective encryption comes as an alternative that aims to provide sufficient security with an important gain in computational complexity and delays [1].

Broadcasting Encryption (BE) aims to distribute the data to all authorized users simultaneously, in an efficient way and securely [15]. This balance is particularly critical with small devices which don't have enough resources to implement complex encryption techniques. Furthermore, this devices usually do not even keep information between sessions – stateless devices. Within stateless receivers, a media server must deliver information securely to the authorized users over a public channel, where the receivers do not update their state from session to session [13].

A typical CAS (Conditional Access System) depends on three level of encryption as shown in Fig. 1; at the sender side, the raw content is encrypted using a Control Word (CW), which is encrypted by a Service Key (SK). SK is embedded into an Entitlement Control Message (ECM), which is encrypted using a Personal Distribution Key (PDK) assigned to each authorized user. The PDK is embedded into an Entitlement Management Message (EMM). The EMM is specific to each subscriber, as identified by the smart card in their receivers, or to groups of subscribers, and are issued much less frequently than ECMs, usually by monthly intervals. Encrypted content (both ECM and EMM) is transmitted through broadcast networks. SK is renewed at intervals of hours or days, while PDK is static and known only by the service provider and the user's terminal, being embedded into firmware. At the receiver side, SK is decrypted and used to decrypt ECM, which allows getting CW, necessary to decrypt the content [14]. For pay TV, many authors preferred building a separate key tree for each multicasting program [10].

As stated above, considering low power receivers, the cipher system should be both robust and low demanding, considering computational resources. The mechanism proposed in this paper tries to respond to those requirements.

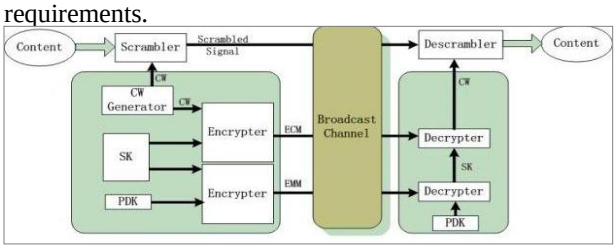


Figure 1. typical Conditional Access System [14].

 ISBN: 978-1-61208-195-3

This paper is organized as following: in section II we overview the related work; in section III, we present our approach, and conclude in section IV.

II. LITERATURE REVIEW

Eagle et al. [5] studied the number of encryption operations necessary to revoke keys. They studied the well-known used trees based broadcast encryption schemes. They proved that the mean number of encryption processes for the complete sub-tree scheme and the subset-difference scheme studied by previous studies were good estimates for the number of encryption processes used by this scheme. Their study focused on proving a normal limiting distribution for the number of encryptions as the number of users became large. They took into consideration the combined number of encryptions and number of privileged users in a random privileged set [5].

In [4], Kirkels et al. described a security architecture for a pay-TV CAS. They focused on the design constraints related to a conditional access client in the design of the architecture and maximum amount of bandwidth available for the transmission of conditional access messages. They presented the design and analysis of their efficient injector model based on queuing theory, conditional access messages into the broadcast stream. To demonstrate the effectiveness of the presented approach, they presented a numerical example with real-world values.

In [1], Massoudi et al. introduced the selection encryption of image and video scheme to reduce the amount of encrypted data, keeping the security goals. Their protocol consists of two parts: public part, where there is no encryption; and protected part, encrypted and only accessible to authorized users.

In [14], Zhang et al. introduced a novel way to solve the tradeoff problem about communication, storage and computation overhead of BE scheme. They suggested getting rid of the computation overhead that come from broadcast key generation. They constructed a scheme based on Subset Difference (SD) and RSA accumulator. Their idea of separating the user-side device into two different function parts (private and public parts), taking advantage of the public device's functionality, minimize the storage and computation overhead of the private device, and make BE scheme more implementation-oriented.

In [9], Shirazi et al. presented and described Mobile Integrated Conditional Access System (MICAS). They demonstrated the various architectures to deliver key information at an arbitrary located device, at the surrounding area of the subscriber. They described the advantages of the system. Their proposed system included the message handling subsystem with a so-called 'Follow-Me' service, which extends mobility and personalization concepts on pay-TV services. Subscriber Management and Subscriber Authentication Subsystems would respond to the subscribers interaction (via mobile phone) issuing them the corresponding access rights. Their system is supposed to reduce the cost for service provider and end-users by respectively cutting down the service deployment cost and eliminating the requirement of additional receiver as changing the service provider.

Abdalla et al. [10] discussed how to communicate securely with a set of users (the target set) over an insecure broadcast channel for application domains: satellite/cable pay TV and the Internet MBone. They concerned about the number of key transmissions and the number of keys held by each receiver. They suggested maintaining single key structure such that receiver should keep a logarithm number of establishment keys, for the entire life time [10].

Zhang et al. [14] presented a CAS model using an encryption scheme for one way broadcasting and protection application. They compared it with traditional Conditional Access Systems. They discussed the advantages and challenges of BE [14].

In [8], Koo et al. presented a key refreshment management scheme for CAS in DTV broadcasting. They concluded that their scheme perform dynamic entitlement management securely and efficiently and reduced a key generation and encryption load for CAS. In [7], authors discussed the problem of a server sending a message to a group of the stateless receivers, assuming that a subset of the keys have been revoked and should not be able to obtain the content of the message. They provided sufficient conditions to guarantee the security of revocation schemes.

In [13], Hwang et al. presented an efficient revocation scheme for stateless receivers. They used a logical hierarchical key tree. They considered Asano's schemes [12] very efficient with respect to key storage. They used hierarchical key based on a binary tree, and they found that it requires the same message length as the SD scheme. Asano proposed two efficient revocation methods for stateless receivers. He used the Master Key technique and the 'Power Set Method' with an a-ary key tree structure in order to reduce the number of keys each receiver stores and the number of ciphered messages broadcasted, respectively. The first method required receivers to store only one key. The second method was supposed to reduce the computational overhead imposed to receivers, but with an increase in the number of master keys they have to store. He discussed the security of his methods and some techniques used in his methods [12].

In spite of all the work already done, key management is still a main concern in multimedia CAS, particularly considering small devices with computational resource constraints and real-time demands. Key exchange and revocation must be very efficient tasks in order to achieve fast and low power consummation operations. None of the previous solutions seems to be an optimal solution, justifying additional research efforts.

III. PROPOSED SYSTEM

The approach proposed in this paper aims at protect copyright, without requiring substantial architecture modifications, and avoiding the need to store or exchange encryption keys; every block is encrypted using a different schedule key, which is scrambled within the message

The proposed protocol assumes that:

 ISBN: 978-1-61208-195-3

1. Registration: users should register and would be granted the personal distribution key (PDK) – this distribution mechanism is not a main concern here.
2. Compression: raw content would be compressed.
3. Private Cipher key (CK) generation: key is generated and expanded to gain schedule key.
4. Encryption: expanded CK is used to encrypt the compressed raw data.
5. Scramble: both CK and encrypted data (C) are transmitted together in a scrambled way.
6. Broadcasting the cipher key can be solved by scrambling the cipher key with ciphered block in a special way that only the legitimated receiver can understand, allowing it to extract both cipher key and ciphered block and this way decypher the original message. The scrambling technique is described next.

A. *Scrambling Algorithm: Starting Random*

First, a Boolean number is randomly generated. A zero means to start with the ciphered data block element, whereas a one implies starting with ciphering key, as shown in Fig. 2. After that, the scrambling process proceeds according to the algorithm presented in the listing bellow, for which an output example is given in Fig. 3, and a flowchart is given in Fig. 4.

Scrambling Algorithm
Input: ciphered data block, ciphering key
Output: Scrambled Token (ST).
Function Scrambling
Begin
Generate start, using lookup table
for i=0 to block Cipher size-1 step by 1
ST 2i+1+ start += ciphered data block i
ST 2i+ start += ciphering key block i
end for
ST= the rest of the ciphering key
end function Scrambling

1) *Illustrative Example Using AES algorithm*

To illustrate how the algorithm works we will show next an example for a symmetric block ciphering (e.g., AES). Assuming Nb represents the block length in bytes (C0-C15), 16 bytes in this example. The ciphering key should be the same length, denoted by Ck0 -Ck15.

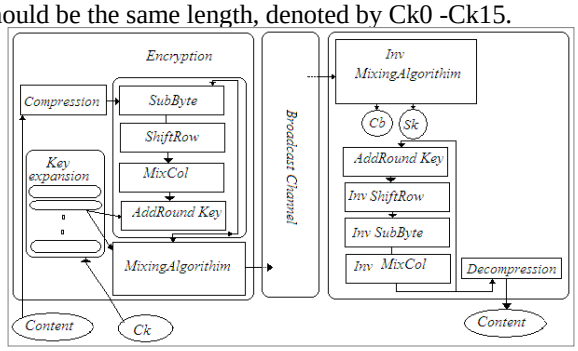


Figure 2. Proposed System

0	1	2	3	4	5
0	C_0	Ck_0	C_1	Ck_1	C_2
6	7	8	9	10	11
Ck_2	C_3	Ck_3	C_4	Ck_4	C_5
12	13	14	15	16	17
Ck_5	C_6	Ck_6	C_7	Ck_7	C_8
18	19	20	21	22	23
Ck_8	C_9	Ck_9	C_{10}	Ck_{10}	C_{11}
24	25	26	27	28	29
Ck_{11}	C_{12}	Ck_{12}	C_{13}	Ck_{13}	C_{14}
30	31	32			
Ck_{14}	C_{15}	Ck_{15}			

Figure 3. Ciphered block and ciphering key

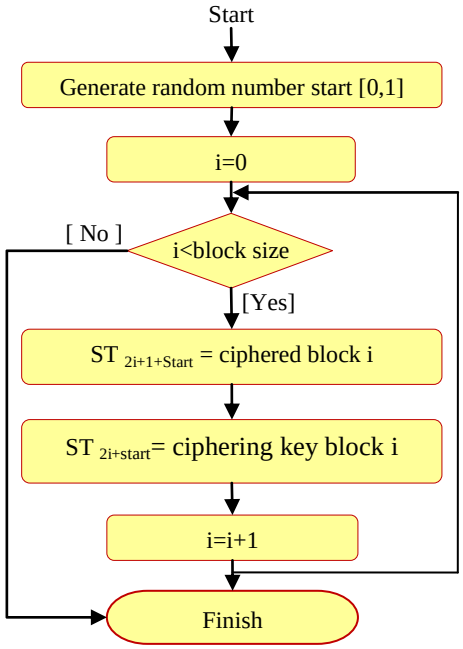


Figure 4. Proposed Algorithm

Following the algorithm proposed, in this case the output would be 33 bytes long, comprising one byte start code, 16 bytes for ciphered data and 16 bytes for cipher key.

Assuming the ciphered data block was: “a6 d9 f3 60 39 53 ff 11 13 6e 03 06 7f 8a 57 fa”, as shown in Fig. 5 and the ciphering key was “41 73 69 6d 20 41 20 45 6c 2d 53 68 65 69 6b 68”, as shown in Fig. 6, and assuming the random generator produced a 0 start code, the output block would be filled starting with ciphered data and will look like the stream showed in Fig. 7.

 ISBN: 978-1-61208-195-3

a6	d9	f3	60
39	53	ff	11
13	6e	03	06
7f	8a	57	fa

Figure 5. Ciphered block

a6	d9	f3	60
39	53	ff	11
13	6e	03	06
7f	8a	57	fa

Figure 6. Ciphered block

0	1	2	3	4	5
0	**a6**	41	**d9**	73	**f3**
6	7	8	9	10	11
69	**60**	6d	**39**	20	**53**
12	13	14	15	16	17
41	**ff**	20	**11**	45	**13**
18	19	20	21	22	23
6c	**6e**	2d	**03**	53	**06**
24	25	26	27	28	29
68	**7f**	65	**8a**	69	**57**
30	31	32			
6b	**fa**	68			

Figure 7. Output of mixing ciphered block & ciphering key

IV. CONCLUSION

This paper introduced a new approach for multimedia Conditional Access Systems (CAS), avoiding the key exchange scheme. This technique is particularly useful for stateless receivers. The proposed approach uses dynamic key generation. Our approach is efficient with respect of the number of keys stored and used to encrypt the data. It reduces the complexity of other solutions as it does not need to encrypt the ciphering keys. It uses fewer keys to reduce the storage and scramble the transmitted data to reduce encrypting the keys. Furthermore, the presented approach eliminates the proposed key refreshment presented in [14] and [8]. However, there will provide a gain in computational complexity and delays. We demonstrated how the proposed technique works using a block cipher like AES, proposed by [3]. A practical example is used to demonstrate the effectiveness of the presented approach.

As future work, we will demonstrate a practical architecture and evaluate the resistance to direct attacks. For future, we will test the algorithm in real CAS environment (IPTV system) to compare it with well-known algorithms. The comparison with similar work in the real systems would be reflected by the benefits that algorithm would introduce to the industry.

ACKNOWLEDGMENT

This work was funded by FEDER through Programa Operacional Fatores de Competitividade – COMPETE, and by national founds through FCT – Fundação para a Ciência e Tecnologia, under project: FCOMP-01-0124-FEDER-022674.

REFERENCES

[1] A. Massoudi, F. Lefebvre, C. De Vleeschouwer, B. Macq, and J. Quisquater (2008), “Overview on Selective Encryption of Image and Video: Challenges and Perspectives”, EURASIP Journal on Information Security, Vol. 2008, (January 2008), Article 5.

[2] A. Rashed (2007), “Using Modified Genetic Algorithm to Replace AES Key Expansion Algorithms”, The International Conference on Information Technology (ICIT’2007) at Al-Zaytoonah University, Jordan on May 9-11, 2007. WWW.alzaytoonah.edu.jo/icit2007

[3] A. Rashed and N. Ajlouni (2004), “An Extended Rijndael Block Cipher Using Java”, the 2004 International Conference on software Engineering Research and practice, Las Vigas, Nevada USA, June 2004, pp. 21-24.

[4] B. Kirkels, M. Maas, and P. Roelse (2007), ”A Security Architecture for Pay-Per-View Business Models in Conditional Access Systems”, ACM Workshop On Digital Rights Management, Proceedings of the 2007 ACM workshop on Digital Rights Management: Alexandria, Virginia, USA:1-9.

[5] C. Eagle, Z. Gao, M. Omar, D. Panario, and B. Richmond (2008), “Distribution of the Number of Encryptions in Revocation Schemes for Stateless Receivers”, Fifth Colloquium and Computer Science, DMTCS proc. AI: pp. 195-206.

[6] D. Dardari, M. Martini., M. Mazzott, and M. Chiani (2004), “Layered Video Transmission on Adaptive OFDM Wireless Systems”, EURASIP Journal on Applied Signal Processing, Volume 2004: pp. 1557 - 1567

[7] D. Naor, M. Naor, and J. Lotspeich (2001), “Revoking and Tracing Scheme of Stateless Receiver”, Proceedings of Crypto01, LNCS 2139, pp. 29-30.

[8] H. Koo, O. Kwon, and J. Kim (2005), “Key Refreshment Management for Conditional Access System in DTV Broadcasting”, International Conference consumer Electronics, Jan 2005 : pp. 29-30

[9] H. Shirazi, J. Cosmas, D. Cutts, N. Birch, and P. Daly (2008), “Security Architectures in Mobile Integrated Pay-TV Conditional Access System”, Networks 2008 - 13th International Telecommunications Network Strategy and Planning Symposium 1.

[10] M. Abdalla, Y. Shavitt, and A. Wool (2000), “Key Management for Restricted Multicast Using Broadcast Encryption”, IEEE/ACM Transactions on Networking (TON), Vol. 8 , Issue 4: pp. 443 - 454

[11] N. Ajlouni, A. El-Sheikh, and A. Rashed (2006), “New Approach in Key Generation and Expansion in Rijndael Algorithm”,

 ISBN: 978-1-61208-195-3

International Arab Journal of Information Technology, vol. 3, no. 1, January 2006.

[12] T. Asano (2002), "A Revocation Scheme with Minimal Storage at Receivers", ASIACRYPT'02, LNCS V.2501: pp. 433-450.

[13] Y. Hwang, H. Chong, and J. Pil (2004), "An Efficient Revocation Scheme for Stateless Receivers", EuroPKI 2004, LNCS 3093, Springer-Verlag Berlin Heidelberg: pp. 322–334.

[14] Y. Zhang, C. Yang, J. Liu, and J. Tian (2009), "Broadcast Encryption Scheme and Its Implementation on Conditional Access System", Proceedings of the 2009 International Symposium on Web Information Systems and Applications (WISA'09) Nanchang, P. R. China, May 22-24, 2009: pp. 379-382

[15] Y. Zhang, C. Yangt, J. Liu, and J. Li (2007), "A Novel Broadcast Encryption Scheme Based on SD Scheme Reconstruction", Communications and Networking in China, 2007. CHINACOM '07, Second International Conference on Digital Object Identifier: 10.1109/CHINACOM.2007.4469408, pp.: 387 – 391.

 ISBN: 978-1-61208-195-3

Multi-Connected Ontologies

Philip Davies
Higher Education
Bournemouth and Poole College
Bournemouth, UK
pdavies@bpc.ac.uk

David Newell
Software Systems Research Group
Bournemouth University
Bournemouth, UK
dnewell@bournemouth.ac.uk

Abigail Davies
St Johns College
Oxford University
Oxford, UK
abigail.davies@sjc.ox.ac.uk

Damla Karagözlü
Software Systems Research Group
Bournemouth University
Bournemouth, UK
damla.karagozlu@gmail.com

***Abstract* – Ontologies have been used for the purpose of bringing system and consistency to subject and knowledge areas. We present a criticism of the present mathematical structure of ontologies and indicate that they are not sufficient in their present form to represent the many different valid expressions of a subject knowledge domain. We propose an alternative structure for ontologies based on a richer multi connected complex network which contains the present ontology structure as a projection. We demonstrate how this new multi connected ontology should be represented as an asymmetric probability matrix.**

Keywords – adaption; semantic; taxonomy; ontology; anthology.

I. INTRODUCTION

A. The present state of ontologies

There has been exceptional growth in the annotation of information prompted by the increasing need to share data and study objects based on their structure and semantics. [1] We now find annotated information in a wide range of areas such as language, biology, computing, medicine, web content, etc. Annotated information is created from structured vocabularies known as ontologies. Many disciplines have now developed their own standardized ontologies to enable the sharing of information in their fields. SNOMED, for instance has been produced in the field of medicine, [2] as well as many others which are now being referenced. [3]

Ontology defines a common vocabulary for researchers who need to share information in a domain. Many subject areas are now developing ontologies so that specialists can share information in their fields not only with other specialists but even with machines. [4] Machine-interpretable definitions of basic concepts in the domain and relations among them enable the widespread use of information on the internet and the construction of expert systems.

An ontology uses relationships to organize concepts into hierarchies or subject domains. [3] This paper investigates the present structure of ontologies and whether they are applicable to describing subject domains in their present form. The basic problem we consider is whether the present structure of ontologies is rich enough to represent subject domains fully. We contend that the concept of ontologies needs to be extended in order to fully realise a complete subject domain and we indicate ways in which this extension might be approached

B. Critique of Ontologies

Our approach to ontology structure originates with the ideas of the German philosopher Martin Heidegger (1889 – 1976). Heidegger was critical of a one-dimensional division of the world into simplistic categories. According to Heidegger, "*The philosophical tradition has misunderstood human experience by imposing a subject-object schema upon it.*" [5]

Heidegger gives the example of a hammer which cannot be represented just by its physical features and functions. To understand the hammer you cannot detach it from its relationship to the nails, to the anvil, to the wood, to the experience and skill of the carpenter or to a hundred other things. Just putting it in a category of tools, in an ontology cannot fully capture the human idea of the object and its role in the world. A more complex structure is needed to capture the representation of reality. [5]

Robert Pirsig [6] has also made the point that there always appears more than one workable hypothesis to explain a given phenomenon, and that the number of possible hypotheses appears unlimited. He has developed the idea that there are two types of thinking, the classical and the romantic. The classical way of thinking is characterised by analysing things into their component parts, whereas the romantic sees things as a whole. Classical thought would analyse an object like a motorbike into its physical components; nuts bolts etc. but you can also analyse the motor bike into its functional parts: heat exchanger, generator, exhaust system etc. Pirsig points out that each analysis is equally valid but produces different results. It depends on how you wield the knife of analysis to separate part from part. For example if you take a cylindrical chunk of clay you can cut it straight down and the product is circles, but if you decide to cut at an angle the result is ellipses, if you cut horizontally you obtain rectangles. The result of any analysis is also the product of what you decide to do and how you decide to cut, as much as it is a product of the artefact you are looking at. No analysis is unique.

This directly affects the construction of ontologies as these are the results of detailed analysis of a subject area. Since different analyses lead to different ontologies and each may be equally valid, it has become necessary to agree on a convention as to what the structure of any given ontology may be and this agreement by subject experts is the method chosen to determine an agreed ontology. But, we contend here that agreement by convention on the structure of a subject ontology is not

 ISBN: 978-1-61208-195-3

sufficient, as there are intrinsic differences between representations which cannot be reconciled because the subject domain is richer than any single ontology can capture. Different ontologies result from the way the knife of analysis has been wielded as much as from the subject area itself.

David Bohm in his book Wholeness and the Implicate Order [7] presents a critique of the fragmentation that classical thought has introduced into our description of the world. He says that it has always been necessary and appropriate to divide things up and separate them in order to reduce problems to manageable proportions but in so doing we lose sight of the whole. In dividing things up we make the mistake of thinking that the fragments we produce are a proper description of the world as it is. The problem is that there are many different ways of thinking about something and of categorising concepts and ideas. And no one way is better than another. He uses the field of quantum mechanics to illustrate this with its wave picture and particle picture of reality which are at the same time incompatible and indivisible. "*All our different ways of thinking are to be considered as different ways of looking at the one reality*" says Bohm. [7] Each view gives only one appearance of the object in some respect. "*The whole object is not perceived in any one view but rather it is grasped only implicitly as that single reality which is shown in all these views*." [7]

This has direct application to the way we use ontologies. These are constructed on the premise that in order to communicate about a particular subject domain unambiguously we need to have an agreed reference point, the ontology, which fixes precisely and unambiguously the component of the subject domain and its fixed relationships to other points. What Heidegger, Pirsig and Bohm are telling us is that this approach may be wrong from the outset and ultimately unachievable in the long term. A single ontology to describe the whole of reality is not something that exists. Rather many incompatible ontologies will exist that are equally valid descriptions of reality. And merely agreeing on one of them for the sake of convention will not enable a full picture of the reality to be represented. What is needed is a larger concept which contains all possible ontologies in a single undivided structure implicitly and from which they can be explicated. We can liken the new structure to a three dimensional object that casts different shadows depending on which way the light falls and each shadow represents the ontology while the object is the reality.

II. A New Approach to Constructing Ontologies

We propose to adopt a new approach to describing knowledge systems based on the idea that there is no one correct method of organising a subject domain in an ontology but rather there are many different ontology structures that adequately and correctly represent a body of knowledge. Each ontology gives only one appearance of the subject in some particular respect.

C. Taxonomy, Ontology or Anthology?

We will use a particular example throughout this paper in order to illustrate the extension to ontologies. However, it should be clear that the approach we employ is applicable to both ontologies and taxonomies. There is debate as to the precise relationship between ontologies and taxonomies while the the distinction between an ontology and a taxonomy is often blurred. [8]. We use here the distinction that a taxonomy is a hierarchical structure to classify information in a subject domain while an ontology is a hierarchical structure which in addition assigns and defines properties and relationships between concepts. An ontology is a richer structure than a taxonomy describing all aspects of the world and its parts [9]. Very simple ontologies would reduce down to a taxonomy in practice. Consequently whatever is true of a taxonomy is also true of an ontology, since an ontology is a taxonomy plus extra information, however the reverse is not the case.

In addition to taxonomies and ontologies we include also anthologies [10] which are collections of information arranged in a hierarchical order. The following example may help to illustrate the difference.

TAXONOMY: Cats
Broader term: Pets
Narrower term: tabby cat, black cat, kitten
Related term: Dogs

ONTOLOGY: Cats
LivesIn: House
Chases: Mice
Eats: Fish, Rats
Colours: Black, White, Ginger, Tabby

ANTHOLOGY: Cats
History of Cats
    Egyptian Cats [including information]
    Persian Cats [including information]
How to Breed Cats
    Types of Breed [including information]
Looking after Cats

The anthology should be understood as also containing the information for each section along with the headings. The data in each section can take the form of text, as may be found in a textbook, or a media file, video presentation etc., where the content that is stored is useful for teaching or other purposes.

Thus we see taxonomies as a subset of ontologies and ontologies as a subset of anthologies; see Fig. 1.

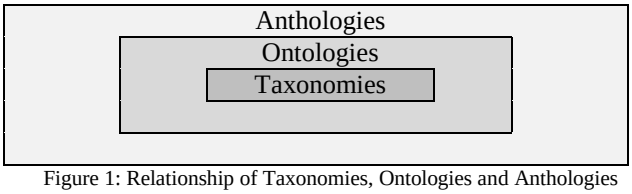


Figure 1: Relationship of Taxonomies, Ontologies and Anthologies

 ISBN: 978-1-61208-195-3

Our discussion of ontologies will apply also to taxonomies and anthologies as the extension is based on the network structure that is common to all. We will for the sake of this paper take as an example an anthology which is a collection of information such as may be seen in the contents page of a textbook which is organized in taxonomic form and contains subject knowledge of a richness which brings it into the definition of an ontology. The addition of information of the subject domain will further extend this to an anthology which contains content that is suitable for teaching. For the sake of convenience and familiarity we will refer to ontologies throughout the remainder of this paper but the reader should understand that our example and procedures are applicable to taxonomies, ontologies and anthologies equally.

We will consequently seek to develop an approach to multi-ontologies and suggest a way in which they can be connected together in to a larger multi connected ontology. We will consider four stages in the systematization of any knowledge system.

Stage 1	Introducing Order
Stage 2	Introducing Coherence
Stage 3	Introducing Proximity
Stage 4	Introducing Co-Requisites and Pre-Requisites

These four stages will lead to a larger concept for ontologies that encompass the present understanding of ontologies as a subset.

III. STAGE 1 INTRODUCING ORDER

Ontologies specify the structure and relationships within a body of knowledge. Usually ontologies are represented as knowledge hierarchies with the most general concepts at the top and more detailed and specific concepts at lower levels. [11] Thus, a body of knowledge is divided into sections, sub-sections, sub-sub-sections etc.

The structure of these knowledge hierarchies is naturally representable as networks, where each node on the network represents a unit of knowledge and where the relationship of each part to every other is determined and specified within the ontology. An ontology can be represented as a tree network where there is a maximum of one path between any two nodes. [12] [13]

We may adopt an addressing system which corresponds to this knowledge hierarchy where each address is correspondingly specified by sections, sub-sections, sub-sub-sections, etc.; see Fig. 2

The advantage of the simple tree model is that the number of hops from the root provides the level of the node. The disadvantage is that the structure does not contain the ordering of the concepts.

However, because of the hierarchical nature of sections, subsection etc, the ontology has an implicit ordering. The structure of an ontology is built up from fragments of knowledge which have an order determined by their pre-requisites. Consequently, the simple tree depicted in Fig. 2 is not sufficient to model this structure as it lacks the necessary order. We use an ordered tree for this description where the branches from each node are ordered so that the sub-nodes have an order of preference. [14]

Principle 1: Simple ontologies are to be represented mathematically as an ordered tree

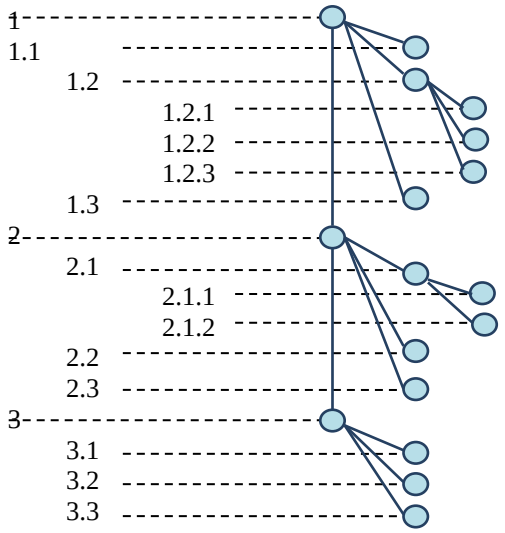


Figure 2: Knowledge hierarchy corresponding to an unordered tree

The ordered tree network is distinguished by

1. there is a maximum of one route from any node to any other node
2. Branches from any given node have an implicit order.

These two properties ensure that the ordered tree network has the necessary properties to represent simple knowledge categorisation and sub-categorisation within an ontology. This structure will also enable a wide variety of knowledge maps to be represented. [15]

IV. STAGE 2 INTRODUCING COHERENCE

We start with the recognition that no one ontology is the correct or the ultimate expression of a subject domain and accept that there are many different ontologies which all adequately represent the knowledge area from different points of view. This is a significant departure from the present understanding of ontologies and we therefore present it as our next principle.

Principle 2: The same structure can be analysed in different ways if it is complex enough

 ISBN: 978-1-61208-195-3

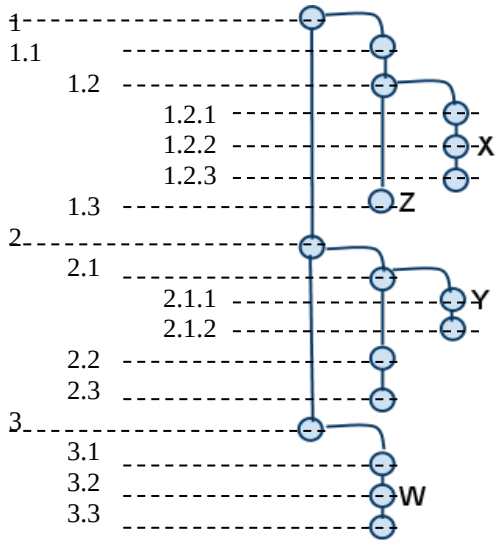


Figure 3: Knowledge hierarchy corresponding to an ordered tree

We next recognise that all these different ontologies are ordered trees which mathematically can be combined into a more complex network containing each of them as a sub network. We therefore introduce a multi connected ontology represented by a mesh network containing multi-connected pathways between nodes. This extends the model for the ontology from that of a tree to a mesh network.

Using Bohm's terminology we would say that the multi connected ontology is the implicate order while a particular decomposition ontology is the explicate order. [7] That means that starting from a larger mesh network we can generate an ordered tree by breaking certain connections in the complex structure effectively decomposing it into a simpler ordered tree. In this way the implicate order of the multi connected ontology becomes the explicate order of the simple ontology or the ordered tree. The breaking of different links in the multi connected ontology will produce a different ontology. [16]

Principle 3: A multi connected ontology can be decomposed into at least one simple ordered tree

Principle 4: Different decompositions produce different but equally valid ontologies

In this way you can unloose or break certain connections in a full multi-connected network which will lead to one decomposition that produces a certain ontology, while another method of breaking connections will lead to another decomposition and a different ontology of the same reality.

Links can be variable because different items of knowledge can be linked together in different ways. What doesn't vary is the items of knowledge themselves. The content of the knowledge must remain invariant but one item can precede another or follow another depending on presentation and other factors.

At a lower level each knowledge item or ontology node may be explained in many different ways. For instance binary arithmetic can be introduced in a variety of ways, but whichever way is chosen it is still teaching the same thing. That is because the content has not varied and the content is determined by the nodes. What is determined by the links is the presentation. Links within ontologies represent the way of explaining the knowledge or packaging the knowledge for student consumption or opening up the subject. All this information is contained in the links. One tutor may adopt a different approach to another by which we mean he will present the nodes in a different order. So each node has a different number of presentations but the content is the same. This is the basic difference between knowledge and education. Knowledge of a subject is the acquisition of a node but the node can be delivered in many different ways and the delivery is concerned with education.

The same relation exists between teaching and learning. Learning is fixed on the acquisition of knowledge nodes, while teaching is involved in the arrangement of the knowledge nodes in a form the tutor presents them. Each tutor may be different and present the knowledge in a different way – yet they are all teaching the same knowledge.

Each presentation may be different and based on different learning styles or learning needs. There may be different degrees of information required where the weak student needs a lot of information and the strong student needs very little. This will define the difference between weak and strong in the student model.

Decompositions

We can formally express decompositions using the adjacency matrix. Let M_{ij} be the adjacency matrix of the multi-connected ontology and let O_{ij} be the particular decomposition tree ontology. Then

$$X_{ij} M_{ij} = O_{ij}$$

where X_{ij} is the decomposition operator. In effect X_{ij} takes the multi-connected M_{ij} into a specific tree O_{ij} which represents the structure and organisation of the knowledge as presented by a particular tutor for a particular student at a particular time with a particular level of subject knowledge. X_{ij} is thus a function of all these parameters.

X_{ij} exists only if M_{ij}^{-1} exists since:

$$X_{ij} M_{ij} M_{ij}^{-1} = O_{ij} M_{ij}^{-1}$$
$$X_{ij} = O_{ij} M_{ij}^{-1}$$

Maximally connected networks (where all nodes are connected to all other nodes) have a simple adjacency matrix K_{ij} in which every component is equal to 1 except for the diagonal components which are equal to 0.

 ISBN: 978-1-61208-195-3

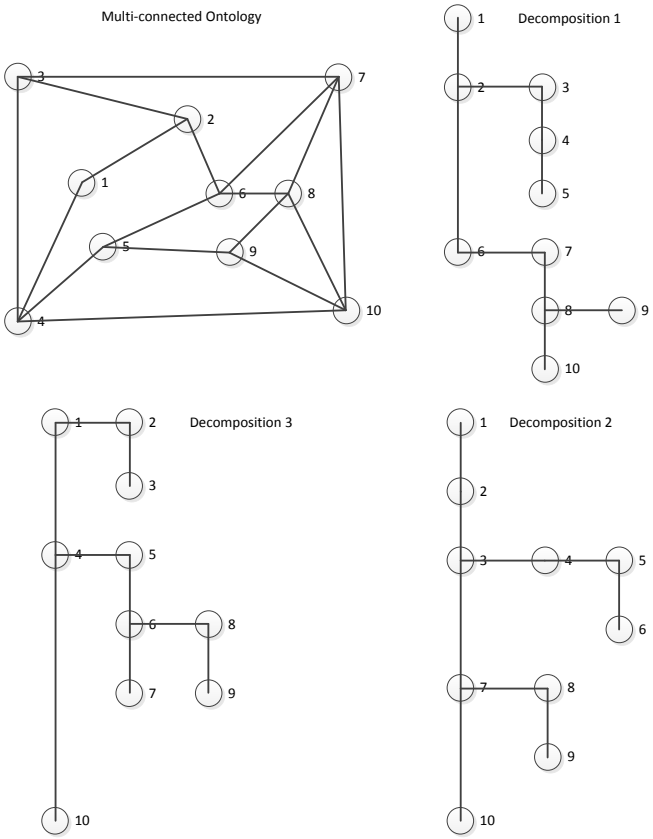


Figure 4: Ten node network decomposition example

$K_{ij} = 1$ (for $i \neq j$) and
$K_{ij} = 0$ (for $i = j$)

All K_{ij} of dimension n have an inverse K_{ij}^{-1} which is given by

$K_{ij}^{-1} = 1/(n-1)$ (for $i \neq j$) and
$K_{ij}^{-1} = -(n-2)/(n-1)$ (for $i = j$)

The existence of the inverse means that every decomposable ontology can be generated from the maximally connected network. The inverse of X_{ij} will be X_{ij}^{-1} which will restore the global multi-connected network from the specific tree

$$X_{ij}^{-1} X_{ij} M_{ij} = X_{ij}^{-1} O_{ij}$$
$$M_{ij} = X_{ij}^{-1} O_{ij}$$

The existence of the inverse is important because it means that given a particular knowledge decomposition we can always get to any other knowledge decomposition via the multi-connected ontology.

This may be clearer if we take a particular decomposition as an example. Consider the ten node network shown in Fig. 4. The multi connected ontology can be decomposed in a number of ways, three of which are illustrated. For clarification, we have numbered the ten nodes consecutively from 1 to 10 and the Adjacency matrix M_{ij} is shown in Fig. 5

	1		1						
1		1			1				
	1		1			1			
1		1		1					1
			1		1			1	
	1			1		1	1		
		1			1		1		1
					1	1		1	1
				1			1		1
			1			1	1	1	

Figure 5: Adjacency matrix Mij

The adjacency matrix M_{ij} has an inverse M_{ij}^{-1} which takes the form shown in Fig. 6

1	3/2	-1/2	-1/2	-1	-1/2	-1	1/2	1	1/2
3/2	0	-1	0	-1/2	1/2	0	1/2	-1/2	0
-1/2	-1	0	1	1/2	1/2	1	-1/2	-3/2	0
-1/2	0	1	0	1/2	-1/2	0	-1/2	1/2	0
-1	-1/2	1/2	1/2	1	1/2	0	-1/2	0	-1/2
-1/2	1/2	1/2	-1/2	1/2	0	0	0	1/2	-1/2
-1	0	1	0	0	0	0	0	0	0
1/2	1/2	-1/2	-1/2	-1/2	0	0	0	1/2	1/2
1	-1/2	-3/2	1/2	0	1/2	0	1/2	-1	1/2
1/2	0	0	0	-1/2	-1/2	0	1/2	1/2	0

Figure 6: Inverse adjacency matrix M_{ij}^{-1}

The adjacency matrix of decomposition 1 $O_{ij}(1)$ is a particular instance of an ontology of the multi-connected ontology M_{ij} given by Fig. 7 where the grey boxes indicate components of M_{ij} which are to be removed.

	1		1						
1		1			1				
	1		1			1			
1		1		1					1
			1		1			1	
	1			1		1	1		
		1			1		1		1
					1	1		1	1
				1			1		1
			1			1	1	1	

Figure 7: The adjacency matrix of decomposition 1 $O_{ij}(1)$

The adjacency matrix of decomposition 2 $O_{ij}(2)$which is another particular instance of an ontology of the multi-connected M_{ij} is given by Fig. 8.

	1		1						
1		1			1				
	1		1			1			
1		1		1					1
			1		1			1	
	1			1		1	1		
		1			1		1		1
					1	1		1	1
				1			1		1
			1			1	1	1	

Figure 8: The adjacency matrix of decomposition 2 $O_{ij}(2)$

The adjacency matrix of decomposition 3 $O_{ij}(3)$ which is another particular instance of an ontology of the multi-connected ontology M_{ij} is given by Fig. 9.

	1		1						
1		1			1				
	1		1			1			
1		1		1					1
			1		1			1	
	1			1		1	1		
		1			1		1		1
					1	1		1	1
				1			1		1
			1			1	1	1	

Figure 9: The adjacency matrix of decomposition 3 $O_{ij}(3)$

It follows from the preceding that not only can multi connected ontologies be decomposed into 'instance ontologies' as we may call a standard ontology but conversely a multi connected ontology can be constructed from instance ontologies and they can be combined into a mesh network. The converse of Principle 3 follows.

Principle 5: A multi connected ontology can be constructed from simple instance ontologies

Thus the three ontologies in Fig. 4 can be constructed from the larger mesh network by the selection of the correct links. However, there must be the same nodes for this to work.

Principle 6 for two ontologies to be identical they must have identical nodes, though not necessary identical links

It is quite easy to prove that any two trees of equal number of nodes but different links could be combined into a single multi-connected ontology. Consider the adjacency matrix of the two complementary trees, call them A and B, then it is always possible to form a new adjacency matrix C such that

$$C = A \oplus B$$

where we have defined $\oplus$ as the operator which adds two matrix elements together according to the rule:

$M_{ij} \oplus N_{ij} = 1$ (if M_{ij} and/or $N_{ij} = 1$)
$M_{ij} \oplus N_{ij} = 0$ (if M_{ij} and $N_{ij} = 0$)

This C will be representable as a new network, which is not a tree.

Indeed we can go further and state that a full maximal multi-connected system of n nodes K_n where every node is connected to every other node can be decomposed into any tree structure of n nodes T_n and that all T_n are subsets of K_n

$$T_n \subseteq K_n$$

In general, every ontology could be decomposed from the maximal multi connected network.

However, we need to be aware that some systems do not yield to this simple analysis as they are not based on different links but on different nodes.

Principle 6 is the fundamental principle that puts a difference between what we are doing and what is being done elsewhere as the structure of an ontology is usually rigidly defined not only by its nodes but also by the fixed links that relate those nodes, so that if the links change then the ontology changes. In Principle 6, we are saying that this is not necessarily so, or that the two ontologies are equivalent, even though they may have different structures. However, the number of nodes must be the same in all cases as they represent knowledge elements and ontologies with different knowledge elements contain different knowledge areas. This is worth restating again.

Two ontologies are the same if they have the same nodes but not necessarily the same links

V. STAGE 3 INTRODUCING PROXIMITY

There are many ways to arrange the nodes of a subject ontology. For instance, if we take the example of computing as a subject area, the knowledge nodes can be arranged in thematic order, logical order, functional order, historical order, geographical order etc. There is no end to the number of ways that knowledge nodes can be linked and presented, other than the mathematical limit of the total number of ways of arranging a finite number of nodes which is n!/2 since the number of ways of arranging n distinguishable objects is n! and we are treating reverse orders as the same arrangement for tree networks.

One way of doing this is to make each of these decompositions dependent on a set of decomposition parameters which determines the ordering. To do this each subject node would need to be tagged with these meta-subject parameters so that each node carries with it the information about its order in history or geography or function etc. However this is not needed if we use the decomposition operator X_{ij} as all the information as to the structure will reside here. Thus, there will be decomposition operators which will represent the different structures. We could speak of a Historical decomposition $X_{ij}(H)$ or a geographical decomposition $X_{ij}(G)$ etc. We can generalise this to $X_{ij}(k)$ In reality there will be a maximum of n!/2 such possible decompositions for a subject area with n nodes.

These decompositions are individually constructed (as are ontologies themselves) by individual subject experts who may be expected to provide their own decompositions very much as different experts would produce different books with different contents structures even though they were writing on the same subject as another expert. Each expert arranges his material in his own way and in a way that suits him and his way of thinking and presenting information. [17] We may speak therefore of individual tutor or expert decompositions

 ISBN: 978-1-61208-195-3

$X_{ij}(E_k)$ corresponding to their understanding of how the subject information should be arranged and presented. Hence

$$X_{ij}(E_k)\ M_{ij} = O_{ij}(E_k)$$

where $O_{ij}(E_k)$ is the ontology produced by Expert E_k

A full determination of E_k will require a tutor model with a full set of identified parameters. Similarly there will be a preferred decomposition for a particular student S who will have his own level of pre-existing knowledge, speed of acquisition of new knowledge etc. The full determination of this will require a student model with a full set of defined parameters. The full details of the tutor model, student model and other models will be dealt with in a separate paper.

The consequence of moving from a tree to a mesh network is that we now have more than one route between any two nodes. Hence within the multi connected ontology M_{ij} there are multiple routes between nodes and not all paths will be equal. Some paths will be very common and chosen by a majority of experts. Some paths may be much rarer and chose by perhaps only one expert. The accumulated frequency of choice may be interpreted as a probability value which indicates the likelihood of one node being linked to another by the creators of the separate ontologies for each decomposable ontology created by an individual expert or tutor.

Consequently, some subject nodes will have a higher probably of transition within the domain than other subject nodes and can be thought of as being 'closer' to each other for that reason. If there is more than one route away from a subject node then each pathway will be weighted according to the probability that an expert may move from one to another. We will model this by introducing probabilities into the adjacency matrix by replacing the 1s with probability values between 0 and 1 where 0 indicates no probability of a transition between two nodes and 1 indicates a 100% probability which means that one node must lead to another.

In this way, the adjacency matrix from Fig. 5 would be transformed, by way of illustration, to Fig. 10.

	.2		.8						
.2		.2			.6				
	.2		.1			1			
.8		.1		.05					,05
			.05		.2			.2	
	.6			.2		1	.4		
		1			1				.1
				.2			.2		.5
			.05			1	.5		

Figure 10: Probability Adjacency Matrix

The problem of finding a suitable pathway through the multi connected ontology which maximises the probability of transition then reduces to a travelling salesman type problem.

VI. Stage 4 Introducing Co-Requisites and Pre-Requisites

Pre-requisite knowledge domains indicate that one area of subject knowledge must be taught prior to another. This is a consequence of knowledge building on previous knowledge [18]. Therefore, within our model a mechanism is required to show which subject knowledge nodes are prior to other nodes. [19] [20] Pre-requisites mean that one subject node must come before another.

	.2		.8						
		.2			.6				
	0		0			1			
.8		.1		0					,05
			.05		.2			.2	
	.6			.2		1	.4		
		1			1		.4		.1
					0	.4		.2	1
				.2			.2		.5
			.05			0	0	0	

Figure 11: Directed Probability Adjacency Matrix

The concept of pre-requisites introduces the notion of direction into the ontology network. Directed networks only allow one route between two knowledge nodes and are usually represented by arrows. To represent this in our adjacency matrix we will introduce asymmetry into the adjacency matrix to show that connections are just one way. In this way, the adjacency matrix from Fig. 5 would be transformed, by way of illustration to Fig. 11 and the consequent directed network is shown in Fig. 12.

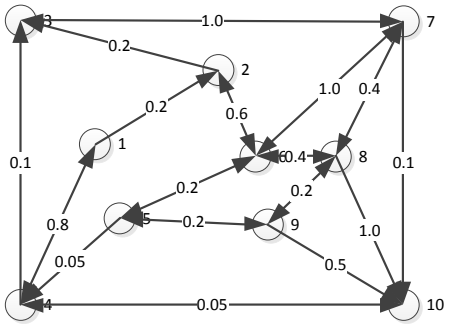


Figure 12: Directed Probability Network

Fig. 12 is our final representation of the multi connected ontology which serves as the complete representation of all the candidate ontologies proposed to represent a given knowledge domain.

VII. Conclusion

We have extended the concept of ontology to include multiple representations of a knowledge domain. The full representation requires an ordered multi-connected

network described by an asymmetric probability adjacency matrix

The ideas presented here provide the underpinning structure for combining all possible ontologies into a single multi-connected ontology of directed probabilistic networks. This solves the problem of how to choose between competing ontologies which have been proposed in any given subject domain. There is now no need to choose between competing candidate ontologies but instead all can be embraced in a single structure.

We liken this to the analogy of many textbooks written on a single subject. Different authors have different approaches to a subject and consequently structure the knowledge and its presentation in different ways which seem suitable to them. Consequently a look at half a dozen different textbooks on any subject will show half a dozen different structures to the knowledge as illustrated by the different contents pages. No two will be alike and yet they will be covering the same subject area.

Does this mean that one is right and all the others are wrong? Not at all. We recognized that there are different and equally valid ways of organizing and presenting knowledge. And consequently there are different ways of organizing and presenting an ontology for a given subject. The problem has always been which ontology should be adopted. The standard approach is to agree on one or decide by international committee. However this does not stop arguments raging as to which should gain universal acceptance. These arguments will never be settled satisfactory as settling them is merely a matter of convention. However by adopting the multi-connected ontology approach we may combine all structures into a single complex network of which the individual ontologies are mere projections.

We contend that this multi-connected ontology has greater claim to being a true representation of the knowledge domain than any individual ontology as it captures all these structures without giving undue prominence to any individual part or favour to any one view. As Goethe says, "*Only everybody knows the Truth."*

VIII. REFERENCES

[1] T. Gruber, "A Translation Approach to Portable Ontology Specifications," *Knowledge Acquisition, ,* vol. 5, no. 2, pp. 199-220, 1993.

[2] C. Price and K. Spackman, "SNOMED clinical terms," *BJHC&IM-British Journal of Healthcare Computing & Information Management,* vol. 17, no. 3, pp. 27-31, 2000.

[3] N. F. Noy and D. L. McGuinness, "Ontology Development 101: A Guide to Creating Your First Ontology," [Online]. Available: http://protege.stanford.edu/publications/ontology_development/ontology101-noy-mcguinness.html. [Accessed December, 2011].

[4] Protégé, "Protégé Ontology Editor," [Online]. Available: http://protege.stanford.edu/. [Accessed January, 2012].

[5] W. Blatner, Heidegger's Being and Time, 2006.

[6] R. M. Pirsig, Zen and the Art of Motorcycle Maintenance: An Inquiry into Values, United States: William Morrow & Company, 1974.

[7] D. Bohm, Wholeness and the Implicate Order, London: Routledge, 1980.

[8] D. L. McGuinness, "Ontologies come of age.," in *Spinning the semantic web: bringing the world wide web to its full potential*, MIT press, 2002.

[9] R. v. Rees, "Clarity in thr usage of the terms ontology, taxonomy and classification," January 2012. [Online]. Available: http://itc.scix.net/data/works/att/w78-2003-432.content.pdf. [Accessed January, 2012].

[10] P. G. Davies and A. R. Davies, "Taxonomies, Ontologies and Anthologies," *(pre-print),* 2012.

[11] A. Bhattachary, A. Bhowmick and A. K. Singh, "Finding Top-k Similar Pairs of Objects Annotated with Terms from an Ontology," *arXiv:1001.2625v2 [cs.DB],* 6 Mar 2010.

[12] S. Cutts, P. G. Davies, D. Newell and N. Rowe, "Requirements for an Adaptive Multimedia Presentation System with Contextual Supplemental Support Media," in *Proceedings of the MMEDIA 2009 Conference*, Colmar, France, 2009.

[13] D. Newell, P. G. Davies, S. Atfield-Cutts and N. Rowe, "Development of a Data Model for an Adaptive Multimedia Presentation System," in *Proceedings of The Third International Conferences on Advances in Multimedia*, Budapest, Hungary, 2011.

[14] M. E. J. Newman, Networks, An Introduction, Oxford: Oxford University Press, 2010.

[15] P. G. Davies, D. Newell, S. Atfield-Cutts and N. Rowe, "An Adaptive Multimedia Presentation System," *International Journal On Advances in Software,* vol. 4, no. 1&2, 2011.

[16] R. McGreal, Ed., Online Education Using Learning Objects., London: Routledge, 2004, pp. 59-70.

[17] P. Chen and L. Y. Wong, Eds., Active Conceptual Modelling of Learning: Next Generation Learning-Base System Development, Springer, 2007.

[18] T. Boyle, "Design Principles for Authoring Dynamic, Reusable Learning Objects," *Australian Journal of Educational Technology*, 2003.

[19] N. Rowe and P. G. Davies, "The Anatomy of an Adaptive Multimedia System," in *Proceedings of The Third International Conferences on Advances in Multimedia*, Budapest, Hungary, 2011.

[20] N. Rowe, S. Cutts, P. G. Davies and D. Newell, "Implementation and Evaluation of an Adaptive Multimedia Presentation System (AMPS) with Contextual Supplemental Support Media.," in *Proceedings of the MMEDIA 2010 Conference*, Athens, Greece., 2010.

 ISBN: 978-1-61208-195-3

Adaptive Virtualisation for Multi Modal Learning Objects

David Newell
Software Systems Research Group
Bournemouth University
Bournemouth, UK
dnewell@bournemouth.ac.uk

Philip Davies
Higher Education
Bournemouth and Poole College
Bournemouth, UK
pdavies@bpc.ac.uk

Suzy Atfield-Cutts
Software Systems Research Group
Bournemouth University
Bournemouth, UK
satfieldcutts@bournemouth.ac.uk

Andrew Yearp
Graduate School
Bournemouth University
Bournemouth, UK
ayearp@bournemouth.ac.uk

Abstract - **Work by the writers has investigated validation methods for creation and manipulation of multi modal learning objects in an adaptive Virtual Learning Environment (VLE) presentation system. This paper investigates the requirements for a robust, autonomous, virtual infrastructure needed to simulate novel adaptive methods based on fragmentation and routing algorithms like OSPF. Evaluation is done of virtualised processes adapted on a software router in a known infrastructure. Adaption is achieved with operations performed on the metadata of learning object fragments rather than link states. Execution of such models in a 'Semantic Ontology Engine' is proposed as an approach to the creation of a cloud computing based semantic multimedia VLE, offering better personalisation. The findings emerge by means of a comparison of simulation results of virtual network components.**

Keywords – ***e-learning, adaptive, semantic, ontology.***

I. INTRODUCTION

Previously, an Adaptive Multimedia Presentation System (AMPS) has been proposed with semi-automated tools for adapting stored computer based learning objects to students' learning needs [1]. It was concluded that a novel, autonomous 'Semantic Ontology Engine' is needed as a key building block to process learning objects by performing decomposition, fragmentation and re-composition. However, a very important research question remained unanswered - how to approach the validation of multimedia structures built by autonomous semantic processes in a VLE, without the services of a human tutor to evaluate 'true' fragments of learning.

An experimental approach will be taken to verify the efficacy of the required semantic ontology function. The operational approach employed in this paper starts with a survey of various pre-existing virtual network simulation tools that are expected to offer at a partial solution to the problematic evaluation and verification of metadata models that satisfy these complex requirements [2]. The resulting tool promises to be an experimental virtual infrastructure capable of executing multiple, proposed semantic models of ontology engines, each capable of manipulating and validating learning objects in a Cloud-based Adaptive Virtual Environment (CAVE) potentially without a human tutor.

Hence, in this paper steps in our research programme are set out to provide robust evaluation of a suitable model for the semantic ontology engine based on an analogy with network routing protocols. In se ction II, a comparison between computer networking routing concepts and the requirements for an ontology mapping based on an ontology calculus is set out. In section III, features of some virtual simulation tools are compared in detail. These are commercial products or open source from educational institutions, with a mixture of local and remotely accessible options. Although the review is far from exhaustive, it includes some well-known and recently introduced packages. After the review of features, one tool is selected for comparison with an actual physical network; Section IV gives the results of this comparison for two scenarios; Section V is an analysis of findings. Finally, Section VI gives conclusions about applications of virtualised networks simulators to learning objects.

II. A COMPARISON OF ROUTING CONCEPTS AND AN ONTOLOGY MAPPING

One of the primary functions of the ontology engine will be to retrieve the learning objects for delivery to the student, in the sequence in which they will be presented. There is unresolved discussion about the most appropriate method to achieve this. One approach is the object oriented modelling approach of Lee & Chung [21]. We propose a new approach based on concepts which are already successfully used in computer networking. It suggests that the concepts used in the selection of the 'best path' determined by router network devices in a computer network between two nodes carrying traffic on a digital network may be used as an appropriate analogy for learning object retrieval from an ontology network.

A. Pathway Determination

A key feature of an adaptive learning delivery system is a process for the selection of learning materials appropriate to the required learning, and suitable for the learning level and style of the individual student. In computer networking, the selection of the best path for traffic delivery is made according to metrics such as 'hop count' and 'bandwidth'. This process is successful at delivering electronic data worldwide and operating at optimum speed within the constraints of the hardware available, whatever that may be. The hop count is the

 ISBN: 978-1-61208-195-3

number of 'hops' to other routers required to reach the destination. As distance is measured in terms of hops, rather than physical distance, the shortest distance is that with the lowest hop count. Bandwidth is the data capacity of a link defined in terms of bits per second that can be transmitted over the medium. Both are useful indicators of speed of delivery. In a virtual learning environment similar metrics can be applied such as the distance attribute described in our developing ontology calculus. In networking, the selection of the path taken by data is determined by a device which connects separate networks together known as a router. This device makes decisions about routes for each packet of data it receives, and that decision making process is completed in fractions of a second. The high speed is made possible by the narrowing of options. Rather than determining the whole path at the beginning of the journey, only the next hop is selected. At each router the options are narrowed only to the other networks which are physically connected.

In our model, the learning objects are likened to the nodes of a network that needs to be traversed by the student who is seeking to learn a particular subject domain. Learning objects are like the routers of a network. Though they have no physical connections they are connected logically through the ontology. In the same way as a network can be mapped, an ontology provides a map of the relationships between topics. A model of this is described in Davies et al. [22]. Rather than sifting through all available learning object segments for related material, using the metadata in the learning objects, their position within an ontology can be determined at the point of implementation. When required, the selection of learning objects for presentation can be narrowed down to other closely related material. Where selection by searching all materials may add a significant time delay, searching only closely related materials should be relatively fast.

B. Delivery Methods

Once materials have been selected, the next stage is delivery to students. E-learning should be extended so that it is deliverable anywhere and everywhere. This is called ubiquitous learning or u-learning. Delivery methods must take into account the destination client device when presenting learning objects for delivery in a virtual environment, for example, a pc or a mobile phone.

In computer networks, routers handle packets containing data. The packets are conceptually an outer wrapper, allowing packets to be unwrapped and rewrapped with new addressing information without disturbing the data itself. In fact, there are many layers wrapped around the data in a networking scenario. Each layer contains different pieces of additional data and at several different layers there may be different kinds of addressing information. Since only the hop to the next router is determined when selecting a route through the network during transit, the outer layer is removed at a router and the data rewrapped with the address of the next destination device. The address of the final destination is kept at another layer undisturbed by this process.

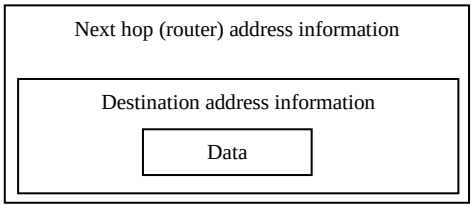


Figure 1. Model of a frame and some conceptual layers in computer networking

The outermost layer contains addressing information and its format is determined by the media on which the frame has to travel. Similarly, in a course delivery system, a wrapper around the learning object would determine to whom it will be presented, when it will be presented and in which order it will be presented. If a learning object is to be presented to a particular student then the student signature represents the address to which that learning object is to be delivered.

TABLE 1. ASSOCIATIONS BETWEEN GRAPH THEORY, NETWORKING AND ONTOLOGY DESCRIPTIONS

Graph Theory	Networking	Ontology
Node	Router	Learning Object
Link	Connection	Relationship
Node location	IP address	Learning object identifier
Algorithm	Protocol	Order of presentation
n/a	Wrapper	Student Signature + other determinants (metadata)

C. Delivery format

The format of the information is determined by the destination client platform. If the page is to be displayed on a pc then a full size web page constructed of html, xml and other web technologies is wrapped around the learning object. If the student is learning on a mobile phone then suitable technologies are required to display a page to suit the small screen size and these will wrap around the learning object before it is sent to the student's learning platform of choice. Connection speeds may also be a metric for changing what is sent.

Connection speeds may go so far as to affect the learning object itself. There is little point in trying to download a high resolution image of great size down a slow connection to a small phone screen. Perhaps enhanced versions (e.g., HD or 3D images) would benefit the student using a larger screen. Therefore, each learning object may be required to consist of different versions of the media file.

Therefore, as when using the Transport Control Protocol in digital networking, an initial exchange of

 ISBN: 978-1-61208-195-3

information between the client and server devices to request, and then supply client platform specification in terms of both hardware and software must take place.

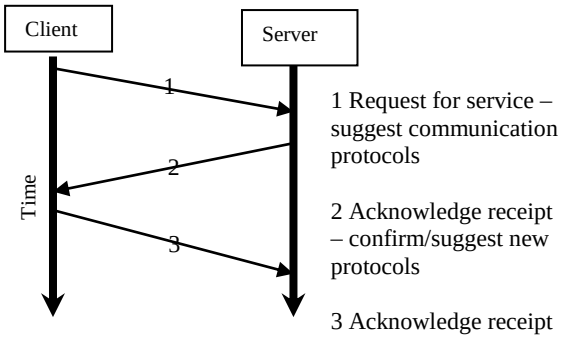


Figure 2. TCP 3 way handshake at start of communication session

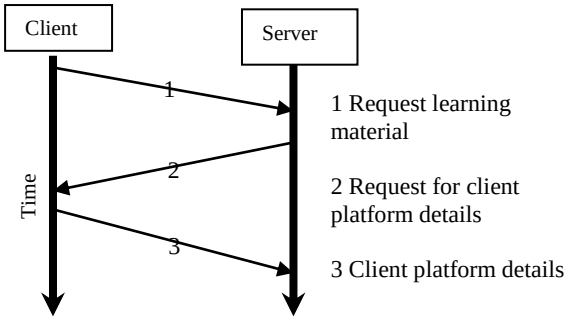


Figure 3. System 3 way handshake at start of learning session

Possibilities for required different formats are enormous and ever changing, such as different resolutions for images and video, different compression rates for audio, and different formatting for text e.g., transforming a document from a word processor into html to improve compatibility. Generating these additional files for each object impacts the authorship work load significantly. The high level of investment required for production of quality learning objects has been an issue since they came about as discussed by Boyle [18]. Dynamically generating suitable versions from high quality originals is a preferred option to increasing storage requirements and second guessing possible future platforms.

The investment in authorship workload will mean that writers are keen to reuse a learning object in more than one area and so its upload to the system requires additional consideration. Contextualization of the learning object becomes an important consideration if re-use is high. This will involve the creation of metadata categories to capture the contextualized data. The IEEE 1484.12.1 - 2002 Standard for Learning Object Metadata [3] is an internationally recognized open standard for the description of learning objects. Attributes of learning objects included could be the type of object, author, owner, terms of distribution, format, as well as pedagogical attributes, such as levels of difficulty or student learning styles. A set of these attributes need extension to include context.

Indzhov et al. [19] explain users of such systems are often poor at completing metadata requirements. Being able to position the object in an ontology map of the knowledge domain would aid this process. Ideally, the metadata for a learning object, where possible should be automatically generated. Bauer et al. [17] discuss the possibility of collaborative tagging relying as it does on a large enough, and knowledgeable enough audience to complete the tagging before use of the semantics within the system becomes essential, and so time is required to carry out 'tagging' before the object itself is useable. Automatic metadata generation is a mature development area. For instance, if an object contains images much work has been done in the area of identifying objects in images by many including very recently Amir et al. [16]. As a result others have studied the composition of the resulting information into metadata that can be used with learning objects. Cardinaels et al. [20] developed a indexing interface for automatic meta data generation, and more recently Bauer et al. [17] surveyed the tools available to do the job and compared them.

Metadata can conceptually be perceived as another layer wrapped around the learning object. Indzhov et al. [19] discuss using the results of tests for calibrating the difficulty levels and usefulness of learning objects, as well as the possibility of assessment question generation from metadata. By using metadata as a wrapper on the outer layer of the learning object, it can be read and updated without disturbing the object itself.

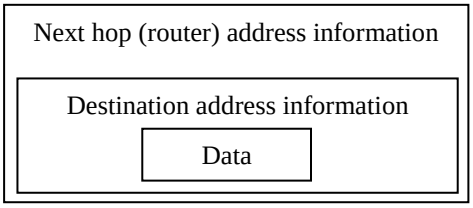


Figure 4. Model of a frame and some conceptual layers in computer networking.

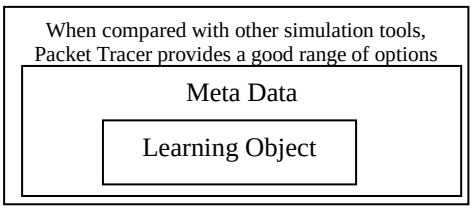


Figure 5. Model of the learning object wrapped in required implementation layers

III. NETWORK SIMULATION AND COMPARISON OF VIRTUAL ROUTERS

We now turn to a network analogy in more detail and consider a closer examination of networking simulation tools provides insights into tools useful for modelling an

 ISBN: 978-1-61208-195-3

ontology engine to process ontology networks and determine the validity of learning pathways.

Due to its nature, discrete event is a method of simulation suitable for modelling systems where processes act on discrete units, for example a data packet in a communications network, a job on a production line. This type of layered operation is important in most types of data communications and networked system. It has been acknowledged that networks such as these are complex in their design and operation. As such, simulation is an important tool for designing and operating these networks.

For modelling computer networks discrete event simulation is the popular choice although other techniques are also used. There are many simulation tools for this task. For this reason there have been many papers written that have reviewed and compared these tools and packages. Most of these papers have studied the tools from the point of view of their usage and suitability for different tasks. This paper intends to look more closely at how some of these tools accomplish what they do, with a view to adapting network simulation techniques to adaptive learning techniques.

Simulation tool packages are well represented in the literature and classified into four main branches shown in Figure 6.

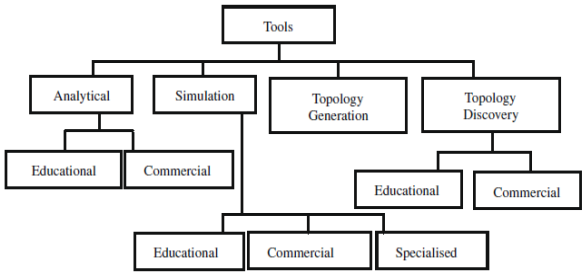


Figure 6. Tool Packages

These categories are quite broad. This paper is concerned with tools for educational and learning activities in VLE. It has been acknowledged by a number of sources that learning the skills required to design and manage computer networks requires practical experience in addition to a theoretical base.

The advantages of visualisation compared with providing physical facilities are well known. Use of simulation tools to create virtual lab environments provides an opportunity to increase access at a lower cost than physical equipment, offering the possibility to carry out more complex experiments than would otherwise be possible.

IV. RESULTS OF THIS COMPARISON FOR SPECIFIC SIMULATORS

We now examine two particular systems to illustrate the range of usage and properties available.

D. Packet Tracer

Cisco Systems has produced Packet Tracer [23] as an educational tool for their network academy program to assist students with their studies for qualifications such as CCNA. It provides many features to assist both students and instructors in the field of network design and maintenance/management. Features include the ability for the instructor to create lab scenarios for students to complete, also included within this is the ability to assess the students. Beyond these preconfigured networks and activities, Packet Tracer also allows the creation of any possible topology that can be built using the available pallet of hardware.

When simulating the network that is being studied there are two options for interaction. The first is real-time; in this network reacts as a real-world system would, for example if you ping one device from another this would occur at realistic speed. The second option for simulation, described as simulation mode by Cisco, allows the user to slow down the operation of the network to see the movement of data packets that are visually displayed on the network diagram (Figure 21). The speed of this animation is controllable as is how quickly it moves to each event. This can occur automatically based on the speed or can be made manually, allowing students to see the movement of data packets within the network.

E. OPNET

OPNET [24] is a commercial research and development package developed by OPNET systems that is popular in both research and commercial applications. It provides the ability to model wired and wireless networks and their interactions using a large library of models provided, and also allows the user to modify or create their own. These models are created using C++ programming language and the source code is included for the models provided.

Additionally, the ability to customise the models when running simulations in OPNET it is possible to vary the level of detail of each simulation run dependent on the requirements of the application. To accomplish this, OPNET provides three methods of simulation.

The first option, giving the highest level, of detail uses discrete event driven simulation OPNET implementation of this comes in two forms. The first being sequential were all tasks performed a linear fashion on a single processor, the second form parallel distributes the tasks over multiple processors which can be part of the same system for distributing over multiple interconnected systems. This latter parallel system improves performance

 ISBN: 978-1-61208-195-3

by spreading the work allowing for faster simulations. There are also additional optimisation options provided for the discrete event simulation. The second option, flow analysis(tm) uses analytical modelling to provide a faster but less detailed simulation, ideal for the use with simulation large networks and repetitive scenarios such as modelling the effect of failures on traffic with the network where it is necessary to run many iterations of simulation. The third option is a hybrid of the first two techniques allowing for balance of speed and detail within one simulation.

These methods coupled with the large library of simulation models allow the OPNET user to create networks varying from a small office all the way up to world-wide communication systems.

F. GNS3

GNS3 [25] is an open source package created to allow users to practice configuring Cisco Systems networking devices in a realistic environment without the need to purchase expensive equipment. This has been accomplished by emulating the heart of a number of such devices. In turn this allows the user to run genuine software from the device on a normal computer system. Although the emulation provides a comprehensive set of hardware features it cannot provide the same speed of response times as the real equipment. The biggest drawback for this package is that it does not support many newer devices as these use proprietary integrated circuits that so far, and probably never will be emulated in software. There are moreover also changes occurring in the newer versions of Cisco's software that will change its licensing mechanism, requiring activation beforehand, thus preventing unauthorised installation and use .

V. ANALYSIS OF FINDINGS.

Although the above systems by no means constitute an exhaustive c exploration of the available solutions, it has considered some of the most popular and new options. Each of these tools has its own advantages and in many cases its own niche in the market. It is not possible to make a sweeping conclusion about which tool is best as each tool has its own place and time. For example, for a beginner to networking Packet Tracer is ideal, but for a researcher studying performance of wireless networks OPNET could be the tool of choice.

Table 2 shows a comparison of simulators including tools for which there was insufficient space to discuss in detail here.

VI. CONCLUSIONS AND FUTURE WORK

Investigations into network simulation tools indicate that there is scope to consider the use of routing algorithms for suggesting analogous models for routing learning objects to determine a specific learning pathway to specific students.

To take this work further we need to construct a full, robust tutor model to automate the learning object segmentation process, an investigation of structure of metadata and a detailed construction of the student model to include the student signature which will directly apply the learning-routing algorithm as a wrapper on the learning object. Our vision is to build this into a novel abstract conceptual data model encompassing all the properties that are needed to make explicit the qualities of an effective adaptive learning system. In this event Critical Success Factors (CSFs) would play a central role

TABLE 2. COMPARISON OF NETWORK SIMULATORS

								Supported technologies		
	Remote access	Vendor specific	Hardware based	Simulation based	GUI	Assessment tools	Network topology	Routing	Switching	wireless
Packet Tracer	No	Yes: Cisco systems	No	Yes	Yes	yes	User defined	Yes	Yes	Yes
OP Net	No	No	No	Yes	Yes	No	User defined	Yes	Yes	Yes
GNS 3	No	Yes:Cisco systems & Juniper	No	Yes	Yes	No	User defined	Yes	Partial	No
VELNET	Yes	No	No	Yes	Yes	No	User defined within limits of simulation	Yes	No	No
Remote Access internet work lab	Yes	No	Yes	No	No**	No	User defined but limited to available equipment	Yes	Yes	No
A Virtual network lab for learning ip networking	Yes	No	Yes	No	Yes	No	Pre set	Yes	Yes	No
An integrated structure for virtual networking lab	Yes	Yes :Cisco systems	Yes	Yes	Yes	Yes	Pre set in hardware for remote lab but user defined for local simulation	Yes	Yes	No*
Virtual Network lab(VNL)	Yes	No	Yes	No	?	No	Pre set	Yes	Yes	No
NetLab+	Yes	Yes :Cisco systems	Yes	No	Yes	No	User defined &some pre set avalible	Yes	Yes	No
IOU	Yes	Yes :Cisco systems	No	Yes	No	No	User defined	Yes	No	No
L2IOU	Yes	Yes :Cisco systems	No	Yes	No	No	User defined	Yes***	Yes	No

Supports Layer three switch functions

 ISBN: 978-1-61208-195-3

in determining the choice of the best network software tool needed for the simulation. The introduction of CSFs on which the best network simulation tool will be chosen is left to a future paper.

It is acknowledged that this work is in its preliminary stages. The next step will involve a simulation for specific software tools and simulation in a real environment.

Finally, although work discussed in this paper answered research questions posed in previous papers, it has indicated further questions with a different emphasis: What is the full specification of the ontology required and how is it captured? How should the ontology engine structure be modelled and evaluated? Can fuzzy logic or data mining techniques be candidates for a useful algorithm? And "What further adaptation features are required and how are they to be evaluated?" We leave these questions to a further paper.

VII. REFERENCES

[1] Cutts, S., Davies, P., Newell, D., and Rowe, N., 2009. *Requirements for an Adaptive Multimedia Presentation System with Contextual Supplemental Support Media*, Proceedings of the MMEDIA 2009 Conference, Colmar, France.

[2] Rowe, N., Cutts, S., Davies, P., and Newell, D. 2010 *Implementation and Evaluation of an Adaptive Multimedia Presentation System (AMPS) with Contextual Supplemental Support Media*. Proceedings of the MMEDIA 2010 Conference, Athens, Greece.

[3] IEEE. 2001. *IEEE Learning Technology Standards Committee* (LTSC) IEEE P1484.12 Learning Object Metadata Working Group; WG12 Home page.

[4] Boyle, T., 2003. Design Principles for Authoring Dynamic, Reusable Learning Objects. *Australian Journal of Educational Technology*.

[5] McGreal, R. (Ed.), 2004. *Online Education Using Learning Objects*. London:Routledge, 59-70.

[6] Protégé (2009) Protégé Ontology Editor, Stanford University California, USA. http://protege.stanford.edu/ [Accessed online April 2012]

[7] Gruber, T., "A Translation Approach to Portable Ontology Specifications", Knowledge Acquisition, 5(2), 199-220, 1993.

[8] Newman, M. E. J. "Networks, An Introduction", Oxford University Press, 2010

[9] Codd, E. (1970). 'Data Models in Database Management, 'ACM SIGMOD Record 11, No. 2

[10] Date C. J. (2000). 'WHAT not HOW: The Business Rules Approach to Application Development' Addison-Wesley. And Date, C. (2004). 'Introduction to Database Systems', 8th Ed., Pearson.

[11] Progress (2010) Objectstore, http://documentation.progress.com/output/ostore/7.2.0/pdf/user1/basicug.pdf (Last Accessed Dec 2010)

[12] Lamb, C., Landis, G., Orenstein, J., and Weinreb, D., (1991). 'The Objectstore Database System', *Communications of the ACM* 34 (10): 50–63.

[13] Date, C., Darwen, H., and McGoveran, D. (1998). 'Relational Database Writings 1994-1997', Addison Wesley.

[14] Chen, P. (1976), 'The Entity-Relationship Model-Toward a Unified View of Data' (1976), *ACM Transactions on Database Systems* 1/1/1976, ACM-Press.

[15] Chen, P. (2007). 'Active Conceptual Modelling of Learning: Next Generation Learning-Base System Development', with Leah Y. Wong (Eds.). Springer.

[16] Amir, A., Amin M., Anang H. and Khan, A. I. (2011) *P2P-Based Image Recognition for Component Tracking in a Large Engineering Domain.* In: The Second International Conference on Parallel, Distributed, Grid and Cloud Computing for Engineering 2011 (PARENG 2011), 12-15 April 2011, Ajaccio, Corsica, France.

[17] Bauer, M., Maier, R., and Thalmann, S. 2010. Metadata Generation for Learning Objects An Experimental Comparison of Automatic and Collaborative Solutions *In*: Physica-Verlag HD, *E-Learning 2010*. 181-195. Isbn: 978-3-7908-2355-4

[18] Boyle, T. 2002. Design principles for authoring dynamic, reusable learning objects. In A. Williamson, C. Gunn, A. Young and T. Clear (Eds), *Winds of Change in the Sea of Learning: Proceedings of the 19th Annual Conference of the Australasian Society for Computers in Learning in Tertiary Education*, pp57-64. Auckland, New Zealand: UNITEC Institute of Technology.

[19] Indzhov, H., Totkov, G. and Doneva, R. 2011. E = MA^2 (e-learning in a Moodle-based adaptive and accumulative system). In *Proceedings of the 12th International Conference on Computer Systems and Technologies* (CompSysTech '11), Boris Rachev and Angel Smrikarov (Eds.). ACM, New York, NY, USA, 498-503. DOI=10.1145/2023607.2023691 http://doi.acm.org/10.1145/2023607.2023691 [Accessed online April 2012]

[20] Cardinaels, K., Meire, M. and Duval, E. 2005. Automating metadata generation: the simple indexing

 ISBN: 978-1-61208-195-3

interface. In *Proceedings of the 14th international conference on World Wide Web* (WWW '05). ACM, New York, NY, USA, 548-556. DOI=10.1145/1060745.1060825 http://doi.acm.org/10.1145/1060745.1060825 [Accessed online April 2012]

[21] Lee, M., Chung, Y. 2010 , "Using object-orientation to conceptualize an adaptive learning content management system modeling," *Advanced Computer Control (ICACC), 2010 2nd International Conference on* , vol.3, no., pp.56-60, 27-29 March 2010 doi: 10.1109/ICACC.2010.5486741 URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5486741&isnumber=5486614 [Accessed online April 2012]

[22] Davies, P., Newell, D., Davies, A. and Karagözlü, D. (2012) Multi Connected Ontologies, (Pre-print) arXiv:1112.6090v1 [cs.DL] [Accessed online April 2012]

[23] Cisco Systems, Packet Tracer, http://www.cisco.com/web/learning/netacad/course_catalog/PacketTracer.html [Accessed online April 2012]

[24] Opnet Technologies Inc. http://www.opnet.com/ [Accessed online April 2012]

[25] GNS3, http://www.gns3.net/ [Accessed online April 2012]

 ISBN: 978-1-61208-195-3

Towards Distributing Multimedia Applications on a Virtualized Cloud Infrastructure

Mak Shama
Birmingham City University
Birmingham, UK
mak.shama@bcu.ac.uk

David Newell
Software Systems Research Group
Bournemouth University
Bournemouth, UK
dnewell@bournemouth.ac.uk

Philip Davies
Faculty of Technology
Bournemouth and Poole College
Bournemouth, UK
pdavies@bpc.ac.uk

Benjamin Todd
Postgraduate Research Student
Bournemouth University
Bournemouth, UK
bwtodd@gmail.com

***Abstract*—We examine some technological aspects of cloud computing, focusing on virtualization applied to various data types including multimedia and identify the benefits & security concerns for a modern IT infrastructure. An experiment to migrate a live company server consisting of Microsoft Exchange e-mail and file server to a cloud infrastructure is conducted. The initial findings are that each process step needed to overcome security issues of server migration.. The writers will propose improved approaches for modelling cloud-hosted multimedia applications, semantics and abstract data models.**

Keywords – Multimedia, Virtualization; Cloud Computing; Server Migration; Security.

INTRODUCTION

In a previous paper [1], we evaluated an adaptive multimedia presentation system with contextual supplemental support media. In this paper we will be considering the requirements for operating this adaptive system from a cloud based infrastructure. 'Cloud Computing' is a term for a multitude of online services allowing in-house computing services to migrate to rented online infrastructures. 'Software, Platform or Infrastructure as a Service' are concepts for processing, developing or hosting application and data on demand. Improved speed and reliability of network connections fuel the explosion of portable, hand-held devices by reducing organisational IT costs.

The Virtualization and Evolution to the Cloud Survey [2] found that 75% of global organisations are positively considering migration to a virtualized /hybrid cloud environment. However, 44% of CEOs and 46% of CFOs still have concerns about migrating their applications and services. It was also suggested that once organisations move their critical applications to virtualized or hybrid cloud environments then they find many benefits.

The current ad-hoc methodology to migrate a company server or application to the cloud uses an infrastructure, storage vendor, or a software service provider. The security aspects are built around the vendor, measures built into the customers own systems and how the data centres operates, for example, on a multi-tenant basis [3]. There are additional security measures that need to be implemented over virtual, public cloud networks that are discussed in this paper.

A small survey of ten local SMEs was undertaken by the writers to find out what services they run. Results confirm the industry trend towards virtualization and migration to the cloud raised security concerns for organisations. The main findings are summarized in Fig. 1.

Most businesses run four essential business applications: e-mail, file storage & printing, application serving and databases. It was found that these businesses have a desire to migrate to the cloud but their primary concerns are confidence in security compared with their on-premises solution, and worry about overall control of business data when required. It is clear from these results that companies, including those involved in multimedia data types, need to be shown how their data is secure in the cloud and the many advantages to be gained from migration.

The structure of this paper is as follows: Section 2 gives a brief survey of current methodological approaches to virtualisation. Section 3 reports on a case study of an actual small company migration. Section 4 discusses findings, including use of a novel multimedia application in an educational setting and finally, Section 5 concludes with a discussion of future work.

METHODS AND APPROACHES

Virtualization is a term given to the creation of a virtual computer within a larger more powerful computer. The more powerful computer is often known as a Virtual Machine (VM) host that runs software known as a hypervisor [4].

The hypervisor is a layer of software that controls and monitors the resource allocation of the host hardware - memory, processor and drive space - to the VMs that run on it. The VMs on the hypervisor are called instances; many of these instances can be run on one hypervisor, limited only by the hardware of the host system, which is usually designed to run multiple guest VMs. This has been illustrated in Fig. 2.

 ISBN: 978-1-61208-195-3

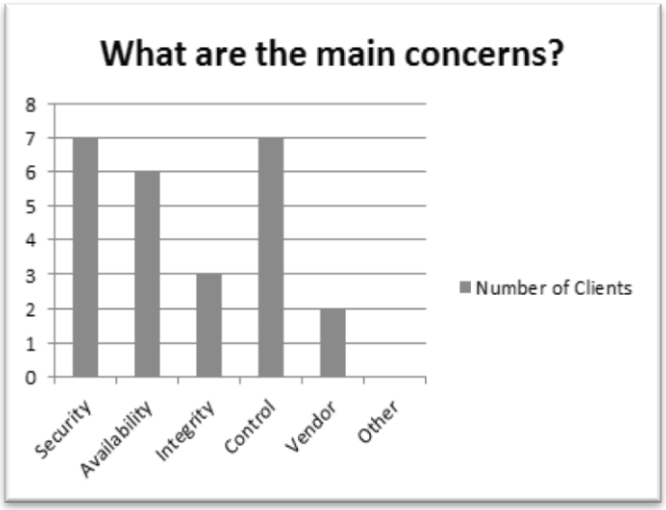


Figure 1. Summary of Survey Results

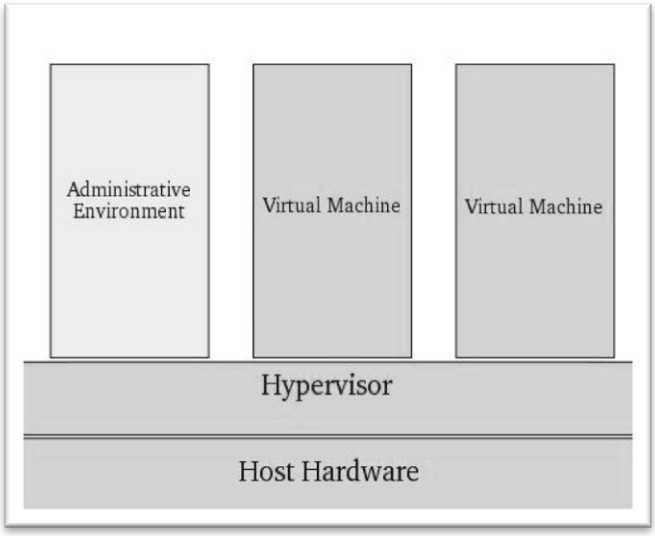


Figure 2. Virtualization Hypervisor Layer (Source: Virtuatopia, 2009)

There are many virtualization platforms available. The most commonly used are compared in Fig. 3. They offer comparable features, run on different specification hardware and are available for use on different physical computer system such as Windows, Linux or Macintosh host server platforms. [5].

Virtualization offers organisations advantages such as reduced operating costs, freeing up resources so users can instantly launch new servers within minutes, improving the flexibility of IT architectures [4].

Modern organisations are moving to a functioning cloud computing model where applications are virtualized and become 'always available' online, rather than the traditional systems, where all IT infrastructures are stored and maintained within a physical location of an organisation.

The general approach to migration to the cloud initially, is for an organisation to migrate to server virtualization followed by other types of virtualization such as storage and desktop/endpoints, and then finally private storage-as-a-service and private/hybrid clouds [6].

Virtualization is an important factor of modern day computing that has three distinct benefits to organisations. Firstly, it offers substantial cost reductions. Using virtualization allows organizations to consolidate their data centres and non-mission critical applications down to less than a quarter of their current data centre requirements, moving servers that are running at 20% capacity to shared servers that run at 70% capacity. Secondly, it offers speed of deployment and scalable resources to an organisation needing to launch IT infrastructure. For example, to go with the launch of a new product, increases in customer demands can easily be achieved in a matter of hours instead of weeks on the old model. Finally, it offers organisations a way to manage expertise and skill sets to easily manage multiple different applications with reduced staff to maintain the network[2].

A small organisation may virtualize just the servers in their office for consolidation, which becomes a small private cloud that they can access internally and externally. Larger organisations can go so far as to create complete virtual networks, using multiple VMs on different cloud infrastructure providers, and create virtual switches and connections with VLANs for their networks. As complexity increases, so may security concerns. With a network that is so complex, there needs to be a way to monitor all VMs that are running correctly [2].

Virtualization Platform	Provider	Host OS
Citrix Zen	XenSource	NetBSD, Linux, Solaris
Virtual Server / Hyper-V	Microsoft	Windows Server
Virtual PC	Microsoft	Windows
Parallels	Parallels	Macintosh
Virtual Box	Sun Microsystems	Windows, Linux, Macintosh, Solaris
VMware / ESX Server	VMware	Windows, Hardware (no host OS)

Figure 3. Comparison of Virtualization Platforms

 ISBN: 978-1-61208-195-3

There are benefits of virtualization in large organisations with thousands of computers otherwise requiring hardware rollouts, patching and maintenance [7]. Each desktop no longer needs individual licences for antivirus or productivity software as this will reside on the main hypervisor desktop virtualization server.

Virtualization is also eco-friendly and saves money in the long run. In the future, organisations will move more towards a virtualized desktop, and some examples of this can already been seen in third-world counties developing Virtual Learning Environments (VLE's) that allow students access to online resources without the need for a powerful local computer [8].

Terminal services such as Citrix have had a multi-user operating system environment for some time. Virtual Desktop Infrastructure (VDI) works in a very different way. Instead of having a server which can be used by multiple users at the same time, the server is running many virtualized single user operating systems which all act independently of each desktop system due to the function of the hypervisor [7].

An example of desktop virtualization using VDI is shown in Fig. 4. It makes use of a thin client at each user work desk. The virtual desktops are stored in a secure environment on the server. Thin clients use less energy, need less maintenance, and updates can all be performed much quicker on the VDI server. VDI is designed to allow the use of desktop operating systems such as Windows as well as Linux in a virtual environment that is easy to deploy, secure and manage with everything stored in the data centre [7].

Virtualisation can be extended to cloud computing. [9] Broadly, there are three levels of service that cloud computing provide:

A. Software as a Service (SaaS)

Applications are available online. For example, e-mail, online file storage, web hosting image banks, online software to perform virtually any task on multimedia data and more. Google Docs, or Sales-Force.com are current examples. There is not much control other than over data, but with no control over routers, firewalls, IPS or WAFs.

B. Platform as a Service (PaaS)

Application platform used for development of online services. For example, client relations management database, online developer tools, web site creation services and more. Combinations of different components managed by someone else allow use of databases or application services. Microsoft Azure is an example.[10]

C. Infrastructure as a Service (IaaS)

Provision of an online infrastructure. For example, online server hosting, virtual servers, direct access to physical servers in the cloud, web hosting and data centre. Amazon EC2 and Savvis (enterprise cloud) are examples that give users a lot more control. Raw virtualization with a good service layer could move a WAF or Gateway to this service. Users themselves start to assert a level of control of the cloud.

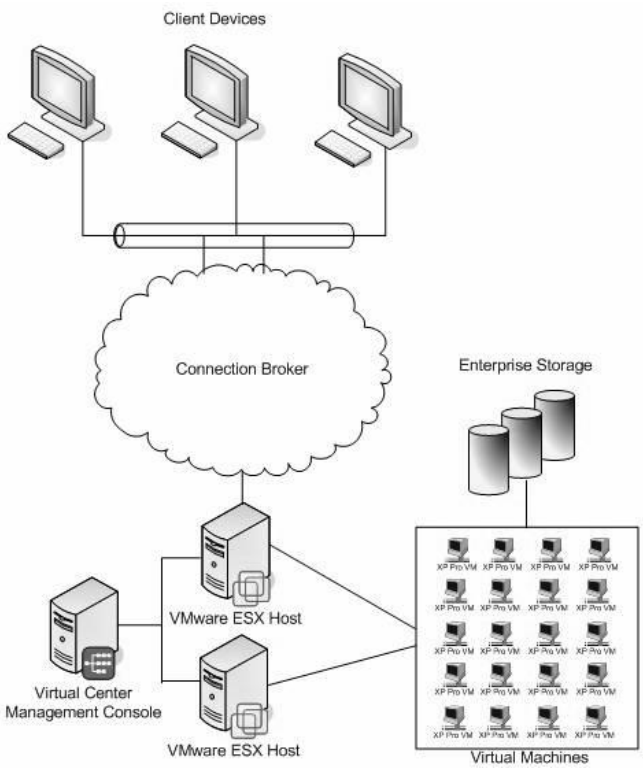


Figure 4. Virtual Desktop Infrastructure (Image Source: [7])

THE CASE STUDY

We perform an experiment to observe the migration of a company e-mail and file server to a cloud infrastructure to evaluate a strategy for transfer and to analyse the security implications. E-mail is moved over to a cloud service to be available on all devices everywhere and reduces costs. The experiment will test the migration of the Exchange Server from a physical server and make a comparison of options.

We consider the case of a local business which has an e-mail server at their head office in Bournemouth and they would like to move to cloud services. Their current server configuration is Microsoft Exchange for e-mail and the same server to store files. This server has anti-virus software for which the licence needs to be renewed each year. The desired result is to remove this server completely. Microsoft offer Exchange online as an alternative to hosted exchange, the web interface to migrate the exchange to online services. A second option is to migrate to Google Apps for business online that offers E-mail, Calendar, Contacts and Notes. It also offers file storage. A third option is to image the entire server, which currently stands at 200Gb of data, to a file and then upload this image to an online virtual server. The same software that was on the physical server will continue to be run on the online virtual server.

Some minor network re-configuration is needed and a redirect of the MX record for the e-mail domain. The shared drives can be mapped using Virtual Private Network (VPN) tunnelling. A drawback with this method is that the 200GB file will take a long time to upload at current upload speeds of 1.5 Mbps.

 ISBN: 978-1-61208-195-3

An alternative solution is an online service with a plug in application for the physical server. [11] This plug in application would selectively upload and replicate the data over to the online vendor with application integrity checking to ensure that the data is transferred correctly. Fig 5 illustrates how this application would function.

Currently, the organisation has Exchange for e-mail and VPN access for file shares. Email access is available through Outlook Web Access, IMAP or Exchange Message Application Programming Interface (MAPI) – one of the benefits of using Microsoft MAPI is that you also can use Calendar, Contacts, Tasks and E-mail all together with one protocol. A cloud provider should be able to support these extra functions. Google Apps for business supports this with an application that synchronises the contacts, calendar and tasks to Google cloud servers. Microsoft Exchange online simply performs all the exchange features that a local Exchange server would perform.

The methods available are to virtualise the server, upload it in its entirety and then make some small adjustments to the network connectivity, or to export the database then import it into an online providers system.

The server migration options are shown in the comparison table in Fig. 6 including security issues.

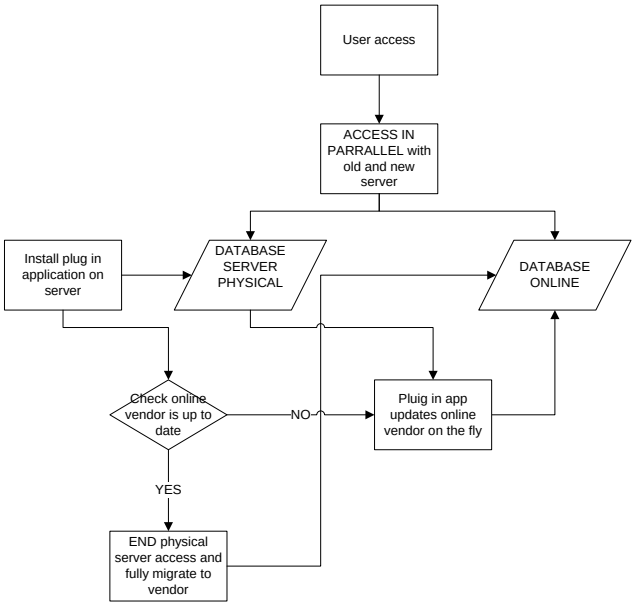


Figure 5. Server Upload Application Flowchart

FINDINGS

The experiment was performed with the server migrated online with all three options evaluated. Although it was found that Exchange Online and Google Apps were good choices, it ultimately depends on individual organisation preference. All e-mail features are available to computers at the office and to all portable devices with very little reconfiguration. However, the virtualized server resulted in a very large data file of 200 GB which takes a long time to upload. The user file is relatively efficient to upload as it is not sending a copy of the entire computer system to the cloud provider but there could still be compatibility issues with the image file when it reaches the online provider. The customer data was exported. For 100 users, the size of the exported file was quite small at 100MB. This file is very fast to upload to the online provider. Import problems could still arise with correct allocation of data fields. The data would need to be looked at to ensure its integrity.

Mission critical applications, as well as data are migrating to the cloud. A recent example of small cloud application is canp.me. [8] Providing a virtual desktop and storage facility using the local desktop processing power, the application provides synchronising data with the application. Hence it is possible to load an application to any PC from the cloud and authenticate the user from data stored on a USB key/dongle. Only one software copy is needed and one license. [12] Deployment times are greatly reduced as the cloud will only contain the latest data and latest copy of the application state. When a cloud based application is called, a minimum set of the appropriate Data Link Libraries (DLLs) are loaded and the remaining application is not loaded until it is required.

The fundamentals of the technology are based around the application's ability to compress, encrypt and transmit desktop applications into smaller parts to the cloud, and then to intelligently rebuild them from the cloud on the virtual desktop in real-time. This saves on bandwidth, because the application and data both reside locally until the data is to be saved and the.

Action	**Method**	**Security Concern**
Hardware Server	Stay at the physical server and update all components.	Security can always be improved: firewalls, IPDS, WAF all cost extra money to run and maintain.
Exchange Online	Install the client exchange transfer connecter module from Microsoft and it will upload automatically.	Security depends on Microsoft, however, they are a large vendor and their reputation is dependent on security for their service.
Google Apps for Business	Install the Google transfer module and exchange mailboxes are sent to Google.	As secure as Google cloud. Google has introduced a 2 step verification to add extra security. Also security features can be managed by Organisation Google accounts over and above individual Google accounts.

Figure 6. Comparison of Migration and Security Issues

 ISBN: 978-1-61208-195-3

application closed. Similar to Citrix XenDesktop, users can appear to travel with their virtual desktop anywhere by creating widgets

In summary, the applications such as cnap.me [8] offers users the possibility of a virtual desktop, synchronised storage, and updated software application, i.e. a low cost managed cloud solution.

EVALUATION

As cloud computing evolves and bandwidth increases, the nature of network connections change. The demand for migrating critical applications including multimedia into a cloud infrastructure will increase rapidly, as will services demanded of online security vendors.

This paper has presented evidence that migrating and storing data in the cloud can be as secure as retention on premises. [13] However, adequate attention to security measures is needed. It is recommended that any organisation considering migration should so do so at the earliest opportunity and take advantage of enhancing security provision.

Users concerned about cloud vendor failure should deploy a solution with failover between different cloud providers to provide access if one provider goes down or has lower costs. This may also be useful for a large organisation that wishes to separate applications in the event of partial failure. [14] Clearly, service level agreements are required for cloud provision. Therefore, each individual case will require evaluation and assessment as a superior service will attract a premium, however, failover, recovery and backup should be provided.

The cloud can offer many advantages over on-premises IT infrastructure. It has the potential to be flexible and cost effective for organisations to use. Security concerns raised can be solved using methods that secure the core aspects of data, identities and devices on an infrastructure. Using this across virtualization platforms will provide a secure network IT infrastructure.

Cloud infrastructure has the potential with appropriate security implementation, to be a better architecture than the old physical model. Virtual security should be much better than physical security.

Security issues are wide ranging and should include securing printer ports that are connected to networks possibly using mac-address based port security authentication. VLANS on virtual networks play a role in the security solution. Access control lists need to function across multiple segments of the network but only allow users access to authorised areas.

NAT, PAT, DNS and DHCP and security are critical network components that still need to be configured locally.

A necessary start for business executives for deployment of virtualized networks is to conduct a full requirements analysis and audit of the current infrastructure, and build a network to provide a future path for migration of physical networks to cloud ones. Plans have to include data, applications and infrastructure.

A major advantage of cloud based solutions is the rapid deployment of a reconfigured service or set of services such as cnap. For example, in the rapidly changing educational environment when a new department is set up, users can be simply added to an existing group policy or set up as a new group. Each user will have access to the applications already present with the inherent security privileges. If a separate server is required for this department, no new hardware is needed; the IT team can request or create a new virtual server within minutes ready to be deployed [15] [16].

CONCLUSIONS AND FUTURE WORK

A variety of methods and approaches have been identified to achieve a successful migration producing a documented strategy for implementation in an organisation with minimum disruption to business functions.

Using physical servers takes up valuable resources and requires a team of technicians to configure monitor and apply patches. There are clear financial and environmental advantages in the use of virtualization and migration to the cloud. A major advantage is rapid time for infrastructure deployment because there is no need to purchase additional hardware.

As more infrastructures are moved into the cloud, organisations can make use of additional services such as multimedia rich media streaming, sharing, and similar services. The reducing costs of infrastructure will open up new markets and opportunities for collaboration.

Plans by the writers will clarify requirements of the network, applications and data to take specific account of ubiquity and the pervasive nature of future multimedia applications & data.

Future goals will be identified for multimedia data models for applications and infrastructure hosted in the cloud. For example, experiments will be undertaken concerned with infrastructure performance to host cloud applications such as cnap.me [8], as a large scale multimedia Virtual Learning Environment (VLE) vehicle in a University. The vision is to clarify data models and requirements for a Cloud-based Virtual learning Environment (CAVE) as a logical evolution of the writers' current research area [1]. Future work will explore migration of a multimedia presentation system to the cloud by combining centralised with cloud-based Virtual Learning Environment (VLE) applications

REFERENCES

[1] Cutts, S., Davies, P., Newell, D. and Rowe, N., (2009) 'Evaluation of an Adaptive Multimedia Presentation System (AMPS) with Contextual Supplemental Support Media'. Proceedings of the MMEDIA 2010 Conference, Athens, Greece.

[2] Symantec. 2011. Virtualization and Evolution to the Cloud Survey: Global Results.https://www4.symantec.com/mktginfo/whitepaper/Virt_and_Evolution_Cloud_Survey_060811.pdf [September 2011]

[3] Morrison, K. S., Chenxi W., Managing the Cloud: An Enterprise Migration Roadmap. Layer 7 Webcast. June 2011.

[4] Rule, D,. and Dittner, R. 2007. The Best Damn Server Virtualization Book Period: Including Vmware, Xen, and Microsoft Virtual Server. Burlington:Syngress.

[5] Virtuatopia. 2009. An Overview of Virtualization Techniques. Available from http://www.virtuatopia.com/index.php/An_Overview_of_Virtualization_Techniques [Accessed December 2011]

[6] VMware,. 2011.VMware Overview. California: VMware. Available from: http://www.vmware.com/files/pdf/VMware-Company-Overview-DS-EN.pdf [Accessed September 2011].

[7] Rouse, P,. 2010. Virtual Desktop Infrastructure (VDI) Overview. Quest Software. Available from: http://www.virtualizationadmin.com/articles-tutorials/vdi-articles/general/virtual-desktop-infrastructure-overview.html [Accessed December 2011].

 ISBN: 978-1-61208-195-3

[8] CNAP. 2011. About cnap.me. Available from: http://www.cnap.me/content/about [Accessed October 2011]
[9] IBM. N.d,.Cloud computing: Paradigm shift or just hype? Available from: https://www-304.ibm.com/businesscenter/cpe/html0/158782.html[Accessed October 2011]
[10] MSDN,. 2011. Definition of Federated Security. Available from: http://msdn.microsoft.com/en-us/library/ms730908.aspx [Accessed December 2011].
[11] InfoTech Spotlight, 2011. Gizmox Debuts Instant CloudMove, an Automated Tool-Based Solution. Available from: http://it.tmcnet.com/topics/it/articles/136729-gizmox-debuts-instant-cloudmove-an-automated-tool-based.htm [Accessed September 2011].
[12] Lombardi, F,. Di Pietro, R,.2010. Secure Virtualization for Cloud Computing. Elsevier Ltd.
[13] Global Knowledge., 2010. Top 10 Security Concerns for Cloud Computing. Available from: https://www.infosecisland.com/blogview/5300-Top-10-Security-Concerns-for-Cloud-Computing.html [Accessed September 2012].
[14] ISO/IEC 27001. 2005. Information Technology Security Techniques: Information security management systems requirements.
[15] ISSAUK. 2011. Information Systems Security Association – UK Chapter. Available from: http://www.issa-uk.org/ [Accessed October 2011].
[16] Miller. 2008. Confidentiality, Integrity and Availability (CIA). Available from: http://it.med.miami.edu/x904.xml [Accessed December 2011]

 ISBN: 978-1-61208-195-3

Nonmetric Earth Mover's Distance for Efficient Similarity Search

Jakub Lokoč and Tomáš Skopal
SIRET Research Group, Faculty of Mathematics and Physics
Charles University in Prague, Czech Republic
{lokoc,skopal}@ksi.mff.cuni.cz

Christian Beecks and Thomas Seidl
Data Management and Data Exploration Group
RWTH Aachen University, Germany
{beecks,seidl}@cs.rwth-aachen.de

***Abstract*—The Earth Mover's Distance is a well-known distance-based similarity measure employed in various domains of data management, especially in computer vision and content-based multimedia retrieval. However, as the computation of the Earth Mover's Distance is a considerably expensive task, efficient processing of content-based similarity queries in large multimedia databases remains a challenging issue. In this paper, we propose to use nonmetric ground distances within the computation of the Earth Mover's Distance in order to speedup its computation, thus improving the efficiency of the entire retrieval process. Moreover, by investigating the inner workings of the Earth Mover's Distance, we show how to balance the trade-off between effectiveness and efficiency in order to adapt the retrieval process to individual user requirements. By making use of metric access methods in combination with the Rubner filter, we empirically show an improvement in efficiency by two orders of magnitude according to the sequential scan, while keeping the retrieval error below 5%.**

***Index Terms*—Earth Mover's Distance; Similarity Search; Indexing.**

I. Introduction

Distance-based similarity search has been successfully utilized in various domains including computer vision and content-based multimedia retrieval, where a database consists of objects represented by nontrivial feature descriptors extracted from unstructured complex data (such as multimedia data). To further improve applicability and effectiveness of such distance-based similarity models, domain experts shift from traditional feature histograms to feature signatures [4], [15], which can flexibly describe the content of multimedia objects. Feature signatures in combination with adaptive distance-based similarity measures [4] then form a powerful tool for effective content-based multimedia retrieval.

However, as the volume of multimedia data grows exponentially, content-based retrieval systems have to provide users with new and more sophisticated exploration facilities. To this end, distance-based similarity models are expected to provide additional trade-off parameters for tuning the precision/performance of the retrieval systems. The system administrators then use the parameters to adapt the retrieval model to better fit user requirements, e.g., more effective but less efficient retrieval, or vice versa. The parameters often affect both quality of the similarity measure and its behavior within an indexing structure for fast retrieval.

In this paper, we focus on the *Earth Mover's Distance* (EMD) [15] – an adaptive distance function for measuring similarity that utilizes a user-defined *ground distance* to penalize some operations within the similarity assessment. In order to speedup the similarity search process using the EMD, we could benefit from indexing by *metric access methods* [6], [20] or by processing queries in a *filter-and-refinement* scheme [2], [3], [16]. Most methods assume *metric* ground distances and some are limited just to feature *histograms*. However, considering also *nonmetric* ground distances could lead to more robust behavior of the distance measure. The arguments for nonmetric ground distances follow from previous image retrieval studies, showing that nonmetric distances performed better than the metric ones [8], [10].

Moreover, using the *parameterized version* of the Earth Mover's Distance [12], we can investigate and tune properties of the distance space, like the "indexability" (i.e., how fast are we able to search/prune that space), measured via the intrinsic dimensionality [6], or the "metricity" that affects the retrieval error exhibited by the metric access methods [18]. In general, for distance spaces suffering from high intrinsic dimensionality, as the traditional (non-parameterized) Earth Mover's Distance [15] on feature signatures [4], [15] does, it is impossible to create an efficient metric index for exact search, and for this reason it is more promising to employ domain-specific filters. Hence, in this paper we investigate the impact of the parameterized Earth Mover's Distance on the distance space properties and point to the promising trade-offs between indexability and retrieval quality. We also combine the distance-based approach with the domain-specific filters.

A. Paper Contribution

The main contribution of this paper is an evaluation of the similarity search under the parameterized Earth Mover's Distance utilizing various L_p ground distances (metric and nonmetric). The findings can be summarized as:

- The parameterized Earth Mover's Distance can be used to tune the retrieval quality and can be efficiently processed by metric access methods or domain-specific filtering techniques. Moreover, we can use the parameter of the distance to guarantee exact searching via more effective fractional L_p distances.

 ISBN: 978-1-61208-195-3

- Although the nonmetric ground distances turn both the filtering and indexing techniques into just approximate methods, the retrieval error caused by the "ground non-metricity" is not significant.
- The parameterized Earth Mover's Distance can be tuned to better fit metric access methods, however, the retrieval error obtained by metric access methods is higher than the retrieval error obtained by the filtering approach.

In the next two sections, we review preliminaries and related work that we utilize and combine in this paper. In section 4, we revisit the filter-and-refinement schema used for efficient EMD processing and we discuss the impact of the nonmetric ground distances. After that, we report and discuss experimental results in section 5 and finally, we conclude contributions of our approach in section 6.

II. Preliminaries

Before we proceed to the contribution and the related work, we briefly summarize the motivation for feature signatures and the Earth Mover's Distance.

A. Feature Signatures

Unlike conventional feature histograms, feature signatures [4], [15] are frequently obtained by clustering the objects' properties [7], [14] within a feature space and storing the cluster representatives and weights. Thus, given a feature space $\mathbb{F}$, the *feature signature* S^o of a multimedia object o is defined as a set of tuples from $\mathbb{F} \times \mathbb{R}^+$ consisting of representatives $r^o \in \mathbb{F}$ and weights $w^o \in \mathbb{R}^+$.

We depict an example of image feature signatures according to a feature space comprising position and color information, i.e., $\mathbb{F} \subseteq \mathbb{R}^5$, in Figure 1. The feature representatives are depicted as circles of the corresponding color, while the weights are reflected by the diameter of the circles. As can be seen in this example, feature signatures adjust to individual image contents by aggregating the features according to their appearance in the underlying feature space.

Although feature signatures are more general than feature histograms, in fact feature histograms are a special case of feature signatures, for similarity search purposes they could be used together with some distances originally designed for histograms, such as the Earth Mover's Distance or the Quadratic Form Distance (generalized to Signature Quadratic Form Distance [5]). In this paper, we put our attention to the Earth Mover's Distance, which we will explain in the next section.

B. Earth Mover's Distance

The Earth Mover's Distance is a distance-based similarity measure originated in the computer vision domain [15]. Its successful utilization, however, gave raise to numerous applications also in different domains. This distance describes the cost for transforming one feature signature (or histogram) into another one. The distance is considered to be a transportation problem and thus is the solution to a linear optimization problem which can be solved via a specialized simplex algorithm.

Fig. 1. Three example images with their corresponding feature signature visualizations.

Given a *ground distance* d that measures the dissimilarity of two features within a feature space $\mathbb{F}$, the *Earth Mover's Distance* (EMD) is defined between two feature signatures $S^q = \{c_i^q, w_i^q\}_{i=1}^n$ and $S^o = \{c_j^o, w_j^o\}_{j=1}^m$ as a minimum cost flow over all flows $f_{ij} \in \mathcal{R}$ as:

$$EMD_d(S^q, S^o) = \min_{f_{ij}} \left\{ \frac{\sum_i \sum_j f_{ij} \cdot d(c_i^q, c_j^o)}{\min\{\sum_i w_i^q, \sum_j w_j^o\}} \right\},$$

subject to the constraints: $\forall i : \sum_j f_{ij} \leq w_i^q$, $\forall j : \sum_i f_{ij} \leq w_j^o$, $\forall i, j : f_{ij} \geq 0$, and $\sum_i \sum_j f_{ij} = \min\{\sum_i w_i^q, \sum_j w_j^o\}$.

These constraints guarantee a feasible solution, i.e., all costs are positive and do not exceed the limitations given by the weights in both feature signatures. In this paper, we assume the ground distance d is the L_p distance over a D-dimensional feature space $\mathbb{F} = \mathbb{R}^D$, defined as

$$L_p(u, v) = \left(\sum_{i=1}^{D} |u_i - v_i|^p \right)^{1/p}$$

While for the parameter $p \geq 1$ the L_p distance is a metric (so-called *Minkowski metric*), for the parameter $0 < p < 1$ it becomes nonmetric (so-called *fractional L_p distance [1]*) as it violates the triangle inequality.

Based on the notion of the EMD and its inherent (non)metric ground distance, we continue with summarizing related work in the next section.

 ISBN: 978-1-61208-195-3

III. Related Work

As there is a minimization problem to solve within the EMD evaluation, the computation time complexity is considerably high (between $O(n^3)$ and $O(n^4)$). Hence, techniques providing efficient similarity search in large multimedia databases using the Earth Mover's Distance are necessary. In the following paragraphs, we shortly summarize the state-of-the-art orthogonal approaches used to efficiently process similarity queries in EMD-based distance spaces.

A. Domain Specific Filters

If we assume the ground distance d as a Minkowski metric (L_p distance, $p \geq 1$), a very simple yet efficient lower-bound for EMD is the *Rubner filter* [15], defined as

$$EMD(S^q, S^o) \geq \sqrt[1/p]{\sum_{k=1}^{D} \left| \sum_{i=1}^{n} w_{q_i} q_{ik} - \sum_{j=1}^{m} w_{o_j} o_{jk} \right|^p},$$

where m, n are the sizes of S^o, S^q, respectively. The Rubner filter holds only if $\sum_{i=1}^{n} w_{q_i} = \sum_{j=1}^{m} w_{o_j}$, otherwise, it becomes an approximate method.

A novel dimensionality reduction techniques for the EMD in a two-step filter-and-refine architecture for efficient exact search can be found in [2], [3]. The authors assume feature histograms and metric ground distances. In [19], Wichterich et al. utilized dimensionality reduction to improve the time of EMD evaluation. They proved that EMD evaluated in a low-dimensional subspace lower-bounds the EMD in the original space. Again, only histograms are applicable to that technique.

B. Transformation to Wavelet Domain

In [17], Shirdhonkar and Jacobs presented a linear-time algorithm for approximating the EMD for low-dimensional histograms using the sum of absolute values of the weighted wavelet coefficients of the difference histogram. Since the EMD computation is a special case of the Kantorovich-Rubinstein transshipment problem, the method can exploit the Hölder continuity constraint in its dual form to convert it into a simple optimization problem with an explicit solution in the wavelet domain. The authors proved the resulting wavelet EMD metric is equivalent to EMD, i.e., the ratio of the two is bounded. The bound estimates were also provided.

C. Metric access methods

Another approach for efficient indexing of the Earth Mover's Distance could be the distance-based indexing, especially the *metric access methods* [6], [20]. These methods utilize precomputed distances stored in a metric index to estimate the lower-bound of the original distance between a query object q and a database object o. In Figure 2, we depict an example of $\delta(q, o)$ lower-bound estimation using one reference point p, where $\delta(p, o)$ is the precomputed distance stored in the metric index and $\delta(q, p)$ is evaluated at the beginning of query processing. In the case $LB(\delta(q, o))$ is greater than the actual query radius r (considering range or kNN query), the original expensive distance $\delta(q, o)$ does not have to be evaluated.

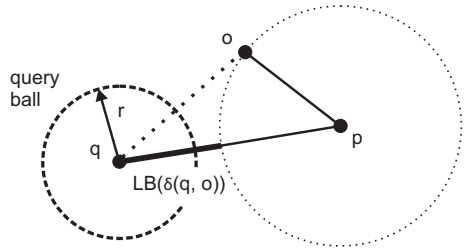


Fig. 2. Lower-bound estimation using a reference point p.

However, the distance spaces based on the Earth Mover's Distance usually suffer from high *intrinsic dimensionality* [6] (low indexability, i.e., $LB(\delta(q, o))$ is mostly lower than actual query radius r), so that only approximate techniques can be used for efficient search. To overcome this limitation, we proposed a parameterized version of the Earth Mover's Distance (pEMD) [12], defined as:

$$pEMD_d(S^q, S^o, w) = \min_{f_{ij}} \left\{ \frac{\sum_i \sum_j f_{ij} \cdot \mathrm{FP}(d(c_i^q, c_j^o), w)}{\min\{\sum_i w_i^q, \sum_j w_j^o\}} \right\},$$

where FP is the *fractional-power modifier* [18] defined as:

$$\mathrm{FP}(x, w) = \begin{cases} x^{\frac{1}{1+w}} & \text{for } w > 0 \\ x^{1-w} & \text{for } w \leq 0 \end{cases}$$

The intuition behind the FP-modifier is rather simple. Depending on the weight parameter w, we can either suppress ($w > 0$) or strengthen ($w < 0$) the transportation costs to outlier features when comparing feature signatures. Hence, the robustness of the measure can be tuned by setting the impact of outliers (noise bins or clusters) on the overall distance. However, the parameterized Earth Mover's Distance is no longer a metric when employing the FP-modifier ($w < 0$), or fractional L_p distance ($p < 1$), or not-normed weights.

All the techniques mentioned above are based on domain-specific solutions for (often low-dimensional) feature histograms, metric L_p distances ($p \geq 1$), or they utilize approximate similarity search to speed up query processing.

IV. Indexing the Earth Mover's Distance

In this paper, we investigate the general problem of EMD-based distance spaces employing feature signatures and non-metric ground distances. Since we consider feature signatures, the original Rubner filter and the distance-based indexing (metric access methods) can only be used to speed up query processing in large multimedia databases.

A. Revisiting the filter-and-refinement scheme

First of all, we would like to reconsider the filter-and-refinement scheme proposed in [2], where Assent et al. used a chain of filters. In general, when trying to exclude irrelevant database objects from the search process, we apply lower-bounding filters that progressively reduce the candidate set of the results. To optimize this process, we first apply the

 ISBN: 978-1-61208-195-3

cheapest filters (e.g., taking less than $O(n)$ time to compute the lower bound, where n is the signature size) and then apply the more expensive ones (which are usually also more effective in filtering). This corresponds to the optimal multi-step search process described in [16].

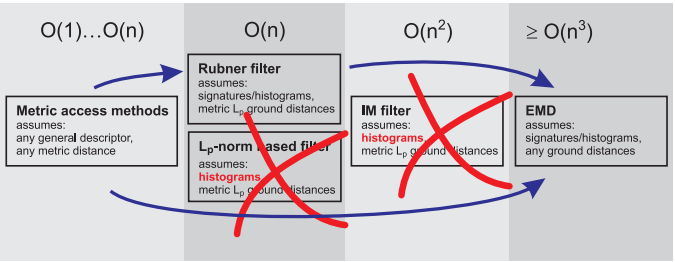


Fig. 3. Filter chain used in the filter-and-refinement scheme for EMD search.

In the aforementioned paper [2], Assent et al. proposed similarity query processing using the cheap Rubner filter ($O(n)$) and subsequently the more expensive independent minimization (IM) filter ($O(n^2)$). The two filters significantly reduce the final set of candidate objects, so that the expensive EMD evaluation is needed only for a fraction of the database. One of the contributions of this paper is the extension of the class of cheap filters by the metric access methods (MAM), which reduce the complexity of lower bound computation from $O(1)$ to $O(n)$. We depict an overview of the proposed filter-and-refinement scheme for efficient EMD-based similarity search in Figure 3. Note, that we do not consider the IM filter and L_p-norm filter anymore, as they only support histograms.

In particular, we combine the Rubner filter and an MAM-based filter (e.g., the pivot tables as detailed in the Section V-C), both cheap and supporting feature signatures. Having a Rubner filter lower-bound LB_{Rub} and the corresponding MAM lower-bound LB_{MAM}, we select the larger estimate $max(LB_{Rub}, LB_{MAM})$. To further improve the performance of the whole filter-and-refinement scheme, we also plan to generalize the IM filter for feature signatures (subject of our future work).

B. Nonmetric ground distances

As another contribution of this paper, we investigate the impact of the nonmetric L_p ($p < 1$) ground distances that have been frequently used in the image retrieval for robust image matching [8], [10]. For instance, consider the three feature signatures depicted in Figure 4, where S^q stands for a query signature and S^x, S^y for database signatures. The first two signatures S^q, S^x consist of very similar distributions of feature clusters, while the third one S^y is slightly different. To provide robust image ranking by the EMD, we have to employ such a ground distance, that will neglect the impact of the outlying noisy cluster c_4 in S^x. The fractional L_p ground distances are a good choice for such purpose, as they decrease the impact of outlying distances to the overall aggregation provided by the EMD. A similar robust behavior can be achieved by using the FP-modifier within the parameterized EMD. Hence, we have tools for fine-tuning the quality of the EMD-based similarity retrieval in a particular database.

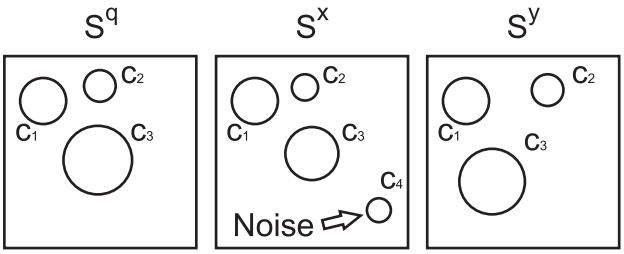


Fig. 4. Comparing a query signature with a database signature.

The main disadvantage of the fractional L_p distances is that they are not metrics. In particular, they lose the triangle inequality, assumed by both the Rubner filter and metric access methods to guarantee exact search. In turn, the nonmetric EMD (incorrectly) employed as a metric brings the possibility of a retrieval error – false dismissals and/or false positives in the query result. Nevertheless, we can control the retrieval error by tuning the parameter w of the parameterized EMD, while as showing in the following experimental section, the retrieval error caused by the "nonmetricity" of the EMD is not significant.

V. Experimental Results

We conducted an experimental evaluation on the MIR Flickr [11] and ALOI [9] databases comprising 25,000 and 72,000 images, respectively. We have extracted feature signatures based on color, position, and texture information, similar to [4], where each image was represented by several 7-dimensional feature centroids and each centroid was assigned a weight. The feature signatures of the ALOI database consist of 12-140 centroids, 54 centroids on average, and those for the MIR Flickr database consist of 8-150 centroids, 57 centroids on average. Although the selected databases were not very large, they provided the ground truth for evaluating the search effectiveness. We have examined 6 variants of the Earth Mover's Distance, each using different L_p ground distances. Three variants were nonmetrics $\{L_{0.25}, L_{0.5}, L_{0.75}\}$ and the other three were metrics $\{L_1, L_2, L_5\}$. As a metric access method, we used simple Pivot tables (using 50 pivots) [6], [13]. In each test, we performed 100 kNN queries ($k = 10$) and averaged the results. As the main observables, we measured both the retrieval efficiency and the retrieval effectiveness. The efficiency was measured in terms of real time as well as in the number of EMD computations. The effectiveness was measured using the mean average precision (i.e., employing the ground truth) and also using the retrieval error defined as the deviation from the referential result obtained by the sequential scan. The tests ran on a workstation 2x Intel Xeon X5660 2.8 Ghz, 24GB RAM, Windows Server 2008 R2 64bit (non-virtualized).

 ISBN: 978-1-61208-195-3

A. Basic tests

The graphs in Figure 5 depict the intrinsic dimensionality and the mean average precision values for the aforementioned databases by changing the parameter w of pEMD (denoted FP weight). The intrinsic dimensionality decreases for both databases with decreasing w, while mean average precision stays at a considerably high level of greater than 0.6 for the ALOI database and 0.26 for the MIR Flickr database. Since pEMD changes the ground distance matrix and thus changes the number of iterations needed to find the optimal solution of pEMD, we have also observed query processing times for the sequential scan.

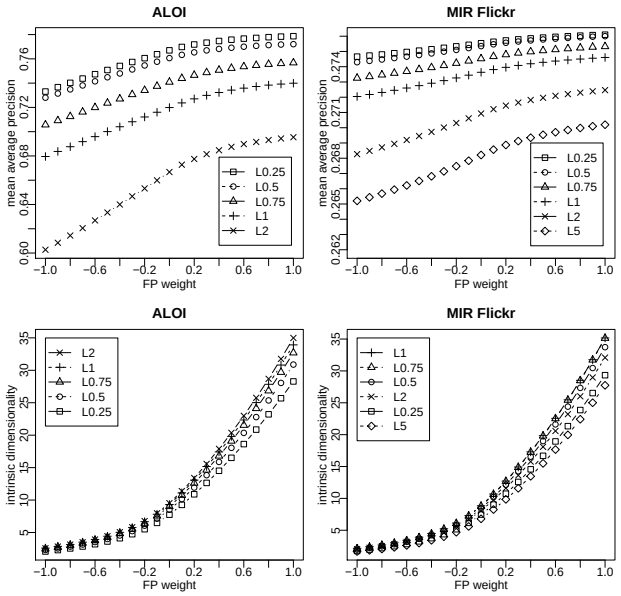


Fig. 5. The intrinsic dimensionality and mean average precision for MIR Flickr and ALOI databases.

As can be seen from the graphs in Figure 6, increasing the parameter w leads to a lower number of iterations, which results in decreased realtime cost. Also note that the L_1 ground distances (i.e., $L_p, p = 1$) led to fastest responses due to the absence of powering/rooting by p in the L_p formula.

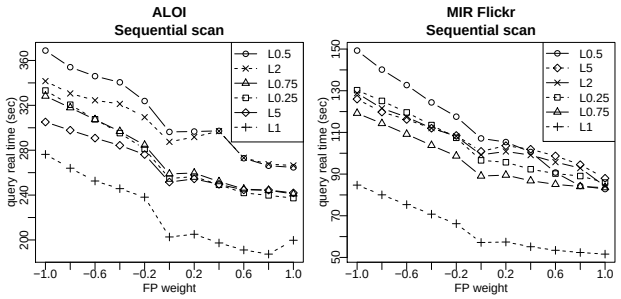


Fig. 6. The realtime of sequential scan for MIR Flickr and ALOI databases.

B. Rubner filter

In the second set of experiments, we evaluated the performance of the Rubner filter for various values of pEMD's parameter w. As we can observe in Figure 7, both realtime and the number of EMD evaluations decreases for higher w. This is caused by the fact, that the Rubner estimation of EMD is more similar to the real EMD value. However, this leads also to a small retrieval error, because the estimation is no longer guaranteed a lower bound and so some relevant objects may be filtered, as shown in Figure 8. We can also observe a good performance of $L_{0.25}$, however this ground distance results also in higher retrieval error.

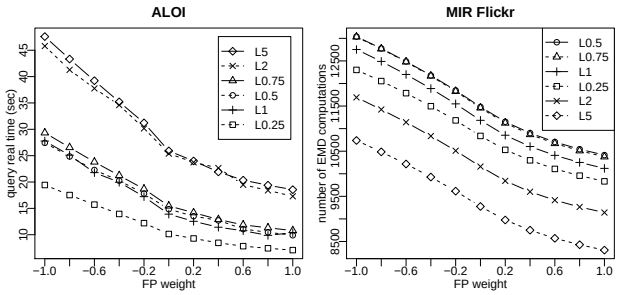


Fig. 7. Effect of the Rubner filter – query real time.

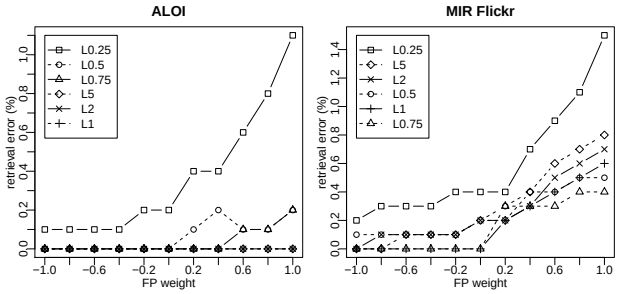


Fig. 8. Effect of the Rubner filter – retrieval error.

C. Rubner Filter combined with the Metric Access Methods

In the last experiments, we combined the Rubner filter with a simple metric access method – the Pivot Table (the original LAESA) [13]. From the Figure 9, we may observe the impact of low intrinsic dimensionality on pivot table filtering. However, the efficiency is coupled with high retrieval error (a lot of false dismissals), see Figure 10. Nevertheless, if the user accepts 5% error then the query is more than twice as fast as using just the Rubner filter. If we require the highest retrieval precision, it is better to utilize the Rubner filter and nonmetric ground distances rather than metric access methods and metric ground distances.

D. Discussion

In the experiments, we have focused on the effective similarity search using nonmetric ground distances employed in the Earth Mover's Distance. To the best of our knowledge, this

 ISBN: 978-1-61208-195-3

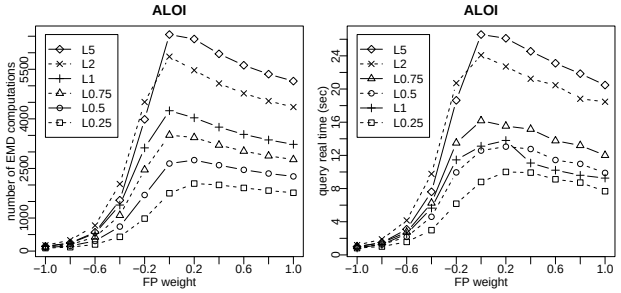


Fig. 9. Effect of the Rubner filter combined with Pivot tables – query cost.

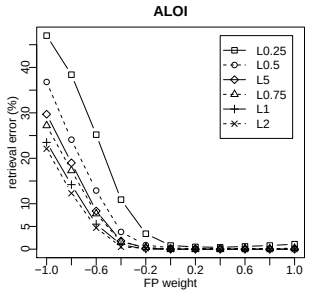


Fig. 10. Effect of the Rubner filter combined with Pivot tables – retrieval error.

paper is the first one, that combines and investigates the effect of two filtering approaches used for efficient query processing – the general metric spaces approach and the domain specific Rubner filter. We also more deeply investigate the effect of the recently introduced Earth Mover's Distance parameter w, that in connection with the Rubner filter can result in more efficient and more effective similarity search. In the case a fast query processing is required and the precision is not preferred, we can employ the metric access methods as an approximate search technique for lower values of w.

Our experimental evaluation reveals that the combination of filter-and-refinement schemes and nonmetric ground distances within the EMD provides the best retrieval results. In particular, the trade-off between efficiency and effectiveness is given by comparatively low query response times and small retrieval errors. Thus, we conclude that our proposed approach is able to outperform state-of-the-art metric indexing solutions for the EMD.

VI. Conclusion and Future Work

We have investigated a parameterized version of the Earth Mover's Distance in combination with nonmetric ground distances. In the experimental evaluation, we showed that nonmetric ground distances can be utilized for effective and efficient similarity search in multimedia databases by the parameterized Earth Mover's Distance. More specifically, using a revisited filter-and-refinement schema, the system administrators can provide user with more sophisticated exploration facilities, e.g., more effective but less efficient retrieval, or vice versa. In the future, we would like to formally describe the observed behavior and investigate other modifying functions that can provide more desirable properties (e.g., preserve metric axioms). We also want to generalize other EMD filters for feature signatures and nonmetrics.

Acknowledgment

This research has been partially supported by Czech Science Foundation (GAČR) projects 202/11/0968 and P202/12/P297, and by the Deutsche Forschungsgemeinschaft within the Collaborative Research Center SFB 686.

References

[1] C. C. Aggarwal, A. Hinneburg, and D. A. Keim. On the surprising behavior of distance metrics in high dimensional spaces. In *Proc. 8th International Conference on Database Theory*, ICDT '01, pages 420–434, London, UK, 2001. Springer-Verlag.

[2] I. Assent, A. Wenning, and T. Seidl. Approximation techniques for indexing the earth mover's distance in multimedia databases. In *IEEE ICDE*, pages 11–22, 2006.

[3] I. Assent, M. Wichterich, T. Meisen, and T. Seidl. Efficient similarity search using the earth mover's distance for large multimedia databases. In *IEEE ICDE*, pages 307–316, 2008.

[4] C. Beecks, M. S. Uysal, and T. Seidl. A comparative study of similarity measures for content-based multimedia retrieval. In *IEEE ICME*, pages 1552–1557, 2010.

[5] C. Beecks, M. S. Uysal, and T. Seidl. Signature quadratic form distance. In *Proc. ACM CIVR*, pages 438–445, 2010.

[6] E. Chávez, G. Navarro, R. Baeza-Yates, and J. L. Marroquín. Searching in Metric Spaces. *ACM Computing Surveys*, 33(3):273–321, 2001.

[7] T. Deselaers, D. Keysers, and H. Ney. Features for image retrieval: an experimental comparison. *Information Retrieval*, 11(2):77–107, 2008.

[8] M. Donahue, D. Geiger, T. Liu, and R. Hummel. Sparse representations for image decomposition with occlusions. In *CVPR*, pages 7–12, 1996.

[9] J.-M. Geusebroek, G. J. Burghouts, and A. W. M. Smeulders. The Amsterdam Library of Object Images. *IJCV*, 61(1):103–112, 2005.

[10] P. Howarth and S. Ruger. Fractional distance measures for content-based image retrieval. In *ECIR 2005*, pages 447–456. LNCS 3408, Springer-Verlag, 2005.

[11] M. J. Huiskes and M. S. Lew. The mir flickr retrieval evaluation. In *Proc. ACM MIR*, pages 39–43, 2008.

[12] J. Lokoč, C. Beecks, T. Seidl, and T. Skopal. Parameterized earth mover's distance for efficient metric space indexing. In *Proceedings of the Fourth International Conference on SImilarity Search and APplications*, SISAP '11, pages 121–122, New York, NY, USA, 2011. ACM.

[13] M. L. Mico, J. Oncina, and E. Vidal. A new version of the nearest-neighbour approximating and eliminating search algorithm (aesa) with linear preprocessing time and memory requirements. *Pattern Recogn. Lett.*, 15(1):9–17, 1994.

[14] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10):1615–1630, 2005.

[15] Y. Rubner, C. Tomasi, and L. J. Guibas. The Earth Mover's Distance as a Metric for Image Retrieval. *IJCV*, 40(2):99–121, 2000.

[16] T. Seidl and H.-P. Kriegel. Optimal multi-step k-nearest neighbor search. In *Proc. ACM SIGMOD*, pages 154–165, 1998.

[17] S. Shirdhonkar and D. Jacobs. Approximate earth movers distance in linear time. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1 –8, june 2008.

[18] T. Skopal. Unified framework for fast exact and approximate search in dissimilarity spaces. *ACM Transactions on Database Systems*, 32(4):1–46, 2007.

[19] M. Wichterich, I. Assent, P. Kranen, and T. Seidl. Efficient emd-based similarity search in multimedia databases via flexible dimensionality reduction. In *Proc. ACM SIGMOD*, pages 199–212, 2008.

[20] P. Zezula, G. Amato, V. Dohnal, and M. Batko. *Similarity Search: The Metric Space Approach (Advances in Database Systems)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005.

 ISBN: 978-1-61208-195-3

File Size Comparisons of Modeled and Pixel-Based Video in Five Scenarios

Juergen Wuenschmann, Christian Feller, and Albrecht Rothermel
Universität Ulm
Institute of Microelectronics, Albert-Einstein-Allee 43, 89081 Ulm, Germany
Email: {juergen.wuenschmann, christian.feller, albrecht.rothermel}@uni-ulm.de

***Abstract*—File sizes have been compared for object modeled video versus pixel-based video with respect to different factors, which influence the data rate. The goal is to simplify the choice, which representation is favorable in terms of file size for varied content. Traditional video compression is very sophisticated and to achieve even lower data rates than the state-of-the-art video codecs while preserving the visual quality is difficult and increases the complexity. Using the object-based representation, possibilities to reduce the amount of data to be stored or transferred are investigated. With our improved object-based encoder a data rate reduction of 67,6-90 percent compared to an uncompressed object-based representation and up to 90.5 percent compared to high quality, variable bitrate, pixel-based video is achieved for the investigated scenarios.**

***Index Terms*—Video signal processing; Video compression; High definition video; Object-based video coding.**

I. Introduction

The reduction of redundancy and irrelevancy has been a major part of digital signal processing since its beginning [1]. With a growing amount of data in video signal processing, due to rising resolution, frame rate, and quality demand, a highly effective compression is essential. The possibility to achieve higher compression rates while preserving a certain quality level becomes harder and the complexity of the implemented algorithms is rising. Most of them are based on the traditional, pixel-based video processing scheme [2].

For modeled and animated material, another path can be chosen. The material can be compressed and transmitted or stored directly, without rendering it to a pixel-based representation. A big advantage of this modus operandi is that it is possible to adapt to any display device property, as, e.g. resolution, frame rate, and number of views, through adjusting a built-in renderer. A basic object-based compression scheme was proposed in [3] and standardized as MPEG-4 pt. 25. This reference implementation has been improved and a comparison to the original implementation is given in [4].

To compare the traditional pixel-based video representation and the object-based video representation, five test scenarios are created. Each isolates the influence of a certain video property on the amount of data and therefore the compression efficiency. With these test scenarios, it is possible to distinguish, which representation is favorable to use for certain video material.

The paper is organized as follows: In Section II, the four pixel-based and four object-based parameter sets are described. The five test scenarios and their influence on the amount of data are explained in Section III. Section IV contains the results and a discussion and Section V summarizes and concludes the paper.

II. Video Representations

The comparisons have been performed between object-based and pixel-based video representations. Both representations have been tested using various parameters. For object-based material these formats have been used:

- *Collada* - XML representation of the scene.
- *Collada rared* - Win Rar compressed Collada.
- *MP4 ref encoded* - encoded using the standard (non-improved) version of the MPEG-4 pt. 25 encoder.
- *MP4 enc xyz OneStream* - encoded using our improved MPEG-4 pt. 25 encoder.

These four formats incorporate settings of the whole efficiency range for object-based material. The most ineffective one is the textual representation, which can be compressed very effectively with the state-of-the-art rar compression. A Collada file has to be decompressed completely to be used. With the object-based video compression MPEG-4 pt. 25, streaming of video clips is possible. Two versions of this encoder are used. The one used as reference is the version proposed in [3], while many improvements are integrated in our updated version. The pixel-based representation is produced rendering a scene in Blender and using the H.264 encoder *x264* to encode the video with a state-of-the-art video encoder. Parameters used were high profile at level 4.1 and two reference frames. Clips were rendered with different resolutions. Due to space limitations, the only resolution shown in this paper is $1920x1080$ Pixel, which is Full HD. The quality settings used are:

- *crf18 1080p* - variable bit rate encoding using constant rate factor 18.
- *crf24 1080p* - as above using constant rate factor 24.
- *10MBit 1080p* - constant bit rate encoding of $10MBit/s$.
- *20MBit 1080p* - as above with $20MBit/s$.

These settings are based on the settings used typically for Blu-Ray discs and represent high quality video encoding.

III. Test Scenarios

Scenario 1: The first analyzed scenario comprises the effect of the scene length on the data rate. To isolate the scene

 ISBN: 978-1-61208-195-3

length from other parameters influencing the amount of data, a generic test sequence was chosen. The sequence is derived from the incrementing cubes scenario described in [4]. It shows an amount of i^2 cubes arranged side by side with random color and random rotation. The difference is, that the

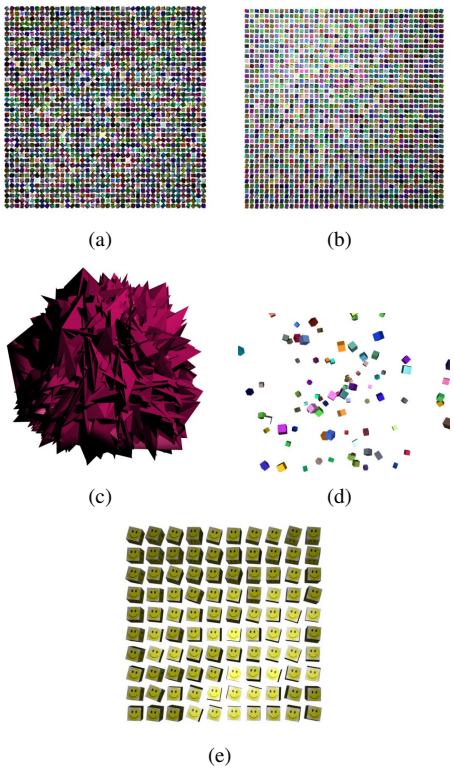


Fig. 1. As example, the first frame of (a) Scenario 1, (b) Scenario 2, (c) Scenario 3, (d) Scenario 4, and (e) Scenario 5 is shown.

number of cubes is kept constant at 1600 Cubes and instead the number of frames is altered.

Scenario 2: The second scenario, derived from the incrementing cubes test described in [4], is the evaluation of how the speed of movement influences the amount of data. It shows an amount of i^2 cubes arranged side by side with random color and random rotation. The number of cubes in the scene is again kept constant at 1600 Cubes and the rotation of all cubes is, despite a random initial orientation, identical. The rotation is defined by the number of turns and is incremented in every iteration by $5 \cdot \pi$. Using this scenario, it is possible to evaluate the quality of a temporal prediction used in video representation encoders.

Scenario 3: The following scenario reveals the effect of the complexity of objects in a scene on the data rate. To investigate this effect, a cube is subdivided with every iteration and the vertices are arranged randomly inside a spherical space. The generated object is rotating. Using simple subdivision, the number of vertices for iteration i is $v = (2^{i-1} + 1)^3 - (2^{i-1} - 1)^3$.

Scenario 4: Camera movement is the fourth evaluated effect on the amount of data. The scenario is build up of a static scene containing 100 Cubes with random orientation and color, randomly arranged in a spherical space. A camera moves around this scene with different velocity. This scenario represents the influence of panning and tilting of a video camera.

Scenario 5: The next analyzed scenario is as well derived from the incrementing cubes scenario of [4]. It has been upgraded to use textures mapped onto the objects. As in the original script, the number of cubes displayed is i^2 for iteration i. Every cube has a random rotation and a texture is mapped onto each side of a cube instead of random coloring for each cube in the original experiment. Using this test, it is possible to evaluate the influence of a more complex surface of moving objects on the data rate of the two video representations.

An example frame for each of the five scenarios can be seen in Fig. 1.

IV. Results

The results for the first scenario are shown in Fig. 2. Despite

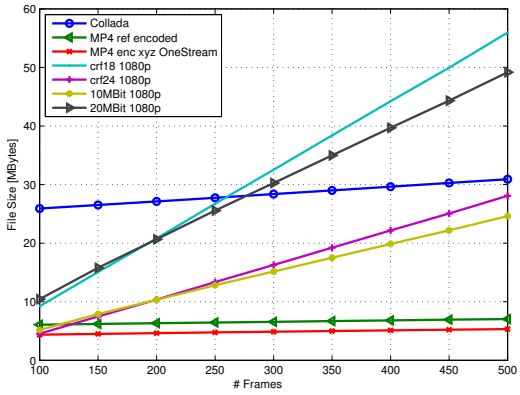


Fig. 2. Scenario 1: Effect of the scene length on the file size of object-based and pixel-based video representations.

the different parameter sets, the pixel-based video representation shows a linear increase in file size with incrementing number of frames. The object-based representation is nearly unaffected by the length of the scene. With growing length, additional key frames for the animations have to be stored, which create little overhead. Consequently, an object-based representation is favorable for long scenes. In Table I the data rate savings for the different compression schemes for object-based material are shown. It is obvious, that the non-streamable *Collada rared* performs best and the MPEG-4 pt. 25 reference encoder performs worst. Our improved encoder lies in the middle, comprehending the benefit of the streaming possibility.

TABLE I
Scenario 1: Data rate savings for compressed object-based representations with respect to uncompressed Collada.

	File Size Saving [%]
Collada rared	95.9 - 95.5
MP4 ref encoded	76.6 - 77.3
MP4 enc xyz OneStream	83.1 - 82.8

 ISBN: 978-1-61208-195-3

The effect of the velocity of animations is shown in Fig. 3. The variable bit rate video files are strongly influenced by

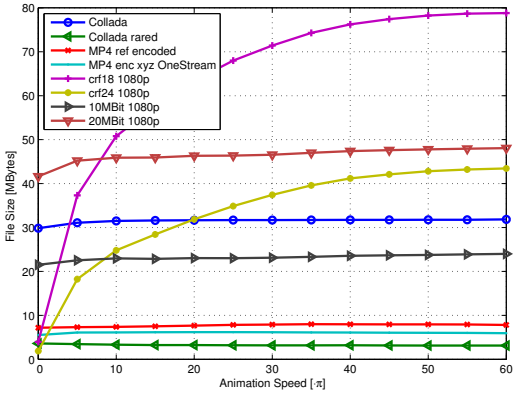


Fig. 3. Scenario 2: Growing animation velocity only influences the variable bit rate video representation.

the animation speed. The temporal prediction is working less efficient for rising speed. For the constant bit rate files, the file size is kept nearly constant by the rate control. Since the number of keyframes is not influenced by a growing animation speed, all object-based representation files have a constant file size. The data rate savings for the different object-based compression versions are shown in Table II. In this scenario, the improved encoder, with up to 81.5% file size saving, performs 6% better than the reference encoder. The non-streamable *Collada rared* performs best with up to 90% saving.

TABLE II
SCENARIO 2: DATA RATE SAVINGS FOR COMPRESSED OBJECT-BASED REPRESENTATIONS WITH RESPECT TO UNCOMPRESSED COLLADA.

	File Size Saving [%]
Collada rared	87.9 - 90.2
MP4 ref encoded	76.6 - 75.5
MP4 enc xyz OneStream	81.5 - 81.4

The results for the third scenario are shown in Fig. 4. The

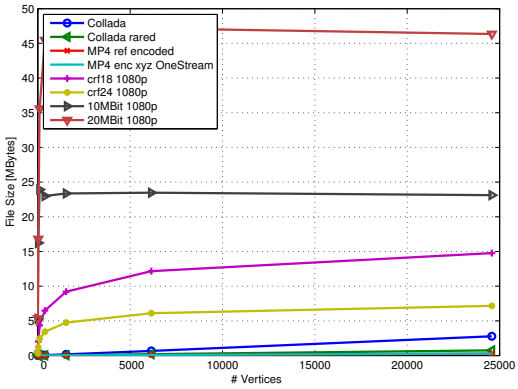


Fig. 4. Scenario 3: Effect of the number of vertices on the file size of object-based representation and pixel-based representation.

effect of the level of detail of an object on the pixel-based representation is low. The constant bit rate files are unaffected except for very small numbers of vertices. The variable bit rate files show a steep incline for a small number of vertices and are reaching saturation for a higher level of detail. The object-based representation shows a linear behavior for an increasing number of vertices. Consequently, for an increasing level of detail a break-even is expected, where the pixel-based representation is favorable depending on the desired quality. In Table III the data rate savings for the different compression schemes for object-based material are shown. The MPEG-4 pt. 25 reference encoder and our improved encoder perform similar with data rate savings up to 90%, comprehending the benefit of the streaming possibility. The reason for the similar performance is that the improvements mainly address the animation encoding, which is a negligible factor in this scenario.

TABLE III
SCENARIO 3: DATA RATE SAVINGS FOR COMPRESSED OBJECT-BASED REPRESENTATIONS WITH RESPECT TO UNCOMPRESSED COLLADA.

	File Size Saving [%]
Collada rared	73.3 - 67.6
MP4 ref encoded	90.0 - 64.0
MP4 enc xyz OneStream	90.0 - 67.6

In Fig. 5 the file size for the different representations is shown with respect to the camera velocity. The constant bit rate video with $20MBit/s$ limit has not reached the bit rate limit, but the file size is increasing with the camera speed as the variable bit rate videos do. In contrast, the object-based representation files are unaffected by increasing camera movement speeds, since the scene and the number of keyframes in the animation does not change. Table IV reveals, that rar compression is the most effective one for the object-based representation. The two MPEG-4 encoded parts have nearly the same compression ratio, because the main improvement is in the animation coding. For the current

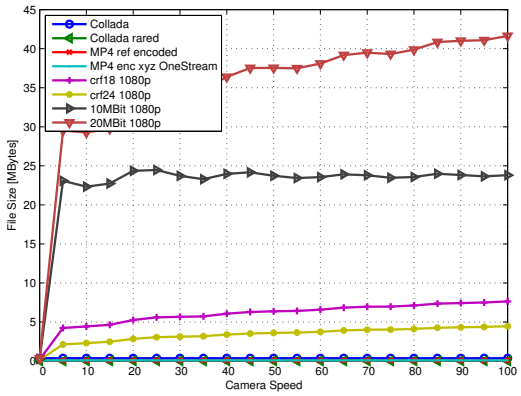


Fig. 5. Scenario 4: The object-based representation is not influenced by camera movement, whereas the pixel-based representation shows growing file size for increasing camera speed.

 ISBN: 978-1-61208-195-3

TABLE IV
SCENARIO 4: DATA RATE SAVINGS FOR COMPRESSED OBJECT-BASED REPRESENTATIONS WITH RESPECT TO UNCOMPRESSED COLLADA.

	File Size Saving [%]
Collada rared	95.9 - 95.0
MP4 ref encoded	75.8 - 75.0
MP4 enc xyz OneStream	75.9 - 75.2

TABLE V
SCENARIO 5: DATA RATE SAVINGS FOR COMPRESSED OBJECT-BASED REPRESENTATIONS WITH RESPECT TO UNCOMPRESSED COLLADA.

	File Size Saving [%]
Collada rared	94.5 - 85.9
MP4 ref encoded	89.0 - -458.6
MP4 enc xyz OneStream	88.0 - 80.1

scenario, only the camera and the according scene illumination is animated, which represents only a small part of the whole scene graph.

The results for the last scenario described are shown in Fig. 6. The experiment is comparable to the one shown in

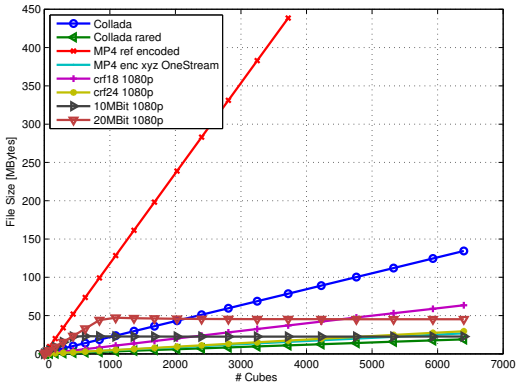


Fig. 6. Scenario 5: Using Textures, the file size grows for all representations. *MP4 ref encoded* is affected the most, because a texture is encoded once for every object it is used on.

[4] despite that textures are used on the cubes. With textures on the surfaces, the complexity of an object is risen compared to simple coloring. No effect can be observed for the constant bit rate video files due to the rate control. The variable bit rate files of the pixel-based representation show higher file sizes than the ones reported in [4]. The increase for *crf18 1080p* is up to 55%. The reason for the increase is the more complex picture content, which degrades the outcome of the temporal prediction. In contrast, the object-based files have a rise of the file size with the size of the texture and a certain overhead for texture mapping and the texture library. A texture itself can be seen as an offset to the object-based representation. For *Collada*, the texture is uncompressed and has a size of $1.02MB$. The other object-based representations use their native compression scheme or known picture compression schemes as JPEG2000 to encode the texture. It is remarkable, that the reference encoder does not recognize the multiple use of one texture. It encodes and writes it to the MPEG-4 stream as often as it is used. Instead of lowering the amount of data to transmit or store, it gets multiplied with that precedure. Our MPEG-4 pt. 25 encoder, however, recognizes multiple use of textures and is able to work effectively when using textures. Comparing the different object-based compressions to *Collada* (Table V), *Collada rared* shows the highest saving. The efficiency of *MP4 ref encoded* is highly dependent on how often a texture is used, as mentioned, and is therefore not comparable to a real encoder. The encoder stopped working at 3969 cubes and the last iteration generated a file, which has 5.586 times the size of the uncompressed Collada file including the texture. If the experiment had run to the last iteration of 6400 cubes, the file size would have been ≈ 9 times the size of *Collada*. *MP4 enc xyz OneStream* shows a remarkable saving between 88.0% and 80.1%.

V. CONCLUSION

Five scenarios have been evaluated to show the effect of different parameters on pixel-based representation and object-based representation. The first scenario shows, that the pixel-based representation has a linear relation to scene length. This is also true for the object-based representation, but the slope is negligible. It has been proven that the two scenarios, increasing animation and camera velocity, show no effect on the object-based representation schemes, as expected, whereas the effect on the variable bit rate video is obvious. A linear relation between the number of vertices and the object-based representation is observed for the third scenario, whereas the pixel-based representation is unaffected or reaches saturation for higher levels of detail. The last scenario shows, that the use of textures does not change the general behavior for a rising number of objects, but the texture itself and its representation inside the object-based file introduces an offset to the file size. Thus, the effect of the discussed factors can be easily predicted for both video representations, when having to choose one of them. Moreover, the compression gain for the object-based representation has been evaluated. Compression ratio improvements of up to 6.5% could be reached for our version of the MPEG-4 pt. 25 codec compared to the reference encoder, not mentioning the data-generating texture compression of the reference encoder. Compared to uncompressed Collada, $67,6\% - 90\%$ file size saving is possible for the investigated scenarios. Up to 90.5% file size saving were observed for long scenes compared to high quality, variable bit rate video.

REFERENCES

[1] R. Steinmetz, *Multimedia Technologie*, Springer, 2nd edition, 1999.
[2] G. J. Sullivan and T. Wiegand, "Video compression - from concepts to the h.264/avc standard," vol. 93, no. 1, pp. 18–31, Jan. 2005.
[3] B. Jovanova, M. Preda, and F. Preteux, "Mpeg-4 part 25: A graphics compression framework for xml-based scene graph formats," *Signal Processing: Image Communication*, vol. 24, pp. 101–114, 2009.
[4] J. Wuenschmann, T. Roll, C. Feller, and A. Rothermel, "Analysis and improvements to the object based video encoder mpeg 4 part 25," in *Proc. IEEE 1st International Conference on Consumer Electronics - Berlin*, September 2011.

 ISBN: 978-1-61208-195-3

A Q-Learning Approach to Decision Problems in Image Processing

Alexandru Gherega, Radu Mihnea Udrea
University 'Politehnica' of Bucharest
Bucharest, Romania
alex.gherega@gmail.com, mihnea@comm.pub.ro

Monica Rădulescu
R&D Softwin
Bucharest, Romania
mradulescu@softwin.ro

Abstract—Decision making in a sequential and uncertain manner is still one of the major problems that all computer science research fields relate to, either by trying to find the optimal solution for it or needing such a solution to obtain optimal results for some other stated problem. This paper's objective is to address the possibility of using reinforcement learning for decision problems in data processing environments. A short review of current uses for reinforcement learning solutions into different fields is depicted as well as a particular case in image processing methods. A solution is proposed for a specific decision problem in the field of image processing. Our implementation shows the results of a reinforced parameterization method for edge detection using Q-Learning.

Keywords-reinforcement learning; Q-Learning; computer cognition; adaptive and learning systems.

I. INTRODUCTION

Reinforcement learning [1][2] is a way of discovering how to reach a specific desired state in order to maximize a numerical reward signal, i.e., how to map situations to actions based on the interactions with the environment. The learning entity is not told which actions to take, as in other forms of machine learning, but instead must discover which actions yield the best reward by random exploration. Actions may affect not only the immediate reward but also the next selected action and through that all subsequent rewards. These two characteristics, trial-and-error search and delayed reward, are the two most important features of reinforcement learning.

In the standard reinforcement learning model, described in Figure 1, an agent is interconnected to its environment via perception and interaction. On each interaction step the agent receives some information regarding the current state of the environment. The agent then chooses an action to generate as output. The action changes the state of the environment and the value of this state transition is communicated to the agent through a scalar reinforcement signal. The agent's desired behavior is to choose actions that tend to increase the long-term sum of reinforcement signals. An action selection policy is used by the reinforcement learning agent in order to choose the appropriate action to change the current state. The agent must find a trade-off between immediate and long-term rewards. It must explore the unknown states, as well as the states that maximize the reward based on current knowledge. A balance between the exploration of unknown states and the exploitation of known, high reward states is needed. There is a wide variety of reinforcement learning algorithms: SARSA, TD-Gammon, SARSA(λ) [1], Q-Learning [3], etc.

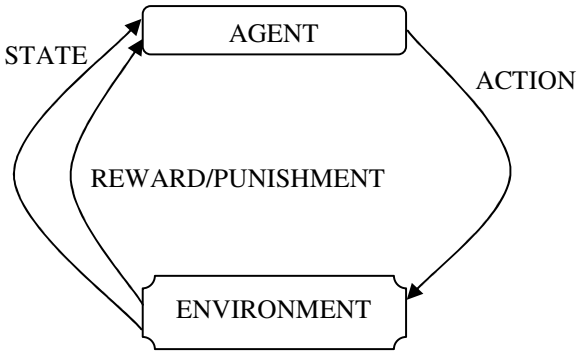


Figure 1. The components of a reinforcement learning agent

In this paper, an edge detector is automatically parameterized using the popular Q-Learning algorithm [3][4][5], which is a reinforcement learning technique that memorizes an action-value function and follows a fixed policy thereafter. One of the biggest advantages of Q-Learning is that it is able to compare the expected rewards without requiring a prior well known model of the environment. One of the policies used by Q-Learning is the Boltzman policy [3], which estimates the probability of taking an action with respect to a certain state. Other well known policies for Q-Learning are the ε-greedy and greedy policies [5]. In the greedy policy, all actions may be explored, whereas the ε-greedy selects the action with the highest Q-value in the given state with a probability of $(1 - \varepsilon)$ and the rest with probability of ε. For the approach proposed, in this paper a greedy policy was used.

Reinforcement learning serves both as a theoretical tool for studying a way entities learn to act under a certain environment as well as a practical computational tool for constructing autonomous systems that improve themselves based on cumulated experience [6][7]. These applications range from robotics and industrial manufacturing to

 ISBN: 978-1-61208-195-3

combinatorial search problems such as computer game playing [8][9][10]. Practical applications provide a test for efficiency and utility of learning algorithms. They are also an inspiration for deciding which components of the reinforcement learning framework are of practical importance.

Image processing represents one of the particular applications of the more general field of two-dimensional signal processing. Due to multimedia market explosion, digital image processing has become an interest area and attracts a lot of researchers from other computer science fields.

Modern digital technology has made it possible to manipulate multi-dimensional signals with systems that range from simple digital circuits to advanced parallel computers.

Digital image processing methods can be classified as follows: image processing (i.e., a bundle of operations such as filtering, labeling, segmentation, object detection, etc.), image analysis (i.e., a bundle of all measurements based on the information contained in the image itself, e.g., PSNR) and image understanding (i.e., collected regions and objects from an image are interpreted and classified, based on their content, for future processing). From these, image processing and image understanding involve solving decision making problems.

Decision making for image processing is a cumbersome process. Because an image's data may be large and highly redundant, it can be subject to multiple interpretations. In order to reduce the dataset size used in image processing, feature extraction methods are used. The use of certain features, that select the relevant information from the input data, simplifies the processing tasks and reduces the resources' costs. Feature extraction is used for a large range of image processing algorithms: face recognition, segmentation, image database indexing and retrieval, watermarking, medical imaging, image searching, etc. Some works divide the used features into different groups: color features, texture features and shape features [11].

Some of the most commonly used features are object size (area, volume, perimeter, surface), object shape (Fourier descriptors, invariant moments, shape measurements, skeletons), object color (space, integrated optical density, absolute and relative colors), object appearance/texture (co-occurrence matrices, run lengths, fractal measures, statistical geometric features), distribution function parameters (moments: mean, variance, median, inter-quartile range) [11]. The main issues with feature extraction algorithms are lack of flexibility and lack of adaptability with respect to input images and user requirements.

The rest of this paper is organized as follows. Section 2 presents a brief overview of reinforcement's learning success in different fields. In Section 3, a short overview of reinforcement learning usage in Image Processing and Computer Vision is reviewed. In Section 4, the proposed solution for automatic parameterization using Q-Learning shows the integration of reinforcement learning with a basic image processing technique (i.e., Sobel edge detector). Section 5 depicts the obtained results and the encountered implementation problems. The final section elaborates on the reached conclusions.

II. DIVERSITY OF REINFORCEMENT LEARNING APPLICATIONS

Current research reveals the potential of reinforcement learning techniques, in distributed environments with a high level of dynamism, for resources' allocations that induce a balanced utilization of the system's capabilities and maximizes the number of served applications [8]. Due to the fact that reinforcement learning has low scalability and the performances from the online training can be extremely reduced, a hybrid method of reinforcement learning and neural networks was proposed in [12] to address this issue.

In [13], a robot navigation model based on environment reference points is proposed. Some of the subcomponents of the navigation system are competing for the image acquisition system. Busquets et al. [13] use an auction based mechanism for solving competition and a reinforcement learning based method for tuning the auction functions' parameters. A doubled performance was achieved compared with the case of manually coding a navigation policy.

Reinforcement learning is useful in applications for control and administration systems as well. An implementation model for memory controllers is proposed in [14] based on the idea of self-optimization. The model's goal is to achieve efficient utilization of the transmission bus between a DRAM memory and a processor unit. Ipek et al. [14] use a reinforcement learning based implementation of an adaptive self-optimizing memory controller able to plan, learn and permanently adapt to processing requests. Ipek's et al. results show a performance improvement around 15%-20% over one of the best planning policy – First-ready first-come-first-serve.

Reinforcement learning techniques are successfully used in different medical areas such as medical imaging [15][16], individualization of pharmacological anemia management [17] or brain stimulation strategy for the treatment of epilepsy [18]. Usually, the information gathered from medical processes forms an excellent dataset for reinforcement learning. For this reason, more and more methods are being developed for decision making processes, feature extraction, separation policy's confidence interval calculation, etc.

The reinforcement learning paradigm was successfully used in autonomic systems for human interaction. Such an example, of human subject interaction, is a system for establishing a verbal dialogue with human subjects [19]. Henderson et al. [19] propose a hybrid learning method based on reinforcement learning and supervised learning (SL), which produced better results compared to other dialogue systems (40%). Just using SL offers a 35% improvement, where as just using reinforcement learning achieves performance under the level of other systems.

 ISBN: 978-1-61208-195-3

III. REINFORCEMENT LEARNING INTEGRATION IN IMAGE PROCESSING AND COMPUTER VISION

In the work of Sahba et al. [16], a reinforcement learning agent based segmentation technique is proposed for medical imaging. Due to a high level of noise, missing or diffuse contours and poor image quality, it is quite difficult to apply segmentation onto medical images.

The reinforcement learning agent's goal is to receive the highest reward by discovering the optimal parameters for each processing stage. The algorithm proposed by Sahba et al. consists of two stages: an offline stage and an online stage. During the first stage, the agent acquires knowledge – stored as a matrix – using a training set of images and segmentations of these images obtained by other methods. The behavior learned during the offline stage is applied by the reinforcement learning agent during the online stage in order to segment images from a test set. The algorithm implemented by Sahba et al. in [16] is a basic segmentation approach. The authors clearly state that their approach should not be compared with existing segmentation methods as the proposed method is just a prototype meant to show the possibility of using reinforcement learning with segmentation techniques.

Another approach for the same image segmentation methods is presented in [20]. In this paper, the results of segmentation are used in a complex object recognition algorithm. Current segmentation and object recognition computer vision systems are not robust enough for most real-world applications. In contrast, the system proposed in [20] achieves robust performance by using reinforcement learning to induce a mapping from input images to corresponding segmentation parameters. Using the reinforcement learning paradigm in an object recognition system is the major contribution in [20]. Through their research, Peng and Bhanu [20] show that, for a set of images, adaptive systems could be used for autonomous extraction of the segmentation criteria and for automatic selection of the most useful characteristics. The result is highly accurate recognition system.

The results of using reinforcement learning in the field of image processing do not stop to segmentation mechanisms. Decisions processes are an important part of edge detectors algorithms as well. In [21], the authors show the experimental results obtained by using a reinforcement learning based neural network for an edge detection algorithm. Similar to standard edge detection algorithms, a part of the image is used as input, in this case for the neural network. A key difference from standard edge detection algorithms is that the neural network doesn't use specific per-image parameters. The experimental results were compared with results from Canny and Sobel edge detection operators. The comparison underlines advantages in accuracy and efficiency.

Reinforcement learning is also used in text detection applications. In [22], ten parameters of text detection in video images are optimized using reinforcement learning.

Reinforcement learning was successfully used to discover the parameters needed in an image retrieval system. The research carried out by Srisuk et al. [23] uses a template comparison mechanism together with a reinforcement learning agent endowed with a Bernoulli policy in order to extract the necessary parameter for such a system.

The main issue with all methods depicted in this section is to eliminate the need of ground truth-based images. Although useful when evaluating the proposed frameworks, their use may be questioned when using online learning methods. In conclusion, the rewards must come directly from result evaluation (e.g., an optical character recognition (OCR) engine, image quality measurements). This will allow for online parameter optimization, which fully utilizes the benefits of reinforcement learning.

In some cases, the rewards come from the system's user. The work presented in [24] describes an approach to a custom image search engine, designed for e-learning. The proposed system takes the user's response as the reward for a reinforcement learning agent. The agent selects a set of images and suggests them to the user. The user chooses from this set the ones which bare any significance to it's search. As a result of this interaction the system learns a user's profile. On a future search the agent will compare the learned set with the returned result set and will chose only those images that match the user's preferences. In order to achieve a seamless interaction between the proposed system and the end users, the application in [24] is endowed with a graphical user interface as well.

A reinforcement learning agent is also used in face recognition algorithms [25] or in multiple filtering applications [26]. Most of all, applications for medical imagining and image retrieval were the first to use reinforcement learning due to the fact that a lot of bad quality images have to be processed, using variable parameters and human feedback.

IV. THE PROPOSED SOLUTION OF AUTOMATIC PARAMETERIZATION USING Q-LEARNING ALGORITHM FOR EDGE DETECTION

An edge detector is a simple algorithm to extract edges from an image. Applying an edge detector to an image generates a set of connected curves which follow the boundaries of objects (real objects, surfaces, texture changes, etc.). Edges play quite an important role in many image processing applications, particularly for machine vision systems that analyze scenes of man-made objects.

Most edge detection methods use either first-order or second-order derivative. The first order derivative of a point is considered to give a measure of the fortitude of an edge at that point while its local maximum's direction estimates the edge's local direction. The Sobel, Canny and Prewit edge detectors are just a few examples using the first-order derivative.

With the second-order derivative edge detection methods search for zero crossings which are local maxims of the gradient, usually the zero-crossings of the Laplacian or the zero-crossings of a non-linear differential expression (e.g., Marr-Hildreth) [27]. Most edge detection methods are based on a threshold that says how large the intensity change between two neighboring pixels must be in order to say that

 ISBN: 978-1-61208-195-3

there should be an edge between these pixels. Most of the times, the threshold depends on the image, its noise level or its details. As a result, the threshold is chosen manually, which makes these methods difficult to apply on large image datasets. The advantages of existing methods are simplicity and fast computing. Since they are part of other complex algorithms most of the times, it's important they are as fast and independent as possible.

In this section, an automatic parameterization for the Sobel edge detector is proposed (Figure 3). The application was developed using our in-house built Q-Learning framework, through which we try to attain more flexibility and scalability with our exploration of Q-Learning based applications. The general architecture for the Q-Learning framework is depicted in Figure 2.

The actual framework application is represented on the Framework Layer section and contains the following major components: a general learning agent based on the Reinforcement Learning paradigm (RL Agent), the Q-Learning implementation (QL Impl), a dictionary data structure (Policy) containing the parameterization for the Q-Learning algorithms – α (learning rate) and γ (discount factor), an extensible set of objective functions used when updating the Q-values. The usual updating function is the *maximum* function, as stated in the general theory – see equation (2), yet other functions could be used, e.g., *minimum*, *summation*, *mean*. Based on the previous components, the most important element of our framework is defined and implemented: the Q-Learning agent (QL Agent).

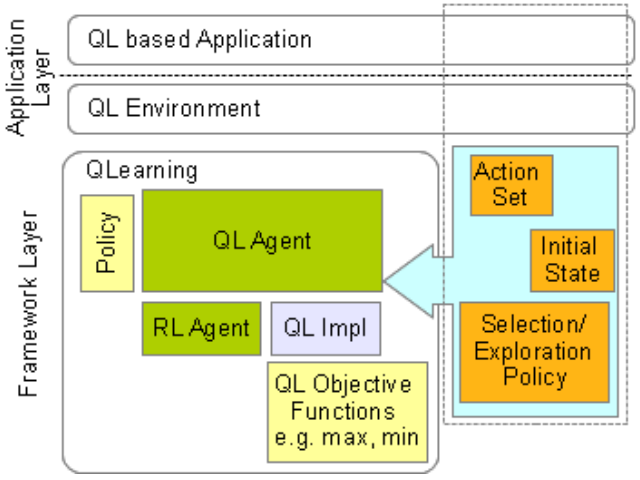


Figure 2. Q-Learning framework architecture

In order for the agent to interact with the environment, it needs to know a set of actions, start from an initial state and be endowed with an exploration/exploitation strategy. These latter components cannot be implemented in the framework and they are left as abstract elements, since they are tightly coupled with the application and environment particularities. This is suggested, in the architecture depicted in the Figure 2, by enclosing the action set, initial state and exploration policy into a dashed square that protrudes into the application layer.

The environment is both part of the framework as well as the application in the following sense:

- the way the Q-Learning agents are initialized, introduced and interlinked in/with the environment induces certain capabilities that the environment component must support (i.e., agents are an active part of the environment);
- each application depends on and exposes a certain environmental universe.

Using these components, the upper applications' instances are implemented based on the problem they try to approach, e.g., in this paper we used the framework to implement the application for finding an optimal edge detection threshold.

The model depicted in Figure 3, for the proposed automatic parameterization application, contains a Sobel edge detector, an evaluation module based on Pratt Figure of Merit [28] and a reinforcement learning agent that learns to find the optimal edge detection threshold for multiple images. A greedy policy Q-Learning algorithm is used. States are represented as values for the searched parameter (i.e., the edge detection threshold), the only two possible actions are increment and decrement of state's value and rewards are computed using Pratt Figure of Merit [28]. The Sobel edge detector extracts edges from the input images using a threshold value provided by the reinforcement learning agent. The extracted edges are assessed by the evaluation module and the result is used as a reward for the reinforcement learning agent, which takes the corresponding action and provides another edge detector threshold. The learning process continues until the optimal threshold is found.

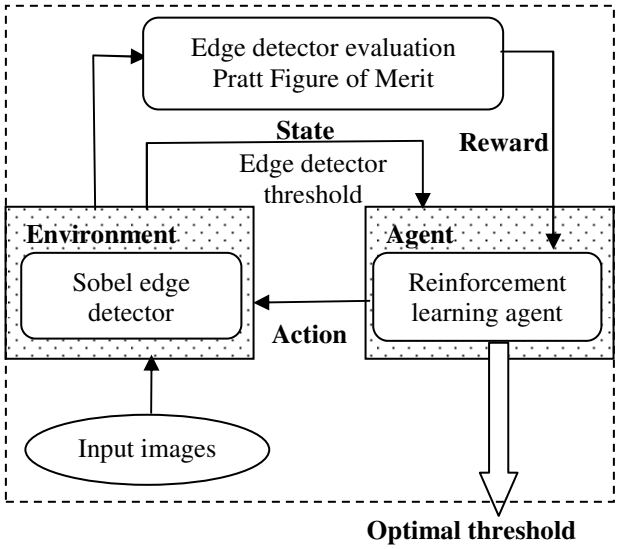


Figure 3. The proposed automatic parameterization for the Sobel edge detector

Pratt Figure of Merit is a well-known measure used to compare edge detectors' outputs. It attempts to balance three types of errors that can produce erroneous edge mappings: missing valid edge points, disconnected edge points and

 ISBN: 978-1-61208-195-3

misclassification of noise fluctuations as edge points. Pratt Figure of Merit uses a known edge for comparison. We consider a known edge the edge resulted from the output of a Canny edge detector based on a manual chosen threshold. For different edge detection thresholds, we visually analyze the resulting edges for the Canny detector and we chose the threshold that achieves the optimal possible edge.

The Figure of Merit is defined as:

$$R = \frac{1}{I_N} \sum_{i}^{I_A} \frac{1}{1 + \alpha d_i^2} \quad (1)$$

In the above equation, I_N is the maximum of I_A and I_I. The I_A value represents the total number of test edge pixels. The I_I value represents the total number of known edge pixels in the image. α is a scaling constant, while d_i is the distance from an actual edge point to the nearest known edge point. The scaling factor is used to penalize edge points that are clustered but away from the known edge pixels. After experimenting with other values we decided to use $\alpha = 0.9$, the value coined by Pratt in [28], a value that drastically penalizes pixels faraway from the known edge.

Q-Learning algorithm updates its Q-values using:

$$Q(s_t, a_t) = Q(s_t, a_t)(1 - \alpha) + \alpha [r_{t+1} + \gamma^{\max}_{a} Q(s_{t+1}, a)] \quad (2)$$

where r_t is the reward given at time t, α ($0 < \alpha <= 1$) the learning rate, may be the same value for all pairs. The discount factor γ is such that $0 \leq \gamma < 1$.

In this paper, a greedy policy is used. We use $\alpha = 1$ and $\gamma = 0.85$. The learning rate is used to establish the influence of the new acquired information against past experiences. A small factor will make the agent learn very little and rely solely on past experience – acquired or given as input premises, while a big factor would make the agent consider only the most recent information. The discount factor establishes the impact of future rewards. A null factor will make the agent consider just current rewards, while a big factor will make it achieve a high long-term reward.

V. EXPERIMENTAL RESULTS

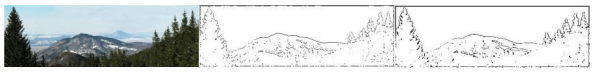
During the offline stage, the reinforcement learning agent is trained using three input images. In accordance with the Q-Learning algorithm, a total of 512 Q-values are computed during this stage with respect to each reward received for each taken action. For this stage the actions and the initial state are randomly selected. During the online stage we use the computed Q-values to determine the optimal threshold necessary for the edge detection algorithm. Starting from a randomly selected initial state, the agent chooses the action that will give a maximum reward.

The obtained optimal threshold is 56. This threshold can only be evaluated by analyzing the results produced by basic edge detectors with respect to it. We visually test the threshold by using it with Canny and Sobel edge detectors first on the training images (Figure 4b, 4c, 5b, 5c, 6b, 6c) and then on a test image (Figure 7). We used different kinds of real life images (landscapes, cartoons, portraits) to test the optimal threshold in vary conditions. Because edge detection methods are widely used in algorithms which address natural as well as artificial, medical and/or binary images, they must prove efficient in real life working scenarios.

One can notice that the edges, obtained using the automatically computed threshold, are continuous and have a small number of offset pixels. It can also be observed that the edges are not smeared.

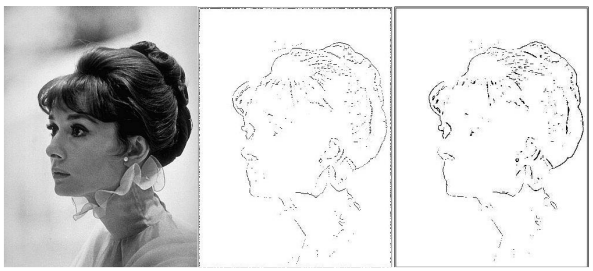
a b c

Figure 4. a. Original image. b. result of edge detector Canny with a 56 threshold. c. result of edge detector Sobel with a 56 threshold



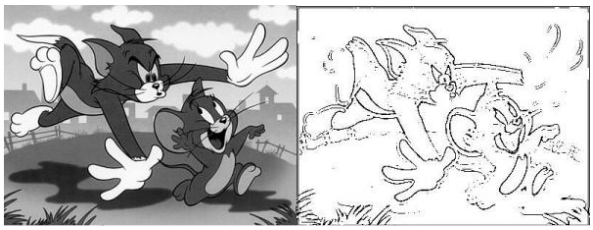
a b c

Figure 5. a. Original image. b. result of edge detector Canny with a 56 threshold. c. result of edge detector Sobel with a 56 threshold



a b c

Figure 6. a. Original image. b. result of edge detector Canny with a 56 threshold. c. result of edge detector Sobel with a 56 threshold

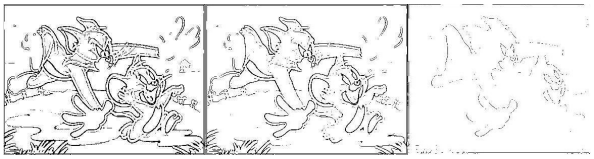


a b

Figure 7. a. Original image. b. result of edge detector Sobel with a 56 threshold

For additional evaluation of the detection threshold, we applied the Sobel algorithm to a test image and manually

selected the optimal threshold value. The obtained results are depicted in Figure 8. The optimal value chosen for the threshold was 52, a much close value to the one automatically achieved using Q-Learning algorithm.



a b c

Figure 8. Result of edge detector Sobel with a a. 29 threshold. b. 52 threshold c. 173 threshold

The proposed parameterization is an example of how reinforcement learning can be used to find optimal parameters for image processing algorithms. The presented results indicate that this solution can be efficient. Computing automatically the detection threshold for an edge detector provides the added benefit of being independent from the input image. As such, it can be used on multiple images with comparable performance. Although the threshold computed by our method may not be optimal for every image, it is sufficiently close enough to the optimal value. As such, the computed threshold provides at least a very good hint to what that optimal value should be.

The proposed algorithm is limited by the use of ground truth images as reference, images that must be known prior. For computing the set of rewards we used a set of reference contours extracted with the Canny detector. For natural images though, choosing the optimal threshold value by hand can be quite difficult to achieve. This issue could be solved by using an interactive system such that the reward, for a specific action, is given as a user's input.

VI. CONCLUSIONS

The paper's goal was twofold: to give a general overview of reinforcement learning as an optimization technique and to ascertain an insight over the benefits that can be drawn for decision making problems in the image processing field.

A short review of current research and the success of reinforcement learning techniques in various fields were presented and we studied the use of a Q-Learning approach to decision making with respect to an image edge detection problem.

We developed a reinforcement learning framework application using the Python programming language. Based on this framework the integration of Q-Learning algorithm was developed for the automatic parameterization of Sobel edge detector.

The obtained results show the potential of the proposed approach, while strongly indicating an improvement in speed, resources' utilization and usability as opposed to Sobel method.

We implemented the threshold discovery agent such that it learns - through the Q-Learning algorithm - to find an optimal threshold by using a greedy policy and a set of reference images. The results are good for test images of the same type as the references (e.g., nature, synthetic images, cartoons, etc.).

The use of reinforcement learning in image processing and computer vision is extending as a way of replacing human assistance with intelligent agents.

ACKNOWLEDGMENT

The work has been co-funded by the Sectoral Operational Program Human Resources Development 2007-2013 of the Romanian Ministry of Labour, Family and Social Protection through the Financial Agreement POSDRU/89/1.5/S/62557.

REFERENCES

[1] R. Sutton and A.G. Barto, "Reinforcement learning: An introduction", The MIT Press, London, England, 2005, <retrieved: February, 2012>.

[2] L.P. Kaelbling and M.L. Littman, "Reinforcement learning: A survey", Computer Science Department, Brown University Providence, USA, 1996, <retrieved: February, 2012>.

[3] L. Busoniu, R. Babuska, and B. De Schutter, "Multi-agent reinforcement learning: An overview", Innovations in Multi-agent systems and applications, Studies in Computational Intelligence, Issue 310, 2010, pp. 183-221, <retrieved: February, 2012>.

[4] E.C. Mariano, P. Cuauhnahuac, and E.F. Morales, "Distributed reinforcement learning for multiple objective optimization problems", Congress on Evolutionary Computation, 2000, pp. 188 – 195, <retrieved: February, 2012>.

[5] T. G. Dietterich, "An overview of MAXQ hierarchical reinforcement learning", Proc. 4th Symposium on Abstraction, Reformulation and Approximation (SARA 2000), Lecture Notes in Artificial Intelligence, New York, 2000, pp. 26–44.

[6] F. Samreen and M. Sikandar Hayat Khiyal, "Q-Learning scheduler and load balancer for heterogeneous systems", Journal of Applied Sciences, 2007, pp. 1504 – 1510.

[7] M. Launer and M. Riedmiller, "An algorithm for distributed reinforcement learning in cooperative Multi-agent systems", Proc. 17th International Conference on Machine Learning (ICML 2000), 2000, pp. 535 – 542.

[8] S. M. Thampi and C. Sekaran, "Review of replication schemes for unstructured P2P networks", Proc. IEEE International Advance Computing Conference (IACC 2009), 2009, pp. 194 – 800, <retrieved: February, 2012>.

[9] A. Galstyan, K. Czajkowski, and K. Lerman, "Resource allocation in the grid using reinforcement learning", Proc. 3rd International Joint Conference on Autonomous Agents and Multi-agent Systems (AAMAS 2004), vol. 3, 2004, pp. 1314 – 1315, <retrieved: February, 2012>.

[10] K. Verbeeck, J. Parent, and A. Nowé, "Homo egualis reinforcement learning agents for load balancing", Proc. Workshop on Radical Agent Concepts (WRAC 2002), 2002, pp. 81-91.

[11] P. Vallotton, "Image analysis – Feature extraction", Commonwealth Scientific and Industrial Research Organisation (CSIRO 2008) Mathematical and Information Sciences, Australia, 2008.

[12] G. Tesauro, N.K. Jong, R. Das, and M.N. Bennani, "On the use of hybrid reinforcement learning for autonomic resource allocation", Cluster Computing 10(3), 2007, pp 287-299.

[13] D. Busquets, R.L. de Mantaras, C. Siera, and T.G. Dietterich, "Reinforcement learning for landmark-based robot navigation", Proc. 1st International Joint Conference on Autonomous Agents and Multi-

 ISBN: 978-1-61208-195-3

agent Systems: part 2 (AAMAS 2002), 2002, pp. 841 – 843, <retrieved: February, 2012>.

[14] E. Ipek, O. Mutlu, J.F. Martinez, and R. Caruana, "Self-Optimizing memory controllers: A reinforcement learning approach", Proc. 35th International Symposium on Computer Architecture (ISCA 2008), 2008, pp. 39 – 50, <retrieved: February, 2012>.

[15] S.B. Magalhaes, B. Netto, V.C. Rodrigues, A.S. Correa, A. Cardoso de Paiva, and N. Areolino de Almeida, "Application on reinforcement learning for diagnosis based on medical image", Reinforcement Learning, cap 20, I-Tech Education and Publishing, 2008, pp. 379 – 398, <retrieved: February, 2012>.

[16] F. Sahba, H.R. Tizhoosh, and M. Salama, "Application of reinforcement learning for segmentation of transrectal ultrasound images", Biomed Central Medical Imaging, 2008.

[17] A.E. Gaweda, K.M. Muezzinoglu, and G.R. Aronoff, "Individualization of pharmacological anemia management using reinforcement learning source" Special issue: Joint Conference on Neural Networks (IJCNN 2005), 2005, pp. 826 – 834.

[18] A. Guez, R.D. Vincent, M. Avoli, and J. Pinea, "Adaptive treatment of epilepsy via batch-mode reinforcement learning", The Association for the Advancement of Artificial Intelligence (AAAI 2008), 2008, pp. 1671-1678, <retrieved: February, 2012>.

[19] J. Henderson, O. Lemon, and K. Georgil, "Hybrid reinforcement/ supervised learning of dialogue policies from fixed data sets", Computational Linguistics, vol. 34, Issue 4, 2008, pp. 471 – 486, <retrieved: February, 2012>.

[20] J. Peng and B. Bhanu, "Closed-Loop object recognition using reinforcement learning", IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 20, Issue 2, 1998, pp. 139 – 154, <retrieved: February, 2012>.

[21] N. Siebel, S. Grunewald, and G. Sommer, "Created edge detectors by evolutionary reinforcement learning", Evolutionary Computation, IEEE World Congress on Computational Intelligence (WCCI 2008), Hong Kong, 2008, pp. 3553 – 3560, <retrieved: February, 2012>.

[22] G.W. Taylor and C. Wolf, "Reinforcement learning for parameter control of text detection in images from video sequences", Proc. International Conference on Information and Communication Technologies (ICICT 2004), Lyon, 2004, pp. 517 - 518, <retrieved: February, 2012>.

[23] S. Srisuk, R. Fooprateepsiri, M. Petrou, and S. Waraklang, "A general framework for image retrival using reinforcement learning", Proc. Image and Vision Computing (IVC 2003), New Zealand, 2003, pp. 36 - 41, <retrieved: February, 2012>.

[24] M. Shokri, H. Tizhoosh, and M. Kamel, "Reinforcement learning for personalizing image search", LORNET Annual E-Learning Conference on Intelligent Interactive Learning Object Repositories, 2006, <retrieved: February, 2012>.

[25] M. Harandi, M. Ahmadabadi, and B. Araabi, "Face recognition using reinforcement learning", Proc. International Conference on Image Processing (ICIP 2004), vol. 4, 2004, pp. 2709 – 2712, <retrieved: February, 2012>.

[26] F. Sahba, H.R. Tizhoosh, and M. Salama,"Using Reinforcement Learning for filter fusion in image enhancement", Proc. Computational Intelligence (CI 2005), 2005, pp. 262 – 266.

[27] D.K. Sharma, L. Gaur, and D. Okunbor, "Image compression and feature extraction using Kohonen's self-organizing map neural network", Journal of Strategic E-Commerce, 2007, <retrieved: February, 2012>.

[28] W.K. Pratt,"Digital image processing", John Wiley and Sons, New York,1991.

 ISBN: 978-1-61208-195-3

A Video Semantic Annotation System Based on User Attention Analysis

Jin-Young Moon and Changseok Bae
BigData Software Research Laboratory
Electronics and Telecommunications Research Institute
Daejeon, KOREA
{jymoon, csbae}@etri.re.kr

Wan-Chul Yoon
Department of Industrial and Systems Engineering
Korea Advanced Institute of Science and Technology
Daejeon, KOREA
wcyoon@kaist.ac.kr

***Abstract*—Automatic semantic annotation of videos is a crucial to the success of video search and summarisation based on content semantics. In contrast to broadcast news and sports, automatic semantic annotation for non-commercial videos generated by ordinary people suffers from lack of semantic data like subtitles and webcast text. It is, however, impracticable to expect that most video shooters or owners of the non-commercial contents do semantic annotation of their videos by using annotation tools manually before video dissemination. This paper proposes a video semantic annotation system that automatically analyzes and annotates a video element with the user attention state. The attention state includes the attention target, attention degree and emotional states by using gaze and Electroencephalography data from a user watching the video. To show the benefits achieved by the proposed system, the paper describes a promising application scenario of video summarisation using semantic annotations based on user attention. The use of annotations generated by the proposed system enables the summarisation system to enrich the possible summary types.**

Keywords-video annotation; semantic analysis; user attention; human factors; video summarisation

I. INTRODUCTION

Due to the huge rise in smartphone usage, shooting and sharing videos by ordinary people have been dramatically increasing nowadays. For example, the number of videos uploaded on YouTube has increased from 35 hours per minute in 2010 to 48 hours per minute in 2011. The number of videos uploaded in two months exceeds that of videos created by big three U.S. television networks (ABC, CBS, and NBC) in six decades [1]. Therefore, automatic or semi-automatic semantic annotation of videos that assigns semantics to various video elements, which can be a whole video, a scene, a shot, an objects or a regions, without a large burden to users, is vital to success of semantic video search and reconstruction among videos. Using the semantic annotation, the videos can be effectively shared by social media or re-created by the users to satisfy their personal needs.

Several techniques have been proposed to extract semantics automatically from broadcast news and sports by combining semantic data like subtitles, text from score box and webcast text [2]-[8]. It is infeasible not only to apply these techniques to non-commercial videos with the lack of those kinds of semantic data as it is but also to expect ordinary people who generate videos to do voluntarily semantic annotation of the videos by using semantic annotation tools before the users upload the video to social media, like YouTube or Facebook.

Therefore, we propose a video semantic annotation system that automatically assigns semantics related to the state of user attention to a specific object or region contained in a video, which a user focuses on, while the user watches the video. The proposed system analyzes an attention target by calculating the gaze position and recognizing the attended target, the attention degree, and emotional state by processing Electroencephalography(EEG) data from an EEG headset. The proposed system generates an annotation element on an attended target per user gaze together with the attention degree and emotional state when the user is attending to the target.

The rest of this paper is organized as follows. Section II provides an overview of the related work. In Section III, we show an overall architecture of the proposed system and describe main functionalities of system modules individually. Section IV focuses on an application scenario adopting the proposed system in order to show its feasibility and anticipated benefits. Finally, we conclude this paper with future work.

II. RELATED WORK

There have been proposed several manual semantic annotation tools for videos. They were introduced in detail and summarized in [6]. The tools enable users to do annotation delicately on the whole video, video segment during a specific time interval, a frame at a specific time, or a region in the form of a rectangle or a polygon within a frame in order to express its concept or the relationship with other elements. They are, however, inadequate for ordinary users because annotating with the manual tools is a difficult and time-consuming process to them without a definite advantage. Therefore, only automatic or semi-automatic semantic annotation techniques will be presented for consideration. In addition, we also examine user-related semantics like emotion and attention as well as content-related semantics.

A. From the perspective of content-related semantics

Extracting semantics for any videos only by computer vision and image processing is very challenging up to the present. Therefore, there have been some proposed

 ISBN: 978-1-61208-195-3

techniques to detect events and do annotation automatically for sports and news videos because those kinds of videos have relatively affluent source of content-related semantics, for example video/audio channels and webcast text. Liu *et al.* [2] made use of periodicity of the video shot content and audio keywords of a racket sports video to detect rally events. Xu and Zhang *et al.* [3][4] and Refaey *et al.* [5] utilized a webcast text feature as well as video/audio features to increase the accuracy rate of event detection in soccer and baseball videos. Bagdanov *et al.* [6] created subtitles by using linguistic and dynamic visual ontologies with reasoning for a soccer video. Messina *et al.* [7] segmented a story by speaker and shot clustering and classified its subjects by using speech-to-text in the audio/video stream of news videos. Mezaris *et al.* [8] proposed a fusion technique to combine visual, audio and text analysis results. However, the proposed techniques are insufficient for most non-commercial videos, which do not follow standard patterns, like the rules of the games, and do not contain easily extractable semantic data, for example score boxes, subtitles and headlines, which are located in given positions in a sports or news video.

B. *From the perspective of user-related semantics*

Both limitation of extracting content-related semantics and necessity of user-related semantics for personalized contents search gave rise to work on extracting on user-related semantics and utilizing them on video reconstruction like emotion annotation on videos and user response-based video summarisation. While watching videos, users show their feelings unconsciously through facial expressions, eye movement and brain waves as a response of the video. Joho *et al.* [9] devised three pronounced levels of facial expression and the change rate of the expressions by analyzing a recorded video of users for affective video summaries. Money and Harry [10] analyzed physiological user responses including electro-dermal response, respiration amplitude, respiration rate, blood volume pulse, and hear rate. They revealed that most entertaining sub-segments within a video bring on an intense response of a user. Peng *et al.* [11] proposed Interest Meter to measure the viewing interest of a user by analyzing facial expressions, saccadic movements, blink, and head movement of the user for home video summarisation. Davis *et al.* [12] recorded EEG signals from EEG headsets as an emotional user feedback of video content related to explicit user tagging. Although the analyzed user-related semantics in [9]-[12] can be used for affective video summarisation, they are unable to be widely used for video reconstruction because the user state has no relationship with content-related semantics.

III. THE PROPOSED SYSTEM

To overcome shortcomings of both content-related and user-related semantics extraction, this paper proposes a video semantic annotation system that automatically analyzes user attention and annotates the recognized attention target within a video with the degree of the user attention and the emotional states. The system is largely composed of the module of annotation generation that analyzes user response related to attention and the module of annotation provision

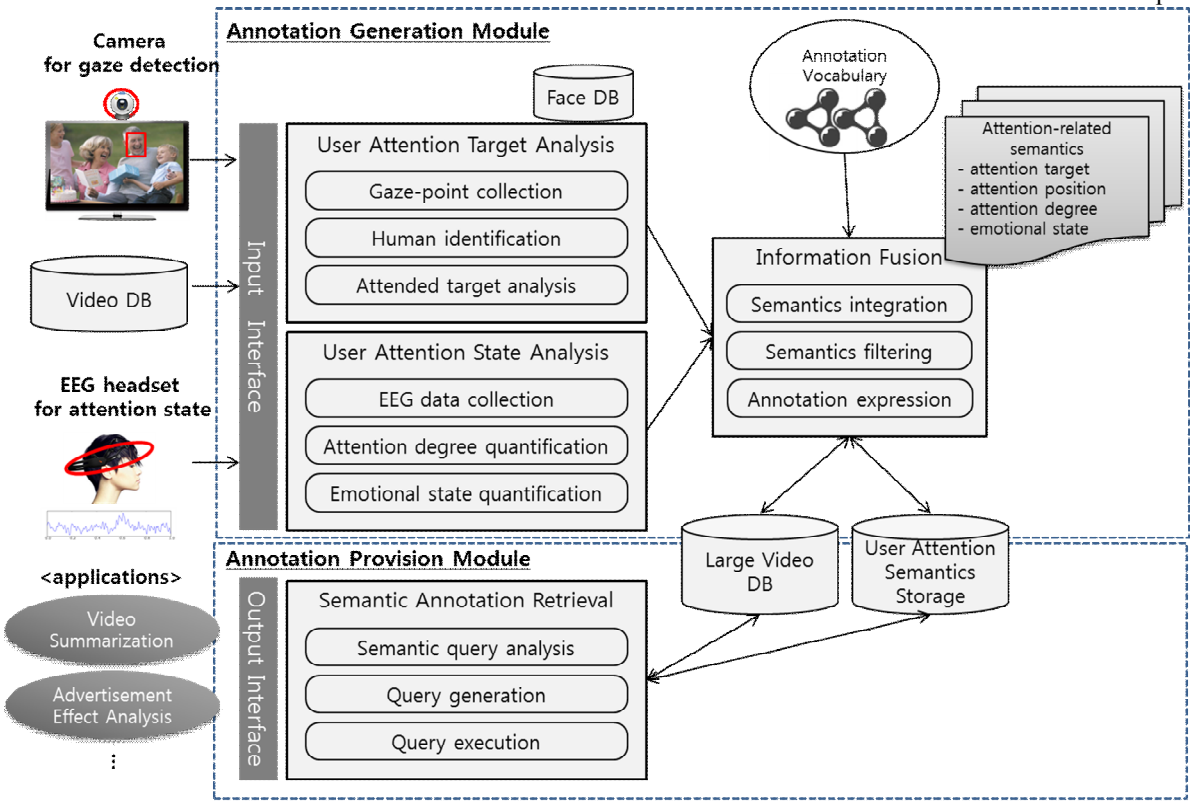


Figure 1. Architecture of the proposed system

module that provides a retrieval method of semantic annotation, as depicted in Figure 1.

The system selects the gaze point measured from a camera observing a user and the attentional and emotional user states analyzed by EEG data from a commercial EEG headset. There are other various user-related semantics, which include eye movement data, like fixation, saccadic movement and blink, and facial expressions recorded by a camera to observe the user. There are three reasons to use both gaze and brainwaves to analyze user attention. First, gaze and brainwaves reflect the delicate mental state of a user. For example, the emotion state of a user is not shown clearly in the face of the user generally. Second, in contrast to only using eye movement, the combined use of gaze and brainwaves can increase the accuracy of user attention analysis. Because mental imagination without external physical stimuli from videos can arouse eye movement, the other attention measure is necessary to screen the false alarm of attention data. Last of all, in contrast to only using brain waves, the combined use of gaze and brainwaves can assign a specific video element to user-related semantics by recognizing the attended target within videos.

In the module of annotation generation, the sub-modules of user attention target analysis and user attention state analysis collect raw data from a camera observing a user and an EEG headset in real-time and analyze the collected data in parallel. The sub-module of information fusion integrates attention data obtained from the two sub-modules of analysis to generate meaningful semantics related to the user attention and to filter redundant or false attention data. By using the annotation vocabularies, like ontologies or MPEG-7 metadata, the generated semantic annotations are expressed and stored with structural constructs in the annotation storage.

In the module of annotation provision, the stored annotations are provided to external applications to search for a video or a video segment according to the video semantics or to reconstruct videos, like video summarisation. The sub-module of semantic query analysis transforms user requests for annotations, annotated videos, or video segments into native queries that can be executed directly in the annotation storage.

IV. APPLICATION OF THE SYSTEM

To differentiate the effects of generated semantic annotations related to user attention combining content-related and user-related semantics from previous semantic annotation techniques related to either of them, this paper describes an application scenario that adopts the proposed system and summarizes a video or multiple videos from a video DB.

Shown in Figure 2, the application scenario is followed. A user watches a video wearing an EEG headset in front of a monitor equipped a camera observing the gaze of the user in order to generate semantic annotations by the proposed system. The every annotation includes the time-stamp to watch the video, the play-time within a video, an attended identified target of an object or a person within a frame, and the degree of attention and emotional state. We assume that the user owns the user's own face DB that contains friends or family who are shown frequently in the home video or public figures like celebrities. After automatic semantic annotation, the annotated video is analyzed to obtain overall user attention in it, like the general feelings of the video and the list of the most attended objects or people that draw interest of the user. The analyzed significant abstract information on user interest can be stored separately in the user interest DB.

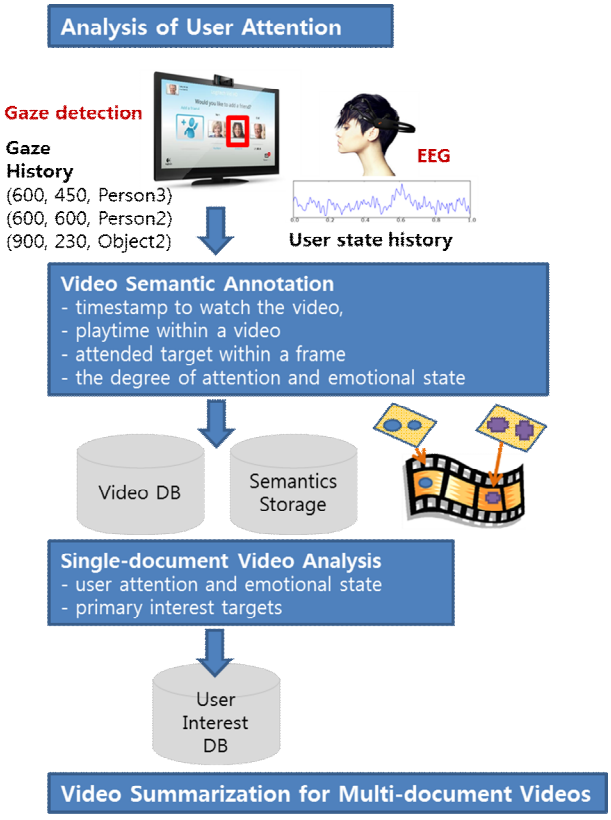


Figure 2. Application Scenario of Video Summarisation

The generated semantic annotations and analyzed information on a single video enable the video summarisation system to provide the new types of video summaries. Previous techniques for emotion annotation [9]-[12] can support summarisation of summary type 1-1 because their techniques of emotion annotations do not have any relationship with a person shown in the videos. The accumulated relationship with a person and the user state as a response corresponding to the person enriches the possible types of the video summary. First, summarisation condition can be set exquisitely by combining a person and its corresponding emotional properties like summary type 1-3. Second, the annotated video can be re-summarized at the current viewpoint, like summary type 2-1 and 2-3. Last but not least, the video can be summarized without annotating process by current interest obtained from annotations recently generated by the proposed system if some people are already recognized in advance, like summary type 3-1 and 3-3. Table 1 enumerates all the possible types of video

 ISBN: 978-1-61208-195-3

summary by adopting semantic annotations generated by the proposed system.

TABLE I. POSSIBLE TYPES OF VIDEO SUMMARY

Time	**Condition by user attention**		
	Attended target	***User attention state***	***Both of them***
$T_A \approx T_S$	summary reflecting user state when watching the video		
	Type 1-1: Containing attended people that attracted the interest of the user at that time	Type 1-2: Containing parts where the user in a particular mood with concentration at that time	Type 1-3: Containing attended people that attracted the interest of the user in a particular mood at that time
$T_A \Rightarrow T_S$	summary reflecting current user state		
	Type 2-1: Containing attractable people included in the current interesting targets	N/A	Type 2-3: Containing attractable people included in the current interesting targets in a particular mood now
only T_S	summary reflecting current user state without annotation		
	Type 3-1: Containing attractable people included in the current interesting targets if the people are identified in advance	N/A	Type 3-3: Containing attractable people included in the current interesting targets in a particular mood if the people are identified in advance

(T_A: Time to annotate, T_S: Time to summarize, $\Rightarrow$: precede in time, N/A: Not Applicable)

If the video summarisation system is used for home video that archives the daily life of a user with the user's family or special events for the family, the system can generate various video summaries of videos that were attracted at that time or are attractable now.

V. CONCLUSION AND FUTURE WORK

The proposed system was designed to analyze automatically the state of user attention with attended target, the degree of attention, and the emotional state. Compared to previous related work, the proposed system produces the relationship with a person shown in a video and its properties related to the user attention. The use of annotations generated recently by the proposed system enables the summarisation system to create a new type of a video summary from the current viewpoint. In addition, it enables the summarisation system to generate a new type of a video summary that the user has never watched if some people shown in video are identified in advance. That is how the proposed system enriches the summary types.

In the future, we will propose a novel algorithm to integrate and filter data related to user attention state and show its feasibility by the experiment of human subjects.

ACKNOWLEDGMENT

This work was supported by the Global Frontier R&D Program on <Human-centered Interaction for Coexistence> funded by the National Research Foundation of Korea grant funded by the Korean Government(MEST) (NRF-M1AXA003-2011-0028371).

REFERENCES

[1] Search Eingine Watch arcitlce, "New YouTube Statics: 48 Hours of Video Uploaded Per Miute; 3 Billion Vies Per Day", <http://searchenginewatch.com/article/2073962/New-YouTube-Statistics-48-Hours-of-Video-Uploaded-Per-Minute-3-Billion-Views-Per-Day> , 14. 02. 2012.

[2] Chunxi Liu, Qingming Huang, Shuqiang Jiang, Liyuan Xing, Qixiang Ye, and Wen Gao, "A framework for flexible summarization of racquet sports video using multiple modalities," Computer Vision and Image Understanding, Volume 113, Issue 3, pp. 415-424, ISSN 1077-3142, March 2009.

[3] Changsheng Xu, Jinjun Wang, Hanqing Lu, and Yifan Zhang, "A Novel Framework for Semantic Annotation and Personalized Retrieval of Sports Video," Multimedia, IEEE Transactions on, vol. 10, no. 3, pp. 421-436, April 2008.

[4] Yifan Zhang, Changsheng Xu, YongRui, Jinqiao Wang, and Hanqing Lu, "Semantic Event Extraction from Basketball Games using Multi-Modal Analysis," Multimedia and Expo, 2007 IEEE International Conference on, pp. 2190-2193, 2-5 July 2007.

[5] Mohammed A. Refaey, Wael Abd-Almageed, and Larry S. Davis, "A Logic Framework for Sports Video Summarization Using Text-Based Semantic Annotation," Semantic Media Adaptation and Personalization, SMAP '08. Third International Workshop on, pp. 69-75, 15-16 Dec. 2008.

[6] Andrew D. Bagdanov, Marco Bertini, Alberto Del Bimbo, Giuseppe Serra, and Carlo Torniai, "Semantic annotation and retrieval of video events using multimedia ontologies," Semantic Computing, 2007. ICSC 2007. International Conference on, pp. 713-720, 17-19 Sept. 2007.

[7] A. Messina, R Borgotallo, G. Dimino, D. Airola Gnota, and L. Boch, "ANTS: A Complete System for Automatic News Programme Annotation Based on Multimodal Analysis," Image Analysis for Multimedia Interactive Services, 2008. WIAMIS '08. Ninth International Workshop on, pp. 219-222, 7-9 May 2008.

[8] Vasileios Mezaris, Spyros Gidaros, Walter Kasper, Jörg Steffen, Roeland Ordelman, Marijn Huijbregts, Franciska de Jong, Ioannis Kompatsiaris, and Michael G. Strintzis, "A system for the semantic multimodal analysis of news audio-visual content," EURASIP J. Adv. Signal Process 2010, Article 47, February 2010.

[9] Hideo Joho, Joemon M. Jose, Roberto Valenti, and Nicu Sebe, "Exploiting facial expressions for affective video summarisation," In Proceedings of the ACM International Conference on Image and Video Retrieval (CIVR '09). ACM, New York, NY, USA, Article 31, 2009.

[10] Arthur G. Money and Harry Agius, "Analysing user physiological responses for affective video summarisation, Displays, Volume 30, Issue 2, pp. 59-70, April 2009.

[11] Wei-Ting Peng, Wei-Ta Chu, Chia-Han Chang, Chien-Nan Chou, Wei-Jia Huang, Wen-Yan Chang, and Yi-Ping Hung, "Editing by Viewing: Automatic Home Video Summarization by Viewing Behavior Analysis," Multimedia, IEEE Transactions on, vol. 13, no. 3, pp. 539-550, June 2011.

[12] S. Davis, E. Cheng, I. Burnett, and C. Ritz, "Multimedia user feedback based on augmenting user tags with EEG emotional states," Quality of Multimedia Experience (QoMEX), 2011 Third International Workshop on, pp. 143-148, Sept. 2011.

 ISBN: 978-1-61208-195-3

Video Retrieval by Managing Uncertainty in Concept Detection using Dempster–Shafer Theory

Kimiaki Shirahama
Graduate School of Economics, Kobe University
2-1, Rokkodai, Nada, Kobe 657-8501, Japan
shirahama@econ.kobe-u.ac.jp

Kenji Kumabuchi and Kuniaki Uehara
Graduate School of System Informatics, Kobe University
1-1, Rokkodai, Nada, Kobe 657-8501, Japan
kumabuchi@ai.cs.kobe-u.ac.jp, uehara@kobe-u.ac.jp

***Abstract*—This paper focuses on *concept-based* video retrieval which examines whether a shot is relevant or irrelevant to a query based on detection results of concepts, like *Person*, *Building* and *Car*. One key problem is *uncertainty* in concept detection. Even for state-of-the-art methods, it is difficult to accurately detect concepts present in a shot. Relying on such uncertain concept detection results degrades retrieval performance. To overcome this problem, *Dempster–Shafer Theory* (DST) is used to represent the probability that a concept is possibly present in a shot. By incorporating DST into maximum likelihood estimation, our method estimates the probabilistic distribution of concepts' presences which characterize shots relevant to the query. A preliminary experiment on TRECVID 2009 video data supports the effectiveness of our method.**

Keywords*-*Video retrieval; Concept Detection; Uncertainty; Dempster-Shafer Theory; Evidential EM Algorithm

I. INTRODUCTION

Video retrieval can be treated as a machine learning process, one that constructs a classifier for discriminating between shots that are relevant or irrelevant to a query. A large number of example shots are required to construct a classifier that can accurately retrieve relevant shots, irrespective of object appearance, environment and camera technique. However, it is impractical to prepare enough example shots to suit all possible queries. This insufficiency of example shots is a key factor in the challenging problem of the *semantic gap* between low-level features computed automatically and high-level semantics perceived by human.

To bridge the semantic gap, one promising approach is *concept-based* video retrieval which retrieves shots, where concepts (*e.g.*, *Person*, *Building* and *Car*) related to a query are detected. This approach utilizes *concept detectors* that detect the presence of a concept in a shot. These are constructed using a large number of training shots that are annotated to indicate the presence or absence of a concept. Hence, the concept can be detected robustly, irrespective of its size, position and direction on the screen. A large number of researchers reported that using such concept detection results as 'intermediate' features significantly improves retrieval performance [1], [2], [3].

Figure 1 outlines concept-based video retrieval. First of all, a shot is associated with *concept detection scores*, each of which represents the probability of a concept's presence (Figure 1 (d)). Given a query represented using text and example shots (*i.e.*, 'multimodal query' in Figure 1 (a)), concepts related to the query are selected (Figure 1 (b)). A classifier is then constructed to discriminate between relevant and irrelevant shots to the query, using detection scores of selected concepts (Figure 1 (c)).

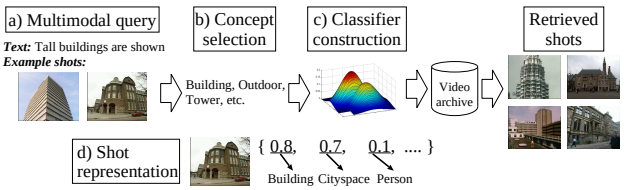


Figure 1. An overview of concept-based video retrieval.

We will now summarize the tasks necessary for concept-based video retrieval. The first is how to define a vocabulary of concepts. The most popular vocabulary is the Large-Scale Concept Ontology for Multimedia (LSCOM) [4]. LSCOM defines a standardized set of $1,000$ concepts in the broadcast news video domain. These are selected based on their 'utility' for classifying content in videos, their 'coverage' for responding to a variety of queries, their 'feasibility' for automatic detection, and the 'availability' (or 'observability') for large-sized training data.

The second task is how to select concepts related to a query. Several concept selection methods have been proposed so far. For example, concepts can be selected based on their lexical similarity to query terms and on detection scores in example shots [1], [7]. We have also developed a concept selection method using various concept relationships (*e.g.*, generalization/specialization, sibling, part-of *etc.*) defined in a knowledge base [8].

The last task that this paper addresses is how to construct a classifier that discriminates between relevant and irrelevant shots to a query. In this task, detection scores for multiple concepts are fused into a single *relevance score*, which represents the relevance of a shot to the query. However, even for most effective methods, it is difficult to accurately detect

 ISBN: 978-1-61208-195-3

any kind of concept. For example, TRECVID is an annual competition where concept detectors developed all over the world are benchmarked using large-scale video data [5]. At TRECVID 2010, the top-ranked methods achieved high performances for concepts such as *Mountain* and *Vehicle* (with average precisions greater than 0.2). On the other hand, the detection of concepts like *Bus* and *Sitting_down* was difficult (with average precisions less than 0.05). Thus, relying on such 'uncertain' concept detection results significantly degrades retrieval performance.

We introduce a method which constructs a classifier based on uncertain concept detection scores using *Dempster–Shafer Theory* (DST) [9]. DST is a generalization of Bayesian theory, where a probability is not assigned to a variable, but instead to a subset of variables. Such a probability is called a *belief mass*. We consider two variables P and A which represent the presence and absence of a concept in a shot, respectively. In addition, we consider $\{P, A\}$ which represents the uncertainty of whether the concept is present or not. Based on these variables, we define three belief masses $m(\{P\})$, $m(\{A\})$ and $m(\{P, A\})$. Here, $m(\{P\})$ and $m(\{A\})$ denote the probability that the concept is definitely present in a shot, and the probability that it is definitely absent, respectively, while $m(\{P, A\})$ denotes the probability that the concept is possibly present in the shot. By incorporating belief masses into maximum likelihood estimation, we can construct a classifier that can account for the uncertainty in concept detection.

II. Related Work

We will review existing methods for constructing classifiers based on concept detection scores. These classifiers can be roughly grouped into four categories: *linear combination*, *discriminative*, *similarity-based* and *probabilistic*. Linear combination classifiers compute the relevance score of a shot by weighting detection scores for multiple concepts. Popular weighting methods use the lexical similarity between query terms and a concept, their correlation (co-occurrence), and the detection scores of the concept in example shots [1], [7].

Discriminative classifiers consider a shot as a multi-dimensional vector, where each dimension represents the detection score of a concept. Based on this, a discriminative classifier, typically an SVM, is constructed using example shots [1], [3]. The relevance score of a shot is obtained as the classifier's output.

Similarity-based classifiers compute the relevance score of a shot as its similarity to example shots in terms of concept detection scores. Li *et al.* used the cosine similarity and a modified entropy as similarity measures [10].

Finally, probabilistic classifiers estimate a probabilistic distribution of concepts using concept detection scores in example shots, and use it to compute the relevance score of a shot. Rasiwasia *et al.* computed the relevance score as the similarity between the multinomial distribution of concepts estimated from example shots and the multinomial distribution estimated from the shot [11].

Our method constructs a classifier that is an extension of probabilistic classifiers. Specifically, ordinal probability (or Bayesian) theory cannot represent the uncertainty of a concept's presence in a shot. The only way to represent the uncertainty is to assign 0.5 to probabilities of the concept's presence and absence. Compared to this, DST can represent the uncertainty using $m(\{P, A\})$. Therefore, the representation of concept detection scores in our method is much more powerful than that of existing methods. To the best of our knowledge, such a representation has not been used in any previous methods.

III. Concept-based Video Retrieval based on Evidential EM Algorithm

In this section, we present a classifier construction method that accounts for the uncertainty in concept detection. First, we describe a method that computes the *plausibility* of a concept's presence (or absence) by combining belief masses for the concept. The plausibility represents the upper bound of probability that the concept is present (or absent) in a shot [9]. Thus, the plausibility of the concept's presence is useful for recovering false negative detections, while the plausibility of its absence is useful for alleviating false positive detections. We then present a probabilistic model based on plausibilities and *Evidential EM* (E^2M) algorithm, which estimates parameters of the model based on maximum likelihood estimation [9].

Plausibility computation based on DST: Let s_i^j be the detection score of the i-th shot ($1 \leq i \leq N$) for the j-th concept ($1 \leq j \leq M$). Based on s_i^j, we consider three belief masses $m_i^j(\{P\})$, $m_i^j(\{P, A\})$ and $m_i^j(\{A\})$, where the superscript j and the subscript i represent the j-th concept and the i-th shot, respectively. DST offers various combinations of belief masses based on set-theoretic operations. The following combination is used to compute the plausibilities pl_i^{j1} of the j-th concept's presence, and pl_i^{j0} of its absence:

$$\begin{aligned} pl_i^{j1} &= \sum_{B \cap \{P\} \neq \phi} m_i^j(B) = m_i^j(\{P\}) + m_i^j(\{P, A\}), \\ pl_i^{j0} &= \sum_{B \cap \{A\} \neq \phi} m_i^j(B) = m_i^j(\{A\}) + m_i^j(\{P, A\}) \end{aligned} \quad (1)$$

where the presence and absence of the j-th concept are represented by '1' and '0', respectively. B indicates any subset of variables overlapping $\{P\}$ or $\{A\}$. For pl_i^{j1}, $\{P\}$ and $\{P, A\}$ are referred by B as shown in the right-hand side. Thus, by defining pl_i^{j1} as the sum of $m_i^j(\{P\})$ and $m_i^j(\{P, A\})$, almost all shots where the j-th concept is present have relatively large pl_i^{j1}. The same is true of pl_i^{j0}.

To compute pl_i^{j1} and pl_i^{j0}, we extract three types of intervals using s_i^j. The first type characterizes $m_i^j(\{P\})$

 ISBN: 978-1-61208-195-3

where the number of shots annotated with the j-th concept's presence is much larger than the number of shots annotated with its absence. In the second type of interval characterizing $m_i^j(\{A\})$, the number of the latter type of shots is much larger than the number of the former type of shots. The last type of interval for $m_i^j(\{P,A\})$ does not have a distribution that is biased towards shots annotated with the j-th concept's presence or absence. However, directly estimating $m_i^j(\{P\})$, $m_i^j(\{A\})$ and $m_i^j(\{P,A\})$ is difficult, since we have no priori knowledge about their probabilistic distributions. Thus, following the spirit of DST in equation (1), we compute pl_i^{j0} and pl_i^{j1} based on the *lower and upper bounds* of s_i^j, which are defined as the minimum and maximum of the interval for $m_i^j(\{P,A\})$, respectively. Thereby, almost all shots where the j-th concept is presence have s_i^j larger than the lower bound, while almost all shots where it is absent have s_i^j smaller than the upper bound.

To implement the above idea, we construct a linear SVM using shots annotated with the j-th concept's presence and absence. In Figure 2, the former and latter cases are represented as $\times$s and $+$s, respectively, where each shot is represented using s_i^j (*i.e.*, a real number). As shown in Figure 2, we use the left and right support vectors as the lower and upper bounds. It can be considered that if s_i^j is larger than the lower bound, the probability of the j-th concept's presence in the i-th shot is at least greater than 0; that is, $pl_i^{j1} > 0$. It is also reasonable to assume that a larger pl_i^{j1} is computed for a larger s_i^j. Hence, pl_i^{j1} is computed using *Line 1* in Figure 2, where pl_i^{j1} is 0 at the lower bound and 1 at $s_i^j = 1$. Similarly, *Line 0* is used to compute pl_i^{j0}, where pl_i^{j0} is 0 at the upper bound and $1/\rho$ at $s_i^j = 0$.

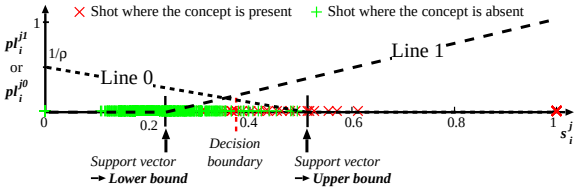


Figure 2. Illustration of the plausibility computation using support vectors.

The following two points are important for the computation of pl_i^{j1} and pl_i^{j0}. The first is that, in order for the interval between the lower and upper bounds to include shots annotated with the j-th concept's presence as well as shots annotated with its absence, we tune the SVM parameters, C^+ and C^-, which penalize mis-classification of the former and latter types of shots, respectively [12]. The second point is that since the number of shots where the j-th concept is present is much smaller than the number of shots where it is absent, putting the same priority on pl_i^{j1} and pl_i^{j0} leads to a classifier that favors the latter type of shot. Thus, pl_i^{j0} is decreased using ρ.

E^{2}M algorithm: Assume $x_i = (x_i^1, \cdots, x_i^M)$ as the vector representation of the i-th shot where the j-th dimension represents the 'complete' presence (or absence) of the j-th concept with no uncertainty, *i.e.*, $x_i^j \in \{P, A\}$. Clearly, obtaining x_i is impossible because we only have uncertain concept detection scores. The best we can do is to estimate the probability and plausibility of $x_i^j = P$ or $x_i^j = A$. The plausibility is modeled as either pl_i^{j1} or pl_i^{j0}, based on DST. We use the likelihood $L(\theta; pl_i)$ for a given pl_i, which is the set of plausibilities computed for x_i [9]:

$$L(\theta; pl_i) = \sum_{x_i \in \Omega} p(x_i; \theta) pl_i(x_i), \tag{2}$$

where Ω is the domain in which x_i is defined as an M-dimensional vector, and $p(x_i; \theta)$ is the probability that x_i is observed (or generalized) based on the probabilistic distribution with the parameter θ. Equation (2) shows that $p(x_i; \theta)$ represents the imprecision resulting from the population of concept detection scores, while $pl_i(x_i)$ represents the uncertainty related to the error in concept detectors. By assuming that each shot is independently and identically distributed, equation (2) can be extended to N shots:

$$L(\theta; pl) = \prod_{i=1}^{N} L_i(\theta; pl_i) = \prod_{i=1}^{N} \sum_{x_i \in \Omega} p(x_i; \theta) pl_i(x_i), \tag{3}$$

E^{2}M algorithm proposed in [9] computes θ that maximizes $L(\theta; pl)$ given plausibilities for N example shots $(pl_1, \cdots, pl_N)$ based on the Expectation Maximization (EM) algorithm. For reasons of space, we will only describe how E^{2}M algorithm is applied to concept-based video retrieval; please refer to [9] for a complete description. We denote $x_i^{j1} = 1$ if the j-th concept is present in x_i and otherwise $x_i^{j1} = 0$. Similarly, $x_i^{j0} = 1$ if it is absent in x_i, and otherwise $x_i^{j0} = 0$. Assuming that x_i^j is independent from each other, $p(x_i; \theta)$ is written as follows:

$$p(x_i; \theta) = \prod_{j=1}^{M} \prod_{h=0}^{1} (\alpha^{jh})^{x_i^{jh}}, \tag{4}$$

where $\theta = \{\alpha^{jh}\}$ is the set of parameters representing the probability of the j-th concept's presence (α^{j1}) or absence (α^{j0}). By applying Equation (4) to Equation (3), we get:

$$L(\theta; pl) = \prod_{i=1}^{N} \prod_{j=1}^{M} \sum_{h=0}^{1} pl_i^{jh} \alpha^{jh}, \tag{5}$$

where the derivation of $\sum pl_i^{jh} \alpha^{jh}$ is based on the fact that $\sum p(x_i; \theta) pl_i(x_i)$ in Equation (3) can be considered to be the expectation of pl_i [9]. E^{2}M algorithm extracts θ that maximizes the likelihood in Equation (5) as follows:

E-Step: Using θ^q which is the estimate of θ at the q-th iteration, compute $Q(\theta, \theta^q)$, which is the expectation of the log-likelihood of $x = \{x_1, \cdots, x_N\}$:

$$Q(\theta, \theta^q) = \sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{h=0}^{1} \gamma_i^{jh(q)} \log \alpha^{jh(q)}, \tag{6}$$

 ISBN: 978-1-61208-195-3

where

$$\gamma_i^{jh(q)} = \alpha^{jh(q)} pl_i^{jh} / \sum_{h'=0}^{1} \alpha^{jh'(q)} pl_i^{jh'}. \quad (7)$$

M-Step: Update θ so as to maximize $Q(\theta, \theta^q)$:

$$\alpha^{jh(q+1)} = \sum_{i=1}^{N} \gamma_i^{jh(q)}. \quad (8)$$

Finally, given pl_t of a test shot x_t, we compute its relevance score as the following likelihood $L(\theta; pl_t)$:

$$L(\theta; pl_t) = \prod_{j=1}^{M} \sum_{h=0}^{1} pl_t^{jh} \alpha^{jh}. \quad (9)$$

This likelihood represents the agreement between plausibilities pl_t of x_t and the probabilistic distribution, with θ estimated by E^{2}M algorithm. The set of $1,000$ test shots with the largest $L(\theta; pl_t)$ is returned as a retrieval result.

IV. Preliminary Experimental Results

To test our method, we used TRECVID 2009 video data [5]. This data consists of 219 development and 619 test videos, comprising $36,106$ shots and $97,150$ shots, respectively. We used concept detection scores provided by the City University of Hong Kong, where detection scores for 374 LSCOM concepts are assigned to every shot [6]. Our method was tested for the query "A view of one or more tall buildings and the top story visible". Ten example shots were selected from the development videos. Based on the text description and the example shots, 20 concepts related to the query (*e.g.*, *Building*, *Tower*, *Sky*, *etc.*) were selected using the method in [8].

We conducted a preliminary experiment to examine the effectiveness of using plausibilities, instead of directly using concept detection scores. The following three methods were compared: (1) *DST:* A probabilistic classifier is constructed using plausibilities modeled based on DST, (2) *Sum:* The relevance score of a shot is computed as the sum of concept detection scores (*i.e.*, linear combination with no weights), (3) *Prod:* The relevance score is computed as the product of concept detection scores [2]. Fig. 3 shows a comparison of retrieval performances between *DST*, *Sum* and *Prod* in terms of their average precision. *DST* is superior to the other two. We are now testing *DST* for different queries, and implementing a probabilistic classifier construction method that directly uses concept detection scores.

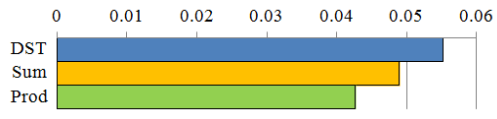


Figure 3. Performance comparison between *DST*, *Sum* and *Prod*.

V. Conclusion and Future Work

In this paper, we introduced a probabilistic classifier construction method which can account for the uncertainty in concept detection by modeling plausibilities of a concept's presence and absence based on DST. We plan to explore the following points to improve the retrieval performance of our method. First, in addition to example shots representing shots that are relevant to a query, we plan to use *counter-example* shots representing irrelevant shots and incorporate them into E^{2}M algorithm. Thereby, irrelevant shots that are retrieved based only on example shots may be excluded from the retrieval result. Second, even for the same query, relevant shots show different combinations of concepts due to varied camera techniques. Thus, we aim to incorporate a mixture model into E^{2}M algorithm. Third, we will explore a method which refines plausibilities of a concept's presence and absence by considering other concepts based on the knowledge base.

References

[1] Natsev A., Haubold A., Tešić, Xie L. and Yan R., "Semantic Concept-based Query Expansion and Re-ranking for Multimedia Retrieval," in *Proc. of ACM MM 2007*, 2007, pp. 991–1000.

[2] Snoek C. et al., "The mediamill TRECVID 2009 semantic video search engine," in *Proc. of TRECVID 2009*, 2009, pp. 226–238.

[3] Ngo C. et al., "VIREO/DVM at TRECVID 2009: High-level feature extraction, automatic video search and content-based copy detection," in *Proc. of TRECVID 2009*, 2009, pp. 415–432.

[4] Naphade M., Smith J., Tešić J., Chang S., Hsu W., Kennedy L., Hauptmann A. and Curtis J., "Large-scale concept ontology for multimedia," *IEEE Multimedia*, vol. 13, no. 3, pp. 86–91, 2006.

[5] Smeaton A., Over P. and Kraaij W., "Evaluation campaigns and TRECVid," in *Proc. of MIR 2006*, 2006, pp. 321–330.

[6] Jiang Y., Ngo C. and Yang J., "Towards optimal bag-of-features for object categorization and semantic video retrieval," in *Proc. of CIVR 2007*, 2007, pp. 494–501.

[7] Wei X., Jiang Y. and Ngo C., "Concept-driven multi-modality fusion for video search," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 1, pp. 62–73, 2011.

[8] Shirahama K. and Uehara K., "Constructing and utilizing video ontology for accurate and fast retrieval," *International Journal of Multimedia Data Engineering and Management*, vol. 2, no. 4, pp. 59–75, 2011.

[9] Denœux T., "Maximum likelihood estimation from uncertain data in the belief function framework," *IEEE Transactions on Knowledge and Data Engineering*, (PrePrint).

[10] Li X., Wang D., Li J. and Zhang B., "Video search in concept subspace: A text-like paradigm," in *Proc. of CIVR 2007*, 2007, pp. 603–610.

[11] Rasiwasia N., Moreno P. and Vasconcelos N., "Bridging the gap: Query by semantic example," *IEEE Transactions on Multimedia*, vol. 9, no. 5, pp. 923–938, 2007.

[12] Hsu C., Chang C. and Lin C., *A Practical Guide to Support Vector Classification*, http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf retrieved February 2012.

 ISBN: 978-1-61208-195-3

A Database of Artificial Urdu Text in Video Images with Semi-Automatic Text Line Labeling Scheme

Imran Siddiqi
Department of Applied Science & Graduate Studies
Bahria University
Islamabad, Pakistan

Ahsen Raza
Department of Computer Software Engineering
National University of Sciences & Technology
Islamabad, Pakistan

imran.siddiqi@bahria.edu.pk, ashen.raza@mcs.edu.pk

Abstract—**This paper describes a novel database of video images containing artificial (superimposed) Urdu text with a semi-automatic text line labeling scheme. The main objective of this study is to provide the community with a standard dataset together with an auto-labeling scheme for algorithmic development and evaluation of textual content based indexing and retrieval systems. We have specifically focused on Urdu text which is increasingly gaining research interest in recent years. The data set comprises 1000 video images collected from 19 different channels of 5 different categories. An attempt is made to capture the maximum possible variation in the text in terms of size, location, appearance and background. The data set is completely labeled by finding the bounding rectangle of each text occurrence facilitating the evaluation of text detection and localization systems. Based on our previous work on text localization, an automatic text labeling scheme is also proposed and the obtained results are compared with manual labeling. Ground truth data, supporting tasks like text recognition and word spotting will be considered in the next version of the data set.**

Keywords-Data Set; Artificial Urdu Text; Text Detection; Text Localization.

I. INTRODUCTION

The availability of data sets is one of the fundamental requirements for development and evaluation in any research domain. Over the recent years, standard databases are becoming increasingly popular in all the scientific research fields. The availability of such data sets not only saves researchers the task of compiling and labeling the database but also provides the possibility of objectively comparing different systems on the same data set. This is further complemented by organization of evaluation campaigns [16, 22] allowing comparison of different techniques under the same experimental conditions as well. Like other research areas, the document analysis and recognition community has also developed a number of standard databases addressing different problem areas. The most popular of these are the databases for handwriting recognition like CEDAR [12], NIST [13], CENPARMI [14], IAM [11] and RIMES [16] for offline while IAM-OnDB [18, 19], UNIPEN [23] and IRONOFF [24] for online recognition. In addition to character and word recognition, some of these data sets have also been used in evaluating tasks like document layout analysis, document segmentation and writer identification/verification.

Another significant research area in the document recognition paradigm is the detection, localization and recognition of artificial and scene text appearing in video images. Scene text recognition finds its applications in autonomous navigation and assistance; ICDAR [1] and KAIST [21] being the two widely used data sets in this domain. Artificial text on the other hand is more useful for applications like semantic indexing and retrieval of video archives. An important component of such keyword based retrieval systems is the detection and localization of textual regions [10]. It has attracted a number of researchers over the last decade [1- 4] and is in fact the subject of our study as well. More specifically, we focus on the artificial Urdu textual content in video images which is relatively a young and unexplored research area as opposed to text in other languages.

Urdu, the national language of Pakistan and a major language of India, has speakers allover the world. Analysis of Urdu documents and recognition/processing of Urdu text is attracting research interest in the recent years [5, 6, 15, 17, 20]. As the research in these areas matures, the need to evaluate the proposed techniques on standard data sets will naturally arise. Contributions have already been made towards the development of handwritten Urdu text data sets [7, 8, 9]. However, despite more than 65 Urdu news, entertainment, sports and religious channels around the world, no attempt has yet been made on the development of an artificial Urdu text data set to the best of authors' knowledge.

In this paper, we present the first version of a collection of video images containing artificial Urdu text. The database is mainly targeted towards the evaluation of artificial text detection and localization systems but may also be extended for Urdu word recognition and word spotting systems. The database comprises a total of 1000 video images captured from 19 different Urdu channels. The ground truth text regions in each image are manually extracted allowing quantitative evaluation of any text localization system. A semi-automatic text labeling is also proposed and compared with manual labeling. The main contributions of this work are:

- A completely labeled artificial Urdu text data set.
- A semi-automatic text labeling scheme.

 ISBN: 978-1-61208-195-3

The rest of the paper is organized as follows. In the next section, we discuss data acquisition followed by some characteristics and statistics of the collected data. We then present the manual ground truth labeling followed by the proposed automatic labeling methodology. The results of automatic labeling are then compared with manual labeling. Finally, we give the concluding remarks and discuss some possible future enhancements to the present database.

II. DATA ACQUISITION

Videos from 19 different Urdu channels were captured using Pinnacle Studio Movie board. All videos were recorded at a resolution of 720x576 and stored in 'avi' format. In an attempt to have natural and unconstrained content, multiple videos from the same channel were recorded at different times on a given day. Individual images from videos were then extracted in such a way that there is no repetition of textual content in different images and the maximum variation of text positions, sizes, colors and backgrounds is captured. All images are stored in 'png' format.

The major part of textual content in each image is in Urdu. In some cases however, the images also contain some occurrences of text in other languages (for example, English, Pashto, etc.). These occurrences are inevitable in some of the Urdu channels we have considered. All such occurrences were separately identified and recorded as well.

In the next section, we discuss in detail the different aspects of the data set as well some useful statistics.

III. CHARACTERISTICS AND STATISTICS

The data set comprises a total of 1000 video images extracted from 19 different channels which are grouped into 5 different categories. These categories are chosen to be news, entertainment, sports, business and religious channels. The number of images in each of the categories is summarized in Figure 1. Naturally, the number of news channels and consequently, the number of images in this category is more than any other category due to a large number of Urdu news channels operating around the world. These images are also rich in textual content due the presence of a continuous ticker text. A more detailed distribution of words in images can be found in Figure 2.

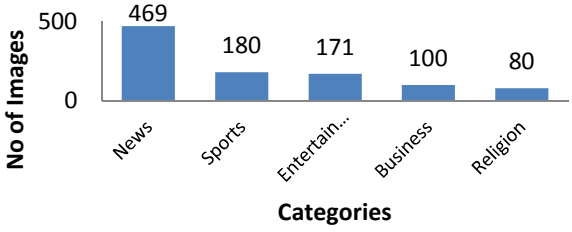


Figure 1. Distribution of images in categories

There are a total of 23833 Urdu words in the collected images. As already discussed, some of the images contain occurrences of English text as well, which make up a total of 5324 English words. A small number (120) of Pashto words also exist in the collected images. In addition to words, the images contain 3339 numerals as well. Table 1 gives an idea of the number of words per image in the data set and some other detailed statistics of the database. On the average, each image contains about 25 Urdu words, 4 words in another language and 4 numerals.

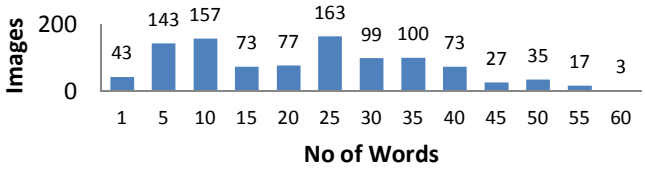


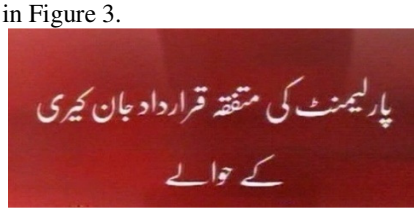
Figure 2. Distribution of words in images.

IV. NAMING & MANUAL LABELING

Once the images are collected, each category as well as each channel is assigned a three digit code. Each image is also given a three digit identification number. These codes are then used to name the images using the convention:

CategoryCode_ChannelCode_ImageID

Some of the images along with example names are illustrated in Figure 3.



(a) 002_003_004

(b)004_016_010

Figure 3. Example images containing artificial Urdu text

For quantitative evaluation of any system using these images, the ground truth data must be labeled. This, naturally, is time consuming, expensive and error prone task [1]. In the first version of the database, we have targeted text localization systems which require the coordinates of all text regions as ground truth data. Labeling of text regions in images is carried out manually using simple software (Figure 4) that allows opening an image and drawing rectangles over the textual content. The x and y-coordinates and width, height of each rectangle are stored in a data file. We have also proposed an automatic labeling scheme the results of which are compared with manual labeling as will be discussed in the subsequent sections.

 ISBN: 978-1-61208-195-3

TABLE I SOME STATISCTICS OF THE DATABASE

Category	Number of Images	Average Urdu words/ image	Second language(s)	Average second language words/ image	Average numeral(s) /image	Total Urdu words	Total second language words	Total numerals
News	469	17	Pashto	5	2	7860	120	1000
			English				2430	
Sports	180	30	English	6	3.5	5450	1070	615
Entertainment	181	26	English	5	3.5	4650	960	635
Business	100	31	English	5	6	3076	532	632
Religion	70	40	English	3	6.5	2797	212	457
Overall	**1000**	**23.8**	**-**	**5.3**	**3.3**	**23833**	**5324**	**3339**

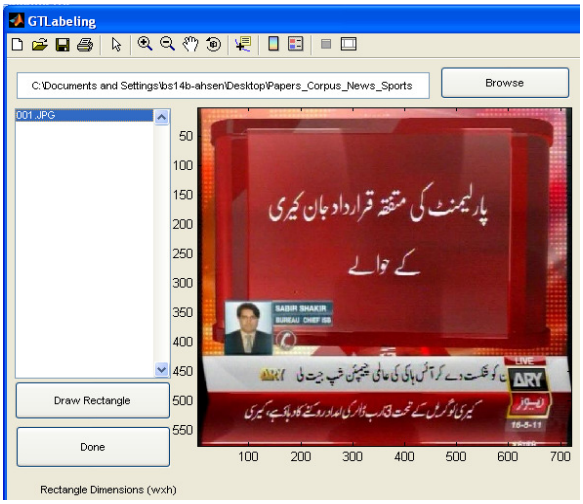


Figure 4. The user interface of the ground truth labeling software

An important factor in labeling is when to start a new rectangle. This will consequently affect the evaluation performance of tested systems depending upon the metric used. We investigated the ICDAR labeling methodology [1] but it cannot be replicated for Urdu text due to different characteristics of the script, for example, non uniform alignment within the same line of text (Figure 5). We therefore devised a labeling methodology that is based on the following heuristics:

- A single rectangle is drawn over a line of text that belongs to the same semantic unit (for example a sentence), when the words in the line are horizontally algined, are of the same size and, do not have a significant inter-word distance (Figure 5a).
- If different blocks of text in the same line have non-uniform size/alignment, they are split into different rectangles so that minimum background becomes part of the rectangles (Figure 5b and Figure 5c).
- In case of overlapping words, separate (overlapping) rectangles are drawn for each of the words (Figure 5d).

Naturally, these heuristics are subjective and the definitions of terms like 'alignment' and 'size difference' may vary from one individual to another. This however is an inherent limitation with such manual labeling. A solution could be to shift from block level to pixel level where each individual pixel is identified as being text or non-text. This, however, is an extremely time consuming job and is not considered in the present version of the database.

Figure 5. Sample labeled images showing the labeling methodology

The ground truth data for each image is stored in the accompanying data file. The data for the entire set of images is also stored in a single file. A part of the the ground truth data file is illustrated in Figure 6. Each line in the data file contains the image name, the language of textual content and the coordinates of the bounding rectangle. The complete data set along with ground truth data is publically available for download [35].

```
#--- Groud Truth Data---#
#
# Urdu Artificial Text Data Set
#
# format: 001_003_100 Urdu 11 435 372 43
#
#      001   -> Category Code
#      003   -> Channel Code
#      100   -> Image ID
#      Urdu  -> Language of textual content
#      11    -> x-coordinate of bounding rectangle
#      435   -> y-coordinate of bounding rectangle
#      372   -> width of bounding rectangle
#      43    -> height of bounding rectnagle
#

001_003_100 Urdu 11 435 372 43
001_003_100 Urdu 412 441 54 23
001_003_100 Urdu 477 436 122 39
001_003_100 Urdu 120 490 401 49
```

Figure 6. A snapshot of ground truth data file.

 ISBN: 978-1-61208-195-3

V. AUTOMATIC TEXT LABELING

In this section we present the proposed text localization scheme for automatic labeling of text regions in the database. Naturally, such automated techniques do have inherent problems with the accuracy and precision of the labeled regions and human assistance is required to correct these labeling errors. In our case, the automatically labeled regions are also compared with the manually labeled regions as will be discussed shortly.

The labeling scheme is primarily based on a series of image processing operations. The complete flow of the proposed scheme is shown in Figure 7. It is to be noted that the labeling algorithm operates on a single image and does not use any temporal features (e.g., redundancy of textual content in the video). This allows localization of textual occurrences on individual frames as well where the complete video is not available.

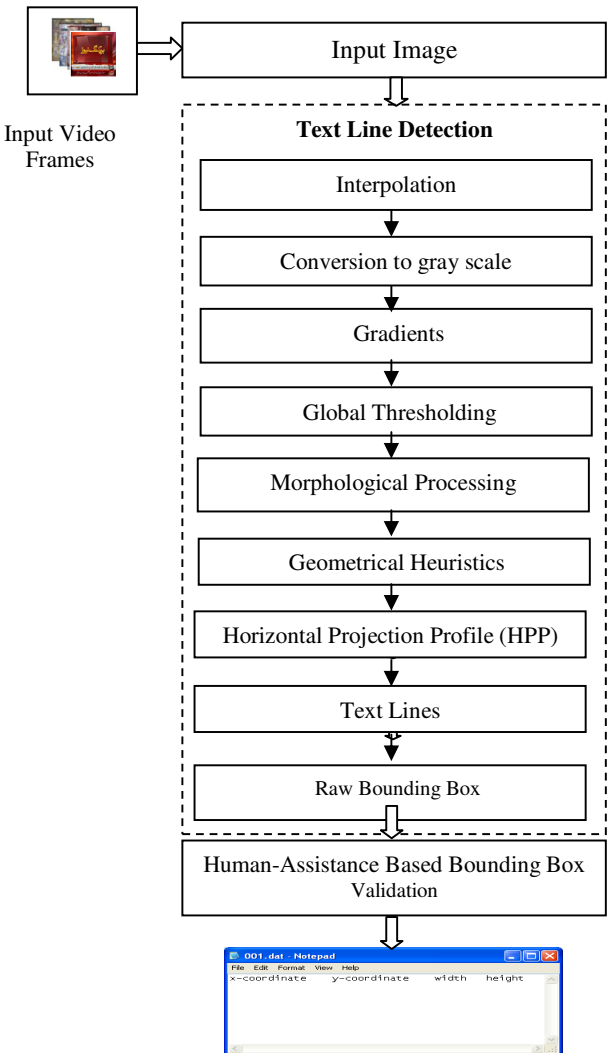


Figure 7. A general flow of text line labeling scheme.

In comparison to English, Urdu lexicon detection is much more challenging with main difficulties being the different shapes of alphabets depending upon their position within the words, the high frequency of diacritics, non-uniform inter and intra word distances and the occurrence of strokes in all directions. These factors make the detection of Urdu text more difficult as many non-text regions may also posses these text-like characteristics.

Our automatic labeling methodology is inspired from [2] with modifications for Urdu text and is mainly based on our previous work on text localization the details of which can be found in [31]. Similar methods have been used for detection of Farsi and Arabic texts [32-34] as well, which are quite similar in nature to Urdu text.

For localization of textual Urdu content in an image, as a first step, the image is converted to gray scale so that further processing is independent of the color information of image. Then, we conditionally resize the image using bi-cubic interpolation to an experimentally known size of 720x576. This gives a smooth estimate of the gray level at any desired point in the image [25]. Since the text is supposed to be readable on the screen, there is a high contrast between the textual content and its background which can be exploited to extract the boundaries of the text regions. We evaluated a number of standard edge detection filters, and finally, chose the Sobel filter for boundary detection as it preserves most of the edges and gives a strong boundary lining for isolated words.

In order to separate the text boundaries from the background, we next binarize the gradient image. This is one of the most critical steps as the subsequent steps are very sensitive to the binarization threshold. We experimenetd with a number of local [26, 27, 28] as well as global [29] thresholding algorithms and finally employed Otsu's thresholding [29] to binarize the gradient image.

As a result of binarization, most of the background is suppressed and the likely text boundries appear as connected components in the proximity of one another. These isolated components need to be merged together into words and ultimately text lines. This is implemented using the standard morphological operations of dilation and erosion. Dilation is carried out to merge all horizontally aligned components together which effectively is the merging of loosley connected characters into words. Dilation is followed by erosion which eliminates the falsely merged components. As a final step we employ the traditionally used geometrical constraints on the identified textual regions to eliminate the ones that do not satisfy the geometrical properties of text. These constraints are based on the aspect ratio and minimum height and width of the detected rectangles giving a set of rectangles which are likely to contain textual content.

Since the labeling is done at line level, the text lines are extracted from the identified textual regions using the well-known horizontal projection profiles [30]. Naturally the localized text lines are not always accurate/precise and need human intervention for validation.

The auto-labeled image is presented to the user with rectangles on potential text regions and the possibility the resize, move, add or delete the text rectangles. Once

 ISBN: 978-1-61208-195-3

validated, the coordinates of each text rectangle in the image are saved to a file. This semi-automatic labeling greatly reduces the effort involved as compared to a total manual labeling. The results of the proposed labeling are also very promising as discussed later in the paper. The detected textual regions can also be used for content based image retrieval (CBIR) applications [2]. The steps involved in labeling are illustrated on an example image in Figure 8. These steps are similar to those as in [2, 31] with adjustment of parameters for Urdu text.

Figure 8. Various steps of proposed labeling scheme. (a) Original image.(b) Grayscale image (c) Sobel filter (d) Binarized gradients (e) Morphological processing (f) Geometrical constraints (g) Detected text lines (h) Manual validation/correction (i) Ground truth data saved to file.

VI. Evaluation Metrics

The performance of a text localization system is traditionally quantified using the precision and recall. The problem however is that the text rectangles detected by a given system will not have a 1-1 correspondence with the text rectangles in the ground truth data. In addition, the size of these rectangles will also vary. To handle these issues, area based definitions of precision and recall are generally used. If G represents the ground truth text area in an image and D the detected area, we have:

$$Recall = \frac{(G \cap D)}{G}$$

$$Precision = \frac{(G \cap D)}{D}$$

More sophisticated metrics have also been proposed in the literature. For example, the ICDAR scene text database [1] defines a metric that searches for the true match of a detected rectangle in the set of ground truth rectangles. Wolf and Jolion [20] improved the ICDAR measure by introducing a novel performance measure that takes into account one-to-one, one-to-many (splits) and many-to-one (merges) scenarios as well.

All these metrics are equally applicable in case of Urdu text as well. Since the only information required in these metrics is the coordinates of the bounding rectangles, they can be easily calculated on the developed data set.

VII. Experimental Results

To evaluate the effectiveness of the proposed automatic labeling, we compared the results of auto-labeling (without human intervention) with manual labeling. On the data set of 1000 images, we achieved an overall precision of 71% and recall of 80% as summarized in Table II. We also studied how the performance of the localizer varies with the size of the image. These results are presented in Figure 9 and indicate that the performance is not very sensitive to the resolution of the image. The errors in terms of false positives, false negatives and misplaced rectangles can then be corrected by human intervention which naturally is much efficient as opposed to a complete manual labeling. Some results of the labeling scheme on a variety of backgrounds are presented in Figure 10.

TABLE II Performance of the Proposed Method

Data Set	Precision	Recall	F-measure
1000 Images	0.71	0.80	0.75

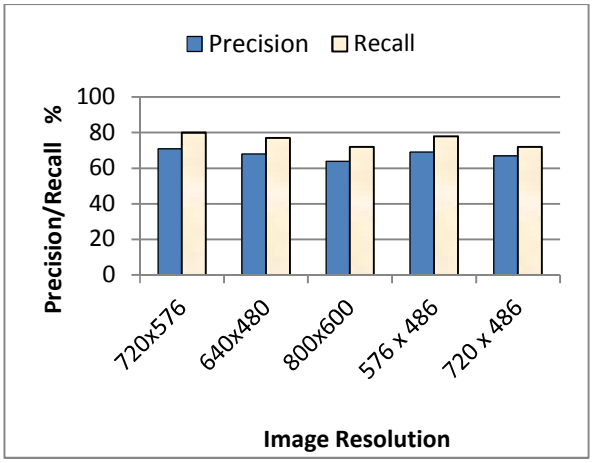


Figure 9. Performance of the proposed method on different image resolutions

VIII. Conclusion and Perspectives

In this paper, we presented the first version of a novel data set for Urdu artificial text along with a semi-auto text labeling scheme. This first version of the database is specifically targeted towards the evaluation of Urdu text localization systems. The ground truth data for each textual occurrence is saved to a file and can easily be used for evaluating such systems using any of the standard metrics. The present ground truth data is based on manual labeling

 ISBN: 978-1-61208-195-3

but we intend to use the proposed labeling scheme with human assistance to generate the ground truth data in the next version of the dataset.

We also plan to include the actual transcription of text in the next version, which will also allow the evaluation of Urdu text recognition and word spotting systems. The size of the data is also likely to double in the next version with additional channels in each of the categories. A similar dataset with text in languages based on the Latin alphabet is also under development finally leading to a huge collection of video images with unconstrained multilingual text. The authors expect that these data sets will prove to be useful for the document recognition community.

Figure 10. Auto-text labeling results (without manual validation) on a variety of images present in the database.

REFERENCES

[1] S. M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, and R. Young, "ICDAR 2003 robust reading Competitions". Proc. 7th International Conference on Document Analysis and Recognition (ICDAR'03), 2003.

[2] C. Wolf and J. Jolion, "Extraction and recognition of artificial text in multimedia documents", Pattern Anal. Appl., 2004, pp.309-326.

[3] K. C. Kim, H. R. Byun, Y. J. Song, Y. M. Choi, S. Y. Chi, K. K. Kim, and Y. K. Chung, "Scene text extraction in natural scene images using hierarchical feature combining and verification," Proc. 17th International Conference on Pattern Recognition, vol. 2, pp. 679–682, 2004.

[4] M. Abdel-Mottaleb, N. Dimitrova, R. Desai, and J. Martino, "CONIVAS: Content-based image and video access system" ,Proc. of ACM Multimedia, pp 427-428, Boston 1996.

[5] M. Humayoun, "Urdu morphology, orthography and lexicon extraction". Masters Thesis, Department of Computing Science, Chalmers University of Technology, 2006.

[6] N. Durrani and S. Hussain, "Urdu word segmentation", Proc. 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics, Los Angeles, June 2010.

[7] M. W. Sagheer, C. L. He, N. Nobile, and C. Y. Suen "A new large Urdu database for off-line handwriting recognition", Proc. ICIAP 2009, LNCS 5716, pp. 538–546, 2009.

[8] D. Becker and K. Riaz, "A study in Urdu corpus construction", Proc. of the 3rd Workshop on Asian Language Resources and International Standardization,Taipei, Taiwan. 2002.

[9] M. Ijaz and S. Hussain, "Corpus based Urdu lexicon development". Proc. Conference on Language and Technology, Peshawar, Pakistan, 2007.

[10] C. Lui, C.Wang, and R. Dai. "Text detection in images based on unsupervised classification of edge based features", Proc. International Conference on Document Analysis and Recognition (ICDAR 2005), pp 610–614, 2005.

[11] U.-V. Marti and H. Bunke. "A full english sentence database for off-line handwriting recognition", Proc. International Conference on Document Analysis and Recognition (ICDAR'99)", pp 705–708, 1999.

[12] J. Hull, "A database for handwritten text recognition research"; IEEE Trans. on Pattern Analysis and Machine In- telligence, 16(5):550–554, May 1994.

[13] R. Wilkinson, J. Geist, S. Janet, P. Grother, C. Burges, R. Creecy, B. Hammond, J. Hull, N. Larsen, T. Vogl, and C. Wilson, "The first census optical character recognition systems", Proc. Conf. #NISTIR 4912, The U.S. Bureau of Cen- sus and the National Institute of Standards and Technology, Gaithersburg, MD, 1992.

[14] C. Suen, C. Nadal, R. Legault, T. Mai, and L. Lam, "Computer recognition of unconstrained handwritten numerals", Proc. of the IEEE, 7(80):1162–1180, 1992.

[15] M. W. Sagheer, N. Nobile, C. L. Hev, and C. Y. Suen, "A novel handwritten Urdu word spotting based on connected components" Proc. International Conference on Pattern Recognition(ICPR) ,2010.

[16] E. Augustin, M. Carré, E. Grosicki, J. M. Brodin, E. Geoffrois, and F. Preteux, "Rimes evaluation campaign for handwritten mail processing", Proc. of the Workshop on Frontiers in Handwriting Recognition, pp. 231–235, 2006.

[17] U. Pal and A. Sarkar, "Recognition of printed Urdu script", Proc. of the Seventh International Conference on Document Analysis and Recognition (ICDAR'03), 2003.

[18] M. Liwicki and H. Bunke, "IAM-OnDB - An on-line English sentence database acquired from handwritten text on a whiteboard", Proc. of the Eighth International Conference on Document Analysis and Recognition, 2005.

[19] E. Indermühle, M. Liwicki, and H. Bunke, "IAMonDo database: an online handwritten document database with non-uniform contents", Proc. of the 9th IAPR International Workshop on Document Analysis Systems, 2010.

[20] C. Wolf and J.-M. Jolion, "Object count/area graphs for the evaluation of object detection and seg-mentation algorithm", International Journal on Document Analysis and Recognition (IJDAR), 8(4) 280–296, 2006.

[21] http://ai.kaist.ac.kr/home/DB/SceneText, 12-02-2012.

[22] D. Pallett, "A look at NIST's benchmark ASR tests: past, present, and future", Proc. IEEE Workshop on Automatic Speech Recognition and Understanding, 2003.

[23] I. Guyon, L. Schomaker, R. Plamondon, M. Liberman, and S. Janet, "Unipen project of on-line data exchange and recognizer benchmarks", Proc. of the 12th International Conference on Pattern Recognition, 1994.

[24] C. Viard-Gaudin, P. M. Lallican, P. Binter, and S. Knerr, "The ireste on/off (ironoff) dual handwriting database", Proc. of the Fifth International Conference on Document Analysis and Recognition, 1999.

[25] C. R. Gonzalez and R. E. Woods, Digital Image Processing. (2nd edition), 2001.

[26] W. Niblack, An introduction to digital image processing, pp. 115-116. Prentice-Hall, Englewood Cliffs (NJ), 1986.

 ISBN: 978-1-61208-195-3

[27] J.Sauvola, T,Seppanen, S.Haapakoski, and M.Pietikainen, "Adaptive document binarization", Proc. 4th Int. Conf. On Document Analysis and Recognition, Ulm, Germany, pp.147-152 (1997).

[28] C. Wolf, J-M. Jolion, and F. Chassaing, "Text localization, enhancement and binarization in multimedia documents", Proc. of the 16th International Conference on Pattern Recognition ICPR'02, Quebec, Canada, pp. 1037-1040, 2002.

[29] N.Otsu, "A threshold selection method from gray-level histograms", IEEE Transactions on Systems, Man and Cybernetics, 1979.

[30] M. Ben Halima, H. Karray, and A. M. Alimi, "A comprehensive method for Arabic video text detection, localization, extraction and recognition", PCM 2010, Part II, LNCS 6298, pp. 648–659, 2010.

[31] A. Jamil, I. Siddiqi, F. Arif, and A. Raza, "Edge-based features for localization of artifiical Urdu text in video images", Proc. of the 11th Int'l Conference on Document Analysis and Recognition (ICDAR 2011), China, pp 1120-1224, 2011.

[32] M. Moradi, S. Mozaffari, and A.A. Orouj, "Farsi/Arabic text extraction from video images by corner detection", Proc. 6th Iranian Machine Vision and Image Processing Conference, Oct 2010.

[33] M. Ben Halima, H. Karray, and A.M. Alimi, "Arabic text recognition in video sequences", Proc. International Conference on Informatics, Cybernetics, and Computer Applications, July 2010.

[34] M. Ben Halima, H. Karray, and A.M. Alimi, "AVOCR: Arabic video OCR", Proc. 5th International Symposium on I/V Communications and Mobile Network (ISVC), Sept 2010.

[35] https://sites.google.com/site/artificialtextdataset/, 12-02-2012.

 ISBN: 978-1-61208-195-3

Text Driven Recognition of Multiple Faces in Newspapers

Nicola Adami, Sergio Benini, and Riccardo Leonardi
Department of Information Engineering, Signals & Communications Lab
University of Brescia, via Branze 38, 25123, Brescia, Italy
Email: nicola.adami@ing.unibs.it, sergio.benini@ing.unibs.it, riccardo.leonardi@ing.unibs.it

***Abstract*—Face recognition is still a hard task when performed on newspaper images, since they often show faces in non-frontal poses, prohibitive lighting conditions, and too poor quality in terms of resolution. In these cases, combining textual information derived from the page articles with visual information proves to be advantageous for improving the recognition performance. In this work, we extract characters' names from articles and captions to restrict facial recognition to a limited set of candidates. To solve the difficulties derived from having multiple faces in the same image, we also propose a solution that enables a joint assignment of faces to characters' names. Extensive tests in both ideal and real scenarios confirm the soundness of the proposed approach.**

Keywords-Face recognition; Newspapers; Text analysis; Visual information; Multimodal.

I. INTRODUCTION

Over the last decades researchers have been making attempts to solve the problem of machine recognition of faces [1]. Algorithms proposed during the years can be coarsely classified into two categories: holistic approaches, such as Principal component analysis (PCA) [2] or Linear discriminant analysis (LDA) [3], and local feature-based ones (see e.g., [1] and [4]). Most recent developments of the holistic approaches include the Marginal Fisher analysis (MFA) [5], Eigenfeature regularization and extraction (ERE) [6], the sparse representation [7] and asymmetric PCA [8] and LDA [9], while Elastic bunch graph matching (EBGM) [10] and Active appearance model (AAM) [11] can be considered among the most performing feature-based algorithms. Despite the advances brought by these recent methods, there are still challenging problems to be tackled in face recognition, such as variations in pose, different facial expressions, make-up, lighting conditions as well as occlusions and cluttered background.

The recognition task is even harder when targeting newspaper images, where human faces are usually pictured at low quality and/or resolution. When visual data are not enough informative, one possible solution is combining natural language and visual information for improving semantic understanding of images. Newspapers in fact provide text that can be used to help the recognition process: each image with characters is commonly connected to an article on the same page, or at least described by a textual caption.

The idea has been relatively unexplored until the work in [12], which first proposes to use captions to locate faces in the accompanying photographs, thus with no recognition aims. A few years later, the rule-based PICTION system [13], trained on a dataset of 50 pictures was able to recognise human faces with a success rate of 65% by combining captions and photographs, even without employing a face recognition system.

In this paper, we target automatic recognition of human faces appearing in real newspapers by combining visual and textual information. As an advance with respect to previous work, we exploit all textual information coming from the newspaper page, thus not limiting the analysis to captions only, as done in the recent investigations in [14], [15], and [16]. Concerning face detection, we first apply an improved version of the Viola-Jones classifier [17] nowadays considered as a standard baseline for face detection. The recognition phase is then performed by using the standard method provided by Principal Component Analysis (PCA) [2]. These traditional approaches usually guarantee acceptable performance in not too complex contexts. However, these specific editorial products are mined by two major impediments that make harder the recognition process: one is related to the often too poor quality of images chosen for publishing in terms of resolution, contrast, illumination and pose; the second is due to the dimensions of the characters' database, which are potentially unlimited. These issues, which may lead to a difficult recognition, are here also addressed. Finally, as an additional contribution, recognition performance in case multiple faces are present in the same image are improved by a mechanism that jointly assigns identities to detected human faces.

Figure 1 describes the workflow: the textual analysis module extracts from newspaper pages potential characters' names to restrict the recognition phase to a few candidates. Results of recognition are inspectable by a human operator, who also takes care of the cases when a name (respectively, a face) found in the page has no associated face (resp. a name) stored in the databases, so that the system is able to learn from previous recognition processes.

Helping the automatic understanding of newspaper articles, such a tool could find application in complex tasks such as news segmentation, or in supporting professional frameworks for production of new multimedia content, such as news aggregators or feeds.

The document is organized as follows: in Section II, we

 ISBN: 978-1-61208-195-3

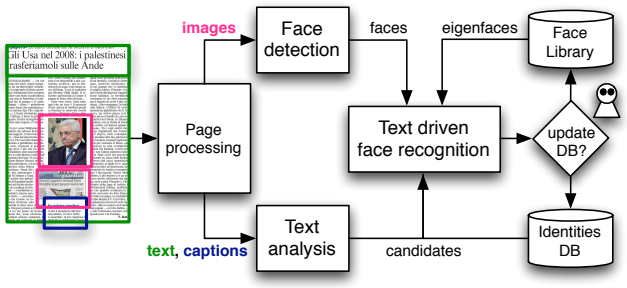


Figure 1: Diagram describing the workflow for the combined visual and text-based face recognition process.

explain how a segmented page looks like. Face detection and text analysis processes are described in Sections III and IV, respectively. In Section V, we focus on the multiple face recognition algorithm that employs a weak supervision in the form of text found in articles and captions. Experiments are conducted in Section VI, while conclusions are finally drawn in Section VII.

II. Page Preprocessing

Newspaper pages are first segmented by a process whose description is beyond the scope of this paper. As shown in Figure 2, the output of the segmentation stage consists of two separate pages (the *image-page* and the *text-page*) provided with two related *structure files* describing all page elements (*images*, *articles, titles*, *captions*, etc.) and their positions in the original page.

Figure 2: Example of how a) a newspaper page is segmented into b) an *image-page*, and c) a *text-page*.

III. Face Detection

In order to ensure a correct recognition of characters, a robust face detection process on the image-page becomes a fundamental prerequisite. Once segmented from the rest, the image-page is fed into a classifier to isolate regions belonging to human faces. This is achieved by relying on the well-known Viola-Jones method [17] endowed with the following adjustments to boost the detection performance.

On the one side we increase the robustness to pose change: despite the fact this classifier is known to work well for frontal faces only, the last implementation in [18] comes with several cascade files for detecting profile faces, even if with slightly lower performance. On the other hand, before being processed, images undergo a process of histogram equalization as in [19] to increase the algorithm invariance to complex lighting conditions. These adjustments are important in the specific domain of newspaper images, where characters are usually pictured in arbitrary poses, under different illumination conditions and often at low resolution.

Last but not least, since the whole system aims at recognising characters, each missed face during detection (i.e., each true negative) results in a final missed recognition. On the contrary, false positives are not that relevant, since they will not match with any candidate face in the database, therefore not producing errors. Thus the final tuning of the cascade face classifiers has been carried out in order to privilege *recall* rather than *precision*. Examples of *non-relevant* and *relevant* errors during the detection phase are given in Figure 3-a and Figure 3-b, respectively.

Figure 3: Examples of errors in detection: a) non-relevant error: the false positive on the tie will not find any matching face during recognition; b) relevant error: the missed detection of the face will determine a missed recognition.

IV. Text Analysis

Since in newspapers face recognition is a hard task, it is necessary to exploit all information we have at hand to support the identification of characters. Luckily, the text-page usually contains a direct reference to the characters' names pictured in the image-page. Even if this might not always happen, from our experiments (see Section VI) the case where a pictured character is not cited in the text is so rare that we can assume this hypothesis as reasonable.

To emulate the process of human comprehension while reading a newspaper is a hard task: a few decades of research dealing with the problem of natural language processing (see e.g., [20]) still have not solved the problem. For our purposes however, it is sufficient that the text module is able to extract names that might correspond to characters contained in images. To achieve this goal, the module relies on two databases: the *identity database* containing several characters' data, and one *common names database* [1].

[1]Lists of names and surnames are easily available for each country. In some languages their extraction is easier since they start with capital letters.

Whenever a name is found in the text, if that identity is already present in the characters' database, then that person is proposed as a *candidate* for the face recognition process. Conversely if a name is recognised as such but the related identity is not in the database, a new identity is proposed for approval to a human operator for insertion. In this case, a new database record is created and few images retrieved on the web (with Google's images search [21]) are used to populate the character's *face library*.

The identity database also manages name variations referring to the same person (e.g., "George W. Bush", "G. W. Bush", "George Bush", etc.), so that when any of these variants is found in the text, a unique *candidate* is selected for the recognition phase. When only the surname (e.g., "Bush") appears in the text, all characters with the same surname (e.g., the singer "Kate Bush") are chosen as candidates. Even if one surname is very popular, this will restrict anyway the candidates to a limited pool of characters.

In the common case when a caption accompanies the image, the importance of the text module increases, since it is sufficient to use the caption text as the only input (instead of the whole related article) to restrict the pool of candidates to a very short list, as shown in the example of Figure 4.

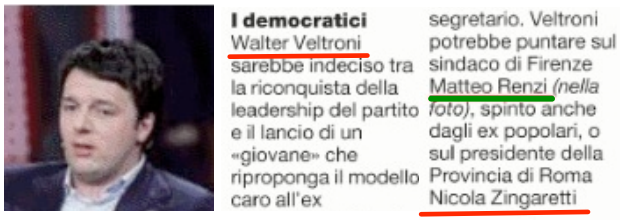


Figure 4: Caption returns a restricted pool of candidates: "Walter Veltroni", "Matteo Renzi", and "Nicola Zingaretti".

The module is not error free: for example it fails in case the text refers to "G. W. Bush" only as "President", or in general, in every case when a person is referred to by means of his/her title, such as "Prime Minister" or similar ones.

V. Text driven face recognition

Due to the fact that the application must be able to update the database during execution, the training phase on face recognition should be as automatic as possible. In fact, in case there is the need to add new faces, it is not reasonable to involve the human operator in too complex update operations. As a consequence, all local recognition approaches that require a manual labelling of points of interests (such as EBGM [10] or and AAM [11]) are not applicable in this context.

On the contrary, traditional holistic methods such as PCA [2] or LDA [3], if provided even with few image examples, are able to build sufficiently accurate models for each identity in an automatic way. Despite more modern approaches exist, performance offered by Principal Component Analysis in its original formulation are sufficient for implementing the visual part of the proposed supervised recognition approach.

A. Training

The initial face library contains a limited set of famous characters belonging to politics, sports, economy, science and culture. As seen in Section IV, each time that a new name pops up from newspaper pages, related face images retrieved on the Internet can be dragged in the face library. To ensure an acceptable level of automation, inserted pictures are automatically cropped and resized without forcing the user to extract faces manually: to do this the face detection algorithm described in Section III on equalized images first extracts face bounding boxes; after, image dimensions are normalized to 64×64 pixels. If the image quality is acceptable, faces are rotated and centered by using an eye detection algorithm, thus obtaining the final training image. Examples of processed images are shown in Figure 5.

Figure 5: 64×64 normalized faces are extracted for training.

Obtained images are then used for the training phase, which is performed by a standard PCA. Eigenfaces are calculated from the training set by keeping the M-images that correspond to the highest eigenvalues, so that they contain at least the 80% of the total energy, as suggested in [22]. These M eigenfaces define the M-dimensional "face space" employed during recognition. As new faces are added to the face database, eigenfaces can be updated or recalculated.

B. Recognition

Face recognition is treated as a pattern recognition task: each detected face Γ is projected onto the "face space" by transforming it into its eigenface components

$$\Gamma = [\gamma_1, \gamma_2, \ldots, \gamma_M]$$

which describe the contribution of each eigenface in representing the input face. The feature vector Γ is then used in a standard pattern recognition algorithm to find which of a number of predefined face classes best describes the face.

Since face recognition is a particularly difficult task, especially in case of numerous possible identities as in newspapers, to improve recognition performance we reduce

 ISBN: 978-1-61208-195-3

the number of face candidates only to those N identities found in the same text-page. Since it is rare that characters in pictures are not referred in the article body or in the caption itself, we accept the risk connected to an excessive candidate reduction.

The advantage of this supervised approach is most apparent when only one face is detected and one name is extracted from the text-page. As shown later in the experiments, this event is frequent when captions are associated to images: in such a situation, there is no need to perform a projection on the face space, but the image face is directly associated to the uniquely extracted name.

In case of multiple candidates instead, classification is performed by comparing the feature vector Γ of the test face with the face classes, which are the average representations

$$\overline{\Phi} = \left[\overline{\phi}_1, \overline{\phi}_2, \ldots, \overline{\phi}_N\right]$$

of each candidate over a number of face images. Comparison is based on the Mahalanobis' distance between Γ and $\overline{\Phi}$ to find which candidate best describes the test face:

$$k = \underset{i}{\operatorname{argmin}} \left\{ \sqrt{\left(\Gamma - \overline{\Phi}_i\right) S_i^{-1} \left(\Gamma - \overline{\Phi}_i\right)^T} \right\} i = 1, 2, \ldots, N$$

where S_i is the covariance matrix of each candidate class. By attributing more relevance to components with larger associated eigenvalues, this metric removes the problems related to scale and correlation that are inherent with the Euclidean distance, and provides in fact, superior experimental performance; in particular the average value is considered in order to better exploit each class distribution, and not relying only on a minimal distance that often leads to misclassification due to the presence of noisy samples. The value of the Mahalanobis' distance is also considered as a confidence level on the classification result: the user has the possibility to choose whether the confidence level is enough high for him not to check the recognition results, or conversely, if too low, to classify the face as *unknown* and add it manually to the face library for later use, so that the system learns to recognize new face images.

The same mechanism allows also for removing false positives introduced during face detection mentioned in Section III. By dividing the training space in two regions ("face" and "non-face" hemispaces) if the distance exceeds the space region boundaries, the detected face is probably a false positive, so that it is removed and recognition is not performed at all.

C. Joint recognition of multiple faces

The system as proposed so far is quite robust, especially until up to one face is detected in each image. In the presence of multiple faces in the same picture however, it is possible that two or more faces are at minimum distance to the same candidate. In this case, the algorithm as defined before, would label both faces with the same identity, thing that is evidently not possible. There is then the need to increase the algorithm robustness and elaborate a strategy to manage the recognition of multiple faces in the same image.

In order to best describe this problem, let us consider the example in Figure 6-a) where, due to his face orientation, Barack Obama is not recognised. In this picture, two faces are assigned to the same candidate "T. Geithner", since both are at a minimum average distance from Geithner's training samples (as shown in Figure 6-b).

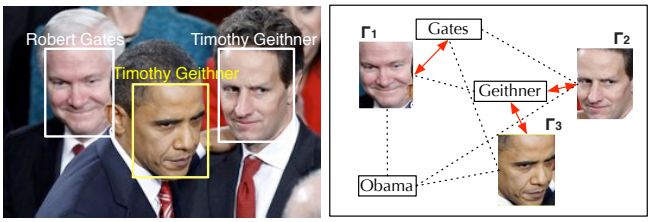


Figure 6: a) Independent recognition on three faces assigns the same identity to two characters, since b) both Γ_2 and Γ_3 are at minimum distance with the same candidate.

In order to increase the algorithm's robustness and correct results in analogous situations, we propose to consider also all other classes' distributions: to be assigned with an identity, a test image should not only be at minimum average distance from the candidate samples but also, at the same time, at maximum average distance from other classes.

This objective is achieved by assigning an heuristic score q to each candidate as the difference between the distance of the test image Γ and all other classes $\overline{\Phi}_i$ and the distance of Γ and the best candidate class $\overline{\Phi}_k$, that is

$$q(\Gamma) = \sum_{i=1}^{N} \left[d\left(\Gamma, \overline{\Phi}_i\right) - d\left(\Gamma, \overline{\Phi}_k\right)\right]$$

where score q is a real positive number. Once defined the heuristic score, the face recognition algorithm for multiple faces works as shown in Figure 7.

```
1. perform face recognition independently on all faces;
2. assign to each face the best candidate Φk;
3. If no conflicts,
4.     then exit;
5. else
6.     compute score q for each conflicting face Γ;
7.     assign face with highest q to identity Φk;
8.     remove Φk from the candidate list;
9.     repeat from line 2;
```

Figure 7: Joint face recognition algorithm.

To correct Obama's identity as in Figure 8-a, when both test faces Γ_2 and Γ_3 are at minimum distance with the same candidate "Geithner", the test face with higher q is

assigned to the best candidate, while the face with lower q is reassigned to the next closest candidate (Figure 8-b). Please notice that $q(\Gamma_2) > q(\Gamma_3)$ means that Γ_2 is on average further from other possible candidates than Γ_3, so that it is more likely that the latter was incorrectly assigned.

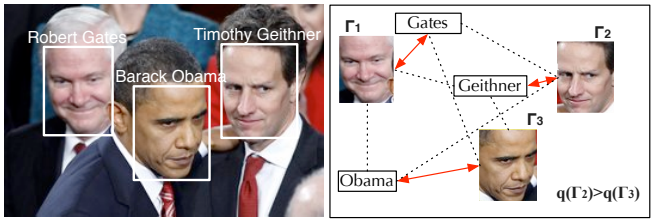


Figure 8: a) Obama's identity is corrected via joint face recognition; b) the face with higher q, i.e., Γ_2 is assigned to its best candidate, while Γ_3 is reassigned to the next candidate.

The algorithm is able to solve an arbitrary number of conflicts and works also when the number of faces is larger than the database possible identities. In this case, faces with worst q scores are labelled as *unknown*. Please notice that the label *unknown* does not necessarily imply that the identity is not in the database, but only that the face was not coupled with any of the candidates extracted from text.

VI. Experiments

The experimental part is subdivided in four progressive steps, so that to appreciate the beneficial brought by different modules of the supervised approach.

In the first part of the experiments, the performance of the face recognition algorithm are tested alone on a public available database "Olivetti-Att-ORL" [23], which contains 400 human faces belonging to 40 unique people, where each individual is pictured 10 times from a frontal view or with a slight tilt of the head (see Figure 9). Training

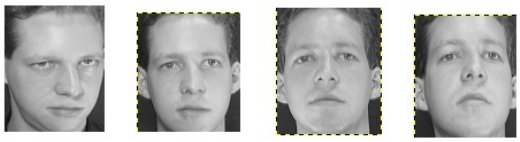


Figure 9: First test: recognition is performed on images from the Olivetti-Att-ORL dataset.

is performed on 325 randomly chosen images, while the recognition task is performed on the remaining part of the dataset. Classification results averaged on 5 runs returns 75% of correct recognition, which is in line with state of the art PCA performance [1].

In the second part of the experimental phase, we test the recognition algorithm on real newspaper pictures. For this aim a face library from newspaper images has been built as follow: around 200 identities have been extracted from different copies of the italian newspaper "Corriere della Sera" and for each identity, an average of four face images has been retrieved from the web, for a total training set of approximately 800 images. The test set instead is built up by using 200 images extracted from 15 different issues of the same newspaper, which contain faces in arbitrary conditions of pose, illumination, and quality, as shown in the examples of Figure 10. In this real application scenario, recognition performance collapse, as expected, around 50%, thus confirming that recognition of characters in newspaper images is far more challenging than on a standard dataset.

Figure 10: Recognition on newspaper images is more challenging due the variety of image poses, light conditions, or quality.

In the third step of the experiment, the approach combining face recognition with the supervision of text has been tested on the newspapers data. For each test image, we constrain recognition only to those candidates whose names are found in the same text-page or in the corresponding image caption. Results obtained yield a percentage of 83% correctly recognised faces.

Finally, the application to the same dataset of the algorithm for multiple face recognition further improves performance up to 86%. Table I summarises all performed tests and the related performance.

Table I: Performed experiments.

Test	Algorithm	Data	Perf.
1	*Face recognition only*	Olivetti-Att-ORL	75%
2	*Face recognition only*	Newspapers	50%
3	*Text driven recognition*	Newspapers	83%
4	*Joint supervised recog*	Newspapers	86%

The built system is able to learn from previous classifications by including the recognised faces in the face library. Therefore it is commonsensical to believe that performance are expected to improve as long as the system is in use.

As final considerations, we mention three causes of errors that influence the system performance. First, errors might be generated in case of wrong initial segmentation into text- and image-pages. For example, if one image is associated to a wrong caption, this likely leads to a missed recognition. Second, errors can be due to the face detection: true negatives, as seen in Section III are estimated to be around 10%. Third and last, the text module might fail in extracting correct names: this happened three times during the 200 performed tests (1.5%) because the mentioned identity was addressed only by his/her title. All these types of error are not included in results of Table I, which accounts only for the performance

 ISBN: 978-1-61208-195-3

of the supervised algorithm when both the text and the face detection modules return correct results.

VII. Conclusions

In this work, we combine text derived from newspaper articles and captions with visual information to improve face recognition performance. Characters' names are used to constrain facial recognition to a limited set of candidates, which are jointly assigned to the related faces in case multiple characters are present in the same picture. The good performance obtained in the experimental phase demonstrates that this approach allows for high recognition rate on newspaper images, notoriously a difficult benchmark since often showing faces in non-frontal poses, prohibitive lighting conditions, and poor in quality and/or resolution. Future work aims at extending the approach to a wider set of editorial publications (including magazines, satyrical, etc.) as well as to integrate higher performing recognition method.

References

[1] W.-Y. Zhao, R. Chellappa, P. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM Comput. Surv.*, vol. 35, no. 4, pp. 399–458, 2003.

[2] M. Turk and A. Pentland, "Face recognition using eigenfaces," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, vol. 591, 1991.

[3] J. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, "Face recognition using LDA-based algorithms," *IEEE Transactions on Neural Networks*, vol. 14, no. 1, pp. 195–200, Jan. 2003.

[4] X. Tan, S. Chen, Z.-H. Zhou, and F. Zhang, "Face recognition from a single image per person: A survey," *Pattern Recognition*, vol. 39, no. 9, pp. 1725–1745, 2006.

[5] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: A general framework for dimensionality reduction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 1, pp. 40–51, 2007.

[6] X. Jiang, B. Mandal, and A. Kot, "Eigenfeature regularization and extraction in face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 3, pp. 383–394, 2008.

[7] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2009.

[8] X. Jiang, "Asymmetric principal component and discriminant analyses for pattern classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 5, pp. 931–937, 2009.

[9] ——, "Linear subspace learning-based dimensionality reduction," *IEEE Signal Processing Magazine*, vol. 28, no. 2, pp. 16–26, 2011.

[10] L. Wiskott, J. M. Fellous, N. Kruger, and C. von der Malsburg, "Face recognition by elastic bunch graph matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 775–779, 1997.

[11] T. Cootes, G. Edwards, and C. Taylor, "Active appearance models," *IEEE Transactions On Pattern Analysis And Machine Intelligence*, vol. 23, no. 6, pp. 681–685, Jun 2001.

[12] V. Govindaraju, D. Sher, R. Srihari, and S. Srihari, "Locating human faces in newspaper photographs," in *Proc. of IEEE Conf. on CVPR*, San Diego, CA, Jun 1989, pp. 549–555.

[13] R. K. Srihari, "PICTION: A system that uses captions to label human faces in newspaper photographs," in *Press, A. (ed.) Proceedings of the AAAI-91*, 1991, pp. 80–95.

[14] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid, "Automatic face naming with caption-based supervision," in *Conference on Computer Vision & Pattern Recognition*, Jun 2008, pp. 1–8.

[15] T. Mensink and J. Verbeek, "Improving people search using query expansions: How friends help to find people," in *European Conference on Computer Vision*, ser. LNCS, vol. II. Springer, oct 2008, pp. 86–99.

[16] D. Ozkan and P. Duygulu, "Interesting faces: A graph-based approach for finding people in news," *Pattern Recogn.*, vol. 43, pp. 1717–1735, May 2010.

[17] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," *Proceedings of CVPR*, vol. 1, pp. 511–518, 2001.

[18] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[19] L. Tao, L. M.-J. Seow, and V. K. Asari, "Nonlinear image enhancement to improve face detection in complex lighting environment," *International Journal of Computational Intelligence Research*, vol. 2, no. 4, pp. 327–336, 2006.

[20] C. D. Manning and H. Schütze, *Foundations of statistical natural language processing*. Cambridge, MA, USA: MIT Press, 1999.

[21] "Google's image search," http://images.google.com/. [retrieved: April, 2012]. Available: http://images.google.com/

[22] R. Tjahyadi, W. Liu, and S. Venkatesh, "Automatic parameter selection for eigenfaces," in *Proceedings of the 6th International Conference on Optimization: Techniques and Applications*, 2004.

[23] F. Ferdinando Samaria and A. Harter, "Parameterisation of a stochastic model for human face identification," in *Proceedings of 2nd IEEE Workshop on Applications of Computer Vision*, Sarasota, FL, Dec. 1994.

 ISBN: 978-1-61208-195-3

Optimisation of JPEG XR Quantisation Settings in Iris Recognition Systems

Kurt Horvath and Herbert Stögner
School of Communication Engineering for IT
Carinthia University of Applied Sciences
Klagenfurt, Austria
KurtKlaus.Horvath@edu.fh-kaernten.ac.at

Andreas Uhl
Multimedia Signal Processing and Security Lab
Department of Computer Sciences
University of Salzburg, Austria
andreas.uhl@sbg.ac.at

***Abstract*—JPEG XR is considered as a lossy sample data compression scheme in the context of iris recognition techniques. It is shown that by optimising the JPEG XR quantisation strategy, JPEG XR default quantisation as well as JPEG2000 based iris recognition can be improved in terms of EER. The optimised JPEG XR quantisation strategy shows good performance across a wide range of iris feature extraction techniques, but has to be adapted for each target bitrate separately.**

***Keywords*-JPEG XR; iris recognition; quantisation optimisation; EER.**

I. Introduction

In distributed biometric systems, the compression of sample data may become imperative under certain circumstances, since the data acquisition stage is often dislocated from the feature extraction and matching stage. In such environments the sample data have to be transferred via a network link to the respective location, often over wireless channels with low bandwidth and high latency. Therefore, a minimisation of the amount of data to be transferred is highly desirable, which is achieved by compressing the data before transmission and any further processing. See Fig. 1 for an illustration involving JPEG XR for compressed data transmission.

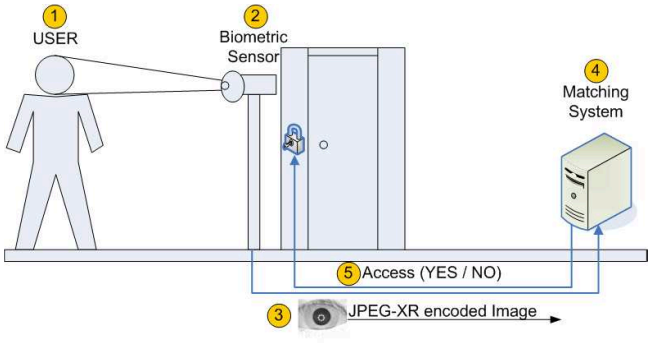


Fig. 1. System View.

As an alternative, the application of feature extraction before transmission looks promising due to the small size of template data but cannot be done under most circumstances due to the prohibitive computational demand of these operations (current sensor devices are typically far too weak to support this while compression can be done e.g. in dedicated low power hardware).

In order to maximise the benefit in terms of data reduction, lossy compression techniques are often suggested. Given the potential impact of lossy compression techniques on biometric recognition performance, it is imperative to carefully select and optimise appropriate codecs and to study their corresponding effect on recognition accuracy.

While current international standards define the application of JPEG2000 for lossy iris sample data compression, we focus in this paper on the optimised application of the recent JPEG XR still image coding standard. We experimentally compare the achieved results to a JPEG2000 based (and therefore standard conformant) environment. In particular, besides reviewing the effects of applying different settings concerning the use of the optional Photo Overlap Transform (POT) as a part of JPEG XR's Lapped Biorthogonal Transform (LBT), we optimise the JPEG XR quantisation strategy with respect to balancing quantisation strength among the three different frequency bands of the LBT. In Section 2, we review related standards and literature in the area of lossy iris sample data compression, while in Section 3, JPEG XR basics and especially the quantisation strategy are briefly explained. Section 4 presents experiments where we first shortly review the four different iris recognition systems employed in this study. Subsequently, the optimisation of the JPEG XR quantisation scheme is explained. Experimental results comparing optimised JPEG XR, different LBT variants in JPEG XR, and JPEG2000 are shown with respect to iris recognition accuracy in terms of EER. Section 5 concludes the paper.

II. Biometric Iris Sample Compression

During the last decade, several algorithms and standards for compressing image data relevant in biometric systems have evolved. The certainly most relevant one is the ISO/IEC 19794 standard on Biometric Data Interchange Formats, where in its former version (ISO/IEC 19794-6:2005), JPEG and JPEG2000 (and WSQ for fingerprints) were defined as admissible formats for lossy compression, whereas for lossless and nearly lossless compression JPEG-LS as defined in ISO/IEC 14495 was suggested. In the most recently published version (ISO/IEC FDIS 19794-6 as of August 2010), only JPEG2000 is included

 ISBN: 978-1-61208-195-3

for lossy compression while the PNG format serves as lossless compressor [1]. These formats have also been recommended for various application scenarios and standardised iris images (IREX records) by the NIST Iris Exchange program (`http://iris.nist.gov/irex/`).

The ANSI/NIST-ITL 1-2011 standard on "Data Format for the Interchange of Fingerprint, Facial & Other Biometric Information" (2nd draft as of February 2011, former ANSI/NIST-ITL 1-2007) supports both PNG and JPEG2000 for the lossless case and JPEG2000 only for applications tolerating lossy compression.

In literature on compressing iris imagery, rectilinear [2], [3], [4], [5] as well as polar [6], [7], [8], [9] iris sample data formats has been considered. With respect to employed compression technology, we find JPEG [3], [4], [5], JPEG2000 [2], [3], [4], [5], and other general purpose compression techniques [4], [5] being investigated. Superior compression performance of JPEG2000 over JPEG is seen especially for low bitrates (thus confirming the choice of the above-referenced standards), however, for high and medium quality, JPEG is found still to be competitive in terms of impacting recognition accuracy. Apart from applying the respective algorithms with their default settings and standard configurations, work has been done to optimise the compression algorithms to the application domain: For JPEG2000, it has been proposed to invoke RoI coding for the iris texture area [10] whereas the removal of the image background before compression has also been suggested (i.e. parts of the image not being part of the eye like eyelids are replaced by constant average gray [3]). For JPEG, an optimisation of quantisation matrices has been proposed to achieve better matching accuracy compared to the standard values for rectangular iris image data [11] as well as for polar iris images [8], [9].

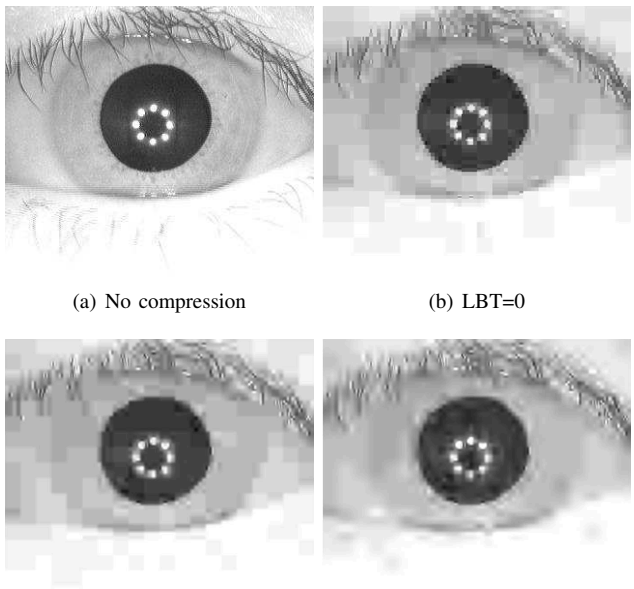
The JPEG XR standard has only recently been investigated in the context of biometric systems [12]. It has been found to eventually represent an interesting alternative to JPEG2000 in iris recognition systems due to its simpler structure and less demanding implementations in terms of memory and CPU resources, while providing almost equal recognition performance.

III. JPEG XR Background

Originally developed by Microsoft and termed "HD Photo", JPEG XR got standardised by ITU-T and ISO in 2009 [13], which makes it the most recent still image coding standard. The original scope was to develop a coding scheme targeting "extended range" applications which involves higher bit-depths as currently supported. However, much more than 10 years after JPEG2000 [14] development and 10 years after its standardisation it seems to be reasonable to look for a new coding standard to eventually employ "lessons learnt" in JPEG2000 standardisation. In particular, the focus is on a simpler scheme which should offer only the amount of scalability actually required for most applications (as opposed to JPEG2000 which is a rather complex scheme offering almost unconstraint scalability).

JPEG XR is a transform coding scheme showing the classical three-stage design: transform, quantisation, and entropy encoding. The transform operates on macroblocks consisting of 16 (arranged in 4 by 4) 4×4 pixel blocks. The first stage of the integer-based transform is applied to all 4×4 pixel blocks of a macroblock. Subsequently, the resulting coefficients are partitioned into 240 "high pass (HP) coefficients" and 16 coefficients corresponding to the lowest frequency in each block. The latter are aggregated into a square data layout (4 x 4 coefficients) onto which the transform is applied for a second time. The result are 15 "low pass (LP) coefficients" and a single "DC" coefficient (per macroblock).

In fact, the transform used in JPEG XR is more complicated as compared to JPEG, it is a so-called "two-stage lapped biorthogonal transform (LBT)" which is actually composed of two distinct transforms: The Photo Core Transform (PCT) and the Photo Overlap Transform (POT). The PCT is similar to the widely used DCT and exploits spatial correlation within the 4 x 4 pixels block, however, it suffers from the inability to exploit inter-block correlations due to its small support and from blocking artifacts at low bitrates. The POT is designed to exploit correlations across block boundaries as well as to mitigate blocking artifacts.



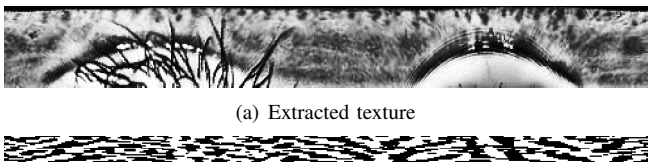
(a) No compression (b) LBT=0

(c) LBT=1 (d) LBT=2

Fig. 2. Rectilinear example images.

Each stage of the transform can be viewed as a flexible concatenation of POT and PCT since the POT is functionally independent of the PCT and can be switched on or off, as chosen by the encoder (this is signalled by the encoder in the bitstream). There are three options: disabled for both PCT stages (LBT=0), enabled for the first PCT stage but disabled for the second PCT stage (LBT=1), or enabled for both PCT stages (LBT=2). In recent work is has been shown that surprisingly, no clear advantage of any of these options with respect to recognition performance can be observed [12].

 ISBN: 978-1-61208-195-3

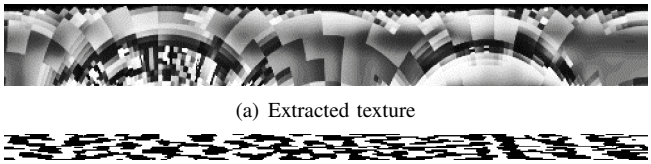


(a) Extracted texture

(b) Iris Code

Fig. 3. No compression applied.

Fig. 2 shows sample images for the uncompressed case and the three transform settings of JPEG XR (LBT=0,1,2) with "uniform" quantisation parameter $q = 100$ (see below).

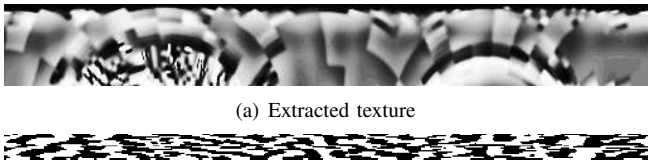


(a) Extracted texture

(b) Iris Code

Fig. 4. LBT=0, HD=0.35.

Figs. 3 - 6 visualise corresponding extracted iris textures as well as computed Masek Iris Codes (see next section) for the four settings shown in Fig. 2. When computing the Hamming Distance (HD) to the iris code derived from the uncompressed image in Fig. 3, we result in 0.35 for LBT=0, 0.403 for LBT=1, and 0.385 for LBT=2.



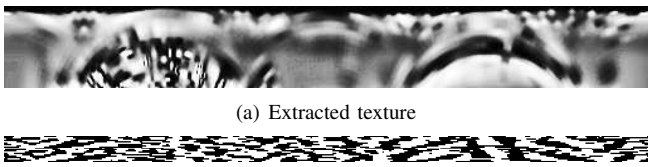
(a) Extracted texture

(b) Iris Code

Fig. 5. LBT=1, HD=0.403.

In this work we specifically focus on the quantisation strategy in JPEG XR. After the LBT transform, the coefficients in the DC,LP,HP bands are quantised by a (integer) value q in the range 1 - 255. In the case of "uniform" quantisation (which is the default setting), all three bands are quantised with the same value. For controlling the amount of compression, q is scaled but can only be of integer type. However, JPEG XR also allows to apply different quantisation parameters for the DC, LP, and HP subbands besides the uniform strategy (in any case, the coefficients within one of these subbands are all quantised with an identical value). This corresponds to giving different emphasis to low frequency (DC band), mid frequency (LP band), and high frequency (HP band) information, respectively.

The aim of this work is to optimise the quantisation parameter settings for the three DC,LP,HP bands in the context of iris recognition instead of applying the default uniform strategy. Results will also shed light on the question which frequency bands do carry the most discriminative information in iris imagery.



(a) Extracted texture

(b) Iris Code

Fig. 6. LBT=2, HD=0.385.

Since our experiments are focused on the evaluation of those quantisation-related questions, we do not describe the subsequent JPEG XR stages in the following, please consult the standard or related publications with respect to these issues [13].

IV. Experiments on Optimising JPEG XR Compression of Iris Sample Data

A. Iris Recognition and Iris Database

It is crucial to assess the effects of compressing iris samples using a set of different iris recognition schemes since it can be expected that different feature extraction strategies will react differently when being confronted with compression artefacts and reduced image quality in general.

Many iris recognition methods follow a quite common scheme [15], close to the well known and commercially most successful approach by Daugman [16]. In our pre-processing approach (following e.g. Ma et al. [17]) we assume the texture to be the area between the two almost concentric circles of the pupil and the outer iris. These two circles are found by contrast adjustment, followed by Canny edge detection and Hough transformation. After the circles are detected, unwrapping along polar coordinates is done to obtain a rectangular texture of the iris. In our case, we always re-sample the texture to a size of 512x64 pixels. Subsequently, features are extracted from this iris texture (which has also been termed polar iris image). We consider the following four techniques in this work, which are selected to represent a broad variety of different template generation concepts:

(1) A wavelet-based approach proposed by Ma et al. [17] is used to extract a bit-code. The texture is divided into N stripes to obtain N one-dimensional signals, each one averaged from the pixels of M adjacent rows. We used $N = 10$ and $M = 5$ for our 512x64 pixel textures (only the 50 rows close to the pupil are used from the 64 rows, as suggested in [17]). A dyadic wavelet transform is then performed on each of the resulting 10 signals, and two fixed subbands are selected from each transform. This leads to a total of 20 subbands. In each subband we then locate all local minima and maxima above some threshold, and write a bitcode alternating between 0 and 1 at each extreme point. Using 512 bits per signal, the final code is then 512x20 bit. Matching different codes is done by computing the Hamming Distance.

(2) Again restricting the texture to the same $N = 10$ stripes as described before, we use a custom C implementation similar to Libor Masek's Matlab implementation (`http://www.csse.uwa.edu.au/~pk/student`

projects/libor/sourcecode.html) of a 1-D version of the Daugman iris recognition algorithm as the second feature extraction technique. A row-wise convolution with a complex Log-Gabor filter is performed on the texture pixels. The phase angle of the resulting complex value for each pixel is discretized into 2 bits. Those 2 bits of phase information are used to generate a binary code, which therefore is 512x20 bit (again, Hamming Distance can be used for similarity determination).

(3) The third algorithm has been proposed by Ko et al. [18]. Here feature extraction is performed by applying cumulative-sum-based change analysis. The algorithm discards parts of the iris texture, from the right side [45^o to 315^o] and the left side [135^o to 225^o], since the top and bottom of the iris are often hidden by eyelashes or eyelids. Subsequently, the resulting texture is divided into basic cell regions (these cell regions are of size 8×3 pixels). For each basic cell region an average gray scale value is calculated. Then basic cell regions are grouped horizontally and vertically (one group consists of five basic cell regions). Finally, cumulative sums over each group are calculated to generate an iris-code. If cumulative sums are on an upward slope or on a downward slope these are encoded with 1s and 2s, respectively, otherwise 0s are assigned to the code. In order to obtain a binary feature vector (to enable Hamming Distance computation for comparison) we rearrange the resulting Iris Code such that the first half contains all upward slopes and the second half contains all downward slopes. With respect to the above settings the final iris-code consists of 2400 bits.

(4) Finally, we employ the feature extraction algorithm of Zhu et al. [19] which applies a 2-D wavelet transform to the polar image first. Subsequently, first order statistical measures are computed from the wavelet subbands (i.e. mean and variance) and are concatenated into a feature vector. The similarity between two of these real-valued feature vectors is determined by computing the corresponding l^2-Norm.

We used the CASIAv3 Interval dataset (http://www.cbsr.ia.ac.cn/IrisDatabase.htm/) in the experiments. It consists of NIR images with 320×280 pixels in 8 bit grayscale .jpeg format (high quality) of 249 persons, where for many persons both eyes are available which leads to 391 (image) classes overall.

For intra-class matches (genuine user matches), we consider all possible template pairs for each class (overall 8882 matches), while for inter-class matches (impostor matches) the first two templates of the first person are matched against all templates of the other classes (overall 2601 matches).

B. Compression Techniques Settings

In JPEG XR quantisation, we aim at optimising the relation among the quantisation parameters for the three subbands DC, LP, and HP, i.e. we look for the triple q:r:s which provides the best solution in terms of recognition performance (measured in equal error rate (EER)). Since it is not obvious that there exists a unique optimal solution independent of target bit rate, we look for an optimal q:r:s triple with respect to a certain target bitrate. Since the number of q:r:s triples is way too large to be tested exhaustively, we have quantised the search space into 18 DC bands, and 15 LP and 15 HP bands, respectively. Still 4050 possible combinations need to be considered, but this is more tractable compared to $255^3 = 16581375$ triples without quantisation.

For enabling a fair comparison between the various quantised triples in the experiments, the same bitrate has to be targeted for all configurations. While specifying a target bitrate is straightforward in JPEG2000, JPEG XR suffers from the same weakness as JPEG being unable to explicitly specify a target bitrate. Therefore we have employed a wrapper-program, continuously scaling the JPEG XR quantisation triples (i.e. multiplication of all three components with the same factor) to achieve a certain target bitrate (given in bytes per pixel bpp). Since q,r,s can attain integer values only, target bitrates are approximated as accurate as possible. In Fig. 7 we show an example of approximating a target bitrate of 0.1968 bytes/pixel for more then 2500 images. On average we get 0.1966 bytes/pixel with a maximal deviation of +5.97% and -6.21%.

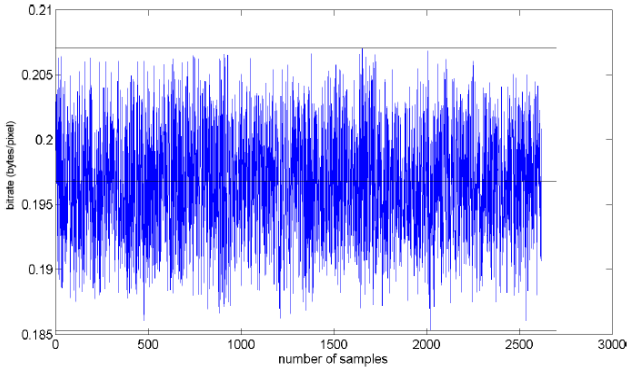


Fig. 7. Rate adaptation approximation.

For experimentation, we use the official JPEG-XR reference software 1.8 (as of September 2009) and for JPEG2000 compression, imagemagick 8.6.6.0.4-3 (employing libJASPER 1.900.1-7+b1) is used with standard settings.

The optimisation is done minimising the EER of the Masek implementation by setting LBT=0 since this is the fastest variant and there are no clear recognition advantages of using LBT=1,2 [12].

The questions we want to answer with our experiments are as follows:

1) Do the optimised settings outperform the "uniform" JPEG XR default settings ?
2) Do the optimised settings outperform JPEG2000 ?
3) Do the optimised settings also generalise to other bitrates (since they have been computed for a single target bitrate) ?
4) Do the optimised settings also generalise to other feature extraction schemes (since they have been computed for the Masek Iris Code) ?

 ISBN: 978-1-61208-195-3

C. Experimental Results

Fig. 8 shows computed tuples r:s, when all triples are normalised with $q = 1$. Out of all considered 4050 r:s(:1) triples, blue dots show configurations when the obtained EER is at least 5% better as compared to uniform q:r:s, and red diamonds depict configurations with at least 15% improvement. The target bitrate for the optimisation has been set to 0.19 bytes/pixel (filesize is 17 kBytes) for all experimental results shown. Note that experiments with different target bitrates lead to highly similar results with respect to the answers to the four questions raised above, but of course not with respect to the actual triples q:r:s computed.

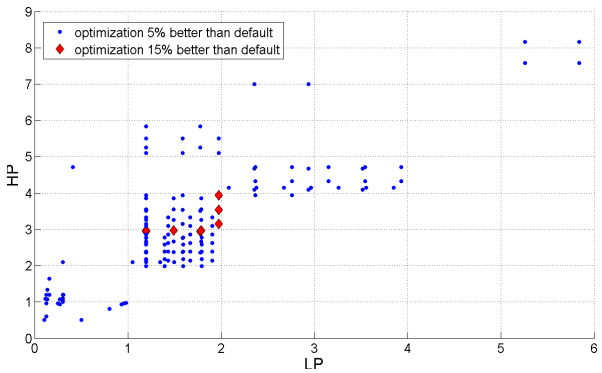
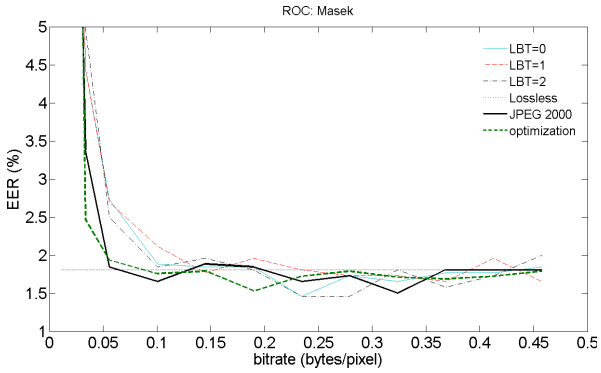
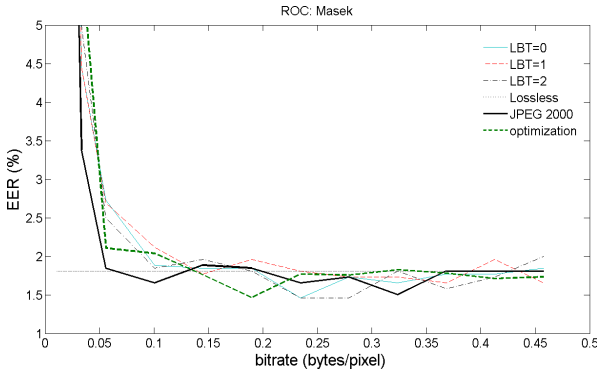


Fig. 8. Result Distribution



(a) q:r:s = 1:1.19:2.93



(b) q:r:s = 1:1.97:3.15

Fig. 9. Recognition with Masek Iris Code.

We clearly note that the best triples are not close to the uniform setting q:r:s = 1 but $1 < r < 2$ and $2.9 < s < 4$. This means that the higher frequency gets, the more severe quantisation should be applied.

Fig. 9 shows the results of two good q:r:s configurations for varying the bitrate in compression (x-axis) and performing iris recognition with the Masek Iris Code EER is plotted on the y-axis). For a comparison, we plot the curves for LBT=0,1,2 with uniform q:r:s and a curve obtained from applying JPEG2000. For both configurations we observe that for the optimisation target bitrate, the optimised q:r:s triple is clearly superior to the "uniform" JPEG XR variants and also superior to JPEG2000.

However, this superiority does not at all extend to other bitrates. The bitrate range where these triples exhibit better performance is quite limited. This means that in an application, specific q:r:s triples need to be optimised for different target bitrates. The behaviour of those two configurations as shown in Figs. 9.a and 9.b is very similar except for the the range of bytes/pixel < 0.15. Here the better preservation of LP and to a lesser extent HP data for q:r:s = 1:1.19:2.93 leads to performance close or even better to JPEG2000 (see Fig. 9.a). Note that for bitrates > 0.05, in many cases EER derived from lossy compression is superior to the values computed from uncompressed data - this effect has been observed in many studies and is due to the de-noising effect of moderate compression settings.

Finally, we want to answer the question in how far the good results of the computed triples do generalise to different types of feature extraction schemes and resulting Iris Codes without explicit optimisation for the respective algorithms.

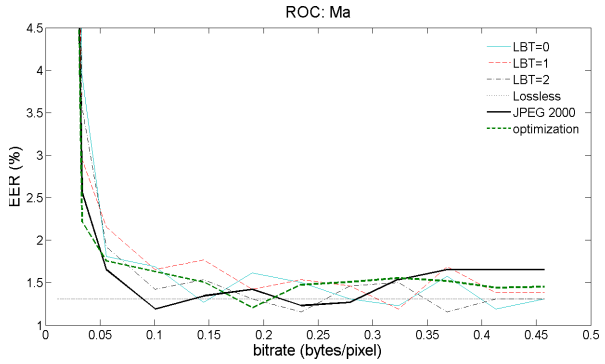
In Fig. 10, we compare the behaviour of the three remaining feature extraction techniques when applied to sample data which have been compressed using the triple q:r:s = 1:1.97:3.15 – which has been optimised for the Masek Iris Code at bitrate 0.19 bytes/pixel. We notice that for the target bitrate, the EER values are fairly good for all three types of iris codes. While for the Ma and Ko variants, the result is better compared to JPEG2000 and all three uniform variants, the Zhu variant is slightly inferior to LBT=2 only, but superior to all other compression schemes including JPEG2000. So it seems that this q:r:s configuration is able to preserve texture information for the targeted bitrate very well, no matter which subsequent feature extraction technique is being applied.

On the other hand, we notice again that the bitrate range where this good behaviour is observed is actually quite limited (except for the Ko Iris Code, where we see good results for lower bitrates also). The specifically good results at bitrate 0.05 bytes/pixel for the Ko and Zhu feature extraction schemes are probably due to optimal denoising behaviour at this compression ratio for these two schemes.

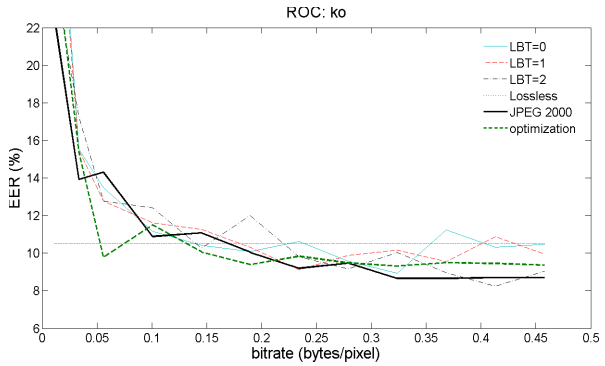
V. Conclusion

We have found that optimising the JPEG XR quantisation strategy leads to improved iris recognition results for a wide range of different feature extraction types. The optimised strategy does not only outperform the default quantisation strategy

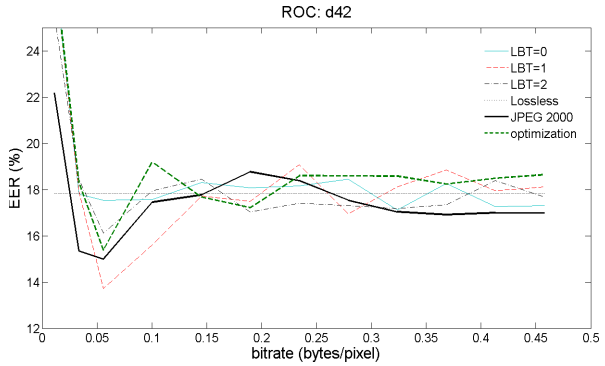
 ISBN: 978-1-61208-195-3



(a) Ma Iris Code



(b) Ko Iris Code



(c) Zhu Iris Code

Fig. 10. q:r:s = 1:1.97:3.15

but also iris recognition relying on JPEG2000 compression. The observed behaviour is only found in a small range of bitrates close to the target bitrate that has been used for optimisation, however, the optimised parameters for a specific feature extraction technique do also provide good results for other types of Iris Codes. The general trend with respect to the importance of different frequency bands is that as opposed to the JPEG XR default configuration, middle LP frequencies and even more pronounced high HP frequencies should be quantised more severely compared to the low frequency DC information.

REFERENCES

[1] K. Horvath, H. Stögner, A. Uhl, and G. Weinhandel, "Lossless compression of polar iris image data," in *Proceedings of the 5th Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA 2011)*, ser. LNCS, J. Vitria, J. M. Sanches, and M. Hernandez, Eds., vol. 6669. Springer Verlag, 2011, pp. 329–337.

[2] R. W. Ives, R. P. Broussard, L. R. Kennell, and D. L. Soldan, "Effects of image compression on iris recognition system performance," *Journal of Electronic Imaging*, vol. 17, pp. 011 015, doi:10.1117/1.2 891 313, 2008.

[3] J. Daugman and C. Downing, "Effect of severe image compression on iris recognition performance," *IEEE Transactions on Information Forensics and Security*, vol. 3, no. 1, pp. 52–61, 2008.

[4] S. Matschitsch, M. Tschinder, and A. Uhl, "Comparison of compression algorithms' impact on iris recognition accuracy," in *Proceedings of the 2nd International Conference on Biometrics 2007 (ICB'07)*, ser. LNCS, S.-W. Lee and S. Li, Eds., vol. 4642. Springer Verlag, 2007, pp. 232–241.

[5] S. Jenisch, S. Lukesch, and A. Uhl, "Comparison of compression algorithms' impact on iris recognition accuracy II: revisiting JPEG," in *Proceedings of SPIE, Security, Forensics, Steganography, and Watermarking of Multimedia Contents X*, vol. 6819, San Jose, CA, USA, Jan. 2008, p. 68190M ff.

[6] S. Rakshit and D. Monro, "Effects of sampling and compression on human iris verification," in *Proceedings of the IEEE International Conference on Acustics, Speech, and Signal Processing (ICASSP 2006)*, Tolouse, France, 2006, pp. II–337–II–340.

[7] ——, "An evaluation of image sampling and compression for human iris recognition," *IEEE Transactions on Information Forensics and Security*, vol. 2, no. 3, pp. 605–612, 2007.

[8] M. Konrad, H. Stögner, and A. Uhl, "Custom design of JPEG quantization tables for compressing iris polar images to improve recognition accuracy," in *Proceedings of the 3rd International Conference on Biometrics 2009 (ICB'09)*, ser. LNCS, M. Tistarelli and M. Nixon, Eds., vol. 5558. Springer Verlag, 2009, pp. 1091–1101.

[9] ——, "Evolutionary optimization of JPEG quantization tables for compressing iris polar images in iris recognition systems," in *Proceedings of the 6th International Symposium on Image and Signal Processing and Analysis, ISPA '09*, Salzburg, Austria, Sep. 2009.

[10] J. Hämmerle-Uhl, C. Prähauser, T. Starzacher, and A. Uhl, "Improving compressed iris recognition accuracy using JPEG2000 RoI coding," in *Proceedings of the 3rd International Conference on Biometrics 2009 (ICB'09)*, ser. LNCS, M. Tistarelli and M. Nixon, Eds., vol. 5558. Springer Verlag, 2009, pp. 1102–1111.

[11] G. Kostmajer, H. Stögner, and A. Uhl, "Custom JPEG quantization for improved iris recognition accuracy," in *Emerging Challenges for Security, Privacy and Trust. Proceedings of the 24th IFIP International Information Security Conference 2009 (IFIP SEC'09)*, ser. IFIP AICT, D. Gritzalis and J. Lopez, Eds., vol. 297. Springer Verlag, May 2009, pp. 76–86.

[12] K. Horvath, H. Stögner, and A. Uhl, "Effects of JPEG XR compression settings on iris recognition systems," in *Proceedings of the 14th International Conference on Computer Analysis of Images and Patterns (CAIP 2011)*, ser. LNCS, P. Real, D. Diaz-Pernil, H. Molina-Abril, A. Berciano, and W. Kropatsch, Eds., vol. 6855. Springer Verlag, 2011, pp. 73–80.

[13] F. Dufaux, G. J. Sullivan, and T. Ebrahimi, "The JPEG XR image coding standard," *IEEE Signal Processing Magazine*, vol. 26, no. 6, pp. 195–199, Nov. 2009.

[14] D. Taubman and M. Marcellin, *JPEG2000 — Image Compression Fundamentals, Standards and Practice*. Kluwer Academic Publishers, 2002.

[15] K. Bowyer, K. Hollingsworth, and P. Flinn, "Image understanding for iris biometrics: A survey," *Computer Vision and Image Understanding*, vol. 110, no. 2, pp. 281 – 307, 2008.

[16] J. Daugman, "How iris recognition works," *IEEE Transactions on Circiuts and Systems for Video Technology*, vol. 14, no. 1, pp. 21–30, 2004.

[17] L. Ma, T. Tan, Y. Wang, and D. Zhang, "Efficient iris recognition by characterizing key local variations," *IEEE Transactions on Image Processing*, vol. 13, no. 6, pp. 739–750, Jun. 2004.

[18] J.-G. Ko, Y.-H. Gil, J.-H. Yoo, and K.-I. Chung, "A novel and efficient feature extraction method for iris recognition," *ETRI Journal*, vol. 29, no. 3, pp. 399 – 401, 2007.

[19] Y. Zhu, T. Tan, and Y. Wang, "Biometric personal identification based on iris patterns," in *Proceedings of the 15th International Conference on Pattern Recognition (ICPR'00)*, vol. 2. IEEE Computer Society, 2000, pp. 2801–2804.

 ISBN: 978-1-61208-195-3

The Design of an Adaptive Multimedia Presentation System

Nick Rowe
Faculty of Technology
Bournemouth and Poole College
Bournemouth, UK
e-mail: nrowe@bpc.ac.uk

Abstract—**The nature of adaptive multimedia presentation system within an e-learning environment is considered and the ability of such a system to add and remove fragments of a presentation discussed. The advantages and drawbacks of using learning objects are described and the ability to map an object to additional data identified as a means to provide links to other objects. The practical design of an adaptive E-Learning system is then considered by first referencing two open source systems, (AHA and MOT), and then considering the requirements of a prototype system. The principle components of such a system are identified and the system structure outlined. Relationships between principle entities in a presentation database are then discussed with the emphasis on adapting the content of the presentation by providing more or less detail. Links between learning segments are then discussed leading to consideration of the different types of link and how this is likely to affect the user experience. Finally, future issues are discussed leading to a consideration of the topics that would need to be evaluated in the use of the prototype system**.

Keywords – multimedia; E-Learning; learning objects; adaptive; ontology; education.

I. INTRODUCTION

Firstly, the paper considers the nature of adaptive E-Learning with reference to the common components found in such systems. The methods of *adaptivity* are then considered and split into two discrete mechanisms: navigational and content-level adaption, the latter being considered in the context of learning objects. The advantages of fragmenting a presentation are investigated and the benefits of interoperability and reuse are weighed against educational context and independence. The concept of wrapping the learning object in meta-data is introduced and using this data to provide links between the objects explained. Different link relationships are then defined and pre-requisite and co-requisite relationships discussed in detail.

The practical design of a prototype E-Learning system is described where a multi-modal interface has been designed to allow different elements to be synchronized under one timeline to allow reinforcement of the content. The structure of prototype E-Learning system is then described and the design of entities and the relationships between them identified. In particular, the relationships between the learning objects, (called segments), link segment and pending question entities are discussed.

The mechanisms for adapting the presentation content are detailed and the formulation of an Adaptive Descriptor identified as a key element in the process of producing presentations with different levels of detail for different users. Next, the authorship of the content is approached with particular regard to populating the Segment Level Identity using a practical editing system for the segment. This editing system allows additional fragments of the presentation to be inserted or fragments of the presentation to be removed. By such action, different levels of detail can be created from a single segment.

Lastly, the structure of these learning segments are considered within the context of knowledge-based ontologies. Characteristics of these segment maps are explored and the salient relationships between linking segments identified; particularly in regard to pre- and co-requisite links. Conclusions are then given and characteristics of the prototype design itemised. Design advantages are listed with a view to being considered against the increased production time of these materials.

II. ADAPTIVE E-LEARNING

Generally, acknowledging the important relationship between individual learners and the material they are viewing has a long history. Shute and Towle, [10], note that the goal of aptitude-treatment interactions, (ATI), research is to provide information about learner characteristics that can be used to select the best learning environment for a particular student to optimise learning outcome. They go on to itemise four components of adaptive E-Learning:

- Content Model, including a knowledge map
- Learner Model, containing information about the user
- Instructional Model, concerned with the presentation of materials
- Adaptive Engine, which uses information from other models to drive the system

Systems can access the learner in terms of domain-dependent information and domain-independent information. The former gains knowledge of the learner through pre-tests and performance data and the latter keeps track of the cognitive abilities and personality traits of the individual. Systems concerned with adaptive instruction tend to base their *adaptivity* on assessments of emergent content knowledge or adjustments of material based on

 ISBN: 978-1-61208-195-3

learner styles. The latter is a less suitable criterion than cognitive abilities for making adaptive instructional decisions.

It is true to say that research into adaptive hypermedia is at the crossroads of multimedia presentation and user modeling. Brusilovsky, [2], defines such systems as giving a presentation that is adapted specifically to the user's knowledge of the subject and suggests a set of most relevant links to proceed further. The second part of the definition is really a type of navigational adaptivity where the learner is given a level of control of over what content to see. So, two distinct areas of adaption are created: content level adaption, often called adaptive presentation, and link level adaption, called adaptive navigational support.

One interesting area that Brusilovsky identifies is the requirement to manipulate a presentation in certain ways according to the user needs. The information is offered in the context of *canned text adaption* and suggests applications can insert and remove text, alter fragments, stretch text, sort fragments and dim fragments. If the concept is extended to multimedia applications then these presentations can be manipulated in a similar manner. The fragments can be manipulated via an adaptive engine. This leads to the practical realization that presentations need to be reduced to fragments to allow these elements to be manipulated. These fragments are generally termed *learning objects* and there has been plenty of research around their use.

A good example of adaptive navigational support offered by an application is AHA!, an open source adaptive hypermedia platform, [6]. The system uses adaptive linking to suggest content for the user. It makes use of prerequisite relationships between the learning objects to link related references ensuring that the user has the required knowledge base to understand a given link. In this manner the user makes decisions about the content they wish to learn.

III. LEARNING OBJECTS

The definition of a learning object is *any entity, digital or non-digital, which can be used, re-used and referenced during technology-supported learning*, [7]. Although the definition is easily understood and widely accepted, the advantages gained by splitting up a lesson into learning objects are somewhat controversial. One of the biggest benefits often sited are that these objects can be reused and repurposed, [1]. However, this interoperability and reusability may have been overstated in the past. McGreal, [9], points out the difficulties in taking a learning object and reusing it in a different environment. This is principally because it is difficult to create learning objects independent of the context it was made in. The likelihood is that the object bears the imprint of the ideology and culture it was produced in.

Consequently, it is difficult to standardize a learning object and an object-oriented approach, as applied to software environments. This is incongruous in the complex context of learning, especially when the learning material is based on narrow technical and specialized concepts. Despite the challenge, the concept persists driven by the joint goals of reuse and adaptivity.

However, whilst learning objects may not be easily reused, segments of a similar content may be referenced and linked to provide the user with additional information.

Boyle, [1], describes the learning object as a wrapper around this object. This wrapper describes the component structure of the object, and includes the descriptive metadata. The learning object is thus packaged in a standard container format. This packaged object can be stored in digital repositories. The metadata permits fast effective searches to retrieve learning objects suitable for a particular purpose. A direct link can be made to the idea of learning objectives in pedagogical theory. This mapping suggests that each learning object should be based on one learning objective or clear learning goal, which links back to the original definition.

The design of the learning objects should be considered carefully to ensure they have minimal bindings to other units, (as well as being as context-free as possible). Even Boyle, [1], admits that this decoupling of learning objects is a considerable challenge and notes that this may be at odds with providing rich, integrated learning experiences. One way round this problem is to create a compound object consisting of two or more independent learning objects that are linked to try to achieve a richness not available to a single object, whilst maintaining a significant basis for re-use.

IV. THE LINKING OF LEARNING OBJECTS

In fact, the linking of learning objects goes further than this and a particular syllabus may be defined as a linked series of these objects. Indeed, much of the research on developing E-Learning systems over the last five years has concentrated on these links. In the design of the open source adaptive hypermedia platform AHA!, (Adaptive Hypermedia Architecture), De Bra et al., [6], describe how the system has been designed to use adaptive linking to suggest content for the user. It uses, what they term, *prerequisite relationships* to link related references and form a path through material. The system is capable of selecting and presenting information content based on the user's previous actions which are processed and stored in a user model. The system selects and annotates the links in a way that guides the user towards the most relevant information. In this way, navigational adaptivity is provided and the system builds concept relationships between the objects.

Once the learning material has been segmented into individual learning objects, two aspects become important for the presentation of these materials. Firstly, a lesson can be considered to be a chosen sequential set of these

 ISBN: 978-1-61208-195-3

segments and secondly that any segment presented may, to a lesser or greater degree, be connected to another segment in the learning repository. These two elements become essential to the development of any E-Learning system. Authoring a lesson to be presented becomes a process of choosing already available segments from the repository and creating new segments for areas not available. The presentation system then needs to be provided with a set of links to other relevant segments that the student may find useful and optional decide to view. The data in the repository needs to be mined to find the relevant links to each segment within the lesson.

Whilst pre-requisite links will allow the user to create a path through a series of segments, it is also useful to provide the user with the ability to view *co-requisite* links as well. These links to segments that are close in content to the viewed segment and provide the user with the ability to view segments in some way connected to the current segment to provide reinforcement and to see the same topic from a different perspective. Figure 1 shows how this can be implemented on a practical system.

Figure 1. An Adaptive Linking Dialog showing Pre- and Co-requisite Connections.

The linked segment entries are ordered by a strength metric, (in brackets). This allows the strength of the link to be graded with some indication of the closeness of the topic covered by the other segment. When linked segments are selected they can be played and are shown within the context of the lesson they were designed in. This context may well be different from the original segment, but as such provides the user with a fuller picture of topic.

V. THE PRACTICAL DESIGN OF AN E-LEARNING SYSTEM

In practice, realization of all these concepts gives rise to two distinct functions of any E-Learning system. These are the authorship of materials and delivery of these materials. Cristea et al., [4], describe an attempt to combine two hypermedia systems, authoring with MOT, (My Online Teacher), and delivery with AHA. MOT uses domain mapping to structure and organize the resources. It uses adaption rules to build an *assembly language* of adaption. Concept weights, (meta-data), are then used to alter the presentation and make it adapt to a particular user. These weights can represent different measurable aspects of a learning fragment like difficulty or importance.

A Common Adaption Format, (CAF), sits between the two systems to convert data from MOT into a form understood by AHA. This is expressed as an XML document. In this manner the systems attempt to establish a common platform and format for the representation of adaptive educational hypermedia: an extremely important goal if learning object re-use is to become a practical reality.

VI. THE DEVELOPMENT OF A PROTOTYPE E-LEARNING SYSTEM

With current computing power there is rarely a problem with delivering rich multimedia content. One of the main issues that requires careful consideration is the synchronization of the different media that go to make up the presentation. Languages like SMIL, [3], seek to remedy this by providing a language to allow multimedia components to be synchronized and presented together. Although the presentations produced this way are impressive, authorship is complex.

A prototype development system was designed, [5], with these requirements in mind allowing multiple units acting together to reinforce the overall delivery.

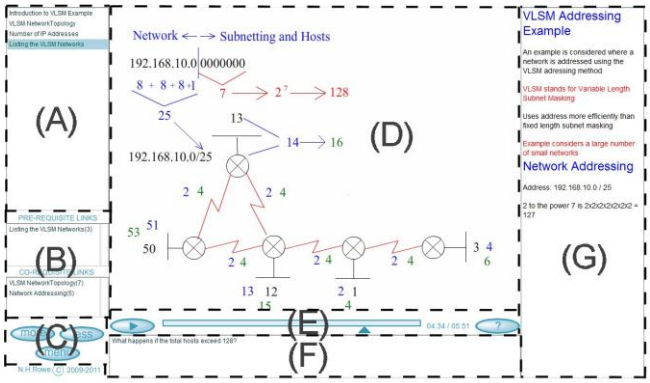


Figure 2. Screen Layout of a Multi-focus E-Learning System.

This has been developed and Figure 2 shows the screen layout of the enhanced system. Here, seven elements are synchronized to act from the same timeline. Element A is a list of segments, (learning objects), that form a lesson and forms a selectable table of contents for the presentation. B provides a selectable display of linked segments that can be optionally viewed by the user. C provides the user with buttons to increase or decrease the level of detail of the presentation. D is the viewing panel for the audio-visual material and E is the timeline control area. This section gives the temporal control of the presentation and also allows the user to ask textual questions. These can be accessed by the segment author and additional presentations created to answer the question. When an answer is published it can be

 ISBN: 978-1-61208-195-3

viewed by all users in area F and thus this area is filled with frequently asked questions for the currently playing segment. When a question is selected, the solution presentation plays temporarily interrupting the current content for the span of the answer. Area G contains incrementally loading HTML, (iHTML). Here the content, text and images, is displayed and reveled in real-time, synchronized with the main presentation in a similar manner to subtitles. Each block, (or paragraph), of the HTML code is not displayed until a specific time is reached in the presentation.

With the publishing of answers to previously asked questions, during the life of the presentation more questions are likely to be asked and therefore the presentation matures over time. This provides more supplementary information useful to a learner viewing the presentation for the first time.

VII. THE STRUCTURE OF AN E-LEARNING SYSTEM

Once the decision in establishing the segment as the heart of an E-Learning system has been made, the rest of the system can be designed around it. Entities including the user and materials, to test the user knowledge, can be included in the E-Learning database.

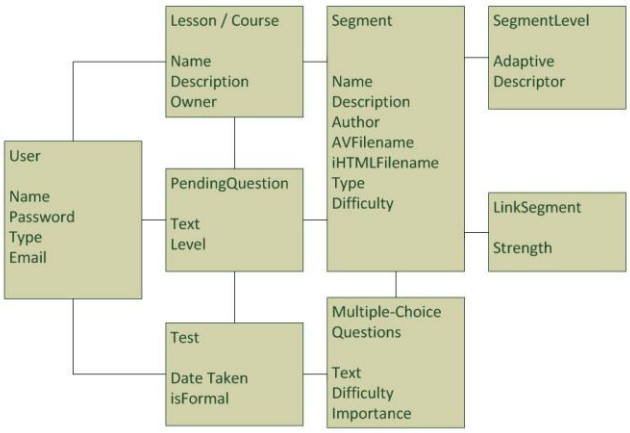


Figure 3. A Prototype E-Learning Structure.

In the development of the materials the educational concepts must be isolated from a lesson of a course and developed into learning objects. The syllabus of a lesson consists of an ordered set of concepts and a course is an ordered set of lesson. Each concept is formed into a segment. Initially, a segment contains audio-visual resources required for its presentation. This entity is then given a set of attributes including the name of the AV file that contains the presentation. This information can be accessed by independent engines that could be looking for links between segments. The algorithm to identify these links could be complex but the end result is simply to record the link between two segments in the *LinkSegment* entity and record a strength attribute that gives an indication of the strength of the link between the two segments. The *PendingQuestion* entity records the student question linked to a segment for access by the author of the segment to provide an answer. Once answered the question becomes another segment linked to the first with a maximum strength value.

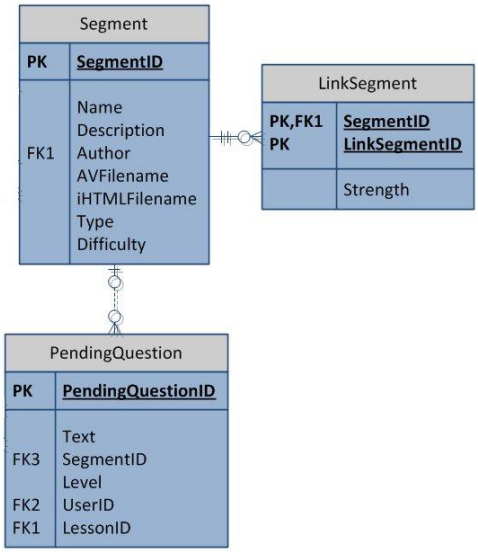


Figure 4. Relationships between a Learning *Segment*, *LinkSegment* and *PendingQuestion* entities.

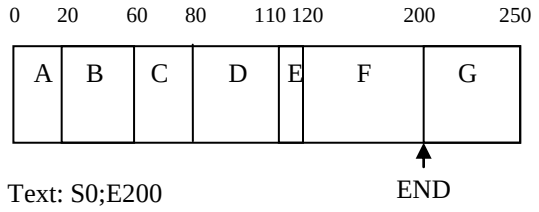
VIII. ADAPTING THE E-LEARNING PRESENTATION

To make a segment adapt to the user's needs during presentation the author must also determine parts of the AV presentation that will be viewed at different levels of detail. By providing these different levels each segment becomes adaptable. During a presentation, the user can be presented with the segment information at a preferred level of detail. The user can then alter this level to provide more or less detail during the presentation. The system can record these levels and change these levels based on other information in the database including the results to tests linked to the segment. Thus, the system adapts to the user needs by presenting the material at the correct level of detail.

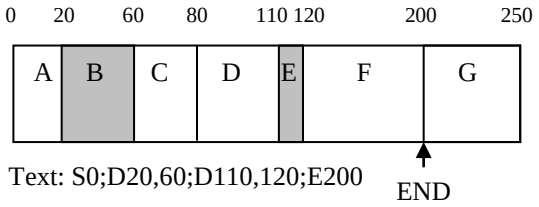
Thus, more or less detail can be created to a standard form and adaptively chosen for the user. A textual code is used to allow the system to piece together the presented form for the level chosen and acts as an adaptive descriptor for the segment. This is shown in Figure 5. Part (a) shows the media file being played as it was recorded from frame 0 to 200. The control text simply gives the end frame so that additional fragments are not played at the end of the file. Part (b) shows fragments of the media file being left out to create a less detailed presentation. Here, fragments B and E are left out of the presented sequence. The control text indicates which frames are to be removed. It also includes the end frame. Part (c) shows more detail being added to the presentation by substituting the larger fragment G in between C and D, (at time 80). Here, more detail can be added to specific parts of the file and therefore particular concepts are elaborated

 ISBN: 978-1-61208-195-3

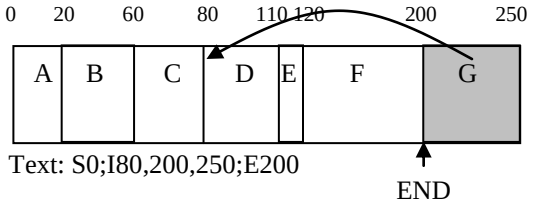
within the segment. These additional fragments are added to the end of the media file and can be additionally recorded at the time the presentation is made. The adaptive descriptor marks the frames to be removed and the frames to be substituted. Thus, a single media file is used for all levels of detail and adaptively presented by use of the set of descriptors for each segment at different levels.



(a) Normal level of detail, (as recorded). Segments A to F are played sequentially



(b) Less detail in presentation. Segments A, C, D and F are played sequentially



(c) More detail in presentation. Segments A, B, C, G, D, E and F are played sequentially

Figure 5. Three levels of detail from a single audio-visual fragment.

The relationship between *Segment* and the *SegmentLevel* entities is shown in Figure 6.

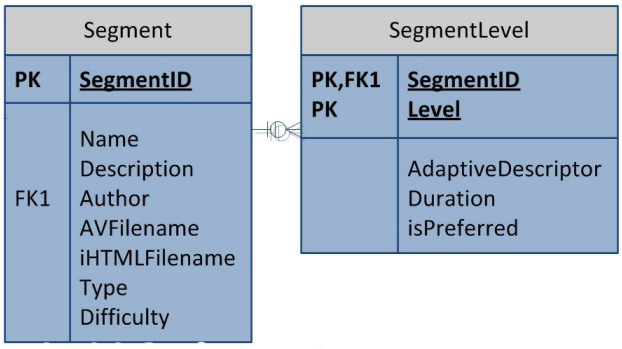


Figure 6. Relationship between a segment and levels of detail.

The attributes of the latter entity consist of the *adaptive descriptor*, a total value for duration of the segment at that level and a marker for the author's preferred starting level. The duration is calculated at the authorship stage in order that, when presented, the total duration of the lesson can be calculated easily by adding all the segment durations at each particular level.

The practical consequence of this editing process is to populate the Segment Level Entity as shown in Figure 7.

SegmentID	Level	AdaptiveDescriptor	Duration	isPrefered
1	0	S10;H20,0;H30,1;H80,2;I120,150,175;E150	165	0
1	1	S10;H20,0;D25,55;H80,2;E150	110	1
1	2	S10;H20,0;D25,55;H80,2;D85,110;E150	85	0
2	0	S5;E55	50	1
2	1	S5;D40,50;E55	40	0
3	0	S0;H10,0;H25,1;H115,2;E150	150	1
3	1	S0;H10,0;D15,45;H115,2;E150	120	0
4	0	S10;H25,0;H80,1;I115,150,165;E150	155	1
5	0	S5;H40,0;E145	140	1
7	0	S0;H1,0;H5.172,1;H14.994,2;H20.819,3;H26.383,4;E33...	33.3844	1
7	1	S0;H5.172,1;H14.994,2;H20.819,3;D21.916,33.266;E33...	22.0344	0
7	2	S0;H5.172,1;D10.84,33.266;E33.3844273	10.9584	0
8	0	S0;H16.718,0;H41,1;H64,2;E112.0389737	112.039	1
9	0	S0;H9.133,0;H13.035,1;H37.666,2;H81.136,3;H109.478...	260.806	0
9	1	S0;H9.133,0;H13.035,1;D14.341,97.541;H109.478,4;D1...	111.256	1

Figure 7. Display of data within the Segment Level Entity.

Here, in the left hand column, 9 segments are shown with an ID from 1 to 9. For each segment, levels are shown from 0 to a maximum value in the next column to the right. The middle column shows the adaptive descriptor with elements separated by a semi-colon. Duration, (in seconds), for each level of detail are given in the next column and it can be seen that the higher the level, the less time the content lasts for. This is because at each increasing level part of the presentation is removed according to the editing process used by the author that formed the data. The right hand side column shows which level is preferred by the author as a default level. This level may be changed by the user, and remembered by the database, but this will be the level viewed by this particular user the first time they encounter this content.

The highlighted second row of the table indicates the preferred level of segment ID 1. The segment will be shown at this level of detail by default. The user can reduce the level of detail, for example if the material is familiar to them, by selecting the 'Less' button. In this case, the segment will now only be 85 seconds long and a portion of the presentation will have been removed. If the user pressed the 'More' button instead, the level would decrease from 1 to 0 extending the presentation to 165 seconds. The difference in each level of the segment is contained in the adaptive descriptor and in each case as the level increased, part of the presentation was removed. For example, in the case of moving from level 1 to 2 an additional element 'D85,110;' is added which has the effect of removing 25 seconds from the presentation and therefore the duration reduces by 25 seconds. By editing the 25 seconds out of this level the author has made the decision that this portion of the presentation was not required at this level. The user can alter

 ISBN: 978-1-61208-195-3

the level of detail, by selecting the more or less buttons, if this turns out to be important, but in this manner the presentation is less likely to contain unnecessary detail, but be pitched at the correct level.

IX. AUTHORSHIP OF AN ADAPTIVE PRESENTATION

Authorship of such a system relies on the choosing fragments on a temporal basis and marking sections to be excluded or included at a particular level. A particular lesson presentation is driven from a sequential set of segments. Each of these segments has additional data connected to the AV file and an adaptive descriptor allows these additional elements to be synchronized with the original AV file. Once the AV files have been collated, the author must mark on a time scale fragments of the segment to be deleted and inserted at a particular level. The adaptive descriptor is then created for each level and the *SegmentLevel* entity populated. The adaptive descriptor can also be used to mark times within the segment presentation to display the iHTML elements to allow them to synchronise with the main audio-visual element. Playback is then controlled by the adaptive descriptor with the playhead being moved ahead in real-time to skip over a section or moved to a completely different time to allow more information to be inserted.

Creating the descriptor consists of the author selecting the segment and this can be played in the system. In this creation mode a new set of editing functions are accessible to allow sections to be marked for insertion and deletion together with the ability to mark in time when iHTML paragraphs are to be displayed. A mechanism for doing this is shown in Figure 8.

These activities are done at each possible level and the author can create new levels, change the current level and mark a preferred level. This is used as a default level for a presentation if no additional information is in the database.

Generally, the original presentation will contain the maximum level of detail and new levels will delete selective sections to provide less detail, but the author may add addition sections with more detail if required. Once the process is complete the author publishes the segment and the adaptive descriptor is created for each level and the *SegmentLevel* entity populated for each level.

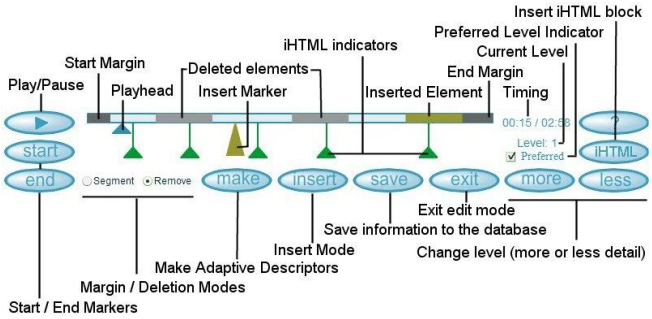


Figure 8. Editing functions for a segment.

The duration of each level is also stored so that the overall duration of a presentation can be calculated quickly. This value will change due to the choice of levels at each segment. The time bar reflects these changes in real-time as changes of level are made by the user.

X. THE STRUCTURE OF LEARNING SEGMENTS

There has been much work done on the structure and classification of knowledge. Since the Dewey Decimal system we have tried to position the vast and varied segments of knowledge into some kind structured network. Just like placing a book on a particular shelf in a library, ontologies seek to place nodes of knowledge onto a two dimensional map.

Holohan et al., [8], note that the combination of these ontologies with learning objects provide a powerful means to creating adaptable educational presentations. Authors can select and customise new or existing subject ontologies and employ a certain teaching and learning strategy in the generation of learning objects. They can also configure systems to offer strictly sequenced presentations to students, or to allow also varying degrees of free student navigation, based on the runtime incorporation of domain ontologies. Students in turn can take the generated courses in the preconfigured delivery environment, and this delivery is dynamically customised to the individual students' preferences and constantly monitored learning track.

There have been many different approaches to developing a structure and framework for this and mathematical systems have been and are being developed to map segments into a meaningful array providing information on the relative proximity of neighbouring segments. This, in turn, leads to providing an automatic metric for linked segment strength.

In this context, it is useful to define some characteristics of these segment maps. Figure 9 shows four types of relationships between learning segments and identifies different relationships between these objects.

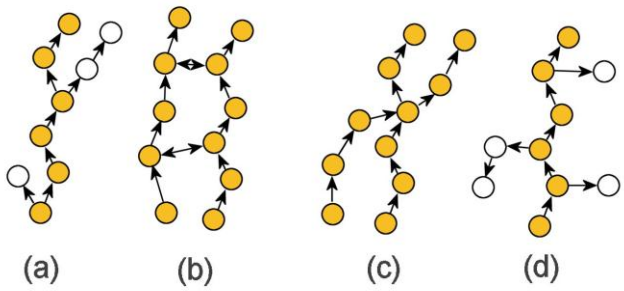


Figure 9. Relationships between segments in a knowledge database. (a) a lesson, (b) co-requisite linking, (c) twisting lesson paths and (d) question segments.

Part (a) shows that a lesson can be defined as a pathway through a set of segments, where the previous segment is a pre-requisite of the next. Branches off this path may exist, (and be given to the user as alternatives), but, at some stage,

 ISBN: 978-1-61208-195-3

a decision is made about the "preferred" route. Part (b) shows co-requisite linking between segments belonging to different lesson paths. By giving the user a link to the other lesson path the idea of lesson cross-talk occurs, where a user jumps across to another lesson path. This can carry on indefinitely provided enough lesson paths contain links. Part (c) shows that lesson paths may cross each other and the segment belonging to both lessons may have an elevated status as a result. Lastly, part (d) shows the relationship between a lesson path and segments created to provide solutions to questions raised by these segments. Effectively, a single solution segment is an orphan as it does not belong to any independent lesson path. However, is very tightly bonded to the original segment and may be part of a path created by users asking further questions about solutions given. This is, in fact, how an on-line discussion forum already works. It can be noted too that pre-requisite links, (including solutions to questions), are unidirectional, while co-requisite links are bidirectional.

XI. CONCLUSIONS AND FUTURE ISSUES

A method for allowing different levels of detail of a single learning segment to be adaptively displayed has been described together with an interface to allow the user to optionally select linked segments and frequently asked questions. The user can also submit questions to be answered by the segment authors to add to this set. Different algorithms can then be experimented with to provide these links with an appropriate strength metric. The diversity of approach and context of closely related segments can only serve to provide the user with different learning approaches to similar subjects. By splitting the presentations in smaller segments the relevancy to linked segments is likely to increase. The downside is that it is unlikely that presentations will be produced by simply selecting segments from the database and will generally require a larger presentation to be split into smaller units based on pedagogy. In this manner, continuity between segments is preserved. However, if lessons are formed from a series of co-requisite segments selected from the database by some engine then problems with continuity may arise if the segments are created by different authors or different contexts. Even this aspect is not simple as discontinuity of delivery within a lesson is not always a bad thing. It is often used as a tool to assist concentration and refresh the learner.

The separation of longer audio-visual files into smaller segments, adding data to these segments and then marking sections to provide different levels of detail all add to the production time of a presentation. There needs to be clear benefits that offset this extra effort. These immediate benefits are likely to be felt by users of the system in following areas:

- the ability to ask questions,
- to gain access to answers for other users' questions,
- to be able to increase and decrease the content detail,
- the ability to branch to other linked segments,
- to allow engines to trawl through the learning database and create links between segments automatically.

The secondary benefits are likely to include the advantages of designing a lesson or course with access to a central coordinated digital repository of segments that are conceptually linked and the ability to use data held in the user model to make decisions about the level of detail required for a segment for a particular user. These benefits increase with an increase in the number of users: using the system to learn or to author segments.

A prototype design has been developed and the next stage is to evaluate its effectiveness with a view to determining if the time taken to produce the materials in this manner is justified.

REFERENCES

[1] Boyle, T., 2003. Design Principles for Authoring Dynamic, Reusable Learning Objects. *Australian Journal of Educational Technology*.

[2] Brusilovsky, P., 2001, Adaptive Hypermedia. *User Modeling and User-Adapted Interaction* 11. 87-110.

[3] Bulterman,D.C.A, Rutledge, L., 2009. SMIL 3.0 Flexible Multimedia for Web, Mobile Devices and Daisy Talking Books. 2nd Ed. Berin:Springer-Verlog.

[4] Cristea, A.I., Smits, D., De Bra, P., 2005. Writing MOT, Reading AHA! - converting between an authoring and a delivery system for adaptive educational hypermedia. *A3EH Workshop*, AIED'05 (2005).

[5] Cutts, S., Davies, P., Newell, D. and Rowe, N., 2009. *Requirements for an Adaptive Multimedia Presentation System with Contextual Supplemental Support Media*, Proceedings of the MMEDIA 2009 Conference, Colmar, France.

[6] De Bra, P., Smits, D., Stash, N., 2006. Creating and Delivering Adaptive Courses with AHA! *Proceedings of the first European Conference on Technology Enhanced Learning*, EC-TEL 2006, Springer LNCS 4227, 21-33, Available from: http://aha.win.tue.nl/ publications.html. [Accessed 13 Feb 2012].

[7] IEEE. 2001. *IEEE Learning Technology Standards Committee* (LTSC) IEEE P1484.12 Learning Object Metadata Working Group; WG12 Home page.

[8] Holohan, E., Melia, M., McMullen, D., Pahl, C., 2005. Adaptive E-Learning Content Generation Based on Sematic Web Technology. *International Workshop on Applications of Sematic Web Technologies for E-Learning*, 12th International Conference on Artificial Intelligence in Education. 2005 Amsterdam. Netherlands

[9] McGreal, R. (Ed.), 2004. *Online Education Using Learning Objects*. London:Routledge, 59-70.

[10] Shute, V., Towle, B., 2003. Adaptive E-Learning. *Educational Psychologist*. 38(2), 105–114

 ISBN: 978-1-61208-195-3

Image Rotation Rectification

in steroscopic 3D on multi-core architectures

Ivan Velciov

"Politehnica" University
Timisoara, Romania
velciov.ivan@gmail.com

Cormac Brick, Marius Predut, Valentin Muresan

Movidius Ltd.
Dublin, Ireland
cormac.brick@movidius.com,
predutionut@yahoo.co.uk,
valentin.muresan@movidius.com

***Abstract*---In this paper, a rectification procedure necessary for obtaining a quality stereoscopic 3D effect, is presented. Rotation rectification is necessary due to misalignment of the pair of image sensors. In particular when the misalignment angle is, greater than 0.7 degrees, the perceived quality of the final 3D stereoscopic image is degraded. This paper offers a low-power, mobile, multi-core solution.**

Keywords - stereoscopic 3D; rectification; VLIW; multi-core

I. INTRODUCTION

In this paper, a rectification procedure necessary for obtaining a high quality stereoscopic 3D effect, will be presented. The 3D effect is obtained by taking two identical pictures, of the same scene, with two identical image sensors, which should be coplanar and with parallel axis. The distance between the two sensors should match the average interoccular distance (63 - 65 mm) [1]. Although the problem of rectification and/or calibration of cameras has been well covered in Computer Vision literature [2] [3] [4], this paper offers a low-power, mobile, multi-core implementation. Even though, in theory, the camera sensors are considered to be coplanar, parallel and having the same rotation angle, but due to the manufacturing process certain errors can be introduced.

The rectification of these errors is what this paper addresess. The first set of rectifications needed, have to do with component placement. Even a component placement tolerance of +/- 0.1 mm (or even 0.025 mm) would introduce noticeable vertical offsets, which need to be corrected. The second problem refers to the fact that a certain physical rotation of the sensors is possible. Starting at an angle of 0.7 degrees, the rotation becomes apparent to the viewer, introducing discomfort. The second set of rectifications, are sensor dependent and involve colour gain, white balance, focal range, lens distorsions, to name just a few.

The next section will go into more detail about stereoscopic 3D, continued by a short presentation of Movidius' platform, on which the rotation rectification was implemented.

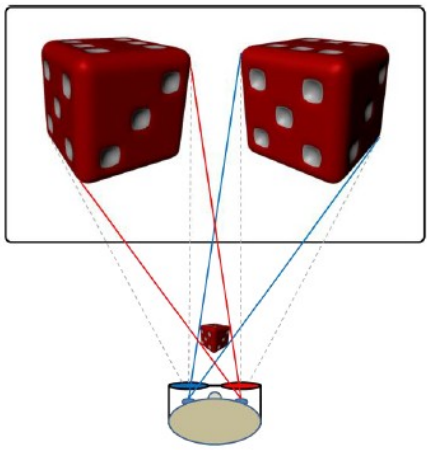


Figure 1. Stereoscopic 3D © 2010 CyberLink Corp.

The results section will describe the achieved performance. The entire test process will be presented afterwards. Finally, conclusions will be drawn.

II. STEREOSCOPIC 3D

Each eye offers a different perspective of the same visualized scene. That is because the eyes are situated at a certain distance, known as interocular distance. This fact will become important later on [1]. Now, an interesting fact is that only one image is perceived, not two. This is due to a process called *stereopsis*, which takes place in the mind. The word *stereopsis* comes from the Greek words *stereo*, meaning *solid*, and *opsis*, meaning *sight*. Besides the two dimensional image depth, distance can also be perceived [6]. How can depth and distance be inferred from two slightly different 2D images of the same scene, captured by two sensors? This is done through depth cues. Depth cues can be monocular or binocular. Monocular cues are perspective, relative size, occlusion, lighting and shadows, relative motion. Perspective refers to the fact that as distance grows objects get smaller. Relative size has to do with the proportions of known objects, for example a mouse

 ISBN: 978-1-61208-195-3

is smaller than a cat. Occlusion is the blocking of view of one object by a second, which is considered to be in the foreground. Lighting and shadows can indicate if an object is sitting on a surface. Objects further away seem to move more slowly than objects in the foreground [7]. This would represent relative motion. When it comes to binocular cues there are three relevant factors: parallax, accommodation and convergence. Parallax refers to the fact that each eye sees a different image. A more detailed explanation will be provided in the next subsection. Accommodation is the muscle tension needed to change the focal length of the eye lens in order to focus at a particular depth. Convergence is the muscle tension required to rotate each eye so that it is facing the focal point [8].

A. *Parallax*

Interocular distance is sometimes referred to as retinal disparity. Parallax and disparity are similar notions, disparity is measured at the eye level while parallax is measured on the display screen, as the distance between two corresponding points in the left and the right view. Parallax can be classified in 3 categories as seen in Fig. 2. Zero parallax, when the eyes converge on the plane of the screen. In other words, the optical axes intersect in a point on the screen. Next category would be positive parallax, when the parallax value is close to the value of the retinal disparity and the optical axes are parallel. In this case, an object would appear to be "inside" the screen. Negative parallax refers to the situation when the optical axes intersect in a point in front of the screen. In this case the objects would appear to be "in front" of the screen [9].

B. *Accomodation / Convergence*

This can be more easily explained through a short example. When you focus on an object 1 meter away from you, two things happen. Your eye changes shape or *accommodates* so that the focal length becomes 1 m and your eyes move so that their axis *converge* on the object. For a graphical representation see Fig. 3 [10].

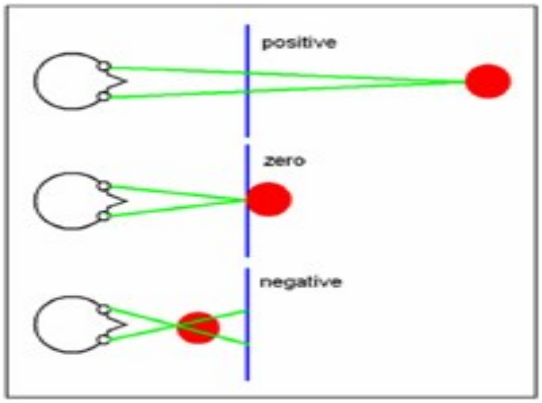


Figure 2. Parallax Classification.

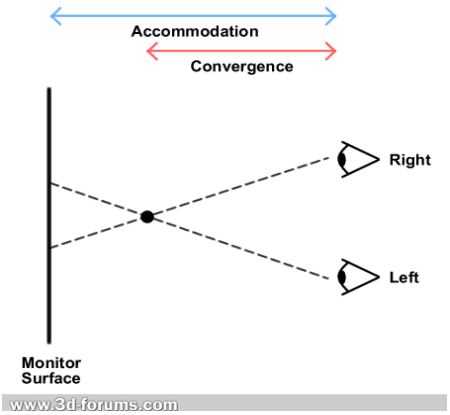


Figure 3. Accomodation / Convergence.

C. *Viewing stereoscopic 3D*

How to display and view stereoscopic images? The simplest method would be to obtain an anaglyph image, which with classic red/cyan glasses, can be viewed on any type of display. Putting it in simpler terms, all that is needed is to overlap the two images, taking into consideration the parallax value. The red/cyan glasses will filter part of the colour spectrum, so that each eye sees the corresponding image, left or right. Because the glasses filter a large portion of the colour spectrum, the viewer experiences a limited colour palette. This problem has been solved with polarized glasses, such as those currently employed in Movie Theatres. The latest commercial technology involves Frame Sequential Displays, paired with active shutter glasses. These displays show in an alternate fashion one frame for one eye and the next frame for the other eye. They need to have a 120Hz refresh rate to avoid flicker. The shutter glasses are synchronized with the display so that the correct frame is displayed at the right time [7].

III. MOVIDIUS PLATFORM

The algorithms for the rectification of the rotated sensor, have been implemented on a Movidius 8 core SoC (System-on-Chip) with a VLIW (Very Long Instruction Word) architecture. The Movidius SoC has 8 DSP (Digital Signal Processor) cores with a VLIW architecture and one RISC (Reduced Instruction Set Computer) core used for control of the DSPs and peripheral system. The SoC is intended for use as a coprocessor to main host processor on a mobile platform. Most Image Processing algorithms display a lot of SIMD (Single Instruction Multiple Data) operations, processing more than one pixel at a time. The Movidius platform really shines when it comes to SIMD operations. Another plus of Movidius' platform, when it comes to Image Processing, is the presence of multiple cores. One could split an image into a batch of lines and apply the same algorithm on them, significantly reducing processing time.

 ISBN: 978-1-61208-195-3

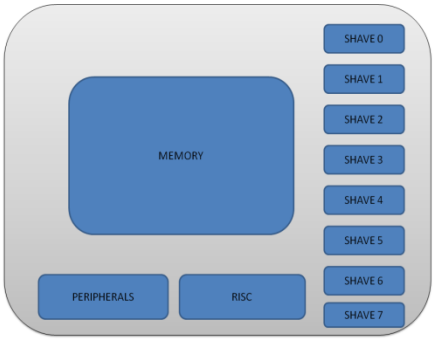


Figure 4. Movidius SOC.

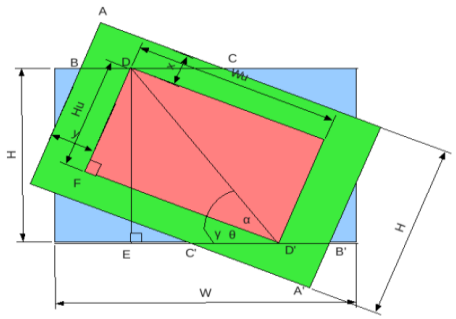


Figure 5. Rotation Compensation.

IV. ROTATION RECTIFICATION

As previously stated, the rectification process requires two main stages. The image - based computation of the rotation angle of the sensor and the actual rectification process. In the next subsections these two stages will be presented in a more detailed fashion.

A. Angle Computation

There are two methods to determine the rotation angle. The two algorithms make use of interest points, the points used to compute the parallax. In other words, using a feature extraction algorithm, such as the Harris corner detector, on one of the images and then use SAD (sum of absolute differences) to find matching points in the other image. Only after these points have been found can the angle computation algorithm begin. The rotation angle is considered to be relative to one of the images. One method requires the selection of 8 interest points: four, with the highest horizontal distance between them, and the other four with the highest vertical distance. Once these eight points have been selected, for each group of four points all the possible lines they can form, are determined. Having the line equations, the angle between this line and the corresponding line in the other image, can be obtained. The obtained angle values are then averaged, with weights, obtaining a single value that represents the angle between the two images. The second approach is based on finding the minimum area convex polygon with the interest points, either being on the polygon edges or inside its area. This can be done with a complexity of O(N * H), where H is the number of points of the polygon. Considering the polygon edges as vectors, they are summed up. The same is done on the second image's polygon. Thus, the relative rotation of the image will be the angle between the two resulting sum vectors.

B. Rotation Rectification

The most basic way to do this would be to rotate the image back. The main problem with this solution is that area loss occurs, in the four corners. The loss of information is directly proportional with the rotation angle. For a sensor rotation not larger than 0.7 degrees this loss is negligible. A useful area is defined as the largest area that can be rotated without area loss. The actual rotation is done only on this area and then the area is resized back to its original resolution. The actual useful area can be determined by finding the width and height of the area and the coordinates of the top left corner, using the following equations.

$$FD' = W_u = \cos\alpha * DD' = \cos\left(arctg\frac{H}{W}\right) * \frac{H}{\sin\left(\theta + arctg\frac{H}{W}\right)} \quad (1)$$

$$FD = H_u = \sin\alpha * DD' = \sin\left(arctg\frac{H}{W}\right) * \frac{H}{\sin\left(\theta + arctg\frac{H}{W}\right)} \quad (2)$$

$$x = \frac{H - H_u}{2}$$
$$y = \frac{W - W_u}{2} \quad (3)$$

, where

H-Height W-Width θ - rotation angle
H_u-useful height W_u - useful width

The equations can be easily obtained by looking at Fig. 5. To facilitate efficient implementation, an algorithm was sought that could be easily parallelized and would work in-place when accessing memory. Alan Paeth's rotation by shear algorithm was chosen [11]. According to this algorithm a rotation can be obtained by three shear operations. With reference to Fig. 6 and equation (7) the first shear is done on one of the axis (shearX(α)), the next one on the other axis

 ISBN: 978-1-61208-195-3

(shearY(β)), and the third one on the first axis, (shearX(γ)). Alternatively one may also construct this operation with two Y-axis operations, and one X-axis operation if more convenient. A two-dimensional shear operation has the following matrix representation, one for each axis [12].

$$ShearX(\alpha)=[1\alpha;01] \quad (4)$$

$$ShearY(\beta)=[10;\beta\ 1] \quad (5)$$

Starting with the familiar rotation matrix,

$$Rot(\Theta)=[\cos(\Theta),-\sin(\Theta);\ \sin(\Theta),\ \cos(\Theta)] \quad (6)$$

, expressing this matrix to the product of the three shear operation matrices equation (8) is obtained.

$$shearX(\alpha)shearY(\beta)shearX(\gamma)=Rot(\Theta) \quad (7)$$

$$[1\alpha;01][10;\beta1][1\gamma;01] = [1+\alpha\beta,\alpha+\gamma+\alpha\beta\gamma;\beta,1+\beta\gamma] \quad (8)$$

Solve for α, β, γ in terms of Θ and obtain:

$$\alpha=\gamma=-\tan(\Theta/2) \quad \beta = \sin(\Theta) \quad (9)$$

Taking into consideration that the maximum rotation angle is 0.7 degrees, for an image with height H, width W and rotation angle α, the total area loss can be determined with the following formula:

$$2*H^2*\tan(\alpha/2) + W^2 * \sin(\alpha) \quad (10)$$

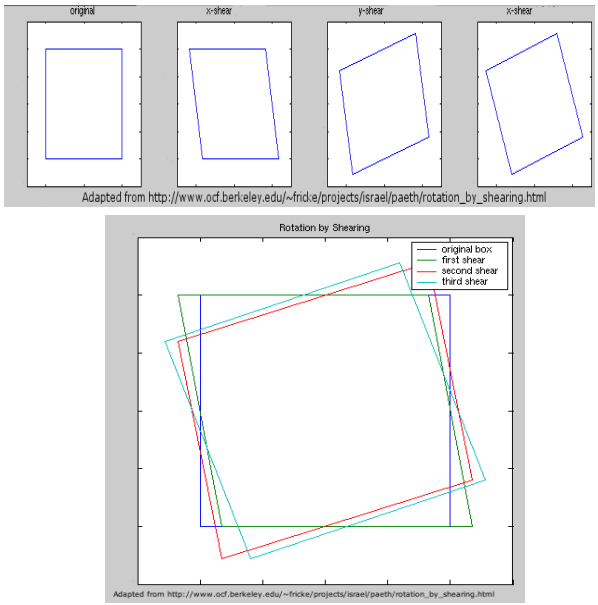


Figure 6. Rotation by Shear.

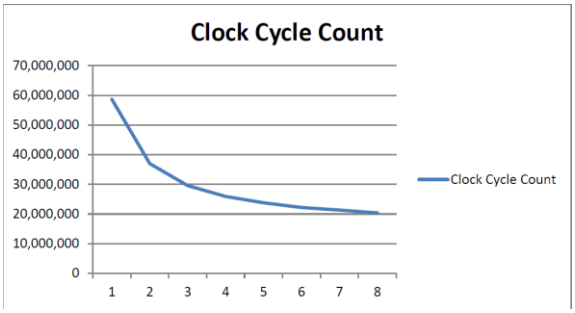


Figure 7. Perfomance Test (clock cycles / no of cores).

V. RESULT

The goal of this solution, was to enable the streaming of HD (High Definition) 3D stereo content at 30 fps (frames per second). The HD resolution achieved was HD 720p (1920x720). The rotation rectification was necessary due to the fact that the 3D stereo algorithm employed, considered, both cameras, to have the same parameters. Due to manufacturing errors this was not possible so this solution became imperative. The implementation and testing was done on Movidius' MV117 development board, equipped with a Myriad SoC, clocked at 180 MHz. The Myriad SoC is a low-power, multi-core, mobile solution, enabling mobile phone manufactures to bring 3D stereo to the mobile world. An in place 3-stage rotation implementation was chosen, for both memory and computational efficiency. This also facilitates data parallel processing so the algorithm may be easily split across a number of processing cores. The input image may be segmented into batches of lines, which can then be processed in parallel. In this case, the number of lines used was 9. A series of tests were performed in order to obtain the optimum number of cores, to use, in a real-time scenario. For this, the clock cycle count was measured as the number of cores was increased. The obtained measurements can be found in Fig. 7.

VI. CONCLUSION

In this paper, it has been shown that many problems must be addressed for a quality stereoscopic 3D image. These problems arise, due to differences in camera sensors and manufacturing placement errors. Although two sensors are identical, because of the manufacturing process they differ in small but essential points for stereoscopic 3D. One such difference was the incorrect rotated PCB (Printed Circuit Board) placement of the sensor. By using Movidius' platform, it becomes clear that a software implementation, on a powerful multi-core, mobile, low-power architecture can handle well stereoscopic 3D video content, at HD 720p resolutions.

 ISBN: 978-1-61208-195-3

ACKNOWLEDGMENT

This research has been supported from the EU Structural Funds Research Project POS-CCE 499-11844 "Falx Daciae – Software Tools and Development Processes for Advanced Multimedia Applications on Mobile Phones".

REFERENCES

[1] L. Kaufman, "Sight and mind: An introduction to visual perception," New York, Oxford University Press, 1974, ISBN-10: 0195017633

[2] J. Zhou, "New image rectification schemes for 3d vision based on sequential virtual rotation," PhD Thesis, June 2009 www.public.asu.edu/~jzhou19/phd_thesis.pdf, last accessed 10/11/2011

[3] R. Hartley, "Estimation of relative camera positions for uncalibrated cameras," Proc. of ECCV-92, G. Sandini Ed., LNCS-Series, vol. 588, pp. 579–587, Springer-Verlag, 1992

[4] C. C. Slama, "Manual of Photogrammetry, Fourth Edition," American Society of Photogrammetry, Falls Church, Va, 1980, ISBN-10: 1570830711

[5] C. Harris and M.J. Stephens, "A combined corner and edge detector," 4th Alvey Vision Conference, pp 147–152, 1988.

[6] D. L. MacAdam, "Stereoscopic perceptions of size, shape, distance and direction," SMPTE Journal, vol. 62, pp. 271-293, 1954

[7] "3D WP Principles of 3D Video and Blu-ray 3D", copyright © 2010 CyberLink Corp. All rights reserved

[8] N. A. Valyus, "Stereoscopy," London, New York: Focal Press, 1962, ISBN-10: 0240387953

[9] L. Lipton, "Binocular symmetries as criteria for the successful transmission of images," Processing and Display of Three-Dimensional Data II, SPIE vol. 507 , 1984

[10] C. Wheatstone, "On some remarkable, and hitherto unobserved, phenomena of binocular vision (Part the first)," Philosophical Transactions of the Royal Society of London, pp. 371-394, 1838

[11] A.W. Paeth, "A fast algorithm for general raster rotation," Proceedings of Graphics Interface, pp. 77-81, 1986

[12] T. Fricke, http://www.ocf.berkeley.edu/~fricke/projects/israel/paeth/rotation_by_shearing.html, last accessed 10/11/2011

 ISBN: 978-1-61208-195-3

CrossTale: Shared Narratives as a New Interactive Medium

Joaquim Colás, Alan Tapscott, Ayman Moghnieh, and Josep Blat
Universitat Pompeu Fabra
C/Tanger 122-140, E-08018 Barcelona, Spain
{Joaquim.colas, alan.tapscott, ayman.moghnie, josep.blat} @upf.edu

Abstract—**Through ages, storytelling has been used as one of the main ways for transferring knowledge and learning. We envision the use of shared narratives as a new kind of social media that empowers the collaborative creation of vast narrative worlds. For this reason, we identified existing information systems related to storytelling, and evaluated how they support multi-authored non-linear narratives. Then, we conducted a pilot experiment to understand the user interaction model with shared narratives more profoundly. This model was consequently used as a premise for designing a prototypical narrative system called CrossTale, which was evaluated to assess the generated user experience. We discuss how the results of these experiments show that shared narratives have the potential of becoming a distinct type of interactive medium supporting a new genre of user experience.**

Keywords-shared narratives; information systems; user experience.

I. INTRODUCTION

Traditionally, storytelling (from mythological parables through literature classics to modern literary fiction best-sellers) has been associated with the oral and written media, the first two channels of information transmission to appear in human history. Since then, different models of cultural expression had appeared, and those modalities had taken profit of technical advances giving birth to the main contemporary narrative vehicles such as novels, cinema, TV series or comic-books. All those kinds of narrative mediums share the trait of linearity, which suits the temporal causality of classic narratives. In spite of that, several experiments about experiencing narratives in a non-linear way were done (e.g., Moholy-Nagy "total theater" [1] and Borges' tales [2]). With the apparition of digital media, new opportunities arise for creating and experiencing narratives in new ways.

Many comtemporary works focus on understanding and modeling storytelling as an interactive experience. Mehan's Talespin [3] is a pioneering approach for automatically generating stories from atomic parts, and is an instigator of a larger body of research focusing on computer-generated narratives. On the other hand, other works studied narrartives from an HCI perspective, placing user interaction at the center: Brenda Laurel's work on interactive fictions, impacting HCI as a discipline by underscoring the properties human interaction with information [4]; and Chris Crawford's work on interactive storytelling [5], which addresses aspects of game design. A wide range of actual works focus on models for creating non-linear narratives [6] [7], but to the extent of our knowledge they don't adress this task from the perspective of user experience and the study of the user's understanding of non-linearity.

We define a shared narrative space as a set of units or scenes each representing a step in a given direction of developments, and connected organically to form a non-linear story. It is a ludic and cultural medium of expression and communication, created, developed, and maintained through the collaboration of multiple users. It is composed of a story and a discourse (storytelling). The story consists of a setting in time and space, characters, and events (or plots). It is usually thematically unified and logically coherent. Its elements are connected through cause and effect relations, thus temporal order is meaningful [8].

This non-linear medium is comparable to the real development of events: multiple stories are happening at the same time, and each can be told from different viewpoints. This points towards the suitability of non-linear narratives not only in developing fiction, but also as a way of sharing information like in online networks (e.g., forums, chats, and communities of creators). Theoretically, the content of social networks could be considered a narrative based on the sequential groupings of threads as scenes. Each forum thread could be regarded as one linear development inside a bigger story, and parts of the thread could belong to different ones as a cause of this inter-relation. However, the relations between threads are usually vague or inexistent, and there is a need for a global connection between them to provide thematic unification and overall coherence.

Our purpose is to define the adequate system concepts and design to represent and interact with non-linear narratives. Therefore, we developed two empirical experiments with paper-based and implemented digital prototypes to extract and understand the user's mental model of interaction with a narrative space, as a basis for the development of modern interactive systems for narratives.

This paper is structured as follows. First, we present six major types of information systems related to storytelling, and evaluate their support for shared narratives as a medium of social interaction and communication. Then, we illustrate a pilot experiment conducted to extract the user model of interacting with a shared narrative space, transduced into requirements for informing the design of supporting systems. Following, we discuss the development of CrossTale, a prototype based on these requirements, and its user evaluation showing the feasibility of supporting new elaborated user experiences with shared narratives. We then discuss how our results deepen our understanding of the characteristics of shared narratives, and argue in support of

 ISBN: 978-1-61208-195-3

their potential as new media for social interaction and communication. Finally, we conclude by summarizing our work and discussing its limitations, and then address their implications on future works.

II. CONTEMPORARY INTERACTIVE SYSTEMS FOR STORYTELLING AND NARRATIVES

In [2], Ryan proposed a classification of interactive narrative types based on the nature of the user participation: users can either experience the narrative acting as an internal character of the story, or as an external agent; they can either alter the ontology of the narrative through interaction, ontologically alter the narrative world through interaction, or explore the narrative without inducing any change. This classification provides a framework to analyze and characterize contemporary systems for interacting with narrative by reflecting on how the user experience is contributing to the narrative, and how the narrative is influencing the user experience.

We have identified six major types of information systems directly related to interactive narratives: The first type are adventure books, which comprise a tale where the reader follows a character and makes choices that lead the story towards distinct developments; the second is tabletop role-playing games (or RPGs), in which the player creates a character and its story, and then devises the character's actions according to a set of rules; adventure videogames are the third type, and they put the player in the role of a character that resolves puzzles in order to advance in the story; the fourth type is role-playing videogames, where the player makes navigation decisions to reach one of several possible endings; the fifth type is Forum or chat-based RPGs, where players collaboratively create a story (usually with a few rules of engagement); the sixth and last type is web communities of fiction writers (fan-fiction), that create stories in the same fiction world, but not always collaboratively. A high number of fan-made wikis can be found on the web, compiling formation about events, characters, and places concerning those worlds. Harrigan gives a wide overview of the complications of maintaining these vast narrative spaces, and how the different fan-fiction communities address them [9]. These systems are described in Table 1 according to Ryan's framework.

TABLE I. CONTEMPORARY SYSTEMS OF INTERACTIVE NARRATIVES

System	Example	Author /reader role	Main role	Author interaction	Reader interaction
Adventure Book	Choose your own adventure	Separated	Reader	-	External Ontological
Tabletop RPG	Dungeons & Dragons	Mixed	Author	Internal, Ontological	-
Adventure videogame	Monkey Island	Separated	Reader	-	Internal Exploratory
RPG-Videogame	Baldur's Gate	Separated	Reader	-	Internal Ontological
Forum / chat RPG	Aelyria.com	Mixed	Author	Internal, Ontological	External Exploratory
Fan-Fiction community	Fanfiction.net	Mixed	-	External, Ontological	External Exploratory

Seeing the particularities of the informative structure of narratives, we point at differences between existing systems and interactive storytelling. In particular, none of these types of systems entirely supports shared narratives as a medium of social interaction. Three of them (books, adventure videogames and RPG videogames) are unidirectional mediums, created by authors and consumed by other people as readers. They support a varied degree of interaction with the content, but they do not allow users to contribute. Tabletop RPGs and forum or chat RPGs, allows user-generated content to be added to ongoing discussions, which together do not constitute a coherent story that can later be consumed as part of the user experience. Only fan-fiction Internet communities fully support both the addition of user-generated content and its consumption as part of the user experience. But the lack of collaboration and cooperation between contributors tends to divide the narrative space into distinct and incoherent flows of events, which only share the original work as a point of reference, resulting in independent narratives [9].

The case of fan-fiction communities is the major exponent of a multi-authored narrative system where usually no one acts as both reader and author to the same shared narrative, but each participant is only the author of his narrative sub-space and reader of others. A similar handicap exists in forum RPGs, where each contribution is by force situated directly after the previous one, and is the only possible type of contributions.

In conclusion, none of these systems helps participants to contribute efficiently to the shared narrative space, nor to collaboratively organize and maintain its overall structure. Therefore, there is still a need for supporting the users' ability to understand and navigate the space, allowing the narrative to grow in an organic way, and extending its contents from any desired point in the narrative flow. This will dissociate the story content of the impositions coming from the way the users elaborate the narrative discourse.

III. UNDERSTANDING DESIGN REQUIREMENTS

The first experiment (Fig. 1) was designed to allow users to freely create a narrative and the rules that operate it. 20 subjects (university students) were provided with paper sheets as a frame to draw and write scenes and a set of elements (fairytale characters and objects) that they could paste in the sheets. The narrative was developed on a large glass wall where the story was created in the form of paper scenes connected by arrows. The subjects proceeded one by one to read the story on the glass, and then modify or expand it by creating new scenes, posting them in the wall, and drawing the connections. Observations were made during this process, and the subjects were later asked to fill a questionnaire of 18 questions. The questionnaire evaluated the story comprehension and congruence as perceived by the subjects, and inquired about the reading or navigation paradigm that they used (narrative elements and concepts they followed throughout the story). It also asked about their contributions (number, content, location, etc…), and if they added scenes to the narratives or contended in modifying existing ones.

 ISBN: 978-1-61208-195-3

Figure 1. The settings of the first experiment.

Observations show that the kind of interaction performed is external, as the users do not assume the role of any particular character. It is also ontological during the creation, and exploratory during the reading. The analysis of the resulting story and the questionnaire answers revealed several aspects about the nature of the user comprehension and interaction with the shared narrative space. These results were traduced as a set of requirements (Table II) for the design of information systems that support interacting with narratives. They are also detailed in the following.

A. *Using storylines and characters to navigate a narrative*

The results revealed that subjects project a "time-space-development" logic on the narrative. The story is mentally situated on a space with a temporal and causal logic, represented in two axes: a temporal relationship between the scenes, and places where these scenes take place. All subjects followed linear sequences (storylines) for reading, being a linear/temporal sequence of connected scenes that track the development of a specific character or plot. 14 out of 20 followed those storylines throughout the narrative space, from the first scene to a finishing one before backtracking. In addition, 12 of them followed character developments, and 10 plot relationships.

Understanding how users navigate the narrative space leads us to consider a visualization that copes with this "time-space-development" logic to facilitate the finding of storylines, and consequently the user interaction.

TABLE II. DESIGN REQUIREMENTS EXTRACTED FROM THE INTERACTION MODEL

Observations	Design requirements
Projection of a logic based on time, space and developments.	Organization of the informative space based on time and space axis.
Reading by following linear sequences about a character or a plot.	Navigation through suggested plot and character storylines.
Unitary and coherent narratives.	Mechanisms for preserving congruence.
Global viewpoint for comprehending the whole story.	One interface mode for a global view of the informative space.
Reading a storyline through a zoom-in viewpoint.	One interface mode for following storylines.
Focusing on a single scene for creating and editing.	Independent interface mode for scene edition.

B. *Preservation of literary consistency*

The results also show that the generated narrative space is unitary, coherent, and with a limited divergence. It is unitary in the sense that all the scenes are interrelated and are part of the same story. In fact, the divergence of the narrative space away from the central topic is limited: subjects found it easier and socially proactive to expand existing storylines instead of creating new ones. This notion of unity is directly derived from the fact that the entire story is predefined and all the storylines are happening simultaneously in the same time stream. This raises consistency issues in the literary fabrics of the narrative, which users thrive to treat by re-ordering scenes or inserting new ones. The literary consistency of the narrative is considered fundamental for understanding the story, and it is one of the main concerns when modifying/ adding scenes to the narrative space: 5 subjects used their contribution only for correcting consistency issues, and in the end of the experiment, only 5% of the scenes were considered inconsistent with the rest of the narrative. For this reason, the system should implement mechanisms for helping to preserve literary consistency without restricting the non-linearity of the narrative. This issue remains outside the scope of this paper.

C. *Three interface modes for three types of interaction*

The subjects' interaction with the narrative space shows that at least three different views for three different purposes are needed for a multi-modal interaction with narratives: a global view of the space is used to approach and comprehend the whole narrative space and its structure, as well as when selecting a point in time and place to add a new scene; a "zoom-in" view for viewing a scene inside a storyline and understanding the other storylines related to it; and a composition view that allows users to create and edit scenes. These views are illustrated in Fig. 2.

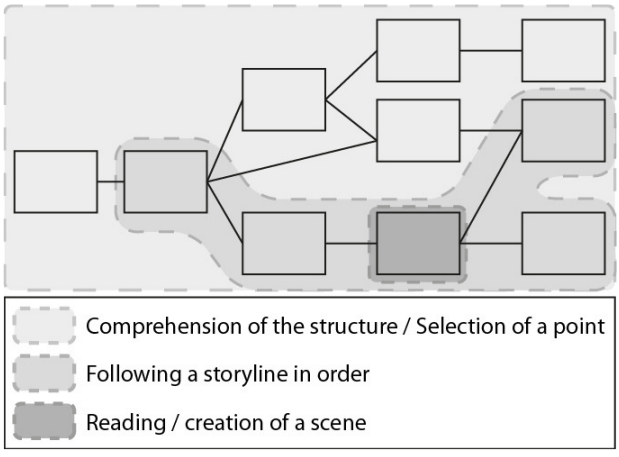


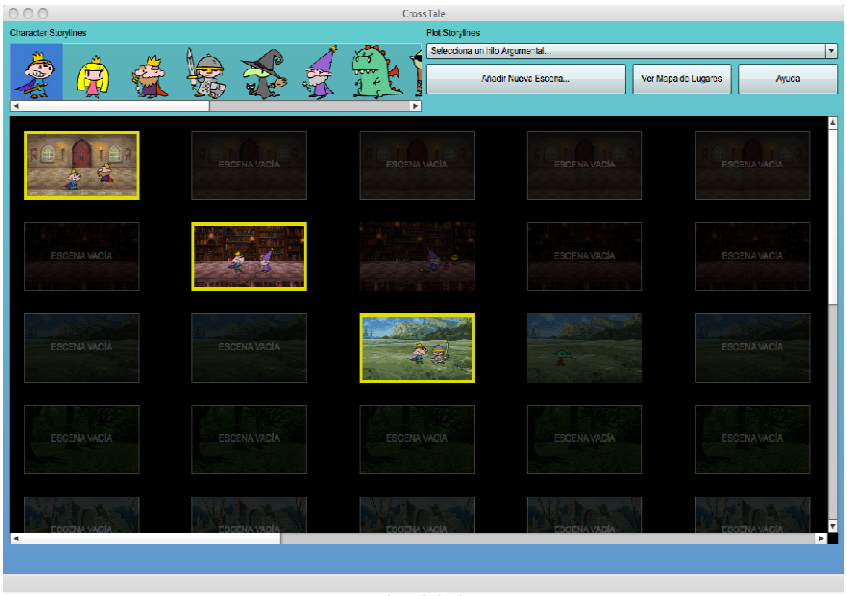
Figure 2. Viewpoints related to the interactions with the narrative space

IV. PROTOTYPING SHARED NARRATIVE SPACES

We developed a prototype named CrossTale based on the design requirements extracted from the first experiment to reproduce the user experience according to them. CrossTale

 ISBN: 978-1-61208-195-3

implements three interface modes defined previously (Fig. 3 a, b and c). The global view lets users explore the whole narrative space, visualized as a grid with the axes of time and places, and select characters and storylines. Selecting a scene/storyline changes the interface into the reading view in which the scene is maximized for reading. In this view the user can also navigate back and forward by the current storyline. Finally, by selecting an empty frame, the user access the creation view where s/he can create a scene by arranging characters and objects, and introducing text, and indicate the related plotline/s.



(a) The global view

(b) The reading view

(c) The creation view

Figure 3. Three interface modes of the CrossTale prototype

The experiment with the prototype consisted of creating a narrative in a similar way to the pilot experiment. A total of 15 subjects (undergraduate students in media studies) were enlisted, and asked to freely use the interface to read and create a shared narrative with their own contributions. Each subject was briefly introduced to the interface controls, and then given an unlimited time to interact and the freedom to add as many scenes as wanted. Then, the subject executed eight interaction tasks provided by the evaluation team, and observations were made. Afterwards, each subject answered a questionnaire to rate the rate the experience on Likert scale, and evaluate the suitability of the design for reading and contributing and the overall user experience.

The results of the task-driven evaluation are summarized in Table III. It describes how many subjects employed each interface view for each task. The results show that 11 out of 15 subjects performed all tasks easily, and the remaining 4 subjects successfully performed 6 out of 8 tasks. The subjects used the global view and/or the reading view to identify and comprehend the narrative elements. Similarly to the first experiment, some subjects concentrated on characters while others on plots, but everyone used one of these two paradigms for finding storylines and navigating the narrative space. During the contribution task all the contributors also used the creation view to compose new scenes, but this view was never accessed for performing the identification tasks. These results indicate that the design supports the modes of interaction identified in the first experiment, and that these modes dispose of adequate functionalities. However, most subjects prefer having more information about the context of scenes while reading them. This means that the dissociation between the global and reading views could be revisited.

TABLE III. RESULTS OF THE DESIGN ADEQUACY EVALUATION

Task NB	Task	Correctly executed	Global View	Reading View	Both Views	Navigating with Storylines
1	**Identify the beginning scenes**	15	13	13	11	11
2	**Identify story end scenes**	15	13	9	7	7
3	**Identify main characters**	15	15	15	15	15
4	**Identify important places**	15	14	14	13	13
5	**Identify simultaneous scenes**	11	13	4	4	4
6	**Identify scenes in the same location**	12	14	2	2	2
7	**Approximate the duration**	15	13	13	11	10
8	**Find any inconsistency**	15	8	12	7	8
9	**Contribute (optional)**	13	13	6	6	5

 ISBN: 978-1-61208-195-3

Table IV shows the evaluation results of the user experience. All subjects appreciated the experience of interacting with narratives through CrossTale. In particular, they found that CrossTale supports reading a non-linear narrative (4.33/5), contributing to it (4.77/5), and finding and correcting inconsistencies (3.92/5).

TABLE IV. RESULTS OF THE USER EXPERIENCE EVALUTION

Question	Average Score
Overall experience	3,93 / 5
Found the system entertaining	4,33 / 5
Design makes reading easy	3,93 / 5
Design helps to maintaining consistency	3,92 / 5
Design facilitates contributing	4,77 / 5

The results of this experiments show that the concepts and design of CrossTale, as a prototype for interacting with narratives, are highly appreciated by the subjects. However, they also point out several issues that need to be addressed in future versions. In particular, it was found that users spend considerable amount of effort (estimated to 20% of the overall activities) to preserve the consistency of the narrative, which is important for understanding and interacting with it. Future versions could provide means to facilitate further the preservation of literary consistency. In addition, social interaction between different authors remains indirect: users cannot communicate directly and the authors' profiles and their contributions are not discernible in the current design. Future versions can include more support for this aspect and study its effects on the user experience and collaboration.

Using Ryan's framework for the classification of interactivity with narrative systems, we can say that the users of CrossTale performed an external interaction during the whole experience, as they took on the role of agents external to the story, and read and contributed in it from outside the fiction world. This interaction is exploratory during reading, in the sense that the readers choose between storylines to follow but the reading itself does not change nor affect the structure of the narrative space. Finally an ontological participation is detected while the user takes the role of author and expands or alters the narrative world.

V. DISCUSSION

The nature of shared narratives presents several challenges over how the inherent information is constructed, presented, and accessed. In a sense, non-linear interactive storytelling has always faced challenges for having to reconcile the sequential nature of narratives with the user ability to choose between different threads of reading (the paradox of coping "storytelling" with the "non-linear"). In this work we provided a first grounding basis for addressing these challenges and developing shared narratives as new social media. Our research is a first step for consolidating a standardized system for sharing and collaboratively constructing narratives, given we extracted, understood, and evaluated the user mental model associated with this interaction.

We can illustrate our vision of shared narratives as new social media through an example such as the October 2011 global protest movements in Europe and North America. Social networks have proven to be an important infrastructure of support for these movements, where participants behaved as authors, actors, and divulgators of the actual events. These participants leave digital imprints of their experiences [10]: many sources of scattered information can be found on Facebook or Twitter as collections of narrations, opinions and images taken. These events can be understood on a high-level by following one single storyline (e.g., watching the news bulletin), but it is more interesting to explore them from the different points of view of participants whom witness interrelated events taking part simultaneously within the movements, starring thousands of different people.

By using tools for shared narratives similar to CrossTale, the participants could collaboratively draw a detailed narrative world around their movements. In other words, what if these participants could compose collaboratively the stories behind these events by describing and positioning their individual experiences in concrete time and place, and relating them with the experiences of others? The resulting structure would suit better the nature of the associated information, and it would provide a new closer way for interacting with it, which virtually could amount to participating in the physical events themselves.

VI. CONCLUSIONS

In modern literature and fiction worlds, it is common to have multiple stories set in a complex chronology inside a common setting, such as in fiction franchises where narratives are constructed through the contributions of multiple professional authors. Tools based on the CrossTale interaction model would be capable of organizing all this encyclopedic knowledge in a structured narrative space that suits better the temporal, causal, and multi-lineal nature of a narrative, empowering the authors to contribute easily to expand the vast fiction worlds and empowering the readers to explore them naturally. With such tools, narrative spaces grow organically and collaboratively; the proactive role of participants consequently diffuses the mono-directionality of the author/audience relation. In that sense, non-linear interactive narratives can become a new kind of media of its own, suitable for collaboration, information sharing, and learning.

By experimentation, we learnt how users perceive and procreate the narrative space into a unitary and congruent way and how they mentally structure the informative space in terms of time and place and navigate it following structured sequences of character and plot-related scenes. This model was used as the basis for designing a functional prototype, CrossTale, which was subsequently evaluated with users. These evaluations show the success of the adopted approach in supporting complex interactions with narrative spaces, which assimilate its non-linearity. It

 ISBN: 978-1-61208-195-3

provides a validation for further investigations on the potential of shared narratives as new media.

This work also has several limitations. In our experiments, the literary traits of the narrative space were somewhat pre-defined, especially the main characters (prince, pricess, witch…) and places (tower, castle, woods…). This discouraged users to think about expanding the literary reach of the narrative space with few exceptions. In online social role-playing games, some people perform the role of content generators, creating more story elements (characters, objects, places…) to the space rather than adding scenes. Such behaviour should be studied further in the future.

There are also some limitations inherent to the nature of the prototype and the experimental settings: the visualizations used have functional limitations (e.g., visualizing all related scenes to a selected one), and several improvable design issues, mostly usability-related, were identified during the evaluation (e.g., the composition view is not user friendly). Finally, the development of the experiment in a controlled environment does not reflect intimately interactions with shared narratives, nor the collaboration phenomenon (performed in an indirect way through te experiment), as ought to take place online and during a greater amount of time.

VII. FUTURE WORK

With this model and prototype as a starting point, our future step consist of shifting the system deployment from our isolated environment to the web. In addition, we will develop features for adding user-generated story elements (as characters and places). Expanding and deploying the system in a network environment would allow us to study how the narrative space grows when users share a common social network, and study the nature of the resulting interaction and narrative structure under those conditions, as well as the potential of shared narrative spaces to empower long-term collaboration.

A complementary part of this research, concerning the mechanisms of consistency preservation, remained outside this study. Our experiments pointed that consistency between all the story elements and scenes is the main conception that readers use to understand the narrative space, and one of the main concerns when expanding the space by adding new scenes. We will explore how a support for the preservation of consistency can be provided through the introduction of rules (e.g. not allowing the same character to be used in two scenes happening at the same time). Currently, several steps have been taken in this direction, and a complete strategy for managing the consistency of shared narratives will be included and evaluated in future prototypes.

REFERENCES

[1] L. Moholy-Nagy, "Theater, Circus, Variety" (1924), in Multimedia, from Wagner to Virtual Reality, R. Packer and K. Jordan, Ed. New York: W. W. Norton & Company, 2001, pp. 16-26.

[2] ML. Ryan, “Beyond myth and metaphor: The case of narrative in digital media”, Game Studies, 2001.

[3] J. R. Meehan, “TALE-SPIN, An Interactive Program that Writes Stories”, in Proceedings of the Fifth International Joint Conference on Artificial Intelligence, 1977.

[4] B. Laurel, “Computers as Theatre”, Addison-Wesley, 1991.

[5] C. Crawford, “Chris Crawford on Interactive Storytelling”, Brkeley, Calif.: New Riders, 2005.

[6] Y. Cao, R. Klamma, and M. Jarke, “The Hero’s Journey – Template-Based Storytelling for Ubiquitous Multimedia Management”, Journal of Multimedia, vol. 6, no. 2, Apr. 2011, pp. 156-169, doi: 10.4304/jmm.6.2.156-169.

[7] Y. Cao, R. Klamma, and A. Martini, “Collaborative Storytelling in the Web 2.0”, in Proceedings of the First International Workshop on Story-Telling and Educational Games, 2008.

[8] S. Chatman, “Story and Discourse: Narrative Structure in Fiction and Film”, Cornwell University Press, 1990.

[9] P. Harrigan and N. Wardrip-Fruin, “Third Person: Authoring and Exploring Vast Narratives”, MIT Press, 2009.

[10] F.Girardin, F. Calabrese, and F.D. Fiore, “Digital footprinting: Uncovering tourists with user-generated content”, Pervasive Computing, 2008.

 ISBN: 978-1-61208-195-3

Silent Voice Elements for Text Input

Peng Teng and Yunde Jia
Beijing Laboratory of Intelligent Information Technology
School of Computer Science, Beijing Institute of Technology
Beijing 100081, China
Email: {tengpeng, jiayunde}@bit.edu.cn

***Abstract*—Speech input systems could not work well in noisy environment, and their usage often makes a leakage of information. To avoid these problems, this paper proposes a concept of Silent Voice Elements, called sivels for short, and a novel articulators-operated text input method with sivels. Sivels are easy-to-recognized phonemes of soft whisper in their tissue-conducted vibration signals. The selection of sivels is a combinational optimization problem which is solved by using a heuristic search algorithm. Encoding text with sivels similarly to Morse code, one can input text accurately by speaking corresponding sivels. Experimental results demonstrate that our method selects a set of sivels with perfect recognizability, and the proposed sivel-based text input also gives an performance with sufficient efficiency.**

Keywords-text input; silent voice element; silent speech interface.

I. INTRODUCTION

Speech input is expected to become the principal text input method replacing keyboards. People have built and deployed numerous speech input systems for various applications. But in noisy environment, these systems have serious degradation in their performances, and can not work well. Besides, speech may be considered as unwanted noise, and often makes a leakage of information as well. Silent Speech Interface(SSI) [1] is a promising technology that can be employed to overcome these problems. For example, [2] captured articulatory movements using Electromagnetic Articulography (EMA) sensors and mapped them into phonemes; [3] aimed to recognize speech from data captured by Surface electromyography (sEMG) on articulatory muscles; [4] investigated an approach which directly recognizes "unspoken speech" in brain activity measured by Electroencephalographic (EEG) signals. Most of SSIs are still on the stage of laboratory research.

There is another SSI called Non-Audible Murmur (NAM) microphone [5], [6], a high-sensitivity contact microphone attached on the skin over the soft tissue in the orofacial region. [7] and [8] reported the NAM enhancement to audible speech for human-human communication. Compared with the sensor data acquired by other SSIs, NAM signal is a tissue-conducted acoustic signal which can provide a more direct and stable representation to the real speech and with insensitivity to noise. However, NAM recognition is difficult to be adopted as the underlying recognition technology for text input, because it could not deliver a low-error performance [9] owing to the poor quality of NAM signal in addition to the intrinsic difficulties for machines understanding human languages.

The goal of this paper is to present an alternative text coding scheme for developing an articulators-operated text input method with high accuracy as well as acoustic environment insensitivity. Specifically, we propose a concept of SIlent Voice ELement called sivel for short, and develop a novel method of articulators-operated text input where text is encoded with sivels. Sivels refer to easy-to-recognized phonemes in tissue-conducted signals of soft whisper, regardless of their linguistic meanings. Similar to Morse code, a user can encode text with sivels using a customized scheme and input text accurately by speaking corresponding sivels in soft whisper. The sivel-based text input method can work without the intervention of hands like speech input, and without noise sensitivity or information leakage. In addition, because of the customized scheme of text coding, sivel-based text input avoids the intrinsic difficulties for machines understanding human languages, and has the potential to provide an articulators-operated text input method for speech disorder people.

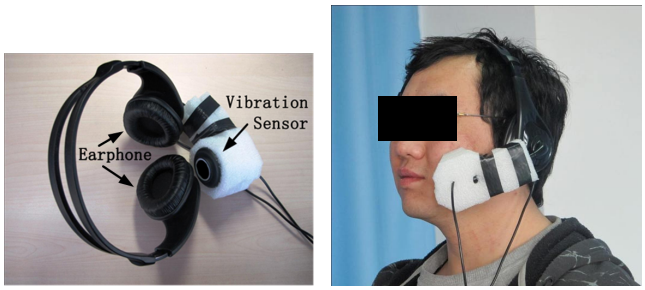
II. SILENT VOICE ELEMENTS (SIVELS)

In this section, we introduce the concept of sivel, and empirically select a set of sivels as an example. Then experiments on sivel recognition are preformed to evaluate the efficiency of these sivels as code elements, and the results will help the development of a general sivel selection method in the next section.

A. Tissue-conducted Soft Whisper

Soft whisper is a kind of low-amplitude sound that people pronounce without the vibration of vocal cords, and is not expected to be heard by others. When speaking soft whisper, one's articulators figure the vocal tract with certain shapes. Airflow out of the lung flows through the vocal tract and generates noise at its constricted segments. Soft whisper is namely the mixture of the noise and its vocal-tract resonance, and also a vibration of air. The vibration stimulates the vocal-tract wall, and some vibration energy transmits to the surface of one's head. With vibration sensors

 ISBN: 978-1-61208-195-3



(a) (b)

Figure 1. The headset of a experiment system(a) and its implementation(b).

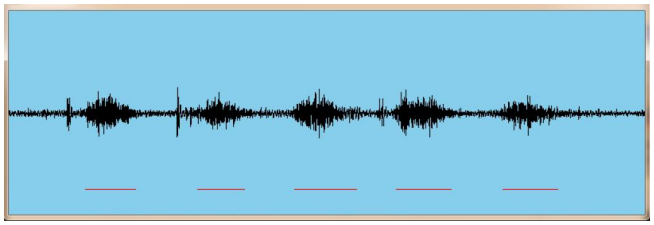


Figure 2. A vibration wave signal recorded by our experiment system where straight line segments denote the corresponding time when soft whisper is being pronounced.

placed on the skin of orofacial regions, the vibration can be detected and recorded, i.e., the tissue-conducted signal of soft whisper. We designed an experiment system as shown in Fig.1. An example of a vibration signal detected by our system is shown in Fig.2 where straight line segments denote the corresponding time when soft whisper is being pronounced.

B. An Example of a Sivel Set

Sivels are easy-to-recognized phonemes in tissue-conducted signals of soft whisper. Served as code elements for text coding, sivels are required to be stable and distinguishable in their signal patterns, so that they can be recognized easily and accurately.

We empirically select a set of sivels from whispered phonemes by considering their pronunciations. When different phonemes are pronounced in soft whisper, different articulatory positions make the vocal tract present different resonance characteristics. Since the source of soft whisper can be seen as the white noise, it can be assumed that signal pattern of a whispered phoneme is mainly resulted by articulatory position when one is pronouncing it, and phonemes with significant differences in their articulatory positions have significant differences in their signal patterns. Consequently, we select whispered phonemes /a/, /ə/, /i/ ,/ɒ/ ,/u/ (in English) from International Phonetic Alphabet (IPA) as sivels initially. Our reasons can be summarized as follows.

- These phonemes are all vowels, i.e., phonemes pronounced with an open vocal tract, so that their tissue-conducted signals can be detected and processed easily due to their relatively high amplitudes.
- These phonemes are all monophthongs, so the articulatory positions are almost unchanged during pronouncing; this makes their signal patterns stable.
- There are significant differences in their articulatory positions according to the IPA vowel diagram which shows the correlation of a monophthong and its corresponding articulatory position, so their signals could be discriminated easily.
- Since these phonemes are all used frequently by speakers who will participate in our experiments later, the pattern of the same phoneme can be generated naturally and with few differences at different time.

Besides, the duration of a whispered phoneme is also a potential discriminative feature which can tell whether it lasts long or short. (The duration threshold of short or long can be determined by analyzing user's individual habit or by a given value, e.g., 0.5 sec.) Therefore, each of the initial sivels corresponding to {/a/, /ə/, /i/, /ɒ/, /u/} can be pronounced in two forms, long and short, and considered as independent sivels. We use `a, e, i, o, u` to denote their short forms , and `A, E, I, O, U` for their long forms, and the sivel set selected empirically is $\mathcal{V}' =$ {`a, e, i, o, u, A, E, I, O, U`}.

C. Sivel Recognition

The speaker-dependent recognition experiments on samples of sivles in $\mathcal{V}'$ were performed to evaluate the effectiveness of the set of sivels as code elements.

The experimental data were collected from 6 speakers (1 female and 5 males). From each speaker, we recorded totally 500 samples (with 8KHz sampling rate) using our experiment system in office environment, that is, 50 samples of each sivel in set $\mathcal{V}'$. These 500 samples were divided into 50 groups, and each group contains one and only one sample for each sivel. Long-time spectral analysis were performed on the whole signal of each sample in order to get stable spectral feature (because these sivels are all monophthongs as we discussed above). Then each sample is represented by a 22-dimension parameter vector which contains 1 energy coefficient, 20 Mel-Frequency Cepstral Coefficients (MFCCs) and 1 duration coefficient. Linear Discriminant Analysis (LDA) were employed in training phase to reduce dimension of parameter vectors and select discriminative features. In testing phase, minimum Mahalanobis distance classifier was used as pattern classifier, labeling a testing sample with the class whose mean vector of training samples has the minimum Mahalanobis distance to that of the testing sample.

To evaluate the accuracy of speaker-dependent sivel recognition, 50-fold leave-one-out cross-validation (LOOCV) was performed on each single speaker's samples. For each run of the validation, one group of samples is used for testing while the rest groups of samples are for training. The final

 ISBN: 978-1-61208-195-3

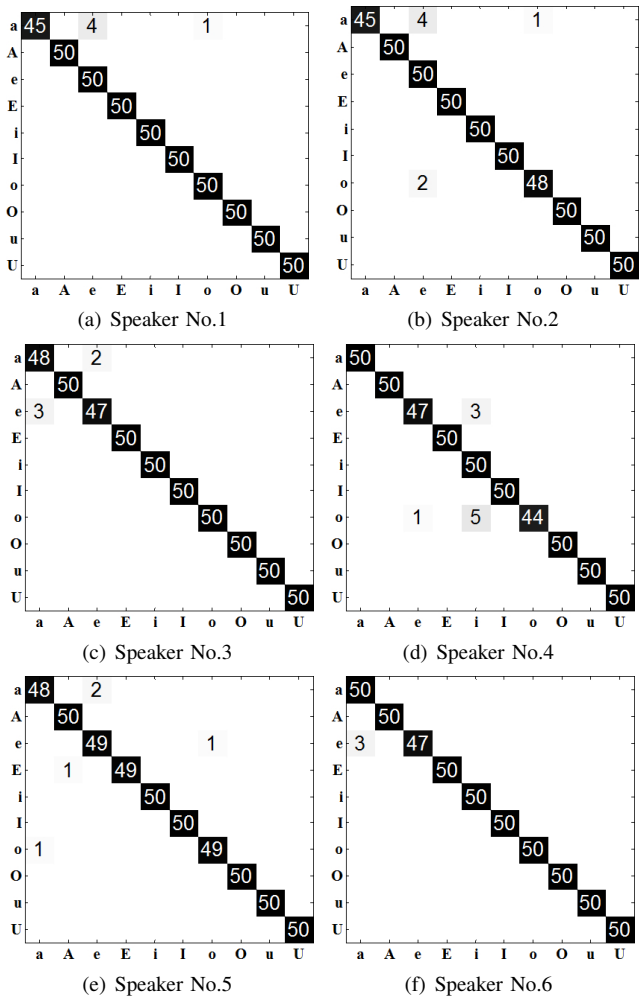


Figure 3. Confusion matrices of the speaker-dependent sivel recognition experiments on $\mathcal{V}'$ of 6 speakers, respectively.

accuracy is calculated by averaging accuracies of all the 50 runs. The experimental results for 6 speakers are illustrated in Fig. 3 and Table I. Observation of the results can be summarized and discussed as follows.

- The results of speaker-dependent recognition experiments all achieve high accuracies, and the mean accuracy is 98.87%. It means the sivels in set $\mathcal{V}'$ could be efficient code elements for text coding.
- There was no confusion between a short sivel and a long sivel, and among long sivels. This demonstrates that duration is a discriminative feature and it could not add confusions among the initial sivels.
- For each speaker, most errors of the recognition are caused by a few certain sivel pairs, such as the confusion between a and e in Fig. 3(a), 3(b), 3(c) and 3(f), and the confusion between i and o in Fig. 3(d).
- Every speaker may have their own pairs of confused phonemes in soft whisper due to their personal pronouncing habits (e.g., their accents).

Table I
ACCURACIES OF THE SPEAKER-DEPENDENT SIVEL RECOGNITION EXPERIMENTS ON V' FROM 6 SPEAKERS.

No.	Gender	Accuracy	No.	Gender	Accuracy
1	Female	99.0%	4	Male	98.2%
2	Male	98.6%	5	Male	99.0%
3	Male	99.0%	6	Male	99.4%
On average		98.87%			

We can see that sivels have the potential to be used as code elements for text input. However, the sivel selection method used in this section is not suitable for every speaker due to personal pronunciation habits. It is necessary to develop a general selection method to select a set of sivels for individuals.

III. SIVEL SELECTION

The sivel selection problem, selecting sivels from a number of phonemes, is described as follows: Given a N-cardinality set $\mathcal{P} = \{p_l | l = 1, \ldots, N\}$ of candidates, find a subset $\mathcal{V}$ which has high recognizability and a proper cardinality . The selection can be accomplished by calculating the recognizability of each subset of $\mathcal{P}$, then picking one subset that has acceptable recognizability and cardinality as the set $\mathcal{V}$, i.e., using so-called exhaustive method. The recognizability of a set is mainly dependent on its classification complexity which characterizes the difficulty of the classification problem on its elements' samples. A number of approaches have been used to measure the classification complexity, such as those mentioned by [10]. Since our aim is to recognize sivels, the LOOCV error rate is an appropriate measure for classification complexity , and further, for the recognizability as well. Unfortunately, the number of the subsets is huge due to the combination explosion, and LOOCV is time-consuming. If LOOCV error rate is used as the measure in the exhaustive method, the computational cost will not be accepted. To reduce the computational cost, we use a heuristic search method to find the set $\mathcal{V}$ of sivels.

As discussed above, most of the LOOCV errors on a set of phonemes are caused by a few certain pairs of its elements, and the pair with highest classification complexity is suggested to contribute the most negativity to the recognizability of the set. Therefore, define $E\big((p_i, p_j)\big)$ where $1 \leq i \neq j \leq N$ to calculate the one-on-one classification complexity of a pair of two different elements in $\mathcal{P}$, then for a subset $\mathcal{Q} \subseteq \mathcal{P}$ where $2 \leq |\mathcal{Q}| \leq N$, we use the maximum of $E(\cdot)$ on $\mathcal{Q}$ as a heuristic estimate of its holistic classification complexity. The heuristic estimate is denoted by

$$H(\mathcal{Q}) = \max_{q_i, q_j \in \mathcal{Q}} E\big((q_i, q_j)\big). \quad (1)$$

The values of $H(\mathcal{Q})$ for all possible $\mathcal{Q}$ while $|\mathcal{Q}| = 2, \cdots, N$, can be calculated in a recurrence way, i.e.,

$$H(\mathcal{Q}) = \begin{cases} E\big((q_1, q_2)\big), & \text{if } |\mathcal{Q}| = 2; \\ \max\Big(H(\mathcal{Q} - \{q\}), \\ \quad \max\limits_{q^- \in (\mathcal{Q} - \{q\})} E\big((q, q^-)\big)\Big), & \text{else} \end{cases} \tag{2}$$

where $q_i, q_j \in \mathcal{Q}$; q is an arbitrary element in $\mathcal{Q}$. We make an assumption that there are no equal values of classification complexity for different pairs of candidates in $\mathcal{P}$. Given a instance of $E(\cdot)$ and k as the desired number of sivels, we can select a set $\mathcal{V}$ of sivels from the set $\mathcal{P}$ of candidates using the sivel selection algorithm summarized in Algorithm 1.

Algorithm 1 Sivel Selection Algorithm

Input:

The set of candidates for sivels $\mathcal{P} = \{p_l | l = 1, \ldots, N\}$; k denoting the cardinality of $\mathcal{V}$, $2 \leq k \leq N$;
The function $E(\cdot)$ to evaluate the classification complexity of two candidates in $\mathcal{P}$, i.e., $E\big((p_i, p_j)\big)$ where $1 \leq i \neq j \leq N$.

Output:

The k-cardinality set $\mathcal{V}$ of sivels.

1: Calculate all the values of $E\big((p_i, p_j)\big)$, then calculate all the values of $H(\mathcal{Q})$ where $\mathcal{Q} \subseteq \mathcal{P}$, $2 \leq |\mathcal{Q}| \leq k$ using Eq. 2;
2: $\mathcal{W} \leftarrow \{w | w \subseteq \mathcal{P}, |w| = k\}$, $M \leftarrow 0$;
3: **repeat**
4: $\quad \mathcal{W}^* \leftarrow \{w^* | w^* = \arg\min H(w)\}$;
5: $\quad M \leftarrow M + 1$, $\mathcal{I}^M \leftarrow \bigcap\limits_{w^* \in \mathcal{W}^*}^{w \in \mathcal{W}} w^*$;
6: $\quad \mathcal{W} \leftarrow \{w | w \leftarrow w^* - \mathcal{I}^M, w^* \in \mathcal{W}^*\}$;
7: **until** $|\mathcal{W}^*| = 1$
8: **return** $\mathcal{V} \leftarrow \bigcup\limits_{m=1}^{M} \mathcal{I}^m$.

IV. Sivel selection Experiment

The experiment was conducted to evaluate the performance of our sivel selection method. The experimental framework is as follows. Given a N-cardinality set $\mathcal{P}$ of candidates for sivels, we selected three sets of sivels for each k ($k = 3, \cdots, N$). One set is selected using the proposed heuristic search method with $E(\cdot)$ to measure the classification complexity of every two candidates. Another set is selected using the exhaustive method with LOOCV error rate to measure the classification complexity of a subset of $\mathcal{P}$. The third set is selected also using the exhaustive method but with $E^+(\cdot)$, the generalization of $E(\cdot)$ for multiple candidates, to measure the classification complexity of a subset. The three sets of sivels are compared on their LOOCV error rates. In the LOOCVs, a same classifier is employed for both the sivel selection and the error rate comparison.

A. Data Collection

We pick candidates for sivels from those phonemes pronounced with fixed articulatory positions. In our experiment, there were 18 phonemes picked into set $\mathcal{P}$ as candidates for sivels, including

- 12 vowels: /ɑ/, /ʌ/, /æ/, /ɔ/, /i/, /ɪ/, /ə/, /ɜ/, /e/, /ʊ/, /u/ and schwa;
- 1 liquid: /l/;
- 5 fricatives: /f/, /θ/, /s/, /ʃ/, /h/.

Their samples were all collected from one male speaker. 60 samples of each phoneme were recorded by our experiment system with sampling rate of 8KHz in office environment. All the samples were divided into 60 groups in the similar way to that in the previous recognition experiment for 60-fold LOOCV. To avoid the influence caused by samples' duration, spectral analysis was only performed on the center 128-millisecond segments of each sample. Each sample is represented by a 21-dimension vector containing 1 energy coefficient and 20 MFCCs.

B. Experimental Configuration

The functions $E(\cdot)$ and $E^+(\cdot)$ are constructed to be positively correlated with the classification complexity. Fisher's discriminant ratio (FDR) is a classic measure of classification complexity for data with two classes [10]. Its form for individual feature values is defined as

$$f = \frac{(\mu_1 - \mu_2)^2}{{\sigma_1}^2 + {\sigma_2}^2}\ , \tag{3}$$

where $\mu_1, \mu_2, {\sigma_1}^2, {\sigma_2}^2$ are the means and covariances of the two classes, respectively. Multiple features, e.g., d features, can form a d-dimensional column vector. Its form for feature vectors is defined as

$$F\big((p_{l_1}, p_{l_2})\big) = \max_{\boldsymbol{w}} J(\boldsymbol{w}) = \frac{|\boldsymbol{w}^t \boldsymbol{S}_{\mathrm{B}} \boldsymbol{w}|}{|\boldsymbol{w}^t \boldsymbol{S}_{\mathrm{W}} \boldsymbol{w}|}\ . \tag{4}$$

$\boldsymbol{S}_{\mathrm{B}}$ and $\boldsymbol{S}_{\mathrm{W}}$ are between-class scatter matrix and within-class scatter matrix given by

$$\boldsymbol{S}_{\mathrm{B}} = (\boldsymbol{m}_1 - \boldsymbol{m}_2)(\boldsymbol{m}_1 - \boldsymbol{m}_2)^t \tag{5}$$

and

$$\boldsymbol{S}_{\mathrm{W}} = \sum_{i=1}^{2} \boldsymbol{S}_i = \sum_{i=1}^{2} \sum_{\boldsymbol{x} \in \mathcal{D}_i} (\boldsymbol{x} - \boldsymbol{m}_i)(\boldsymbol{x} - \boldsymbol{m}_i)^t\ , \tag{6}$$

where $\mathcal{D}_i$ is the vector set of the ith class; $\boldsymbol{m}_i$ is the mean vector of $\mathcal{D}_i$; $\boldsymbol{w}$ is a d-dimensional vector. Multiple discriminant ratio (MDR) is used as the generalization of FDR for data with multiple classes. It is defined as

$$F^+\big((p_{l_1}, \cdots, p_{l_c})\big) = \max_{\boldsymbol{W}} J^+(\boldsymbol{W}) = \frac{|\boldsymbol{W}^t \boldsymbol{S}_{\mathrm{B}}^+ \boldsymbol{W}|}{|\boldsymbol{W}^t \boldsymbol{S}_{\mathrm{W}}^+ \boldsymbol{W}|} \tag{7}$$

 ISBN: 978-1-61208-195-3

where c is the number of classes. The multi-class generalization of $\boldsymbol{S}_{\mathrm{B}}$ and $\boldsymbol{S}_{\mathrm{W}}$ is given by

$$\boldsymbol{S}_{\mathrm{B}}^{+} = \sum_{i=1}^{c} n_i(\boldsymbol{m}_i - \boldsymbol{m})(\boldsymbol{m}_i - \boldsymbol{m})^t \qquad (8)$$

and

$$\boldsymbol{S}_{\mathrm{W}}^{+} = \sum_{i=1}^{c} \boldsymbol{S}_i = \sum_{i=1}^{c} \sum_{\boldsymbol{x}\in\mathcal{D}_i} (\boldsymbol{x} - \boldsymbol{m}_i)(\boldsymbol{x} - \boldsymbol{m}_i)^t \ , \qquad (9)$$

where n_i is the cardinality of $\mathcal{D}_i$; $\boldsymbol{m}$ is the global mean vector of all the c classes. $\boldsymbol{W}$ is a matrix whose size is $d \times rank(\boldsymbol{S}_{\mathrm{B}}^{+})$. Since the values of FDR and MDR are negatively correlated with classification complexity, we constructed $E(\cdot) = -F(\cdot)$ and $E^{+}(\cdot) = -F^{+}(\cdot)$, respectively. Support vector machines (SVMs) were employed as the classifier in LOOCVs, and LIBSVM [11] with Gaussian kernel was used to implement SVMs.

C. Comparison Experiment and Results Analysis

For each given k ($k = 3, \cdots, 18$), three sets of sivels were selected using our method with FDR, exhaustive method with LOOCV error rate and exhaustive method with MDR, respectively. We computed the LOOCV error rate of each of the three sets, as well as the mean error rate of all the k-cardinality subsets of $\mathcal{P}$. Four curves corresponding to the four error rates by different k are drawn in Fig. 4 and named after "Ours+FDR", "Exhau.+LOOCV", "Exhau.+MDR" and "Mean Error Rate". The sets selected by the exhaustive method with LOOCV error rate can be taken as the best solution for the sivel selection problem, and the mean error rate can be seen as the effectiveness of a random solution. From the experimental results we can see that: our method with FDR usually gives the approximate optimal or the optimal solution, whereas the exhaustive method with MDR could not provide such excellent solution; all the error rate curves present upward trend as k is increasing, which means that more sivels will degrade the accuracy of sivel recognition though they can encode characters more efficiently (i.e., with shorter average code length). A tradeoff for a better global performance should be made between number of sivels and accuracy of recognition. After all, it can be summarized that our sivel selection algorithm has selected sets of sivels with low computational cost and the approximate optimal solution. These sets of sivels have rather low LOOCV error rate. With a proper k, sivels are efficient as code elements.

V. Sivel-based Text Input

The sivel-based text input adopts a customized scheme of encoding characters with sivels similarly to Morse code. The "characters" here include 26 letters, space, comma, period, digits 0∼9 and some control commands such as Backspace and Enter. A user can input text to machines by speaking relevant sivels.

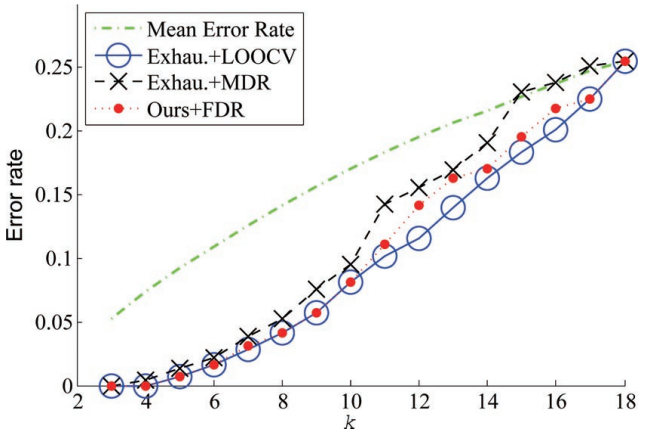


Figure 4. Comparison of the sets of sivels resulted by different methods on their leave-one-out cross-validation error rates .

We describe the usage of sivel-based text input by taking a user's input of a test sentence "hello world, i am a sivel typewriter." for an example. The user is the male speaker who participated in the sivel selection experiment. Using our proposed method, his personal sets of sivels as code elements are selected from the set $\mathcal{P}$ of candidates for sivels. 16 sets of sivels are resulted corresponding to different k ($k = 3, \cdots, 18$). Their 60-fold LOOCV (SVM as the classifier) error rates are computed and drawn as a curve by k, i.e., the curve denoted by "Ours+FDR" in Fig. 4. Owing to the "tradeoff" mentioned above, we choose the set selected when $k = 6$, $\mathcal{V} = \{$ /ɑ/, /ɔ/, /l/, /u/, /s/, /h/ $\}$ with the LOOCV error rate of 98.33%, as the set of sivels initially. Notice that the duration (how long a sivel's signal lasts) was not considered in the selecting phase, and as known from previous experiments, the duration can be used to discriminate between a short sivel and a long sivel. Therefore, the set $\mathcal{V}$ with 6 sivels is extended to the set $\mathcal{V}^{+}$ with 12 sivels by dividing each sivel into the short form and the long form. Similar to what we did previously, the extended set of sivels is denoted by $\mathcal{V}^{+} = \{$`a, o, l, u, s, h, A, O, L, U, S, H`$\}$. To evaluate the recognizability of these sivels, 50 samples for each of these 12 sivels were recorded from the user, and each sample is represented by a 22-dimension vector (1 duration coefficient added). 50-fold LOOCV on these samples achieved an accuracy of 99.67%.

Characters are encoded with these 12 sivels as Table II taking each character's appearance, usage frequency, pronunciation and each sivel's recognizability into account. Our rules are as follows.

- More frequently-used characters have shorter codes, such as space, backspace/enter, e, t, a, o, and i.
- Encode a character with sivel(s) pronounced as more similar to itself as possible, such as r, u, l, h, s, j, c, f, y, and v.

 ISBN: 978-1-61208-195-3

- Make codes look like the appearance of the character when they are put together, such as m, w, b, d, h, q, p, x, 8, k, and g.
- Use easier-to-pronounced combinations of sivels first, then that of easier-to-recognized ones, such as comma, period, z, n, and other digits.

Based on the text coding scheme above, the user can input text any character by speaking its sivel code. To recognize the stream of sivels spoken by the user, beforehand, the samples of these 12 sivels are used as training data of our experiment system. Then with the vibration sensor implemented on the position where the training samples were collected, the user is required to speak every sivel in isolation but with a short interval (within 0.5 sec) for each character. This is helpful to segmenting sivels and avoiding coarticulation effect. After speaking the code of a character the user pauses to wait for the feedback, and at the same time an interval longer than 0.5 sec is detected as the trigger of decoding. The resulted character is obtained and shown on screen (or its corresponding synthetic voice is sent to the user by earphones) as feedbacks. The whole sivel string spoken by the user is "h U l l o O uu o A l ol ou O S O a hh O a O s S uh U l O L ua la U uu A S L U A uo" in which a space means an interval longer than 0.5 sec. Ideally, it takes about 50 sec to input the test sentence with 37 characters by speaking 46 sivels. In practice, the cost time is around 63 sec with an input error rate of around 4%. The extra part of time is resulted by the user's reaction to feedback and the correction of input errors. To improve the inputting efficiency, an auto-correction strategy is used. This strategy allows the user to speak sivel codes of a word without waiting for the feedback of each character. After a space or a punctuation is input, the latest input word is automatically revised according to a dictionary. With the help of the auto-correction strategy, the user is able to input the test sentence within 55 sec.

VI. Discussion

Sivel-based text input (called sivel input for short) holds many advantages which is similar to speech input and NAM input (text input method with NAM recognition as its underlying recognition technology). They are all articulators-operated text input methods. They recognize the time series of code elements from signals generated by users' articulators, then generate text with a certain text coding scheme. Therefore, they can provide a useful channel for human-machine communication in some situations such as when users hands or eyes are busy, when hands-operated methods are difficult to be implemented, and when users can not move their arms or hands reliably due to disabilities. In addition, sivel input uses signals having little interaction with the ambient acoustic environment. This makes sivel input applicable to more situations with challenging acoustic environments than speech input such as where are noisy or silence-needed.

Table II
Characters and their codes with sivels in $\mathcal{V}^+$.

Char	Code	Char	Code	Char	Code	Char	Code
a	a	k	lu	u	u	4	UU
b	lo	l	l	v	uh	5	SS
c	ss	m	hh	w	uu	6	HH
d	ol	n	oh	x	uU	7	LL
e	U	o	o	y	ua	8	OO
f	hu	p	la	z	aH	9	OL
g	oa	q	al	0	AA	,	ou
h	h	r	A	1	aa	.	uo
i	S	s	s	2	oo	(Space)	O
j	sa	t	L	3	ll	(Bs/En)	H

The most significant difference among sivel input, speech input and NAM input is that sivel input adopts a customized scheme of text coding. Sivel input encodes text with only particular phonemes according to a customized scheme which can be optimized for specific users or tasks, whereas speech and NAM input encode text with conventional phonic units (phonemes and syllables) according to knowledge on linguistics. Although sivel input is not such a natural text input method as the other two, it is more effective in some special applications.

Here are two instances illustrating the advantages of the customized text coding scheme. One instance is the potential application of sivel input in secure communications. Encoding the messages with sivels itself is also an encryption process. Using sivel input, users can send messages quietly in various acoustic environments with high accuracy and without the participation of hands (and even eyes, if feedback via hearing), which makes the communication action difficult to be noticed by others. The other instance is that sivel input can enable speech disorder people to communicate with machines using an articulators-operated method. There are many speech disorder people who are not able to sound speech but only some phoneme-like segments of (silent) voice. They can select sivels from those segments of (silent) voice that they can sound reliably regardless of whether these voices have linguistic meanings. After giving user-customized names to these sivels and encoding text with them, speech disorder people are able to input text to machines using their articulators.

VII. Conclusion and Future Work

This paper has proposed the concept of silent voice elements (sivels) for text input. We selected a set of sivels empirically as an example, at first. The experiments of sivel recognition on the example show that sivels have potentials to be efficient code elements and they are speaker-dependent.

 ISBN: 978-1-61208-195-3

Then the selection of sivels for individuals has been accomplished using the sivel selection algorithm, and experimental results demonstrate that the algorithm can select set of sivels with high recognizability. We have introduced the sivel-based text input method in which characters are encoded with sivels according to a customized scheme. Using the sivel-based text input, a user inputted a sentence with 37 characters by speaking 46 sivels within 55 sec. Finally, the comparison among speech input, NAM input and sivel-based text input have been made and discussed to illustrate the advantages of the customized scheme adopted by the sivel-based text input.

For future work, we are currently testing the various robustness of sivel-based input, such as the robustness to the placement of the vibration sensors and how robust the sivel classifier are over time, and to improve sivel-based input for everyday utility.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers and Prof. Petre Dini from IARIA for their valuable comments and suggestions to improve this paper.

This work was supported in part by the Natural Science Foundation of China(NSFC) under Grant No. 90920009 and NSFC-Guangdong Joint Fund under Grant No. U1035004.

REFERENCES

[1] B. Denby, T. Schultz, K. Honda, T. Hueber, J. Gilbert, and J. Brumberg, "Silent speech interfaces," *SPEECH COMMUN.*, vol. 52, no. 4, pp. 270–287, 2010.

[2] M. Fagan, S. Ell, J. Gilbert, E. Sarrazin, and P. Chapman, "Development of a (silent) speech recognition system for patients following laryngectomy," *Medical engineering & physics*, vol. 30, no. 4, pp. 419–425, 2008.

[3] T. Schultz and M. Wand, "Modeling coarticulation in EMG-based continuous speech recognition," *SPEECH COMMUN.*, vol. 52, no. 4, pp. 341–353, 2010.

[4] M. Wester and T. Schultz, "Unspoken speech-speech recognition based on electroencephalography," Master's thesis, Karlsruhe: Universität Karlsruhe (TH), 2006.

[5] Y. Nakajima, H. Kashioka, K. Shikano, and N. Campbell, "Non-audible murmur recognition input interface using stethoscopic microphone attached to the skin," in *ICASSP*, vol. 5. IEEE, 2003, pp. V–708 – V–711.

[6] Y. Nakajima, "Development and evaluation of soft silicone NAM microphone," *IEICE Technical Report*, vol. 105, no. 97, pp. 7–12, 2005.

[7] T. Hirahara, M. Otani, S. Shimizu, T. Toda, K. Nakamura, Y. Nakajima, and K. Shikano, "Silent-speech enhancement using body-conducted vocal-tract resonance signals," *SPEECH COMMUN.*, vol. 52, no. 4, pp. 301–313, 2010.

[8] V. Tran, G. Bailly, H. Loevenbruck, and T. Toda, "Improvement to a NAM-captured whisper-to-speech system," *SPEECH COMMUN.*, vol. 52, no. 4, pp. 314–326, 2010.

[9] D. Babani, T. Toda, H. Saruwatari, and K. Shikano, "Acoustic model training for non-audible murmur recognition using transformed normal speech data," in *ICASSP*. IEEE, 2011, pp. 5224–5227.

[10] T. Ho and M. Basu, "Complexity measures of supervised classification problems," *PAMI, IEEE Transactions on*, vol. 24, no. 3, pp. 289–300, 2002.

[11] C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machines*, 2001, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

 ISBN: 978-1-61208-195-3

Summarization of Real-Life Events Based on Community-Contributed Content

Manfred del Fabro, Anita Sobe, Laszlo Böszörmenyi
Institute of Information Technology (ITEC)
AAU Klagenfurt University
Klagenfurt, Austria
{*manfred,anita,laszlo*}*@itec.aau.at*

***Abstract*—In this paper, we investigate whether community-contributed multimedia content can be used to make video summaries of social events. We implemented an event summarization algorithm that uses photos from Flickr and videos from YouTube to compose summaries of well-known society events, which took place in the last three years. The comparison with a manually obtained ground truth shows a good coverage of the most important situations of these events. We do not claim to produce the best summaries possible, which may be compared to the work of a human director, but we analyze what can be achieved with community-contributed content by now.**

***Keywords*-video summarization. event summarization. social media. real-life events. video retrieval. image retrieval. multimedia entertainment.**

I. Introduction

Twenty years ago, people were informed about a social event, such as a royal wedding, through a few, authorized, professional camera teams and journalists of printed press. Nowadays, a vast amount of additional photos, videos and some text, the latter mainly in form of metadata of the images, are uploaded to social platforms, such as Flickr and YouTube.

If we query these platforms to get informed about a certain social event, like the royal wedding of William and Kate in April 2011, we get a – usually extremely long – list of photos or videos. Even though the list is sorted corresponding to relevance, this is not a proper answer for such a question. We rather preferred to get a compact presentation of a predefined length, which gives us a summary, composed from the views of many people that have witnessed the event. This is not necessarily the best view, but the view that can be created based on the information people provided when uploading the content. Nevertheless, this view is usually very rich and contains a lot of interesting, even surprising elements. Of course, it may also contain garbage and even malicious content, but this is out of scope of this paper.

For this study, we only used social events related to entertainment. However, our approach is also applicable to other events, such as a traffic jam on a highway [23], seen by a number of drivers on the road, or a certain medical event, identified by a group of medical doctors in an arthroscopic surgery video [13].

This paper is organized as follows. Section II describes the related work, in Section III our event summarization algorithm is described in detail, in Section IV the experimental results are presented, and Section V concludes this paper and gives an outlook for future work.

II. Related Work

The summarization of multimedia content is the target of many research projects. Most of them focus on video abstraction and video summarization. Two extensive reviews of key-frame extraction and video summarization approaches are given in [15][22]. The presented algorithms summarize single videos with selected still images or with a short summary video. In our approach, we generate summaries that consist of content that comes from multiple sources.

One approach, which uses multiple videos as input for a summarization algorithm, is introduced in [10]. Videos of a whole basketball season in the USA and the corresponding metadata are used to create summary videos under different aspects, like summaries of the whole championship, of only one team or even of a single player. The authors only consider a single database for the content selection. Furthermore, professionally produced content from TV stations is used. No community-contributed annotations and ratings are available. In contrast, the system presented in this paper takes advantage of all context information that is provided by the community.

Not only the summarization of videos has been extensively studied, but the summarization of image collections has also been a target of research activities [19]. This work defines three aspects of an effective summary and formalizes models to optimize them: (1) quality of the content, (2) diversity of the content and (3) coverage of the whole collection. These aspects are also important for our summaries.

During the last few years, more and more research activities were focusing on real-life events in the context of multimedia data. In [26], a common event model for multimedia applications is proposed. Eight basic aspects are defined to describe an event, but also the relationship of an event to other events.

An event-based clustering algorithm is proposed in [14]. A layered clustering algorithm produces different clusters of videos, where each cluster represents one event.

 ISBN: 978-1-61208-195-3

In [12], information from online event directories is used to get metadata about an event, like the title or the geo information. With the help of this additional information the authors try to gather as many photos and videos as possible from Flickr and YouTube.

A visual-based method for retrieving events in photo collections of community-contributed contents is introduced in [21]. Based on a query image, an image collection is searched for similar photo records that may be of the same event.

In [25], an automatic remixing approach for community-contributed content from music concerts is presented. Users can record and upload videos during live events. Afterwards, the shared content is synchronized based on the creation timestamps and a master audio track is extracted from the single audio tracks of the synchronized videos. In the end, video remixes of a concert are automatically created based on automatically detected regions of interest.

The organization of tagged photo collections based on landmark and event detection is presented in [16]. Photos are arranged on their spatial closeness and their relatedness to events.

In [9], a joint content-event model is proposed, which allows an event-based indexing of videos instead of a concept-based one. A content model that describes videos in terms of scenes and shots is linked to an event model that defines different events and how they may be related to each other.

All these event-related approaches identify events in multimedia content or they index or cluster the content according to events. However, in this paper, we investigate aspects how community-contributed content is suited for the generation of visual summaries of social events, which to the best of our knowledge has not been done before.

A work that is not event-centric but that shows the power of utilizing community-contributed content is presented in [18]. Images of online photo collections are used to generate 3D views of famous places in the world where a lot of photos are taken. The introduced application allows an exploration of places based on the content of people that have really been there.

III. Event Summarization

A summary of a social event should consider the three aspects, how to build a summary [19]: (1) quality, (2) diversity and (3) coverage. (1) Photos and videos of poor visual quality should be not included into the summary. (2) Similar photos or videos should not be included more than once. (3) The resulting summary should cover the event as good as possible showing as many situations that occurred as possible.

As the quality aspect has been intensively studied, we concentrate in this paper on the two other aspects. During the generation of the summaries we focus on the maximum diversity of the content. Our summarization algorithm may not produce the best summary possible, but it creates a representation that emerges from most relevant and most popular contents related to a certain event of social media sharing platforms. In our evaluation we then investigate the coverage of that emerging view.

A. Summarization Algorithm

A summary is built according to search terms, specified by the user, such as: *Royal wedding of William and Kate*. First we cluster the content, based on the available textual descriptions. After that we filter wrongly located content based on GPS information. At last, we create a summary, from the remaining content. A flow chart, which illustrates these steps, is shown in Figure 1.

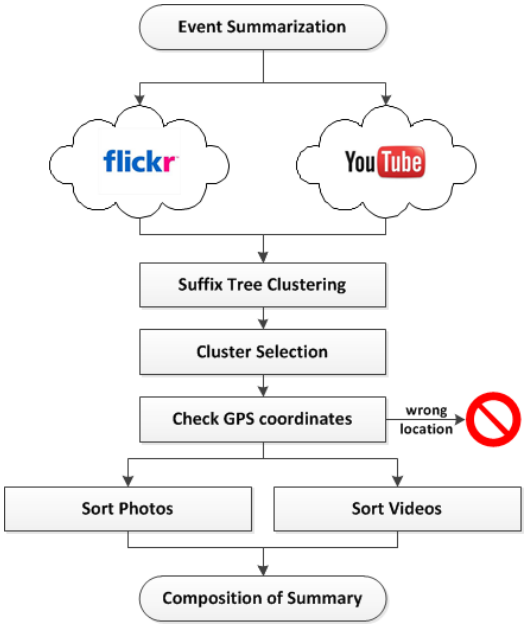


Figure 1. Flow chart of algorithm

The composition of an event summary is influenced by six parameters: (1) search terms to describe the event with keywords, (2) number of photos or videos to be shown in parallel, (3) maximum duration of the summary in seconds, (4) location, (5) start of the timespan the content must have been produced, and (6) end of the timespan the content must have been produced.

The search terms are passed to Flickr and YouTube as text queries. The more comprehensive the query, the better focused the retrieved content will be. Therefore, different queries lead to different summaries. The results of both platforms are sorted by relevance. We rely on the relevance calculations of both platforms and do not perform our own ones. This is the default sorting mode of both platforms. If people use the web interfaces of Flickr or YouTube, they also get the results sorted by relevance. A summary may

 ISBN: 978-1-61208-195-3

consist of more than a single sequence of photos and videos. Figure 2 shows a screenshot of an event summary, which consists of four parallel streams.

We only query for content that has been produced within the indicated timespan. We are well aware that the timestamps of the photos and videos may be wrong or even missing. We are going to pay attention to this fact in our evaluation. The queries do not return the photos and videos themselves, but only their metadata. For runtime reasons, we decided to limit the amount of Flickr results to 5000 per query. The YouTube API limits the amount of results to 1000 per query. The amount of photos and videos we consider for the summary generation is still much larger than a user would manually examine when clicking through Flickr and YouTube results. Therefore, we are of the opinion that this limitation is reasonable.

Figure 2. Screenshot of event summary

B. Clustering

In the next step, we cluster the photos and videos based on their textural descriptions. For that purpose, we use the text suffix tree clustering algorithm introduced in [27]. It has already successfully been applied to web document clustering and shows some interesting properties that can be exploited for our task.

The suffix tree clustering algorithm separates relevant from non-relevant content, even if only text snippets are available. Furthermore, the authors of the algorithm showed that it works fast. Therefore, it is well suited for multimedia content, as for photos and videos typically only short descriptions are available. For each photo and video retrieved from Flickr or YouTube, we extract title, description and tags. This information is the input for the clustering algorithm. At the end, we receive several clusters consisting of photos and videos. For each cluster, a summary in form of a *dominant phrase* is provided by the clustering algorithm. For the content selection, we choose the largest cluster of which the dominant phrase includes the search terms of our query.

C. Content selection and composition of event summaries

Photos and videos often have misleading descriptions regarding their location. We try to overcome this problem by investigating the GPS coordinates of the content. The location indicated in textual form is translated in GPS coordinates using the Google Geocoding API [2]. Using the retrieved GPS coordinates and the level of detail (country, region, city or street) we are able to eliminate content that has been produced in a wrong place. If ambiguities are possible (e.g., Paris, France and Paris, Texas), the location must be specified precisely. Otherwise wrong content may be included in the summary.

The selection of photos for the summary is based on the number of how frequently a photo has been viewed on Flickr. The selection of videos is based on the user ratings (up to 5 stars), the number of views and the number of *likes* a video has on YouTube. For each event summary, we select content in such a way that the amount of time when photos are shown and the amount of time when videos are played are approximately equal. While videos have a natural length we define a default duration of 7 seconds for still images in the summary. In a single sequence this may be too long to show a single image, but as soon as more than one sequence is shown in parallel the viewers need more time to look at all photos. For example, for a video with a duration of 28 seconds we add four photos to the summary. This ratio is automatically adapted if the number of either the photos or the videos is too low. It may happen that no videos are included in a summary, because the selected cluster does not contain videos at all or the length of the contained videos exceeds the maximum duration of the summary.

One important aspect of the summarization of content is to avoid redundancy [22]. We rely on visual image features to identify redundant photos. Each image selected as a candidate for a summary is matched against all other photos that are already in the summary. If the visual similarity to a photo in the summary is too high, the candidate image will not be added. For the estimation of the visual similarity, we extract the Color and Edge Directivity Descriptor (CEDD) [7] from each photo. The CEDD can be extracted fast and it showed good results in an evaluation of different image features for video summarization [11].

Finally, when all photos and videos are selected we sort the whole content based on their creation timestamps. With this simple approach we want to investigate how good timestamps are suited to make a temporal alignment of the content.

D. Summary format and presentation

In the resulting event summary, the videos are played first and then slide shows of the photos are shown. We think that the viewers get a good impression and an overview by watching the videos first, while photos are better suited to cover certain aspects in detail that the videos may miss.

 ISBN: 978-1-61208-195-3

Table I
DETAILS ABOUT COMMUNITY-CONTRIBUTED DATA RELATED TO CERTAIN SOCIAL EVENTS

Search terms	inauguration obama	royal wedding	fifa world cup final 2010	champions league final 2011
Flickr results	59643	47372	2535	1529
YouTube results	15800	52500	547000	186000
Photos/Users selected	1062/182	1516/343	668/81	161/22
Videos/Users selected	1/1	211/211	114/90	83/72
Photos with GPS	333	437	333	42
Videos with GPS	0	7	7	0
Wrong location	81	211	160	1
Photos/Users in summary	168/51	73/28	81/17	83/14
Videos/Users in summary	0/0	5/5	4/4	5/5

We present the generated summaries in our own Video Browser [8], which is depicted in Figure 2. This video browser allows showing of several videos and photos in parallel. The audio playback is selected from one of the presented videos by default or by mouse-over on one of the videos.

To organize the temporal presentation the player interprets a formalism called Video Notation (ViNo) [20]. ViNo is a multipurpose multimedia language, which we use to define the presentations in a short and flexible way. Each event summary is a ViNo expression, which consists of a sequence of videos or photos shown in parallel. E.g., the presentation of four videos as shown in Figure 2 can be expressed in ViNo as $[u1||u2]|[u3||u4]$, where we assume that u*i* is the identifier of a video and $||$ means parallel presentation. Each line is grouped by squared brackets.

IV. EVALUATION

We chose four well-known social events that took place in the last three years for the evaluation: (1) the inauguration of Barack Obama [5], (2) the Royal Wedding of William and Kate [6], (3) the FIFA World Cup Final 2010 [3] and (4) the UEFA Champions League Final 2011 [4]. All four events took place on one single day, were attended by several thousands of people and attracted the attention of millions of people around the world.

The same algorithm was used for all four summaries. We did not tune it according to the events. All event summaries in our evaluation consist of 4 parallel streams and have a maximum duration of 5 minutes. The timespan we used for our queries starts with the day the event took place and ends one month after that. Other investigations showed that even a time interval of 7 days is sufficient [12]. Screen captures of the four composed event summaries are available online [1].

Table I lists the *Search terms* that were used as input for the summary generation and gives details about the retrieved content. We tried to use as few search terms as possible to describe the events, because people also tend to use only a few terms when searching for multimedia content online [24].

The same query, which is used for the summary generation, has also been used to query the Flickr (*Flickr results*) and the YouTube (*YouTube results*) website to get a first impression of the available content. For the first two queries much more photos can be retrieved from Flickr than for the two soccer matches. The reason for that is that more specific text queries were used for the two soccer matches consisting of 4 and 5 terms, compared with only 2 terms for the first two queries. The more specific a query is the less results are returned from Flickr. Interestingly, for the two soccer matches a huge amount of videos is available. A closer examination shows that people played these matches also on their gaming consoles and published videos of that computer games online.

The event summary algorithm originally included the 5000 most relevant Flickr and the 1000 most relevant YouTube results. Finally, even a smaller subset – as produced by the clustering – is used for the content selection. The rows *Photos/Users selected* and *Videos/Users selected* list how many photos and videos were included in the final cluster for the summary generation and how many distinct users uploaded these contents. It can be seen that several photos are selected from each included Flickr uploader, while in most cases the included YouTube videos have different users.

In the created summaries 3 to 6 photos of a single uploader (*Photos/Users in summary*) are included. Each video in these summaries (*Videos/Users in summary*) has a single uploader. The summary of the inauguration of Barack Obama only consists of photos. The cluster selected for the

 ISBN: 978-1-61208-195-3

summary only contains one video of his oath, but its length exceeds the maximum duration of the summary. In general, these summaries include content from a variety of users, thus these summaries are really conveying a broad view of people that witnessed the selected events.

The retrieved data shows that the available GPS data provide only a strongly limited support to estimate the location where the content was produced. For only 25 – 50 % of the selected photos (*Photos with GPS*) are the GPS data available and videos (*Videos with GPS*) hardly having this data associated at all. Nevertheless, many photos could be filtered that were taken in a wrong location. The relatively high amount of photos excluded due to wrong semantic location (*Wrong location*) can be easily explained. The events chosen for the summaries were broadcasted all over the world. The excluded content was produced by people somewhere else on the world. In most cases people celebrated parties to follow the original event in a group on TV. The content produced at those parties was annotated with textual descriptions related to the original event. Therefore, it was initially included in the results sets retrieved by Flickr and YouTube.

The coverage of the created summaries is compared against a manually obtained ground truth. The most important *situations* of the chosen events were figured out with the help of Wikipedia articles [3][4][5][6]. For each event a corresponding set of situations was identified. A situation may be a temporal happening, such as *exchange of the rings*, a location, such as the Westminster Abbey or even persons, such as *Prince Harry*. Table II lists the identified situations for all four events. Further information about these situations can be obtained from the Wikipedia articles. Later in this paper, we refer to these four lists of situations when the evaluation of the coverage of the generated summaries is presented.

We decided to rely on Wikipedia, because it is difficult to find an objective evaluation metric for the quality of summaries. Summaries are always somehow based on subjective opinions as [17] showed. Wikipedia articles usually have several authors, who perform discussions and have to agree on the text of the article. Therefore, Wikipedia articles convey the common opinion of a crowd of people. We take advantage of that common opinion to get a more objective ground truth for the evaluation of the coverage of the generated event summaries.

We compared our event summaries with a standard web search on Flickr and YouTube. As the evaluated summaries have a duration of 5 minutes, we limited the number of Flickr and YouTube results to amounts that could approximately be browsed in that time span. The first 120 photos from Flickr and the first 20 videos from YouTube are investigated for each query. If we compare the coverage of the generated summaries with the Flickr and YouTube results in the following parts of this evaluation, we always refer to result sets of that size (indicated by *Flickr* resp. *YouTube* in the following diagrams).

Table II
INTERESTING SITUATIONS OF THE FOUR SOCIAL EVENTS

Inauguration Obama	**Royal Wedding**
1. United States Capitol	1. Westminster Abbey
2. Music live performances	2. Bride (Kate)
3. Invocation by pastor	3. Groom (William)
4. Aretha Franklin singing	4. Pippa Middleton
5. Oath of Vice President	5. Prince Harry
6. Oath of Barack Obama	6. Queen Elisabeth II.
7. Inaugural address	7. Young bridesmaids
8. Prayers	8. Pageboys
9. Departure of former president	9. Arrival of Kate
10. Signing of first orders	10. Exchange of rings
11. Luncheon	11. Lesson
12. Parade	12. Sermon
13. Inauguration balls	13. Leaving Westminster Abbey
14. National prayer service	14. Return to palace in coach
15. Oath of office	15. Lunchtime reception
	16. Appearing on balcony
	17. Harpist performance
	18. William & Kate leaving with car
	19. Private dinner
	20. Wedding cake
	21. Merchandise
	22. Broadcasting
World Cup	**Champions League**
1. Soccer City Stadium	1. Wembley Stadium
2. de Jong's kick against Alonso	2. Chance Hernandez (ManU)
3. Chance Robben (NED)	3. Chance Villa (Barca)
4. Chance Sneijder (NED)	4. Chance Villa (Barca)
5. Chance Ramos (ESP)	5. Goal Pedro (Barca)
6. Red card Heitinga (NED)	6. Goal Rooney (ManU)
7. Goal Iniesta (ESP)	7. Chance Messi (Barca)
8. Award ceremony	8. Chance Messi (Barca)
	9. Goal Messi (Barca)
	10. Chance Messi (Barca)
	11. Chance Xavi (Barca)
	12. Goal Villa (Barca)
	13. Chance Rooney (ManU)
	14. Chance Nani (ManU)
	15. Award ceremony

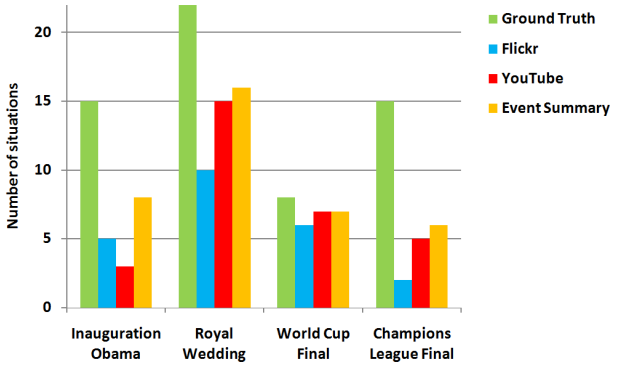


Figure 3. Comparison of situations found

The results are shown in Figure 3. In all cases, the first Flickr results only include few situations of interest. The reason for that is that people tend to photograph themselves

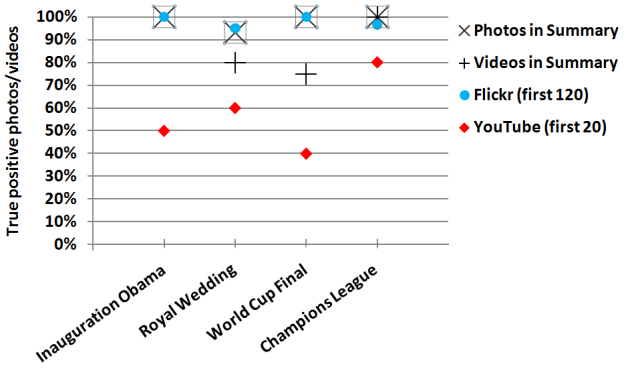


Figure 4. Amount of true positive photos or videos

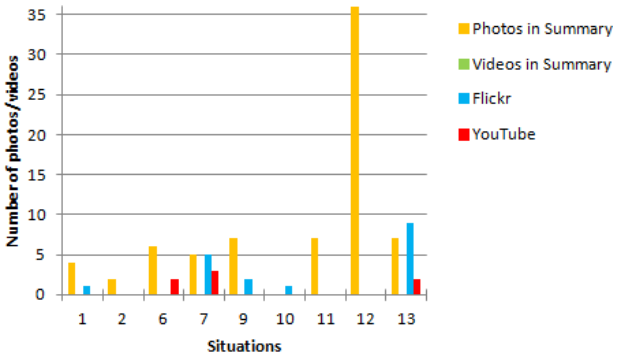


Figure 5. Inauguration Obama

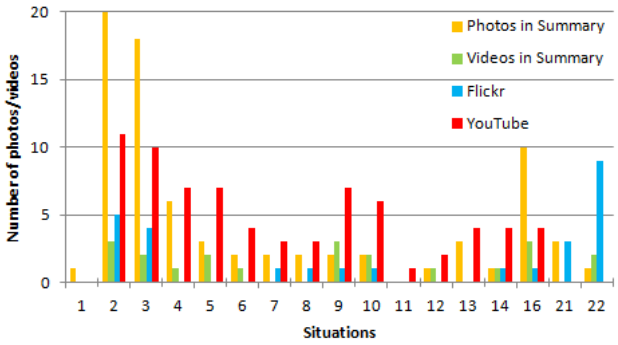


Figure 6. Royal Wedding

when visiting an event. Therefore, a lot of images show visitors of the event and only few photos show situations as they were identified based on the Wikipedia entries. Except for the inauguration of Obama the YouTube results show more interesting situations than the Flickr results. The event summarization algorithm shows in all cases the best performance. It includes as much situations as Flickr or YouTube or even more.

If content is examined regardless of the searched situations, it can be recognized that the precision of the Flickr results is high. They include a high amount of content that is related to the searched events. Figure 4 shows the percentage of true positive photos and videos in the Flickr and YouTube results as well as in the event summaries. For the latter, we distinguish between photos and videos. A photo or video is regarded to be a true positive if it is somehow related to the event. The Flickr results contain a lot of true positives, which also has a positive effect on the photos in the summaries. Except for the Champions League final the YouTube results have a lot of false positives, although only the 20 most relevant results returned by YouTube are considered. The event summarization algorithm also includes false positives in the summaries, but the ratio of true positives is much better than the one of the YouTube results. This is an effect of the suffix tree clustering of the content. As the biggest cluster is chosen, which is related to the query, it is more likely that this cluster includes relevant content. Note that false positives include photos and videos, which are not wrong, rather strange. For example, if some people record the movements of the police at the royal wedding (as they did indeed), this is topic for a non-technical discussion, whether or not these images are misplaced.

The comparison of the coverage shows that quite a lot of the defined situations of interest are not included in the summaries as well as in the Flickr or YouTube results. Therefore, we want to take a closer look at the situations found. Figure 5 shows the situations detected for the inauguration of Barack Obama. It can be noticed that a lot of photos are showing the parade (situation no. 12) after the inauguration. That was somehow expected, because the parade was watched by a lot of people along the track and thus a lot of photos have been made. For the other situations it can be stated that people especially took photos of the highlights, like the oath of Obama (6), his inaugural address (7) or the departure of the former president Bush (9). Also the society events like the luncheon (11) and the balls (13) seem to attract people. The oath of the Vice-President (5), prayers (8) or events that took place in the office of Obama, like the signing of the first orders (10) or his second oath (15) are not covered by the content we received from Flickr and YouTube.

Figure 6 shows the identified situations of the royal wedding in detail. As it can be seen the involved people like Kate (2), William (3), Pippa (4), Prince Harry (5) or the Queen (6) get a lot of attention. Also the appearing on the balcony (16) or situations that took place in the streets or in front of the church (9, 13 and 14) are included often. The reason is again that for public situations a lot of content is produced, while for private ones like the family celebrations (15) or the private dinner (19) in the evening nothing can be found.

We also wanted to investigate events where the interesting situations may be clearer. Therefore, we decided to investigate event summaries of two soccer games that attracted

 ISBN: 978-1-61208-195-3

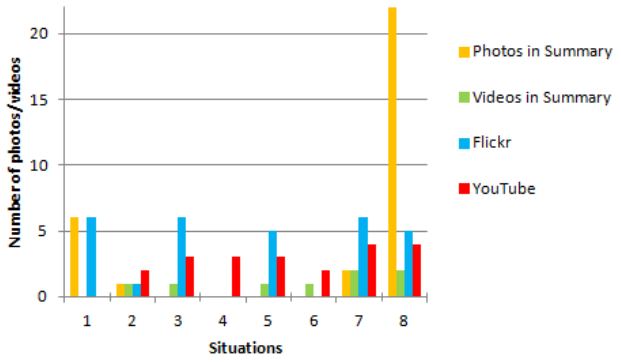


Figure 7. FIFA World Cup Final 2010

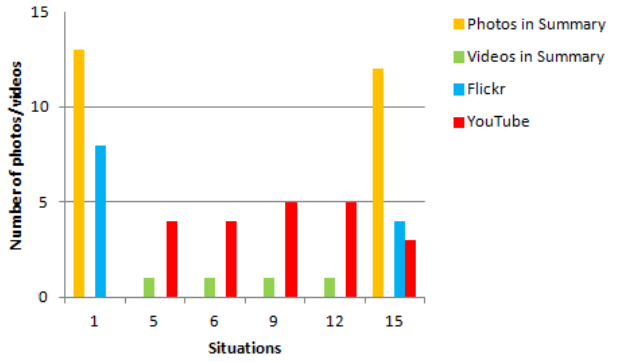


Figure 8. UEFA Champions League Final 2011

the attention of millions of people around the world. The identified situations for the two games are shown in Figure 7 and Figure 8. For both games it can be stated that all goals are identified by our summaries, but nearly all chances that did not result in goals are missed. In addition to the goals both summaries also include a lot of situations showing the venue and the award ceremonies of the winning teams.

Regarding the temporal alignment of the content we must state that the timestamps of the content are not sufficient for good ordering of the content. By simply watching the generated summaries it can be seen that the content is mixed up temporarily in all summaries. It seems that people do not care about their cameras having correct date and time settings. Nevertheless, this could change, if people notice in the future that innovative tools can make good use of this information.

V. Conclusion

In this paper, we presented an algorithm for the summarization of real-life events based on community-contributed multimedia content. We composed four summaries of events that attracted a lot of people during the last three years using photos from Flickr and videos from YouTube. We evaluated the coverage of our summaries by comparing them with Wikipedia articles that report about the corresponding events. This innovative evaluation technique allows us to identify the important happenings of social events without doing manual observations of these events, but by relying on the common opinion of a group of people that created and edited the corresponding articles. Furthermore, we investigated some characteristics of community-contributed content with respect to event summarization. The composed summaries show a good coverage of interesting situations that happened during the selected events. Next, we plan to perform user studies to investigate how the quality of our summaries is perceived by people that watch them.

There are still several open questions that remain in this rather new topic, like the correct temporal alignment of content or the identification of malicious content. In our future work we are also going to incorporate additional sources of information, like textual descriptions of the events, for the temporal alignment as well as for the selection of content. Furthermore, we are going to make investigations regarding the sensitivity of our algorithm to be able to state which results can be achieved under which circumstances.

Further investigations have to be done on events that last longer than one day (e.g., the whole FIFA World Championship), events that have many parallel sub-events (e.g., Olympic Games), and small events (which only attract the attention of a small audience). But not only events that are related to entertainment are of interest. The presented approach can also be applied to spontaneous real-life events like a traffic jam on a motorway or a catastrophe scenario like a heavy earthquake. If summaries of such events are constructed from the content that involved people or witnesses have captured, emergency response teams may profit from that information and may be steered and coordinated in a better way.

Acknowledgment

This work was supported by Lakeside Labs GmbH, Klagenfurt, Austria and funding from the European Regional Development Fund and the Carinthian Economic Promotion Fund (KWF) under grant KWF-20214 17097 24774 and grant KWF-20214 22573 33955.

References

[1] Demo videos of event summaries. http://soma.lakeside-labs.com/?page_id=279. [2012-02-09].

[2] Google geocoding api. http://code.google.com/apis/maps/documentation/geocoding/. [2012-02-09].

[3] Wikipedia: 2010 FIFA world cup final. http://en.wikipedia.org/w/index.php?title=2010_FIFA_World_Cup_Final&oldid=439386816. [2012-02-09 (Permalink)].

[4] Wikipedia: 2011 UEFA champions league final. http://en.wikipedia.org/w/index.php?title=2011_UEFA_Champions_League_Final&oldid=440623020. [2012-02-09 (Permalink)].

 ISBN: 978-1-61208-195-3

[5] Wikipedia: Inauguration of barack obama. http://en.wikipedia.org/w/index.php?title=Inauguration_of_Barack_Obama&oldid=439374433. [2012-02-09 (Permalink)].

[6] Wikipedia: Wedding of prince william and catherine middleton. http://en.wikipedia.org/w/index.php?title=Wedding_of_Prince_William_and_Catherine_Middleton&oldid=440475841. [2012-02-09 (Permalink)].

[7] S. Chatzichristofis and Y. Boutalis. CEDD: Color and edge directivity descriptor: A compact descriptor for image indexing and retrieval. In A. Gasteratos, M. Vincze, and J. Tsotsos, editors, *Computer Vision Systems*, volume 5008 of *Lecture Notes in Computer Science*, pages 312–322. Springer Berlin / Heidelberg, 2008.

[8] M. del Fabro, K. Schoeffmann, and L. Böszörmenyi. Instant video browsing: A tool for fast non-sequential hierarchical video browsing. In G. Leitner, M. Hitz, and A. Holzinger, editors, *HCI in Work and Learning, Life and Leisure*, volume 6389 of *Lecture Notes in Computer Science*, pages 443–446. Springer Berlin / Heidelberg, 2010.

[9] N. Gkalelis, V. Mezaris, and I. Kompatsiaris. Automatic event-based indexing of multimedia content using a joint content-event model. In *Proceedings of the 2nd ACM international workshop on Events in multimedia*, EiMM '10, pages 15–20, New York, NY, USA, 2010. ACM.

[10] R. Kaiser, M. Hausenblas, and M. Umgeher. Metadata-driven interactive web video assembly. *Multimedia Tools and Applications*, 41:437–467, 2009.

[11] M. Kogler, M. del Fabro, M. Lux, K. Schoeffmann, and L. Böszörmenyi. Global vs. local feature in video summarization: Experimental results. In *Proceedings of the 10th International Workshop of the Multimedia Metadata Community on Semantic Multimedia Database Technologies (SeMuDaTe09) in conjunction with the 4th International Conference on Semantic and Digital Media Technologies (SAMT 2009).*

[12] X. Liu, R. Troncy, and B. Huet. Finding media illustrating events. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, ICMR '11, pages 58:1–58:8, New York, NY, USA, 2011. ACM.

[13] M. Lux, O. Marques, K. Schöffmann, L. Böszörmenyi, and G. Lajtai. A novel tool for summarization of arthroscopic videos. *Multimedia Tools and Applications*, 46:521–544, 2010.

[14] J. Makkonen, R. Kerminen, I. D. D. Curcio, S. Mate, and A. Visa. Detecting events by clustering videos from large media databases. In *Proceedings of the 2nd ACM international workshop on Events in multimedia*, EiMM '10, pages 9–14, New York, NY, USA, 2010. ACM.

[15] A. G. Money and H. Agius. Video summarisation: A conceptual framework and survey of the state of the art. *J. Vis. Comun. Image Represent.*, 19(2):121–143, February 2008.

[16] S. Papadopoulos, C. Zigkolis, Y. Kompatsiaris, and A. Vakali. Cluster-based landmark and event detection for tagged photo collections. *Multimedia, IEEE*, 18(1):52 –63, jan. 2011.

[17] G. J. Rath, A. Resnick, and T. R. Savage. The formation of abstracts by the selection of sentences. part i. sentence selection by men and machines. *American Documentation*, 12(2):139–141, 1961.

[18] I. Simon, N. Snavely, and S. M. Seitz. Scene summarization for online image collections. *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8, Oct. 2007.

[19] P. Sinha, S. Mehrotra, and R. Jain. Summarization of personal photologs using multidimensional content and context. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, ICMR '11, pages 4:1–4:8, New York, NY, USA, 2011. ACM.

[20] A. Sobe, L. Böszörmenyi, and M. Taschwer. Video Notation (ViNo): A Formalism for Describing and Evaluating Non-sequential Multimedia Access. *International Journal on Advances in Software*, 3(1 & 2):19–30, sep 2010.

[21] M. R. Trad, A. Joly, and N. Boujemaa. Large scale visual-based event matching. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, ICMR '11, pages 53:1–53:7, New York, NY, USA, 2011. ACM.

[22] B. T. Truong and S. Venkatesh. Video abstraction: A systematic review and classification. *ACM Trans. Multimedia Comput. Commun. Appl.*, 3(1):3+, 2007.

[23] R. Tusch, A. Fuchs, H. Gutmann, M. Kogler, J. Köpke, L. Böszörmenyi, M. Harrer, and T. Mariacher. A multimedia-centric quality assurance system for traffic messages. In J. Düh, H. Hufnagl, E. Juritsch, R. Pfliegl, H.-K. Schimany, and H. Schönegger, editors, *Data and Mobility*, volume 81 of *Advances in Intelligent and Soft Computing*, pages 1–13. Springer Berlin / Heidelberg, 2010.

[24] R. van Zwol, B. Sigurbjornsson, R. Adapala, L. Garcia Pueyo, A. Katiyar, K. Kurapati, M. Muralidharan, S. Muthu, V. Murdock, P. Ng, A. Ramani, A. Sahai, S. T. Sathish, H. Vasudev, and U. Vuyyuru. Faceted exploration of image search results. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 961–970, New York, NY, USA, 2010. ACM.

[25] S. Vihavainen, S. Mate, L. Seppälä, F. Cricri, and I. D. Curcio. We want more: human-computer collaboration in mobile social video remixing of music concerts. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, CHI '11, pages 287–296, New York, NY, USA, 2011. ACM.

[26] U. Westermann and R. Jain. Toward a common event model for multimedia applications. *IEEE Multimedia*, 14(1):19–29, 2007.

[27] O. Zamir and O. Etzioni. Web document clustering: a feasibility demonstration. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '98, pages 46–54, New York, NY, USA, 1998. ACM.

 ISBN: 978-1-61208-195-3

Healthcare Multimedia Application for Multi-modal Mobile Device Interaction

Penhaker Marek, Kijonka Jan
Department of Cybernetics and Biomedical Engineering
VSB – Technical University of Ostrava, FEI, K450
Ostrava, Czech Republic
marek.penhaker@vsb.cz jan.kijonka@vsb.cz

***Abstract*—This study presents the healthcare multimedia application for multimodal device interaction based on mobile cell phone styles. The application provides visualization and user interaction of health care status information. The data are processed from a user interface and from sensors embedded in mobile devices and surrounding personal health care devices. The multimedia application is ready to use for personal health care and also for elderly people self monitoring.**

Keywords-mobile; multimedia; application; interaction

I. INTRODUCTION

A human can be monitored by user adaptive visualization system in home ambient. The information of measured data could be provided to the patient by several ways like visualization on a TV screen. The user would select the kind of vital function data to visualize on the TV screen by wireless driver. The disadvantages are worse portability of TV, necessity of creating a special data channel for connecting on TV video input, special wireless TV driver designing.

The other visualization possibility is on a PC monitor where the user would examine the measured data with a desktop computer or on a notebook screen. The disadvantages are the required computer and the cumbersome PC. The advantages are large screen area and simple handling of PC application by mouse and keyboard. Data would be transmitted by WiFi.

Visualization on PDA is the advantages are compactness and device mobility. It would be necessary to make an application for PDA. The data would be transmitted by WiFi or BlueTooth. Disadvantage is battery consumption and handling the smart phone by elderly people. [9]

Visualization by use of a mobile phone is the new measured values would be transmitted by SMS. The disadvantage is the reduced visualization option.

The best one way is developing a special mobile device (mobile unit for data measuring and visualization) with a user interface (LCD and keyboard) which is suitable to the visualization requirements. The device allows wireless transmission of vital functions data. The advantage is great stability in comparison to PC or PDA as the application doesn't work on any operation system. The device is useful for emergency calls (emergency medical service calling), for example after the given keyboard button push. In the following chapters we describe the mobile unit and user communication interface. [10]

II. MOBILE UNIT

The mobile unit is a device, which can monitor human basic vital functions and provide communication with surroundings by wireless nets.

The mobile unit integrates all the measuring and visualization functions to one compact mobile device, which is capable of ZigBee wireless communication. The device is embedded to a case of mobile phone size. The LCD, keyboard and Power button are used to create the user interface. The output connector is used for measuring and programming wires connection and for the battery charger connection. The device is powered by Li-Ion 3.3V battery. All components of the device are designed for this voltage.

A. Block Diagram

The block diagram of the mobile unit is in Fig. 1. There is symbolized its inner structure, function of individual blocks and communication between them. An individual description is given to the blocks of the user communication interface.

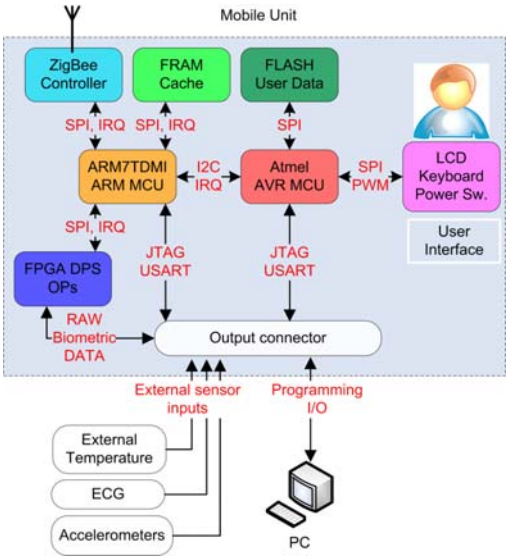


Figure 1. The mobile unit block diagram

B. Mobile Unit Control

The control microcontroller of the mobile unit is an MCU based on the ARM architecture. It transmits all the visualization data to the user interface control unit MCU based on the AVR architecture. The user interface control unit processes the received data.

The communication between ARM and AVR MCUs proceeds through the I2C serial interface. The I2C protocol is described in chapter I2C Communication Protocol. The components of the communication protocol include also an interrupted line allowing for the duplex communication.

C. User Interface Control

The ATmega644P is a low power CMOS 8-bit microcontroller based on the AVR enhanced RISC architecture. It contains enough capacity of Flash program memory (64K Bytes). The clock frequency is set to 12MHz, maximal for 3.3V MCU operating voltage. The power consumption is 7mA in active mode and 0.2µA in power-down mode (used during the device down). Used peripheries are PWM (LED Backlight intensity adjustment), timers/ counters, I2C, USART, JTG and SPI interfaces. [1]

The AVR MCU allows for communication with other components of the user communication interface (LCD, keyboard, external Flash memory) and external devices (MCU ARM, PC). The AVR MCU contains the main program in Flash program memory. The program allows for receiving vital function data,, storing and reading from an external data Flash memory, communication with the LCD driver, keyboard keys status reading and communication with PC (MCU programming, data Flash uploading).

D. User Data Flash

The AT45DB081D with serial SPI interface is used as the external Flash memory. Its 8Mb of memory are organized as 4096 pages of 256 or 264 bytes each. It allows intelligent memory operations, page programming and flexible erase options (page, block, sector, and chip erase). 100 000 program/erase cycles per page are guaranteed. [3]

Content of external user data Flash:

- Video data (icons, user menu panels, battery and signal strength states, the map of monitoring area) is stored in RGB format suitable for LCD writing, or compressed (RLE, JPEG).
- Fonts (3 font styles 6x8, 8x8 and 8x16 pixels)
- Trend data (temperature, pressure, oxygen saturation, weight, and position)
- Settings (back light, curves' speed rate, actual indexes of trends data arrays)

E. LCD

The LCD has relatively the simple 9-bit serial interface, running on 6.6MHz maximum frequency. It uses an Epson S1D15G10 or a Phillips PCF8833 controller driver, which allows driving and data imaging on LCD. LCD driving with the Phillips PCF8833 controller and driver is described in chapter LCD Driving. LCD has powerful white LED backlight - 7V @ 40-50mA (very bright). [5]

It's necessary to use the DC-DC step-up converter for LED backlight power supply. The used step-up converter MC34063 is a switching regulator. Back light intensity is regulated by PWM signal generating. [7]

F. Keyboard

It is used several push buttons of matrix keyboard for user menu controlling. The user menu is described in chapter User Menu.

G. Output Connector Programming I/O

MCU AVR and external Flash memory programming is possible by the USART and JTAG interfaces. USART and JTAG are connected to the output mobile unit connector.

There was the MATLAB GUI used to create an application, which enables uploading the external user data Flash. This application is described in the chapter Interactive User Data Flash Uploading. The mobile unit is connected with PC through the RS232 serial interface (COM port PC). It is possible to use a chip based on MAX232 for RS232 to USART conversion.

The AVR JTAGICE mkII programmer serves for programming the AVR MCU. It supports all types of AVR MCUs. For programming and debugging the AVR MCU we can use the AVR Studio development platform.

III. ADAPTABILITY USER INTERFACE

The mobile unit for measuring and measured data visualization is a compact device, which is able to be configured by the user to present well-arranged results. The visualization system consists of several construct parts described below.

A. User Interface Parts

- LCD
- Matrix keyboard
- Power button
- User interface control unit and external user data Flash
- Output connector programming I/O

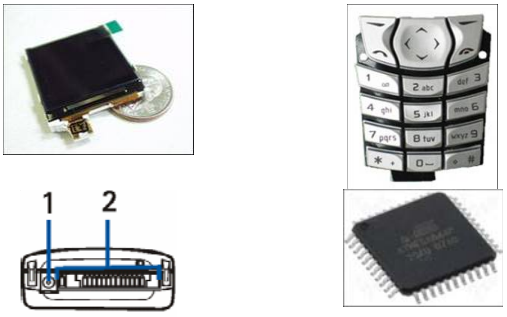


Figure 2. User interface parts a) LCD b) matrix keyboard c) output connector d) MCU and Flash

B. User Interface LCD Imaging

For the user of a mobile unit, the user menu displayed on LCD was created.. Moving and options in the user menu are

achieved by several keys of the keyboard. The user has the possibility to display vital functions and additional data. He can also execute an emergency call, upload user data Flash, or launch the demonstration mode of visualization. Additional options are settings of the parameters as waveform sweep speed, LCD back light intensity.

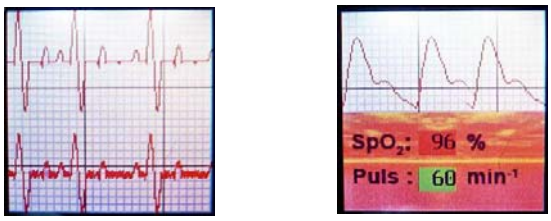


Figure 3. Actual ECG and pletysmography record demonstration on LCD a) ECG channels 1 and 2 b) pletysmography

Actual waveform measurement imaging ECG (channels 1 and 2) and pletysmography (including SpO2 and pulse values). SpO2 values are stored in the user data Flash memory to examine it in the saturation trend.

Trends imaging – the body and ambient temperatures, systolic and diastolic pressure, a pulse, SpO2, weight, the position (the room ID). It is possible to set the common or separate scales for the body temperature and ambient trend, on/off line with particular value displaying, on/off minimum and maximum value of graph displaying.

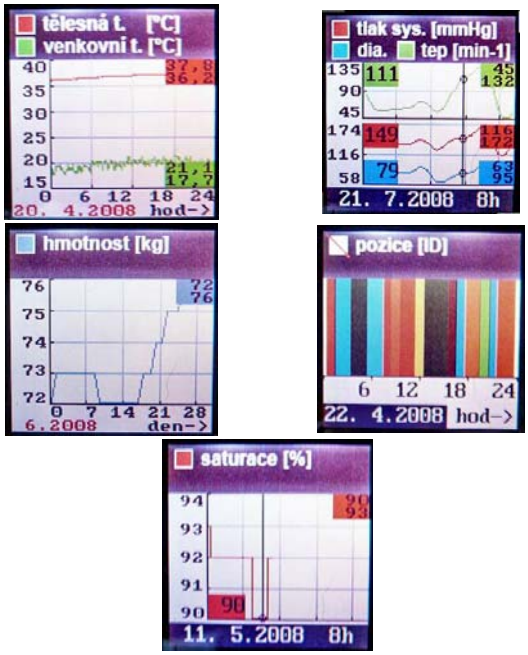


Figure 4. Trends imaging on LCD screen demonstration a) body and ambient temperature curves b) pulse, systolic and diastolic pressure curves c) saturation curve d) body weight curve e) position

There is also other information as actual position displaying, emergency calling, event displaying, DEMO mode, user data Flash uploading, settings - waveform sweep speed, LCD back light intensity.

IV. LCD DRIVING

The used LCD type is driven by Phillips PCF8833 controller driver. The PCF8833 is a single chip low power CMOS LCD controller driver, designed to drive color Super-Twisted Nematic (STN) displays of 132 rows and 132 RGB columns. All necessary functions for the display are provided in a single chip, including RAM which has a capacity of 209 kbit (132 x 12-bit x 132). The PCF8833 offers the 3-line serial interface.

The PCF8833 has 2 access types; the one defining the operating mode of the device (instruction) and the other filling the display RAM (data). Efficient data transfer is achieved by auto-incrementing the RAM address pointers.

A. Write mode – the serial interface

SPI 9-bit communication. Each data packet contains a control bit D/C and a transmission byte (instruction/data). The logic value of control bit D/C interprets the following byte as instruction (D/C is logic 0) or data (D/C is logic 1).

B. Instructions

There are 3 types of instructions; the one defining the display configuration (data format, color inversion, partial mode, rolling scroll mode, etc.), the one setting X and Y addresses, and miscellaneous data.

Different display data formats are available because different color depths are supported by PCF8833. The color depths supported are: 4 Kbyte colors (12-bit/pixel), RGB 4 : 4 : 4 bits input (4 bits for red, 4 bits for green and 4 bits for blue color resolution). The data coming from the interface is directly stored in RAM, or with better color depth, for maximal use affording a realistic imaging, as flat scene imaging it supports 65 Kbyte colors (16-bit/pixel), RGB 5 : 6 : 5 bits input. The 16-bit data coming from the interface is mapped by means of dithering to 12-bit data. Then, the dithered 12-bit data is stored in the RAM.

V. I2C COMMUNICATION PROTOCOL

For communication between ARM and AVR MCUs we define the I2C communication protocol.

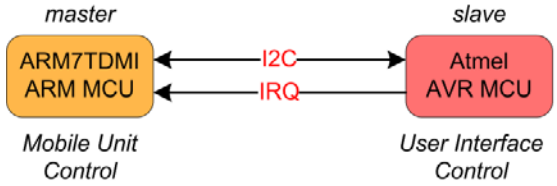


Figure 5. I2C communication protocol block diagram

Mobile Unit Control MCU – it transmits a continual data stream, which contains a vital measured data function, actual time, and actual events' data.

There is defined sampling frequency for ECG, pletysmography, temperature and time; a single-shot value for weight, pressure, position, and events' data.

User Interface Control MCU – it receives a continual data stream; transmits the events' data (emergency calling, power down request). Transmitter mode is activated by the interrupted line.

 ISBN: 978-1-61208-195-3

Data stream – the communication protocol consists of 1 byte opcode, which interprets following data type, or it defines some event in itself. The opcode is followed by several data bytes or by next opcode.

VI. PROGRAMMING AND THE APPLICATION DEVELOPMENT

For programming the microcontroller ATmega644P, we can use the AVR STK500 Flash Microcontroller Starter Kit. The AVR STK500 presents a complete tool kit for the Atmel AVR microcontrollers. It allows development, debugging and testing prototypes applications. [8]

The application was programmed in AVR Studio 4 with AVR GCC programming language. The compiler is included in AVR Studio.

A. *Libraries and Functions*

The program is built-up of considerable functions' libraries, which are necessary to communication with LCD controller driver, primitive graphic objects, text and numeric values rendering, image data rendering (images, flat maps of monitored area), vital functions data receiving and the user interface. The program configures and uses microcontroller peripheries as I/O, SPI, USART, I2C, TC0, TC2, etc.

The video data are stored in the user Flash memory. We access the data by memory page address listed in the fat.h library.

The draw.h library contains the functions for graphic objects on LCD imaging (pixel, line, rectangle, circle, text, and number).

The functions for Flash memory handling, reading image data and displaying images on LCD are in the flash_lcd.h library.

The functions for time handling, trends, ECG and pletysmography waves imaging and user interface options are stored in particulars libraries time.h, trends.h, waves.h, TIMER0_okna.h.

In Figure 6, there is an image stored in the Flash memory with illustrated coordinates (x0, y0) and (x1, y1), which are stored in initialization bytes of data file. Also, this illustrates the width and height of an image in pixels. The part of the image actually displayed on the LCD screen is marked in the red frame with the width x1_LCD0 - x0_LCD0 pixels and the height y1_LCD0 - y0_LCD0 pixels. The parameters x0_LCD0, x1_LCD0, y0_LCD0 and y1_LCD0 determine the part of the LCD display screen used for image displaying.

The appropriate part of the image is displayed on LCD according to the received position coordinate (x_pozice, y_pozice). (x_pozice, y_pozice) is actual position of patient in monitored area. x_pozice has range <x0, x1>, and x_pozice has range <y0, y1>. By mapa_bod argument selection we can on/off displaying of the point of position coordinate. [10]

The appropriate function for image data LCD displaying is executed by the compression type of the stored image data.

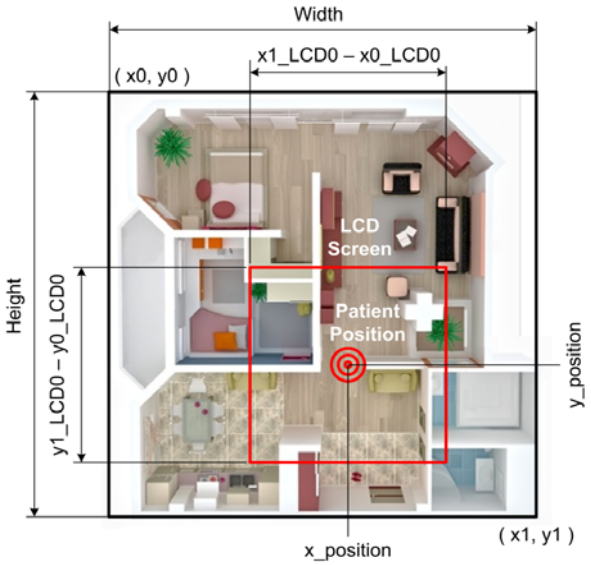


Figure 6. Map of monitored area and image parameters stored in flash memory

VII. PROGRAM FUNCTION DESCRIPTION

The program starts to execute in the int main (void) function, just through the AVR MCU connected to the battery voltage. In the int main (void) function, there are executed the following tasks: the Watchdog reset, the power push button configuration as interruptible input, enable all interrupts, and the power consumption reduction (the ADC shutdown). Finally, the AVR MCU is switched to the power-down mode for the minimal power consumption. From this power mode, the AVR can wake-up only by external interrupt (the power push button). At the end of the int main (void) function, there is the infinite loop placed while (1); to continue the program after MCU wake-up.

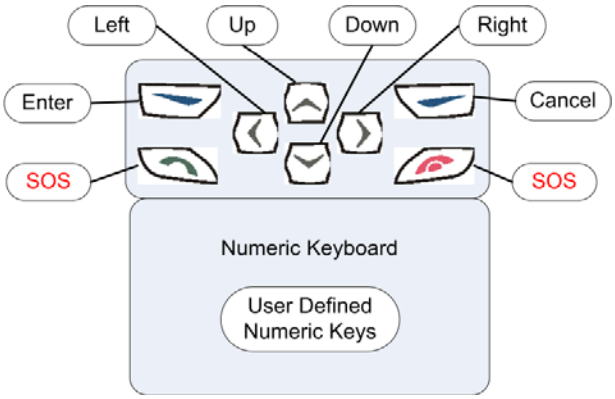


Figure 7. Keyboard use

After the device is switched-on, the void INIT() function is executed. After the device initialization the push-buttons status is periodically read in timer 0 interrupt. There are particular functions executed in the interrupt code of the user

 ISBN: 978-1-61208-195-3

interface menu. The I2C interface is also active for communication between ARM and AVR MCUs.

VIII. User Menu

The user can move in the user menu displayed on a LCD screen by several push buttons.

A. Main Menu

After switching on the mobile device, a logo is displayed on LCD (figure 8).

The ZigBee signal strength status (Figure 9), battery status (Figure 10), and the actual time formatted to HH:MM are displayed in the upper part of the LCD screen. These data are called as "Initiative Line". Below this line this is space for one line message called as "Event Line". It is the last event concerning data receiving: a new measured pressure value begins next and the end of EKG or plethysmography measuring, ALARM – smoke in the room. Below the "Event Line" there is displayed the main menu (Figure 11) namely one of the mentioned option (a, b, c, and d). Moving in the main menu is realized by arrow buttons "Left" and "Right". The selection is approved by the key "Enter".

The similar style of moving in the user menu is in other menus: waves, trends and settings. [13]

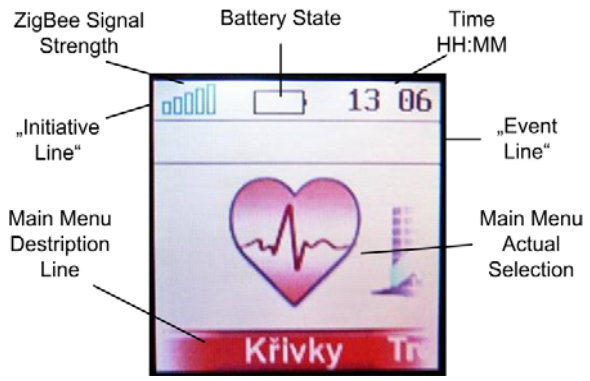


Figure 8. The main menu window

Figure 9. The ZigBee signal strength status

Figure 10. The battery statuses

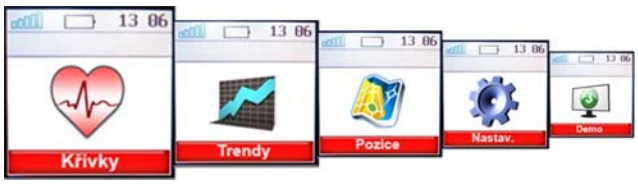


Figure 11. Main menu

IX. Interactive User Data Flash Uploading

There was MATLAB GUI used to create an application, which enables uploading the external user data Flash. You can see the main window of the application in the Figure 12. We use a serial COM port PC for programming. An image with *.bmp, *.ico, *.jpeg, *.png, *.tiff extensions, a data file and others can be stored into the Flash memory.

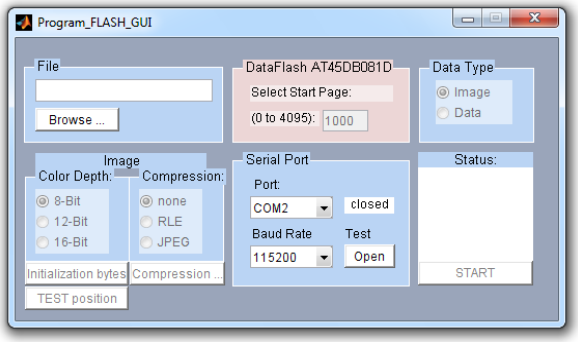


Figure 12. Program_FLASH_GUI

Images must be stored into the Flash memory in a format, which is suitable for displaying on the mobile unit LCD screen. The LCD – PCF8833 driver supports 3 data formats (for 8-bit, 12-bit or 16-bit colors) as stated it the chapter LCD Driving. It is also convenient to compress images in order for them to take as least space as possible in the Flash memory with 8Mbit capacity. [12]

The data file of each image contains 24 initializing bytes in its beginning. We cannot change the first 6 bytes as those contain necessary information about the image (the width and height of the picture in pixels, the compression kind and colors quality). It is not possible to identify an image without this information, to use an appropriate decompression function, to set the format and so on. The byte 7 – 10 serve for entering the scale of the image. These bytes must be entered if the image is a map. Other initializing bytes are not used in this application. They can be configured additionally.

X. GUI Description

The main window of the user interface for programming the Flash memory is shown in the Figure 12.

After clicking on the "Browse" button you can choose the required file containing a picture or data, which is to be loaded into the Flash memory.

You can choose the first memory page, which will be used for storing data, in the "DataFlash AT45DB081D" item in the GUI. If you select a page, which already contains data, these will be lost by loading new data. It is possible to load an image in a selected format or data like types of a font or trends. The loaded are saved into the Flash exactly the same way as it is saved in the initial file.

A picture can be compressed into JPEG or RLE (Run Length Encoding), which is a compression with no losses. It is also possible to select a quality of colors. When loading images you can set the initializing bytes after clicking on the dedicated button in the GUI.

 ISBN: 978-1-61208-195-3

The data is sent through a serial COM port. It is necessary to set a free port and a data rate of transition. The data rate of transition is preset to 15200 Bd/s. The “Open” button checks the availability of the selected port. [13]

When all the parameters are chosen, the data transfer from PC to Flash memory can begin by clicking on the “START” button on the right bottom corner of the GUI. The “Status” window displays the data file size [byte], the page end of the memory, and additional information. When the data type “Image” is chosen, the original and compressed image is displayed in new MATLAB figures.

After that runs the time 10s, during that the mobile unit should be connected. The item “Flash” should be selected in the mobile unit’s user menu. The data transfer begins after pressing the key “Enter”. After the transfer termination, the status message is displayed both on mobile LCD unit and on the PC screen.

XI. CONCLUSIONS

The objective was to create a multimedia user interface for the mobile unit in the form of applications. The generated application works on previously realized hardware. Programming and STK500 communication ports are connected to COM ports on PCs. AVR μC is programmed in the AVR Studio 4. The communication port is used to program the external FLASH memory of the MATLAB GUI. With socket for external development, the board is managing AVR μC connected to an external flash memory, LCD and keypad Nokia. The external field is also involved DC converter generating DC voltage suitable for backlighting LCD and keypad.

Simulation of receiving and sending vital data functions are performed in two ways. The first method is implemented by running the DEMO mode in the user menu. The LCD displays are fictitious measured data. In this case, however, the function of the communication interface is not verified. The second way is by the communication protocol. For this purpose μC AVR Atmega168, which simulates the communication between AVR and ARM AVR Atmega168 is programmed to broadcast a fictitious measured data and that all options according to the protocol. It also has activated external interruption simulating the ARM interrupt. Thus, the functional and messages such as "Off" and "Request for Assistance".

The work was for me a great benefit in terms of programming the AVR μC. It has been used many AVR function and all communication interfaces that offer AVR. Also, capacity SRAM and programming FLASH memory have been largely used. Due to the scale of the program was necessary to abandon the original type AVR Atmega168 and go to ATmega644P with 64kB flash memory / 4 KB SRAM.

Another direction of development and improving the user interface of the communication could be to the user menu of new features and use the whole keyboard for entering numbers of parameters. To view the trends would be appropriate to move the LCD larger. This can select a more readable font size for most users.

ACKNOWLEDGMENT

The work and the contributions were supported by the project SV SP 2012/114 “Biomedical engineering systems VIII” and TACR TA01010632 “SCADA system for control and measurement of process in real time”. Also supported by project MSM6198910027 Consuming Computer Simulation and Optimization. The paper has been elaborated in the framework of the IT4Innovations Centre of Excellence project, reg. no. CZ.1.05/1.1.00/02.0070 supported by Operational Programme 'Research and Development for Innovations' funded by Structural Funds of the European Union and state budget of the Czech Republic.

REFERENCES

[1] ATmega164P/324P/644P [online]. Last revision on the 1st of February 2009 [cit. 2009-04-20].

[2] http://atmel.com/dyn/resources/prod_documents/doc8011.pdf

[3] AT45DB081D [online]. last revision 1st of February 2009 [cit. 2009-04-20].

[4] <http://www.atmel.com/dyn/resources/prod_documents/doc3596.pdf

[5] Color LCD 128x128 Nokia Knock-Off [online]. last revision 5th of April 2009 [cit. 2009-04-05]. http://www.sparkfun.com/commerce/product_info.php?products_id=569

[6] PCF8833 [online]. last revision 20th of April 2009 [cit. 2009-04-20]. www.nxp.com/acrobat_download/datasheets/PCF8833_1.pdf

[7] MC34063 [online]. last revision 20th of April 2009 [cit. 2009-04-20]. http://www.datasheetcatalog.com/datasheets_pdf/M/C/3/4/MC34063.shtml

[8] STK500 User Guide. [online]. last revision 1st of March 2003 [cit. 2009-04-10] http://www.atmel.com/dyn/resources/prod_documents/doc1925.pdf

[9] Dvorak, J., Havlik, J., Data Synchronization for Independent USB Devices. In 2011 International Conference on Applied Electronics. Plzeň: Západočeská univerzita v Plzni, 2011, p. 111-113. ISBN 978-80-7043-987-6.

[10] O. Krejcar, D. Janckulik, L. Motalova, and J. Kufel, “Mobile Monitoring Stations and Web Visualization of Biotelemetric System - Guardian II”. Proc. EuropeComm 2009. LNICST vol. 16, pp. 284-291. R. Mehmood, et al. (Eds). Springer, Heidelberg (2009).

[11] Krajcuskova Z., Kukucka M.: Time-Frequency Representation and Unconventional Reliability Growth Model, In TSP 2011 : Proceedings of 34th International Conference on Telecommunications and Signal Processing, August 18th-20th, Budapest, Hungary, 2011, s. 518-520, ISBN 978-1-14577-1409-2

[12] Kasik, V.: Acceleration of Backtracking Algorithm with FPGA. In 2010 International Conference on Applied Electronics, pp. 149-152, Pilsen, Czech Republic, 2010, ISBN 978-80-7043-865-7, ISSN 1803-7232.

[13] Kijonka, J., Penhaker,M., Kasik,V., Stankus, M. User Adaptive System for Data Management in Home Care Maintenance Systems, In Proceedings of 3rd Asian Conference on Intelligent Information and Database Systems, Springer, Lecture Notes in Computer Science, 2011, pp. 492-501,20. - 22. April, 2011Volume 6592/2011, DOI: 10.1007/978-3-642-20042-7_50

 ISBN: 978-1-61208-195-3

Distribute the Video Frame Pixels over the Streaming Video Sequence as Sub-Frames

Hussein Muzahim Aziz
School of Computing
Blekinge Institute of Technology,
371 79 Karlskrona, Sweden
+46455385876
hussein.aziz@bth.se

Markus Fiedler
School of Computing
Blekinge Institute of Technology,
371 79 Karlskrona, Sweden
+46455385653
markus.fiedler@bth.se

Håkan Grahn
School of Computing
Blekinge Institute of Technology,
371 79 Karlskrona, Sweden
+46455385804
hakan.grahn@bth.se

Lars Lundberg
School of Computing
Blekinge Institute of Technology,
371 79 Karlskrona, Sweden
+46455385833
lars.lundberg@bth.se

***Abstract*—Real-time video streaming over wireless channel has become an important issue due to the limited bandwidth that is unable to handle the flow of information of the video frames. The characteristics of wireless networks in terms of the available bandwidth, frame delay, and frame losses cannot be known in advance. As the effect of that, the user may notice a frozen picture in the mobile screen. In this work, we propose a technique to prevent freezing frames in the mobile devices based on spatial and temporal locality for the video stream, by splitting the video frame into four sub-frames and combining them with another sub-frames from different sequence positions in the streaming video. In case of frames losses, there is still a possibility that one fourth (one sub-frame) of the frame will be received by the mobile device. The received sub-frames will be reconstructed based on the surrounding pixels. The rate adaptation mechanism will be also highlighted in this work, by skipping sub-frames from the video frames. We show that the server can skip up to 75% of the frame's pixels and the receiving pixels (sub-frames) can be reconstructed to acceptable quality in the mobile device.**

Keywords-streaming video; wireless network; frame splitting; sub-frame crossing; rate adaptation.

I. INTRODUCTION

Nowadays mobile cellular networks provide different type of services and freedoms to the mobile users anywhere and at any time while the mobile users on the move. Streaming services become an important application to the mobile user, while streaming video is the classical technique for achieving smooth playback of video directly over the network without downloading the entire file before playing the video [1][5][14].

The unpredictable nature of wireless networks in terms of bandwidth, and loss variation, remains one of the most significant challenges in video communications [9]. In this context, video streaming needs to implement an adaptive techniques in terms of transmission rates in order to cope with the erroneous and time variant conditions of the wireless network [9][10].

Bandwidth is one of the most critical resources in wireless networks, and thus, the available bandwidth of wireless networks should be managed in an efficient manner [7]. Therefore, the transmission rate of the streaming video should be maintained according to the networks bandwidth [2][6][11].

Network adaptation refers to how many network resources (e.g., bandwidth) a video stream should utilize for video content, resulting in designing an adaptive streaming mechanism for video transmission [15]. To stream video, it is desirable to adjust the transmission rate according to the perceived congestion level in wireless networks, to maintain the suitable loss level and fairly shared bandwidth with other connections. Furthermore, it is favorable for the streaming video to be aware of the transmission level in order to obtain good streaming quality by appropriate error protection.

In this paper, we proposed a sub-frame crossing technique based on frames splitting. The video frame will be split into four sub-frames, and combine the sub-frame with another sub-frame from different sequence position and from different spatial data in the streaming video. The crossing frames in the streaming video will carry pixels from four different frames that belong to four different positions and will transmit over a single wireless channel. In case of sequence of frames losses or frames corruption from the streaming video, the losses of the sub-frames will be distributed on the streaming video and there is still a possibility that one of the fourth sub-frames will be received by the mobile device, while the missing sub-frames from the frames will be reconstructed based on the surrounding pixels.

The remainder of this paper is organised as follows. Section II provides background and related work. Section III explain the proposed of streaming the video as sub-frames crossing. The rate adaption mechanism is presented in Section IV. The results are discussed in Section V, while the conclusion is presented in Section VI.

II. BACKGROUND AND RELATED WORK

Various techniques are proposed by many researchers for video frame slicing and reconstruction. The proposed techniques are based on H.264/AVC standard tools[20], where the Flexible Macroblock Ordering (FMO) slicing type dispersed to split the video frames and streaming them over the networks, while adaptive the slices is needed to send the highest priority information.

Huang [13] proposed a scheme for Adaptive Region of Interest (AROI) extraction and adaptation by integrating the visual attention model in the human visual system. The scheme are applied to the Region of Interest (ROI) based on video coding for adaptation and delivery, by embedding the anchor point of focusing Macroblock (MB) in each key frame and motion vectors in other frames in the coded video stream or the sequence parameter set in the Scalable Video Coding (SVC). The error resilience tool FMO can be used to define certain of ROI in SVC, while the slice groups can be used to constitute a number of columns covering the frame by some elaborated tiled partitions in order to meet the mobile terminals with different resolutions.

Wang and Tu [16] introduce an adapter FMO type, which classifies the MBs into important and unimportant slices. The important slice involves the details of the frames which represent the important contents. The complexity of MB content and texture change which are used to judge the importance of the MB. The unimportant MBs are divided into two slices based on edge match rule, which contributes to the error concealment in the decoder. The important slice is protected than the unimportant slice in the receiver so that the subjective quality of the reconstruction frame will be improved greatly. The proposed of adapter FMO scheme is to increase the error resilience of the encoded video stream and contribute to the error concealment realization in the decoder. The adapter FMO strategy is suitable technique for the video transmission over low bandwidth.

Aziz et al. [3] present a technique to overcome the freezing frames problem on the mobile device and providing a smooth video playback over a wireless network. The frames in the streaming video will be splitted into four sub-frames on the server side and transmitted over Multiple-Input Multiple-Output (MIMO) by using the Multiple Descriptions Coding (MDC) technique. Where an initial delay time had been set between different channels to avoid the interruption on the sub-frames that are belong to the same frame. In case of the sub-frames that belonging to any subsequence are lost during the transmission, a reconstruction mechanism will be applied in the mobile device to recreate the missing pixels that are belongs to the missing sub-frames based on the average of the neighbouring pixels.

To overcome the transmission of each frame over MIMO and to increases the ability to handle long losses during the transmission over unreliable network. A splitting technique is proposed to deal with the sub-frames as equally important, by splitting the frames into sub-frames and cross them with another sub-frame from different sequence position.

The initial idea is been proposed in [4], where the frames been splitted into two sub-frames, where one sub-frame contains the even pixels and another contains the odd pixels. The combination of the sub-frame with another sub-frame from different sequences positions within the same transmission rate. The combined sub-frames will be streamed over a single wireless channel. In case of the frame being lost the available sub-frame in the mobile device will be reconstruct based on the surrounding pixels, while the maximum frames sequence lost that can be tolerated is half second.

The work has been extended to tolerate a maximum frame sequence lost up to six seconds (in worst cases), while the adaption mechanism allow us to stream up to one fourth (skipping three sub-frames) of the video frames to the mobile device according to the proposed technique. The reconstruction to the sub-frames in the video sequence will be measured by the Structural Similarity (SSIM) index.

III. THE PROPOSED TECHNIQUE

Mobile video streaming is characterized by low resolutions and low bit rates. The bit rates are limited by the capacity of UMTS radio bearer and the restricted processing power of the mobile terminals. The commonly used resolution is Quarter Common Intermediate Format (QCIF, 176 x 144 pixels) for mobile phones [8].

Mobile real time applications like video streaming suffer from high loss rates over the wireless networks [12] and the effect of that the mobile users may notice some sudden stop during the playing video, the picture is momentarily frozen. The frozen pictures could occur if a sequence of video frames is lost.

Distribute the frame's pixels as sub-frames over the streaming video is considered in this work by splitting each frame into four sub-frames [3], where each sub-frame contains one fourth of the main frame pixels, as shown in Figure 1. The crossing technique will be applied after splitting the frames as sub-frames and it will be crossed with other sub-frames that are from different frame sequence position.

During the interactive mode where the mobile clients request the connection to the video server, the server will start streaming the frames based on the frames splitting and frames crossing technique, as shown in Figure 2, 3 and 4, respectively.

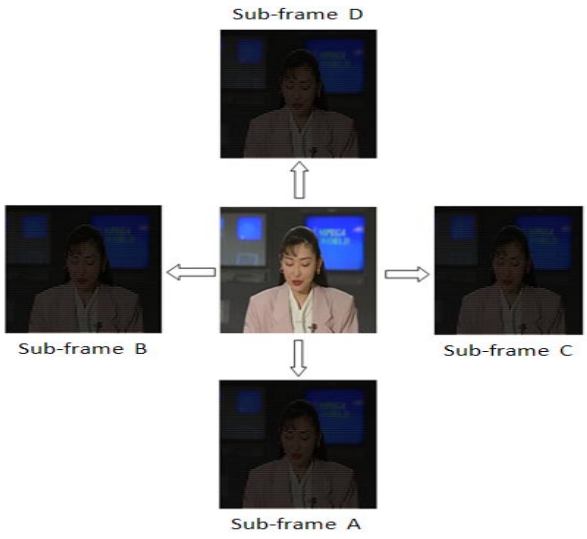


Figure 1. Snapshot of Akiyo frame splitting.

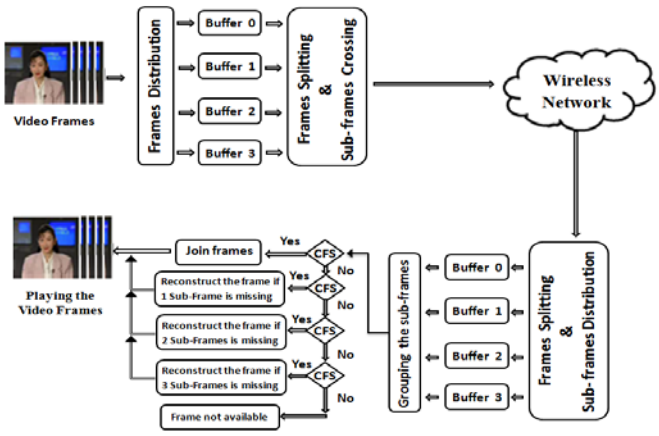


Figure 2. Streaming video as sub-frame crossing over wireless network.

Each video frame is splitted into **s** sub-frames, where **s = 0,…S-1**, where **s** is four sub-frames (**A**, **B**, **C**, **D**), as shown in Figures 1 and 3(a), respectively.

Each sub-frame contains different pixels information which make it possible to implement the frames crossing technique among the frames groups to created the new frames crossing (**FC**).

The sequence of the video frames will be grouped in the streaming server according to the transmission rate per second as a frames group (**FG**), as shown in Figure 3(a), where **g** is the index of the frames group, **g = 0,…, G-1**.

To implement the frames crossing technique between different frames in different group where **i** is the index of the frames group crossing (**FGC**) where **i = 0,…, S-1**, where the sub-frames **s** of group **g** of the **FGC i** is obtained as

$$FGC_{i}(g,s) = sF_{(G.((s+i) \bmod S)+g,\, s)}, \qquad (1)$$

and are illustrated in Figures 3 and 4, respectively.

Crossing the sub-frames among the frames groups is required **s** buffers to queue the **FG**s, where the buffer size is equal to the frames rate, as shown in Figure 2. As an example, the first frames group **FG0** will be queued in buffer 0, and the second **FG1** will be queued in buffer 1, the third **FG2** will queued in buffer 2, and the fourth **FG3** will be queued in buffer 3. During the process of each buffer the next arrival group of frames, which is the fifth **FG4**, will be queued in buffer 0 and so on.

The transmission rate are considered in this work is 30 frames per second, where the frames group (**FG**) size will be 30 frames, during the arrival of the streaming video; the first 30 frames (**FG0**) will be splitted into four sub-frames, as shown in Figure 1 and 3(a). The same technique will be applied to the arrival of the second 30 frames (**FG1**) and so on.

When the first frame from the fourth group (**FG3**) of 30 frames arrived, the frames will be splitted into four sub-frames and the crossing technique will be applied immediately to distribute the frames pixels among the four groups in the streaming video, as shown in Figure 3.

		A	B	C	D
FG0	0	$sF_{0,0}$	$sF_{0,1}$	$sF_{0,2}$	$sF_{0,3}$
	1	$sF_{1,0}$	$sF_{1,1}$	$sF_{1,2}$	$sF_{1,3}$
	:	:	:	:	:
	:	:	:	:	:
	g	$sF_{g,0}$	$sF_{g,1}$	$sF_{g,2}$	$sF_{g,3}$
FG1	g+1	$sF_{g+1,0}$	$sF_{g+1,1}$	$sF_{g+1,2}$	$sF_{g+1,3}$
	g+2	$sF_{g+2,0}$	$sF_{g+2,1}$	$sF_{g+2,2}$	$sF_{g+2,3}$
	:	:	:	:	:
	:	:	:	:	:
	2g	$sF_{2g,0}$	$sF_{2g,1}$	$sF_{2g,2}$	$sF_{2g,3}$
FG2	2g+1	$sF_{2g+1,0}$	$sF_{2g+1,1}$	$sF_{2g+1,2}$	$sF_{2g+1,3}$
	2g+2	$sF_{2g+2,0}$	$sF_{2g+2,1}$	$sF_{2g+2,2}$	$sF_{2g+2,3}$
	:	:	:	:	:
	:	:	:	:	:
	3g	$sF_{3g,0}$	$sF_{3g,1}$	$sF_{3g,2}$	$sF_{3g,3}$
FG3	3g+1	$sF_{3g+1,0}$	$sF_{3g+1,1}$	$sF_{3g+1,2}$	$sF_{3g+1,3}$
	3g+2	$sF_{3g+2,0}$	$sF_{3g+2,1}$	$sF_{3g+2,2}$	$sF_{3g+2,3}$
	:	:	:	:	:
	:	:	:	:	:
	4g	$sF_{4g,0}$	$sF_{4g,1}$	$sF_{4g,2}$	$sF_{4g,3}$
FG4	4g+1	$sF_{4g+1,0}$	$sF_{4g+1,1}$	$sF_{4g+1,2}$	$sF_{4g+1,3}$
	4g+2	$sF_{4g+2,0}$	$sF_{4g+2,1}$	$sF_{4g+2,2}$	$sF_{4g+2,3}$
	:	:	:	:	:
	:	:	:	:	:
	5g	$sF_{5g,0}$	$sF_{5g,1}$	$sF_{5g,2}$	$sF_{5g,3}$
FGg-1	5g+1	$sF_{5g+1,0}$	$sF_{5g+1,1}$	$sF_{5g+1,2}$	$sF_{5g+1,3}$
	5g+2	$sF_{5g+2,0}$	$sF_{5g+2,1}$	$sF_{5g+2,2}$	$sF_{5g+2,3}$
	:	:	:	:	:
	:	:	:	:	:
	g-1	$sF_{g-1,0}$	$sF_{g-1,1}$	$sF_{g-1,2}$	$sF_{g-1,3}$

a. The sub-frames that are related to the original frame sequence.

		A	B	C	D
FGC0	FC0	$sF_{0,0}$	$sF_{g+1,1}$	$sF_{2g+1,2}$	$sF_{3g+1,3}$
	FC1	$sF_{1,0}$	$sF_{g+2,1}$	$sF_{2g+2,2}$	$sF_{3g+2,3}$
	:	:	:	:	:
	:	:	:	:	:
	FCG-1	$sF_{G-1,0}$	$sF_{2G-1,1}$	$sF_{3G-1,2}$	$sF_{4G-1,3}$

b. The crossing frames position for FGC1.

		A	B	C	D
FGC1	FCg+0	$sF_{g+1,0}$	$sF_{2g+1,1}$	$sF_{3g+1,2}$	$sF_{0,3}$
	FCg+1	$sF_{g+2,0}$	$sF_{2g+2,1}$	$sF_{3g+2,2}$	$sF_{1,3}$
	:	:	:	:	:
	:	:	:	:	:
	FC2G-1	$sF_{2G-1,0}$	$sF_{3G-1,1}$	$sF_{4G-1,2}$	$sF_{G-1,3}$

c. The crossing frames position for FGC2.

		A	B	C	D
FGC2	FC2g+0	$sF_{2g+1,0}$	$sF_{3g+1,1}$	$sF_{0,2}$	$sF_{g+1,3}$
	FC2g+1	$sF_{2g+2,0}$	$sF_{3g+2,1}$	$sF_{1,2}$	$sF_{g+2,3}$
	:	:	:	:	:
	:	:	:	:	:
	FC3G-1	$sF_{3G-1,0}$	$sF_{4G-1,1}$	$sF_{G-1,2}$	$sF_{2G-1,3}$

d. The crossing frames position for FGC3.

		A	B	C	D
FGC3	FC3g+0	$sF_{3g+1,0}$	$sF_{0,1}$	$sF_{g+1,2}$	$sF_{2g+1,3}$
	FC3g+1	$sF_{3g+2,0}$	$sF_{1,1}$	$sF_{g+2,2}$	$sF_{2g+2,3}$
	:	:	:	:	:
	:	:	:	:	:
	FC4G-1	$sF_{4G-1,0}$	$sF_{G-1,1}$	$sF_{2G-1,2}$	$sF_{3G-1,3}$

e. The crossing frames position for FGC4.

Figure 3. The position of the sub-frames in the video sequence.

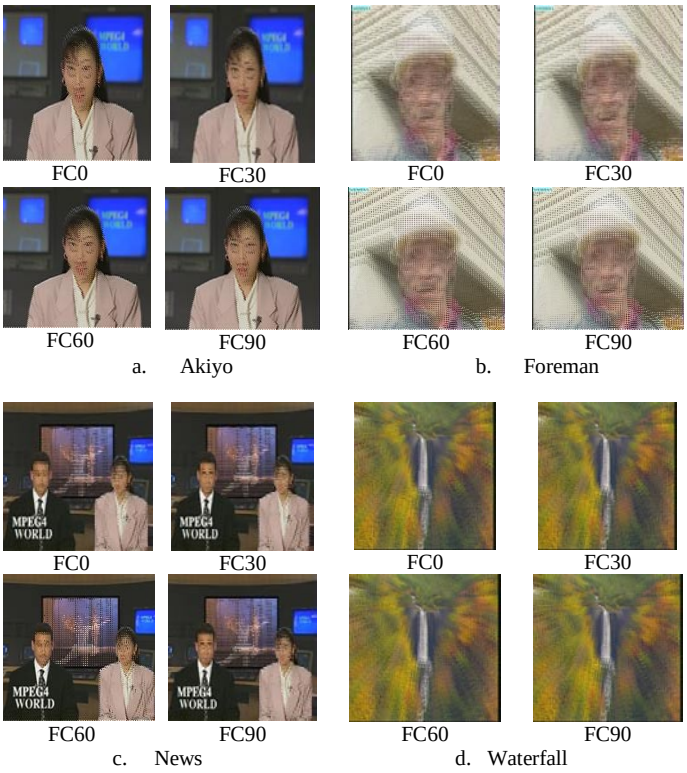


Figure 4. Snapshot for the Sub-frame crossing.

The crossing technique is implemented based on the frames crossing; where the frames crossing (**FC**) contains four different sub-frames from different **FG**s that belong to the same group. As an example, the first frame crossing **FC0** will contains the sub-frame **A** from frame number 0, sub-frame **B** from frame number 30, sub-frame **C** from frame number 60, and sub-frame **D** from frame number 90, while the second **FC1** will contains the sub-frame **A** from frame number 1, sub-frame **B** from the frame number 31, sub-frame **C** from frame number 61, and sub-frame **D** from frame number 91. In another way, the streaming video will be based on the sub-frames crossing and it will be transmitted as;

FC0(A0,B30,C60,D90),**FC1**(A1,B31,C61,D91),…,
FC30(A30,B60,C90,D120),**FC31**(A31,B61,C91,D1),…,
FC60(A60,B90,C120,D30),**FC61**(A61,B91,C1,D31),...,
FC90(A90,B120,C30,D60),**FC91**(A91,B1,C31,D61),…,
FC120(A120,B30,C60,D90),.., and so on, as shown in Figures 3 and 4 respectively.

The cost for implementing the proposed technique will be 3 seconds as an initial delay time, where the delay time is the time to queue **FG0**, **FG1**, **FG2**, for splitting and waiting for the fourth **FG3**, the time of the first frame from **FG3** arrived it will be split and combine them with another frames from **FG0**, **FG1**, **FG2** based on the proposed technique been described early. In this case we manage to distribute the frames pixels from different frame numbers and from different frames positions in the streaming video.

The crossing technique will be applied to all the frames in the video streaming sequence and it will be transmitted over a single channel. The reason behind that, if there is lost or dropped of sequence of frames from the streaming video and under different networks condition. The effect will be on at least one fourth of the sub-frames from the four different sub-frames that are in different positions.
The quality of the video will be affected and it will be distributed on the streaming video frames.

After each frame has been received by the mobile device, a splitting frame technique will be applied. The sub-frames will be held in different buffers and according to the order they been splitted at the server side, as shown in Figure 2. The sub-frames will be distributed to the relevant buffers and the combination of the sub-frames that are related to the same frame and according to their sequence positions based on switching between buffers to create the original frames sequences for the streaming video.

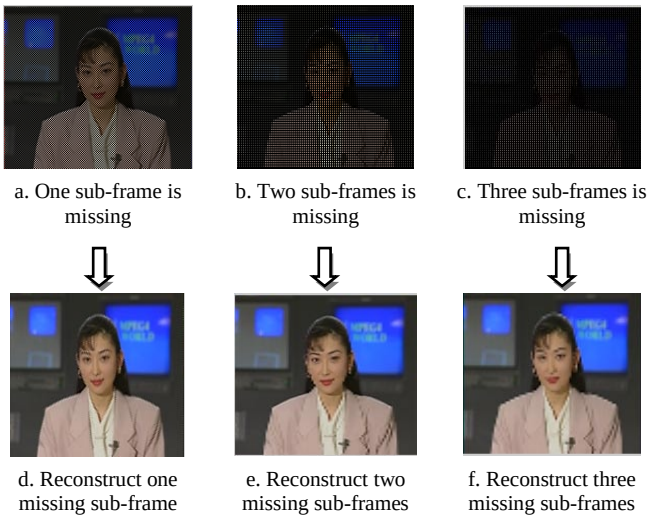


Figure 5. Akiyo snapshots of the missing and the reconstruction to the sub-frames.

The check frame sequence (CFS) procedures will take place in the mobile device, to check the availability of the sub-frames. The CFS and the reconstruction mechanism are used to identify the missing sub-frames and to reconstruct the missing pixels from the frames by considering the following checking procedures [3], and as shown in Figure 5;

- The first CFS will check whether all the sub-frames that are related to the same original frame are available. If the four related sub-frames are available, then a joining mechanism will be applied to return the frame to it is original shape.
- The second CFS will check if at least three sub-frames are available. If one sub-frame is missing, then the average of the neighbouring pixels will be calculated to replace the missing frame pixels.

- The third CFS will check if at least two sub-frames are available. If two sub-frames are missing, then the average of the neighbouring pixels will be calculated to replace the missing sub-frame pixels.
- The fourth CFS will check if at least one sub-frame is available. If three sub-frames are missing, then the average of the neighbouring pixels will be calculated twice, the first time to find the half of the frame and the second time to return the full frame to it is normal shape.

IV. RATE ADAPTION

Rate adaptation for streaming video is regarded as it necessary mechanism to handle the network conditions, and the fluctuations of the network bandwidth.

The adaption rate for the sub-frames crossing technique should be considered carefully to avoid skipping the sub-frames that belonging to the same frame and with the consideration of the available bandwidth and network interruption to the streaming video. The adaption rate can be implemented by not considering the combination of the four sub-frames and transmitting either three or two or one sub-frame to the mobile device and according to the following adjustments cases:

- 25% adjustment, the streaming server will skip only one sub-frame from the video frames sequence, as shown in Figure 5 (a).
- 50% adjustment, the streaming server will skip two sub-frames from the video frames sequence, as shown in Figure 5 (b).
- 75% adjustment, the streaming server will skip three sub-frames from the video frames sequence, as shown in Figure 5 (c).

The rate adaptation mechanism is needed to adjust the transmission rate based on the congestion level. The server will adjust the transmission rate by skipping the sub-frames that are not related to each other and the skipping rate limits shouldn't cross 75% from the frames pixels to avoid discard to the sub-frames that are related to the same video frame. The receiving sub-frames will be reconstruct to it is original frames, as shown in Figure 5.

V. RESULTS AND DISCUSSIONS

In the normal situation, when the streaming video is transmitted over a single channel, the mobile device will start receiving the video frames and it will be held in the buffers until the mount of frames rate are arrived to start playing the video. While real time video streaming suffers from high loss rates over wireless networks [17], the result of that, the users may notice a sudden stop during the video playing. The picture is momentarily frozen, followed by a jump from one scene to a totally different one.

The proposed technique is based on sub-frames crossing for the video test sequences Akiyo, Foreman, News, and Waterfall, as it is a well known professional test sequences [19], with a transmission rate of 30 frames per second.

The quality to the reconstructed sub-frames is expressed in terms of the Structural Similarity (SSIM) Index [18]. The SSIM index will measure the reconstructed video frames to the reference frames, as shown in Figures 6, 7, and 8 respectively.

Considering the same losing frame sequence in [3], where the number of frames are lost are 20 frames as light lost rate from the streaming video, then the affect of losing frames will be distributed on the streaming sequence and the affect will be on 80 frames, as these frames will loss one sub-frame. As an example, if the frame losses is started from frame 121 to 140, then the effect of losing one sub-frame will affect the frames sequence from 121 - 140, 151 - 170, 181 - 200, and from 211 - 230, as the losses of these frames are fall in the same crossing group. The frames that lost the sub-frame it will be reconstructed and therefore, the quality level of the frames will be affected.

If the numbers of frames are lost are 40 frames as medium lost rate from the streaming video, then the affect of losing frames will be distributed on the streaming sequence and the affect will be on 120 frames, as some frames will lose one sub-frame while others will lose two sub-frames. As an example, if the loss of frames starts from frame 121 to 160 then the effect of losing one sub-frame from 131-150, 161-180, 191-210, 221-240. While the following frames sequence will lose two sub-frames will affect the frames 121-130, 151-160, 181- 190, 211-220. Therefore, the quality level of the video will be distributed on the video sequence after been reconstructed as some video frames loss one sub-frame and others will loss two sub-frames.

If the numbers of frames are lost are 60 frames as high lost rate from the streaming video, then the effect of losing frames will be distributed on the streaming sequence and the effect will be on 120 frames. As an example, if the falls of frame losses are started from frame 121 to 180, then the affect of losing sub-frames will affect the frames from 121 to 240 as all the effected frames will loss two sub-frames. The receiving sub-frames will be reconstructed in the mobile devices to return the missing pixels for each frame and played in the mobile screen with less quality than the original frames.

The losses duration can be handle in this technique is up to six seconds, as shown in Figure 3. If the losses occur in the **FGC1**, **FGC2**, **FGC3**, **FGC4**, **FGC5**, and **FGC6**, the mobile device will received the following sequence of one sub-frame from 0 until 239, as these sub-frames are received by **FGC0** and **FGC7**.

The adaption rate is also considered in this paper, where the server can skip either one, or two, or three sub-frames, where the quality level of the video will be affected according to the adaption rate, as shown in Figure 8. Skipping three sub-frames shows low quality than skipping two or one sub-frame. The Waterfall video shows better results as the pixels of the video frames have similar data where the reconstruction mechanism didn't been effected that much, while the News video is been effected highly by the reconstructions mechanism as it is quite motion video and it can be seen clearly in Figure 6.

Original frame

The reconstruction for one sub-frame missing

SSIM : 0.955

Original frame

The reconstruction for one sub-frame missing

SSIM : 0.948

The reconstruction for two sub-frames missing

SSIM : 0.929

The reconstruction for three sub-frames missing

SSIM : 0.911

The reconstruction for two sub-frames missing

SSIM : 0.941

The reconstruction for three sub-frames missing

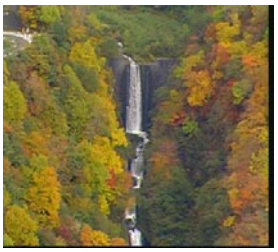
SSIM : 0.907

a. Akiyo b. Foreman

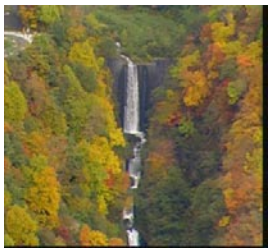
Original frame

The reconstruction for one sub-frame missing

SSIM : 0.939



Original frame



The reconstruction for one sub-frame missing

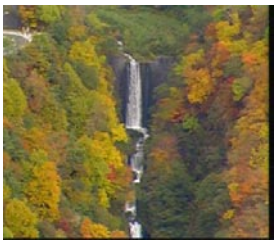
SSIM : 0.982

The reconstruction for two sub-frames missing

SSIM : 0.923

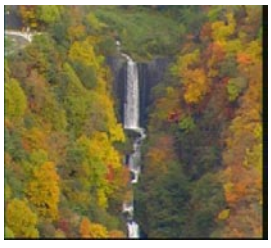
The reconstruction for three sub-frames missing

SSIM : 0.874



The reconstruction for two sub-frames missing

SSIM : 0.972

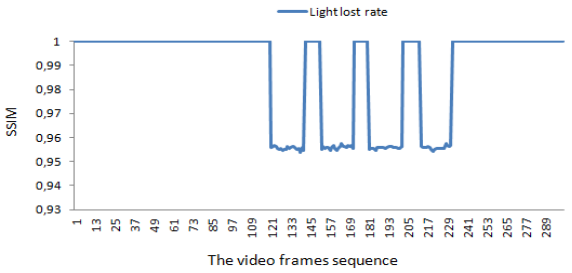


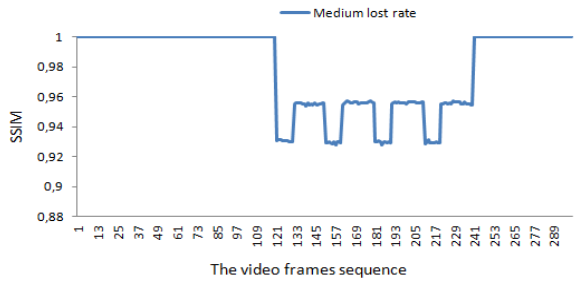
The reconstruction for three sub-frames missing

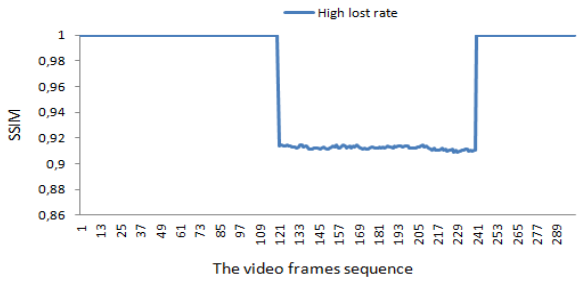
SSIM : 0.945

c. News d. Waterfall

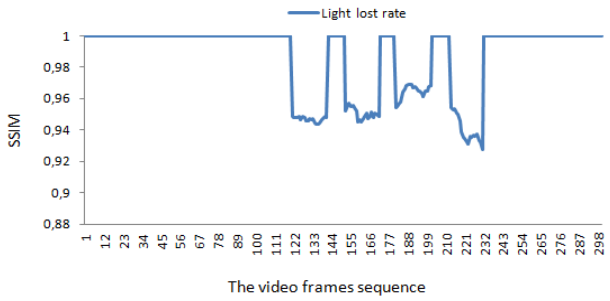
Figure 6. The SSIM for the frame number 140.

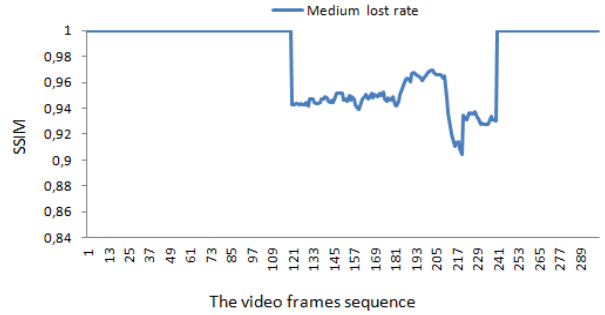


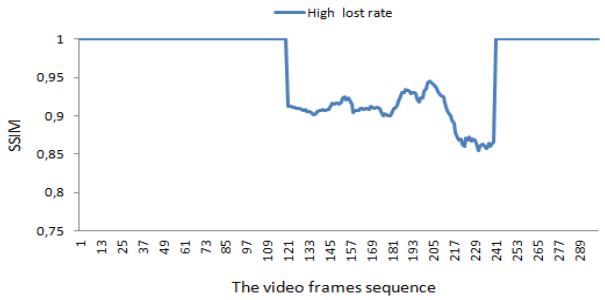




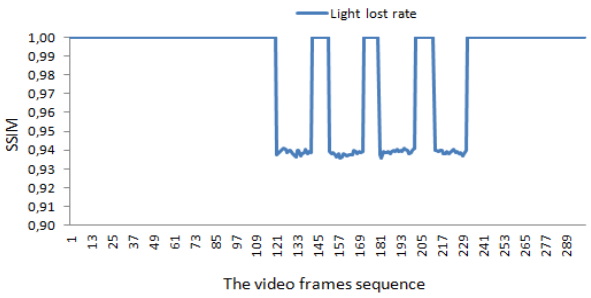
a. Akiyo

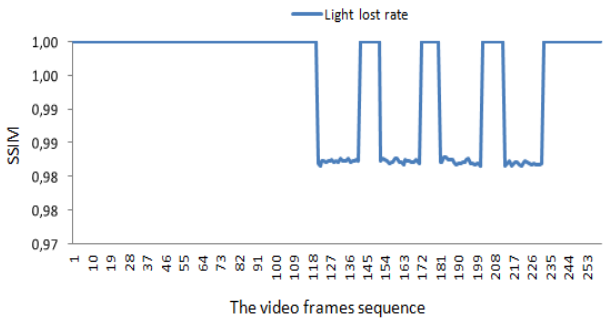


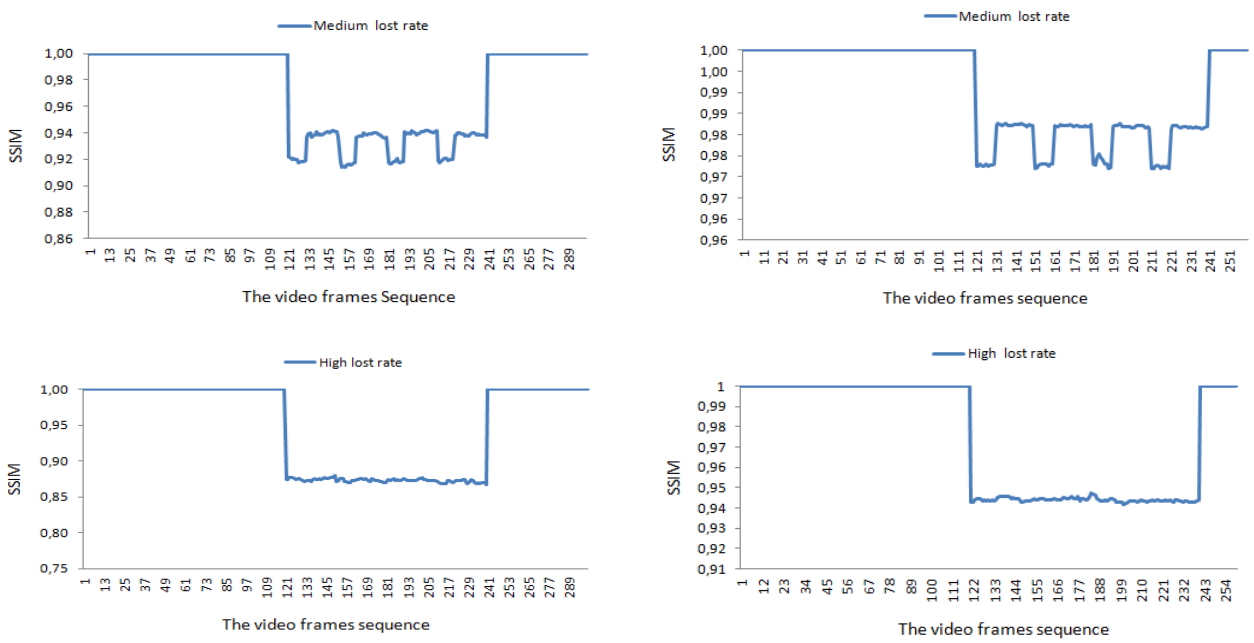




b. Foreman



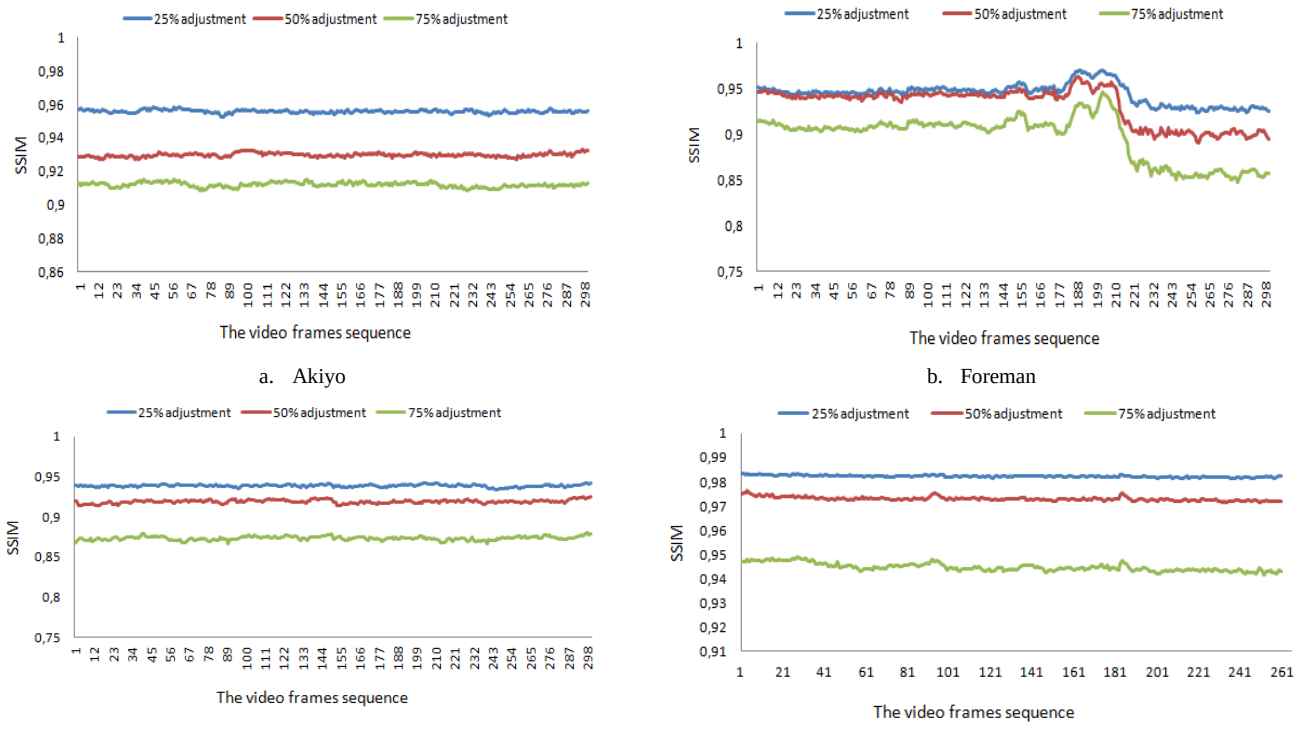




c. News

d. Waterfall

Figure 7. The SSIM for video frames after the lost been distributed and reconstructed to the sub-frames.



a. Akiyo

b. Foreman

c. New s

d. Waterfall

Figure 8. The SSIM for the reconstruction sub-frames for the adaption rate to the video frames sequence.

VI. CONCLUSION

In this paper, we proposed a sub-frames crossing technique to distribute the pixels as sub-frames in different positions in the sequence of the streaming video by combining it with other sub-frames from different positions. The idea behind that is to eliminate the losses of the complete single frame and allow at least one fourth of the frame (one sub-frame) to be received by the mobile device. The receiving sub-frames will reconstruct based on the neighboring pixels to replace the missing pixels.

From the results, it is shown that our proposed technique provides a promising direction for eliminating the frozen picture in the mobile screen, that been caused by missing frames from the streaming sequence. Adjusting the number of frames according to the bandwidth changes is highly needed to reduce the amount of data to be transmitted to the mobile device in a congested network.

However, the quality of the played video is degraded and it depends on the number of frames that are lost or skipped. The numbers of buffers are needed will be equivalent to the number of crossing group, while the initial delay time it needed to implement the crossing technique.

ACKNOWLEDGMENT

We would like to thank Katarzyna Wac from University of Geneva for her helpful discussions. We would like to thank also the Swedish Knowledge Foundation for sponsoring a part of this work through the project QoEMoS (21601420).

REFERENCES

[1] G. Bai and C. Williamson, "The Effects of Mobility on Wireless Media Streaming Performance," Proc. of the Wireless Networks and Emerging Technologies (WNET 04), July 2004, pp. 596-601.

[2] G.-R. Kwon, S.-H., Park, J.-W. Kim, and S.-J. Ko, "Real-Time R-D Optimized Frame-Skipping Transcoder for Low Bit Rate Video Transmission," Proc. of the 6th IEEE International Conference on Computer and Information Technology (CIT 06), Sept. 2006, pp. 177-177,doi: 10.1109/CIT.2006.158.

[3] H. M. Aziz, M. Fiedler, H. Grahn, and L. Lundberg, "Streaming Video as Space – Divided Sub-Frames over Wireless Networks," Proc. of the 3rd Joint IFIP Wireless and Mobile Networking Conference (WMNC 10), Oct. 2010, pp.1-6, doi: 10.1109/WMNC.2010.5678760.

[4] H. M. Aziz, H. Grahn, and L. Lundberg, "Sub-Frame Crossing for Streaming Video over Wireless Network," Proc. of the 7th International Conference on Wireless On-demand Network Systems and Services (WONS 10), Feb. 2010, pp. 53 – 56, doi: 10.1109/WONS.2010.5437132.

[5] H. Zhu, H. Wang, I. Chlamtac, and B. Chen, "Bandwidth Scalable Source-Channel Coding for Streaming Video over Wireless Access Networks," Proc. of Wireless Networking Symposium (WNCG 03), Oct. 2003.

[6] H. Luo, M.-L., Shyu, and S.-C. Chen, "An End-to-End Video Transmission Framework with Efficient Bandwidth Utilization," Proc. of the IEEE International Conference on Multimedia and Expo (ICME 04), June 2004, pp. 623-626, doi: 10.1109/ICME.2004.1394269.

[7] J.-Y. Chang and H.-L. Chen, "Dynamic-Grouping Bandwidth Reservation Scheme for Multimedia Wireless Networks," IEEE Journal on Selected area in Communications, vol. 21, Dec. 2003, pp. 1566-1574**,** doi**:** 10.1109/JSAC.2003.814863.

[8] M. Ries, O. Nemethova, and M. Rupp, "Performance Evaluation of Mobile Video Quality Estimators," Proc. of the European Signal Processing Conference (EUSIPCO 07), Sept. 2007, pp. 159-163.

[9] P. Antoniou, V. Vassiliou, and A. Pitsillides, "ADIVIS: A Novel Adaptive Algorithm for Video Streaming over the Internet," Proc. of the 18th Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC'07), Dec. 2007, doi: 10.1109/PIMRC.2007.4394583.

[10] R. Weber, M. Guerra, S. Sawhney, L. Golovanvsky, and M. Kang, "Measurement and Analysis of Video Streaming Performance in Live UMTS Networks," Proc. of the 13th International Symposium on Wireless Personal Multimedia Communications (WPMC 06), Sept. 2006, pp. 1-5.

[11] S. Cen, C. Pu, and R. Staehli, "A Distributed Real-time MPEG Video Audio Player", Proc. of the 5th International Workshop on Network and Operating System Support of Digital Audio and Video, LNCS, 1995, pp. 142-153, doi: 10.1007/BFb0019263.

[12] T. Nguyen, P. Mehra, and A. Zakhor, "Path Diversity and Bandwidth Allocation for Multimedia Streaming," Proc. of the International Conference on Multimedia and Expo (ICME 03), July 2003, pp. 1-4.

[13] T.Y. Huang, "Region of Interest Extraction and Adaptation in Scalable Video Coding," Proc. of the 7th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 10), Aug. 2010, pp. 2320-2323, doi: 10.1109/FSKD.2010.5569822.

[14] X. Cao, G. Bai, and C. Williamson, "Media Streaming Performance in a Portable Wireless Classroom Network," Proc. of IASTED European Workshop on Internet Multimedia Systems and Applications (EuroIMSA´05), Feb. 2005, pp. 246-252.

[15] X. Zhu and B. Girod, "Video Streaming over Wireless Networks," Proc. of the European Signal Processing Conference (EUSIPCO 07), Sept. 2007, pp: 1462-1466.

[16] X. Wang and X. Tu, "Adaptive FMO Strategy for Video Transcoding," Proc. of the International Conference on Communications, Circuits and Systems (ICCCAS 09), July 2009, pp. 540 – 544, doi: 10.1109/ICCCAS.2009.5250462.

[17] Y. Wang, A. R. Reibman, and S. Lin, "Multiple Description Coding for Video Delivery," Proc. of the IEEE Journal, vol. 93, Dec. 2004 pp. 57-70, doi: 10.1109/JPROC.2004.839618.

[18] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity," IEEE Transactions on Image Processing, vol. 13, April 2004, pp. 600-612, doi: 10.1109/TIP.2003.819861.

[19] http://trace.eas.asu.edu/yuv/index.html (visited, 1/11/2011)

[20] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the Scalable Video Coding Extension of the H.264/AVC Standard," Proc. of the IEEE Transactions on Circuits and Systems for Video Technology, vol. 17, Sept. 2007, pp.1103-1120, doi: 10.1109/TCSVT.2007.905532.

A Robust and Fast Gesture Recognition Method for Wearable Sensing Garments

Ali Boyali
Department of Computing
Macquarie University
Sydney, Australia
ali.boyali@mq.edu.au

Manolya Kavakli
Department of Computing
Macquarie University
Sydney, Australia
manolya.kavakli@mq.edu.au

Abstract—— **There is an increasing demand for motion capture and full body gesture recognition due to fast paced ubiquitous computing developments and their requirements for natural input modalities. Analysis and synthesis of human body movements are significant issues in bio-medical applications for rehabilitation and identification purposes as in post-stroke patient rehabilitation and gait recognition studies. This paper presents the development of a robust gesture and posture recognition algorithm based on an emerging research field Compressed Sensing (CS) and Sparse Representation (SR), in signal processing for the wearable sensing garment which consists of a sensor network having piezo-resistive properties. The gesture recognition algorithm presented in this study is highly accurate regardless of the signal acquisition method used and gives excellent results even for high dimensional signals and large gesture dictionaries. Our findings state that gestures can be recognized with over 99% accuracy rate using Sparse Representation-based Classification (SRC) algorithm. We tested the algorithm using 3 different gesture dictionaries acquired in 3 different gesture domains with user dependent and user independent test gesture and dictionaries. The system gives 100% recognition accuracy for the gestures performed by sensing t-shirt with two different gesture sets.**

Keywords-Gesture recognition; wearable sensing garments; compressed sensing; sparse signal classification

I. INTRODUCTION

Among the wide variety of motion capture tools in Human Computer Interaction (HCI) applications, Motion Capture (MoCap) suits are the most complex ones due to their high dimensional sensor networks producing complex data and the large amount of computations needed to interpret a complex data set.

Motion capture in HCI applications has two aspects. The first aspect is the acquisition of motion information, extraction of parameters for reconstruction of motion in a virtual environment [1], and analysis of motion parameters. The second is the synthesis of captured motion information to extract meaningful context as in gesture recognition studies. Although these two aspects serve for distinct aims the latter cannot be implemented without capturing the data.

The structure of motion capture suits is defined by the type of sensor signals, e.g. inertial, acoustic [2], optical, and magnetic or hybrid signal types which makes use of several signal domains. Every signal domain has pros and cons over other domains. Magnetic motion tracking systems suffer from distortion in their magnetic field [3], whereas optical and accelerometer based systems suffer from occlusion and inherent drift respectively [4, 5]. Other motion capture suits that employ mechanical connections are obtrusive and are not suitable for motion analysis [6, 7] in most cases. In this study, our goal is to develop a robust gesture recognition system for a wearable sensing garment, namely The Sensor T-shirt developed by the research teams at the Electronic Engineering Department of University Pisa, Italy [6, 7]. The sensor t-shirt consists of piezo-resistive sensor threads smeared on an elastic fabric substrate which allows the user to perform motions without any constraint. This feature of the sensor t-shirt makes it a perfect candidate for analysis and synthesis of the motion and many other possible studies. The Sensor T-shirt can be used to aid quadriplegic people to control a wheelchair using their available muscles [9]; or to assist in gesture analysis [10].

The proposed gesture recognition system is based on a new research field, Compressed Sensing (CS) which brings a new insight into signal acquisition and recovery. CS and dimensionality reduction methods such as random projections have been studied intensively in pattern recognition studies. One of the most successful applications of CS and sparse signal recovery is implemented by Wright et al. [11] for face recognition under varying illumination and occlusion. The team simply benefit from the discriminative nature of sparse signal recovery to classify the faces and name their method Sparse Representation-based Classification (SRC).

In this study, we show that gestures can be recognized with an accuracy rate of over 99% using the SRC algorithm without introducing an additional operator in the measurement domain. The adaptation of the SRC method is an advantageous approach in gesture recognition studies.

This paper brings about following advantages in gesture recognition.

- Multi-dimensional gestures can be reshaped (multi dimensional readings are put into vector form) and represented as a one dimensional vector as in the study face recognition implemented by Wright et al. [11] and have a high recognition accuracy.

 ISBN: 978-1-61208-195-3

- No prior clustering algorithm is necessary, however, pre-classification algorithms can be used to reduce the computation time, and there is no upper and lower bound for the number of classes and the number of gesture classes.
- It can be applied in any signal acquisition environment.
- The algorithm achieves high accuracy for rich gesture databases.
- The feature selection is random.

The rest of the paper is organized as follows. Section II gives brief description of the previous sensor garment and gesture recognition and CS and SRC based signal classification studies. Section III presents the contemplated gesture sets for wearable sensor t-shirt and the application of SRC for gesture recognition by sensor garments. The paper ends with conclusion section.

II. PREVIOUS STUDIES

The sensor t-shirt (Fig. 1) consists of a new class of strain sensors network developed at the university of Pisa in Italy [7] which satisfies the user requirements such as wearability, comfort and long term monitoring. The first study on mapping from sensor readings to position and posture domain are conducted by Tognetelli et al in the study [7] who developed and used the sensor garment equipping an elastomeric fabric with piezzo-resistive graphite stripes by smearing them on the garment.

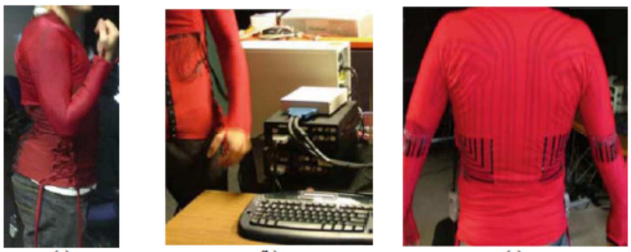


Figure 1. Sensor Shirts and Data Acquisation System Used in Experiments (b-c [20])

Tognetelli et al [7] define posture a geometrical model of body kinematics. According to the definition, when the sensor t-shirt is worn by a user and a posture is performed, the sensor network produces electrical signals strictly related to the posture. The non-linear signal behaviour of the network is modelled by the linear combination of exponential functions. The construct (a mapping function F) between the sensor space and kinematic configuration readings, and corresponding measurements by using a goniometer are stored in a database, and sensor readings are mapped using multi-variate piece-wise interpolation and function inversions. The proposed method in [7] is considered to be time-consuming by the authors as a high number of matrix inversions are necessary.

There are numerous gesture recognition methods available in the literature. Some of these methods require feature extraction and clusterization algorithms peculiar to gestures that are designed for only their gesture domain and cannot be generalized unlike our gesture recognition algorithm which gives highly accurate results for different measurement and signal domains. On the other hand, stochastic Hidden Markov Models, Dynamic Time Warping and Neural Network based classification methods are common and the accuracy rates vary depending on the application. Although we tested our SRC based gesture recognition method for different measurement domains – with a touchpad, an IR camera and piezo-resistive signal measurements, we only focus on gesture recognition performed by using the sensor jacket in this study. Therefore, we only review the research studies using the sensor jacket, CS, and SRC based classification in gesture recognition.

CS and SR methods were first used in the study [12] by Akl and Valee for gesture recognition with a complementary algorithm Affinitiy Propagation (AP) proposed by Frey and Dueck [13] which clusters the data by message passing between the points. In their study, the gestural data which consists of accelerometer signals in x, y and z coordinates acquired by a Wiimote are clustered using the AP algorithm into gesture classes using the Dynamic Time Warping (DTW) similarity measure. Then, the gesture to be recognized is compared with candidate exemplars determined by AP. Final classification is carried out converting the classification problem to SR type by introducing a pre-processor. The CS solution is applied to the finalist exemplars hence recovering the gesture class with a high recognition rate attained.

We tested the proposed algorithm by [12 and 19] for three different rich gesture sets. The critical issue is the pre-classification algorithm which will be detailed after outlining the principles of CS and SRC.

Compressed Sensing (CS) and Sparse Representation (SR) of signals is a new research field which allows signals to be recovered with a few number of samples -much below the well-known Nyquist sampling rate from random/non-adaptive measurements- as long as the signal is sufficiently sparse in measurement domain [13-17].

Mathematically, given a sufficiently k-sparse signal $x \in R^n$ whose members consists of a few nonzero k-elements and zeros in a measurement domain with an orthonormal basis Ψ (such as Fourier, Direct Cosine Transformation or wavelet bases), the whole content of x can be recovered by sampling via a random matrix $U \in R^{mxn}$ which satisfies the Restricted Isometry Property (RIP) by preserving the lengths of k-sparse elements with the condition that $m<<n$. The resulting equation $y = U\Psi x$ is then solved by the linear programming method ℓ_1 minimization.

The SRC algorithm uses a dictionary matrix consisting of training samples. In the algorithm, first, training samples of k classes and the test image are converted to a column vector and projected on a lower dimensional space using a generated random measurement matrix. Then training vectors are stacked into a matrix to construct the dictionary. The resulting equation is solved to recover the sparse signal x by ℓ_1 minimization methods after the columns of the

 ISBN: 978-1-61208-195-3

dictionary and test vector are normalized. The SRC algorithm assumes the test image vector lie in the linear span of training samples (1) associated with the same class of object where the signal *x=[0, 0, 0, 0, α_1, α_2,... α_n, 0, 0, 0, 0].*

$$y = \alpha_{i,1}\,\vartheta_{i,1} + \alpha_{i,2}\,\vartheta_{i,2} + \ldots + \alpha_{i,ni}\,\vartheta_{i,ni} \qquad (1)$$

The pseudo code for the algorithm is as follows:

- Construct the dictionary matrix $A=[\vartheta_{i,1}, \vartheta_{i,2}, .. \vartheta_{i,k}] \in R^{mxn}$ for k classes by reducing the dimension using a random matrix having RIP and converting the samples, and test image vector to one dimensional vectors ϑ_i and y
- Normalize the columns vector of reduced $\tilde{A}$ and the reduced test image vector $\tilde{y}$,
- Solve the $\ell 1$ minimization problem $\hat{x} = \text{argmin}_x \|x\|_1$ subject to $\tilde{A}x = \tilde{y}$
- Compute the residuals $r_i(y) = \|\tilde{y} - \tilde{A}\delta_i(\hat{x})\|_2$ where δ_i is a selection operator for $\hat{x}$ corresponding the ith class span in A
- Identify y by finding the minimum of ri(y)

In the study by Akl and Valee [12], 3 axis gesture traces are divided into the acceleration components of corresponding axes *Rx*, *Ry* and *Rz* and the

$$\bar{y}_x = \bar{R}_x x + \varepsilon_x \qquad (2)$$

where $\bar{y}_x$ is the randomly sampled x component of the gesture to be recognized and $\bar{R}_x$ is the classes of the remaining gesture traces after narrowing the AP results.

They convert the recognition problem to CS and SR recognition problem as shown below.

$$Q_x = Orth(\bar{R}_x^T)^T \qquad (3)$$

$$W_x = Q_x \bar{R}_x^{\dagger} \qquad (4)$$

Where Q_x is the orthonormal basis for $\bar{R}_x$ and $\bar{R}_x^{\dagger}$is the pseudo-inverse of $\bar{R}_x$thus the gesture recognition problem has a new form of

$$h_x = W_x \bar{y}_x = Q_x x + \varepsilon'_x \qquad (5)$$

Equation (5) is solved for all axis components to identify the gesture class by computing (6) and taking the minimum of $\hat{x}_{eq}(i)$.

$$\hat{x}_{eq} = \hat{x}_x^2 + \hat{x}_y^2 + \hat{x}_z^2 \qquad (6)$$

The algorithm gives higher accuracy rates when the AP is applied with this method. The CS based classification with the introduction of the pre-processor operator only gives efficient results for a few gesture classes. If the AP is eliminated with our gesture sets.

III. THE SENSOR JACKET SYSTEM AND SRC FOR GESTURE RECOGNITION

The sensor t-shirt has 52 individual piezzo resistive sensor strips which are located from wrist to shoulders on the right and left side of the t-shirt. The data is acquired by the National Instrument Data Acquisition Unit (Fig. 1.).

There are three gesture classes to be recognized by wearing the sensor t-shirt in the first gesture set. These are; moving the arm from relaxed to front at shoulder level, from relaxed position to side at shoulder level and from side to front keeping the arm straight moving horizontally at shoulder level. However we further expands our gesture recognition study by designing a second gesture set to verify and repeat the study. The second gesture set includes 5 gesture classes (Table 1.).

TABLE I. TABLE 1. SECOND GESTURE SET

	**Right arm up:** Right arm is moved from rest to shoulder level straight, hand points ahead
	**Right arm up at side**: Right arm is moved from rest to shoulder level towards right straigthly
	**Forearm** is moved upper from the rest.
	**Right hand** is put on head from the relaxed position
	**Right hand** is put on heart level from the rest

Each sensor reading (Fig. 2.) is sparse in Discrete Cosine Transform (DCT) domain. Before using sensor readings we apply a light Gaussian smoother to the readings to eliminate jitter on the data .

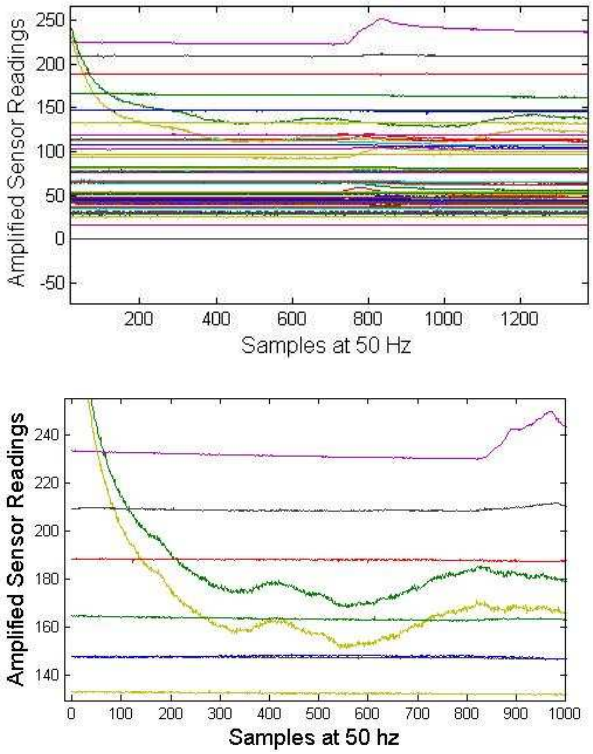


Figure 2. Sensor Data at 50 Hz (Readings from 52 Sensor Threads and a Local View)

All sensor readings are normalized, giving a zero mean and a variance of one. All inactive sensor signals are eliminated defining an absolute total variation threshold that is the ratio of any value in a thread signal to the absolute value of the difference between maximum and minimum values. If this ratio exceeds 10%, it is assumed to be there is a considerable variation in thread readings which contributes to the identification of the activity and gestures (Fig 2).

The remaining sensor readings are concatenated in a vector for an individual gesture. A gesture dictionary is constituted from three gesture classes by stacking the gesture traces as columns of the dictionary. The rest of the solution is to design a measurement matrix and apply ℓ_1 minimization.

CS theory states that if the signal is sparse in any domain, signals can be recovered with an overwhelming probability by random measurement matrices having The Restricted Isometry Property (RIP) condition. Random Gaussian or Achlioptas' matrix [18] can be used for both dimensionality reduction and measurements, since the values having RIP properties preserve distances in the embedding space. We use Achlioptas' matrix since 2/3^{rds} of the matrix is sparse, making it easier to construct than a Gaussian matrix thus saving computation time.

Achlioptas' matrix is defined as

$$U_{ij}\sqrt{3}\begin{cases}+1 & with\ probability\ 1/6\\ 0 & with\ probability\ 2/3\\ -1 & with\ probability\ 1/6\end{cases} \quad (7)$$

The pseudo code for the gesture recognition algorithm are as follows.

- Normalize the data, apply a Gaussian smoother and reshape all the gestures and the test gesture, and take DCT of each
- Find the longest length of gestures (lh_{max}) and make the other gestures of the same length by zero padding, and stack the training gestures into training matrix $A_G \in R^{lhmax\ xn}$

$$A_G = \begin{bmatrix} G_{11} & G_{21} & \dots & G_{k1} \\ G_{12} & G_{22} & \dots & G_{k2} \\ \vdots & \vdots & & \vdots \\ G_{1n1} & G_{2n2} & \dots & G_{knk} \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad (8)$$

- Take m measurement from both the test gesture vector and the training matrix with the designed random measurement matrix U^{mxn} to form the reduced equations $\tilde{y} = \tilde{A}_G x$, where $x=[0, 0, 0, 0, \alpha_1, \alpha_2, \dots \alpha_n, 0, 0, 0, 0]$ consists of a few nonzero coefficients corresponding to gesture class.
- Solve the ℓ_1 minimization problem $\hat{x} = argmin_x \|x\|_1$ subject to $\tilde{A}_G x = \tilde{y}$
- Compute the residuals $r_i(y) = \|\tilde{y} - \tilde{A}_G \delta i\ (\hat{x})\|_2$ where δi is a selection operator for $\hat{x}$ corresponding the i^{th} class span in $\tilde{A}_G$
- Identify y by finding the minimum of $r_i(y)$

There are 10 gesture traces collected for each gesture class using sensor jacket for our gesture recognition study. We construct gesture dictionary by randomly choosing 6 gestures from the database, the remaining 4 gestures from each class are used for testing purpose. The proposed algorithm gives 100% recognition rate for the sensor jacket gesture recognition. This paper gives only a brief presentation of the recognition algorithm for the initial studies of gesture recognition with a sensing garment. However, we used the same method for other two gesture databases.

The first database consists of 23 gestures (Fig. 3.) which are 2D gestures captured by the IR camera of a Wiimote. 15 gestures from each gesture class, 10 for dictionary and 5 for testing are captured from 3 subjects. In total, we have 1035 gesture traces from 3 subjects stored in xml format. These gesture files are then converted to .mat files which are then used as input for the SCR gesture recognition algorithm.

 ISBN: 978-1-61208-195-3

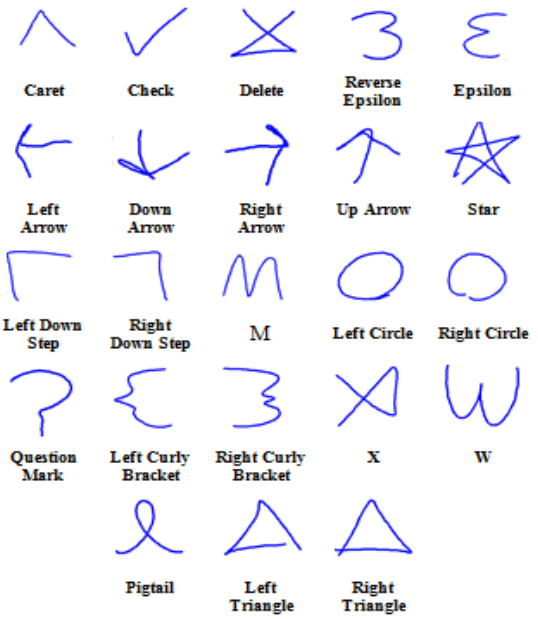


Figure 3. Mixed Gesture Set for Wiimote Gesture Recognition

Our system gives 100% user-dependent recognition accuracy for each person with a full dictionary matrix which consists of 10 gestures for each gesture class including 230 columns in total. To be able to provide a user independent gesture recognition system, the dictionary is built by arbitrarily chosen gestures from each subject with the same number of gestures from each class. Then, we used the remaining gestures for testing purposes. The system misclassified only 2 gestures out of 300 test samples, corresponding to 99.33% recognition accuracy for 20 gestures. 3 gestures were removed from the database, since one gesture class (star shown in Fig 3) was not collected from one of the subjects, and 2 gesture classes which are check mark and left arrow are performed in very different manner.

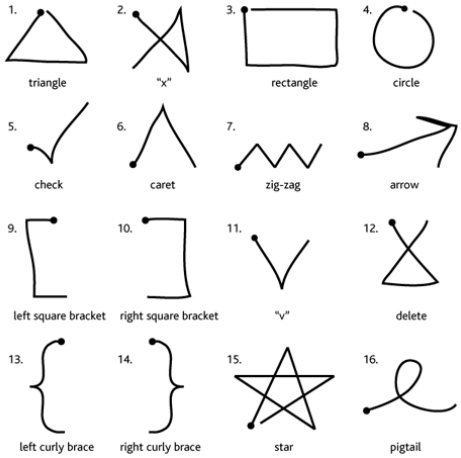


Figure 4. $1 Gesture Set [19]

We tested our system also with the gesture sets of the $1 gesture recognition study [19]. The set consists of 16 different gestures collected from 11 subjects (Fig. 4.).

Each gesture is repeated 10 times in a 3 speed profile (fast, slow and medium speed) by each subject. The gesture dictionary is built by choosing two random gestures. The random gestures belong to any speed profile from the gesture folder of each of 5 subjects, so that every gesture class consists of 10 gestures. The gestures tested are chosen from the remaining subjects' folder randomly. The SRC gesture recognition algorithm misclassified only 2 gestures out of 80 test gestures with 97.5% accuracy. When the two misclassified gestures are analysed (Fig. 5.)

It is seen that the algorithm confuses the circle and rectangle gesture since both of them is unclosed and even may be confused by human brain.

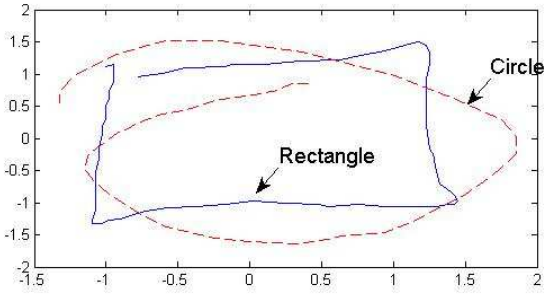


Figure 5. Confused Gestures in $1 Gesture Set

IV. CONCLUSION

In this paper, we presented a robust gesture and posture recognition algorithm based on an emerging research field CS and SR, in signal processing for the wearable sensing garment which consists of a sensor network having piezo-resistive properties. The gesture recognition algorithm we presented is highly accurate regardless of the signal acquisition method used, and gives excellent results even for high dimensional signals and large gesture dictionaries. Our findings state that gestures can be recognized with over 99% accuracy rate using the SRC algorithm.

Gesture and posture recognition studies in which sensing garments are used have been studied in literature both theoretically and experimentally [6-7, 20]. Various algorithms were proposed in those specific applications for gesture and posture recognition. Our algorithm outperforms in the gesture recognition studies realized by using the sensor jacket with an accuracy level 100% in mapping the sensor readings into gesture domain.

This study can be extended for detection of postures in sensing garment based studies. There are several optimization algorithms proposed for the solution of convex optimization problems. We utilized GPRS (Gradient Projection for Sparse Reconstruction) method proposed by Mario et al. (2008) for the ℓ_1 linear programming problem, as it solves the reconstruction problem in a significantly shorter time [21].

Solution of the equations for the sensor jacket gesture recognition study takes less than 0.1 second with a AMD

 ISBN: 978-1-61208-195-3

Turion 2x2.2Hz processor. This time period can be regarded sufficient for real time applications. The gesture recognition method given in this paper is promising and can provide solutions to high dimensional gesture recognition problems. Gesture spotting is the second fundamental problem in this research field. We focus on the development of a new algorithms which make use of recent developments for low-rank and sparse matrix separation methods for robust posture recognition and gesture spotting.

ACKNOWLEDGMENT

This project is sponsored by the Australian Research council Discovery Grant (DP0988088) titled "A Gesture-Based Interface for Designing in Virtual Reality".

REFERENCES

[1] F. J. Perales, "Human Motion Analysis & Synthesis using Computer Vision and Graphics Techniques State of Art and Applications" Proc. of the 5th World Multi-Conference on Systemics, Cybernetics and Informatics (SCI2001), July 2001. doi:10.1.1.90.6176.

[2] C. Einsmann, M. Quirk, B. Muzal, B. Venkatramani, B., T. Martin and M. Jones, "Modeling a Wearable Full-Body Motion Capture System". Proc. of 9th IEEE International Syposium on Wearable Computers, Oct. 2005. pp. 144 – 151. doi=10.1.1.74.1235

[3] J. G. Hagedorn, S. G. Satterfield, J. T. Kelso, W. Austin, J.E. Terrill and A. E. Peskin, "Correction of Location and Orientation Errors in Electromagnetic Motion Tracking". in J. MIT Press Teleoperators and Virtual Environments, 2007, vol 16-4, pp. 352 -- 366.

[4] A. Hornung, S. Sar-Dessai and L. Kobbelt, "Self-Calibrating Optical Motion Tracking for Articulated Bodies". Proc. of Virtual Reality (VR 2005), IEEE Press. March 2005. pp. 75 – 82.

[5] R. Slyper and J. K. Hodkings, "Action Capture with Accelerometers". Proc. of ACM SIGGRAPH/Eurographics Symposium on Computer Animation, ACM Press. July 2008. pp. 193-199.

[6] A. Tognetti, F. Lorussi, R. Bartalesi, S. Quaglini, M. Tesconi, G. Zupone and D. De Rossi. "Wearable Kinesthetic System for Capturing and Classifying Upper Limb Gesture in Post-Stroke Rehabilitation". J. Neuro Engineering and Rehabilitation. June 2005. vol. 2, 1-16.

[7] F. Lorussi, S. Galatolo and D. E. De Rossi. "Body Segment Position Reconstruction and Posture Classification by Smart Textiles". IEEE J. of Sensors. Sept. 2009. vol 9, pp. 1014 -- 1024, doi: 10.1109/JSEN.2009.2024867.

[8] T. Gulrez and M. Kavakli, M. "Precision Position Tracking in Virtual Reality Environments Using Sensor Networks". Proc. of IEEE International Symposium on Industrial Electronics (ISIE2007), Nov. 2007. pp. 1997-2003. doi: 10.1109/ISIE.2007.4374914.

[9] M. Kavakli. "Gesture Recognition in Virtual Reality". In: Special Issue on: Immersive Virtual, Mixed, or Augmented Reality Art of The International Journal of Arts and Technology (IJART), 2008, Vol 1 no 2, pp. 215-229. doi: 10.1504/IJART.2008.021928.

[10] J. Wright, A. Yang, A. Ganesh, S. Sastry and Y. Ma. "Robust Face Recognition via Sparse Representation", IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI). Feb. 2009. vol. 31. issue 2, pp. 210—227. doi: 10.1109/TPAMI.2008.79.

[11] A. Akl and S. Valaee. "Accelerometer-based Gesture Recognition via Dynamic-Time Warping, Affinity Propagation, & Compressive Sensing". Proc. of IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP). March 2010. pp 2270 – 2273. doi: 10.1109/ICASSP.2010.5495895.

[12] B.J. Frey and D. Dueck. "Clustering by Passing Messages Between Data Points". Science. Feb. 200. Vol. 31, no. 5814, pp. 972--976. doi: 10.1126/science.1136800.

[13] E. Candès, J. Romberg and T. Tao. "Robust Uncertainty Principles: Exact Signal Reconstruction from Highly Incomplete Frequency Information". Feb. 2006. IEEE Trans. Inform. Theory, vol 52, issue 2, pp. 489--509. doi: 10.1109/TIT.2005.862083.

[14] E. J. Candès and T. Tao. "Near Optimal Signal Recovery from Random Projections: Universal Encoding Strategies?". IEEE Trans. Inform. Theory. Dec. 2006. vol. 52, issue 12, pp. 5406–5425. doi: 10.1109/TIT.2006.885507

[15] D. Donoho. "Compressed Sensing". IEEE Trans. Inform. Theory. Apr. 2006. vol 52, issue 4, pp. 1289--1306, doi: 10.1109/TIT.2006.871582.

[16] E.J. Candès, and M.B. Wakin "An Introduction To Compressive Sampling". IEEE Signal Processing Magazine, March 2008. vol 25, issue 2. pp. 21-30. doi: 10.1109/MSP.2007.914731.

[17] R. G. Baraniuk, E.J. Candès, E. Nowak and M. Vetterli. "Compressive Sensing". Signal Processing Magazine. March 2008, IEEE press, vol 24(2), pp. 12-13. doi: 10.1109/MSP.2008.915557.

[18] A. Dimitris. "Database-friendly Random Projections". Proc. of PODS '01 Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. May 2001. pp. 274-281. doi: 10.1145/375551.375608.

[19] J.O. Wobbrock, A.D. Wilson and Y. Li. "Gestures without Libraries, Toolkits or Training: A $1 Recognizer for User Interface Prototypes". Proc. of the ACM Symposium on User Interface Software and Technology (UIST '07). ACM Press. Oct. 2007. pp. 159-168. doi: 10.1145/1294211.1294238.

[20] G. Pioggia, M. Ferro, G. Zupone, L. Chirulli, and D. D. Rossi, "Development of a sensing seat for human authentication". Proc. of 3rd IET International Conference on Intelligent Environments. Sep. 2007. pp. 441–446.

[21] M. Figueiredo, R. D. Nowak, and S. Wright, "Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems". IEEE J. of Selected Topics in Signal Processing. Dec. 2007. vol. 1, no. 4, pp. 586–597. doi: 10.1109/JSTSP.2007.910281.

 ISBN: 978-1-61208-195-3

Influence of Culture in Multimodal Interfaces

Gesture behavior between Anglo-Celtic and Latin Americans

Karime Nasser Alvarez
Department of Computing
Macquarie University
Sydney, Australia
Karime.nasser-alvarez@mq.edu.au

Manolya Kavakli
Department of Computing
Macquarie University
Sydney, Australia
manolya.kavakli@mq.edu.au

***Abstract*— In multimodal interfaces, hand gestures often help convey meaning to the spoken word; therefore, the cultural background of the gesturing person might be an influential factor in the interaction with these interfaces. This paper presents an empirical study aimed at singling out basic cultural differences in hand gesture performance between two cultures: Anglo-Celtics and Latin Americans. The focus in this paper is given to the video analysis of the two cultures describing two objects with their hands. The purpose is to use gesture segmentation to define predominant hand gestures by culture. Conclusions are drawn from the experiment and are linked to cultural attributes proposed by theorists like Hall and Hofstede. The findings state that cultural differences exist in the description of the object, which might have implications for the development of gesture-based multimodal interfaces. As Anglo-Celtics are low context cultures, they used more words and gestures in longer time. On the other hand, Latin Americans, which represent the high context culture, had more frequent gestures, but performed fewer ones, in shorter time. We also found that as the complexity of a task increases, so does the use and type of gestures. The performance of a multimodal interface will not only be affected by the task being performed, but by the cultural background and language skills of the user.**

Keywords- Gesture recognition; HCI; culture; Anglo-Celtics; Latin Americans; gesture based interaction; performance; frequency.

I. INTRODUCTION

The aim of Human Computer Interaction (HCI) is making the interactions as natural as possible, as if communicating with another human [1]. Gestures, such as pointing, are where language, culture and cognition meet [2]. Humans have an innate need to use gestures; since they complement our ideas, to such an extent that humans are known to gesture even when talking on the phone [3].

The significance of this study relies on the intention of defining the gesture variances from one culture to another and relating them to cultural traits. Culture has been studied by anthropologists all over the world, and these have arrived to the science behind stereotypes. Our intention is to identify, if any, the cultural influences that may possibly affect the representations of hand gestures. Our approach follows the experiment conducted by Bischel et al. where a designer is required to describe a mechanical device to another designer [4]. In this case, the participants of each sample were recorded depicting two different chairs with their hands. These videos were recorded for later segmentation and used timestamps to assess the cultural influence via metrics such as frequency and the quantity of certain gesture types.

The paper is structured as follows. First, there is a summary of related works. Second, we describe the experiments conducted. Third, we analyze the data collected and conclude with a discussion of the findings.

II. LITERATURE REVIEW

The means to communicate with computers has evolved from classic mouse input, to rich multimodal data [5]. Multimodal interfaces have combined various user input modes beyond the known keyboard and mouse input/output [6], and now include a wide range of possibilities; such as hand gestures, both static and dynamic, speech, head and eye tracking. Apart from usual voice interaction, advances like sensory output have also been developed in videogames.

Games and infotainment are not, however, the only use for gesture based interfaces. The Intuitive Surgical da Vinci surgical system, for instance, is an example of a system for the capture of subtle motions of the surgeon, to teach complex procedures [7]. One may assume that in tasks such as the manipulation of objects, cultural implications might not be of considerable importance, but in the context of cultural and physical differences between surgeons, the subject calls for more attention [8].

Gesture-based interfaces enable freer, more intuitive, and richer digital interactions, than conventional user interfaces [9], leading to better idea generation [10]. When developing multimodal interfaces and applications, developers and designers work together to understand what types of gestures are used for what tasks, as well as the frequency, the importance, and ease of use of the interface. Therefore, there have been many attempts to design an appropriate gesture classification and segmentation "dictionaries".

A. Gesture classification and segmentation

Gesture offers versatility when representing objects, or qualities of these in the scientific field. The main problem

 ISBN: 978-1-61208-195-3

here is that there is no common database of gestures used between developers and scientists. The most recognized gesture classification, and the one referred to from now on, is the one established by McNeill in 1992 [11]. McNeill classifies 4 types of gestures; iconic (resemble what is being talked about, e.g., flapping arms when mentioning a bird), metaphoric (abstractedly pictorial, e.g., drawing a box shape when referring to a room), beat (gestures that index a word of phrase e.g., rhythmic arm movement used to add emphasis), and deictic (gestures pointing to something, e.g., while giving directions).

The iconic ones are of particular interest to HCI and developing technologies as they allow accurate depiction of objects encountered by the user. The cultural background might be an influential factor in the design of gesture-based interfaces. Metrically, culture could be reflected in the interactivity, symbol variety, re-hearsability and pre-processability of gestures.

B. Culture

As defined by Hofstede [12] "Culture is the collective programming of the mind that distinguishes the members of one group or category of people from another". Through the appropriate design of support-focused interfaces how we obtain maximum usability. Technology has been conceived in 'prosthetic' terms, as an extension to the body, or support for tasks [13] and given the global diversity, cultures will perceive these tasks differently. Language and representation are critical elements in the study of culture, because we are locked into our cultural perspectives and mindsets [14].

1) Culture and Interfaces

We communicate and exchange information with a system or a device through interfaces. The more familiar or intuitive an interface is, the higher its usability.

Cultural preferences determine the type of layout, texture, pattern and color [15] in website portals. Certain colors are offensive or uncomfortable for certain cultures, for instance, red is bad luck for Koreans, therefore, Korean websites might avoid the use of red. These examples illustrate the need to adapt interfaces to attract the targeted market, or in this case, culture. Culture does not exist as a computational term in HCI, even though there are efforts like tailored interfaces to a targeted culture. With every use of the technology, the success depends on the capabilities embedded in a persona who is "programmed" in a specific way. The mental "coding" of this persona will affect the usability.

The cultural behavior is visual, but it is not always evident until there is an interaction. One instance is Rehm, Bee, and André [16] try to identify the culture of the user so that the behavior of an interactive system can be adapted to culture-dependent patterns of interaction. This was achieved via a Bayesian network model that based itself on gesture expressivity via speed, power or spatial extent.

In our globalised reality, there is also the implication of remote international collaboration. Here, each participant has their own symbolic, iconic and metaphoric influence on their gestures [16]. As Hofstede concludes in writing about the influence of communication technologies, the software of the machines may be globalized, but the software of the minds that use them is not [12]. Therefore, the dominance of technology over culture is an illusion, and differences between cultures exist.

2) Hofstedes Cultural Dimensions

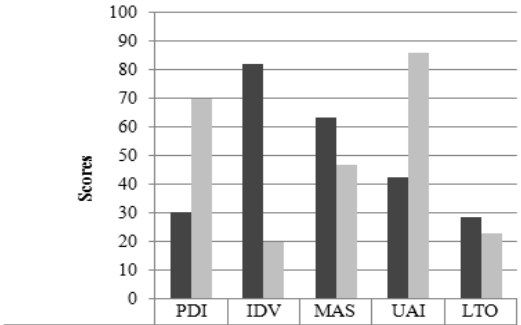
The most renowned cultural study involving the identification of common attributes is the work done by Gert Hofstede [12]. Hofstede developed a set of culture build-ups that describe the way in which national societies are built and the rules by which people think, feel and act. These differences are defined as five dimensions and are measured as indexes. The higher or lower the index, more or less the culture portrays this feature.

The Hofstede's model of dimensions of national culture has been applied predominantly in international business; marketing and consumer behavior works [18]. Now we briefly describe the dimensions by Hofstede.

- Power Distance (PDI): is the acceptance and expectation of power to be distributed unequally.
- Uncertainty Avoidance (UAI) indicates the extent to which the members of society feel uncomfortable or comfortable in an ambiguous or abnormal situation.
- Individualism (IDV) is the extent to which individuals are merged into groups.
- Masculinity (MAS) refers to the distribution of emotional roles between the genders, and also serves to classify a culture as assertive/ competitive (masculine) or modest/caring (feminine).
- Countries with high Long- Term Orientation (LTO), foster pragmatic virtues oriented towards future rewards, in particular saving money, persistence, and adapting to changing circumstances.

Now we present the cultures used in our experiments: Anglo-Celtic (Australian, British, Irish, New Zealanders) and Latin Americans (American countries where Spanish is primarily spoken: Argentina, Chile, Colombia, Costa Rica, Ecuador, Salvador, Guatemala, Mexico, Panama, Peru, Uruguay, and Venezuela).

 ISBN: 978-1-61208-195-3



	PDI	IDV	MAS	UAI	LTO
Anglo-celtic	30	82	63	43	29
Latin American	70	20	47	86	23

Figure 1. Hostedes 5D Model comparing Anglo-Celtic and Latin American countries.

In Fig. 1, we can see a comparison of both samples; an average was taken of the indexes of the countries mentioned above. As we can see in Fig. 1, the Anglo-Celtic culture had a lower PDI (30, 70), and UA (43, 86). On the other hand, they have higher IDV (82, 20), MAS (63, 47) and LTO (29, 23) than the Latin American countries.

Therefore, we assume that due to the greater equality (Low PDI) Anglo-Celtics feel, they are more individualistic (High IDV) and can master new challenges (Low UAI) better than their fellow Latin American colleagues. Hosftede developed a solid foundation for identifying the possible complication of cross-cultural interactions, what makes cultural differences, and how they would act upon this [12].

Even though Hosftede is cited by an extensive amount of sociologists and anthropologists, for the analysis taken in this paper, it is also beneficial to analyze the context classification made by the anthropologist Edward T Hall [19]. Hall identifies a culture's use context in routine communication and classifies them as High or Low. In a high context culture (including much of the Middle East, Asia, Africa, and South America), many things are left unsaid, letting the culture explain. There is more non-verbal communication, a higher use of metaphors, and more reading between the lines. In a lower context culture (including North America and much of Western Europe), the emphasis is on the spoken or written word. They have explicit messages, focused on verbal communication, and their reactions could be visible, external and outward [19].

We can say that Anglo-Celtic cultures (e.g. Australian, British, Irish, and New Zealanders) categorize as low context cultures and Latin Americans (American countries where Spanish and Portuguese are primarily spoken) correspond to the high context cultures. This classification lets us make certain assumptions, like the Anglo-Celtic may predominantly use words, while the Latin Americans would use gestures.

These characteristics identified for each of the samples will be later referred to in order to understand possible influence of these in the gesture performance after the experimentation.

III. EXPERIMENT

In order to explore the influence of culture in gesture performance, an experiment was carried out. As following up on Bischels' experiment the participants will be required to describe two chairs to the camera. They were sat in front of the camera and told to act as if having a video conference with someone. This experiment was chosen because it is not of interest to determine the types of gestures used to draw an object as these may be standardized worldwide, it was of interest to know what the user himself would bring to the table. Bischels' experiment also brings together language and gesture; both of these being important in defining a culture (as stated in Section 2.B). Throughout this study, the observational task analysis method will be used. The observational technique, via the video footage, will permit a careful analysis of gestures occurring at certain timestamps during the interaction. This will be helpful in identifying individual gesture differences in task performance.

A. *Hypothesis*

The hypothesis taken as a base for the analysis is as follows:

"Designers' culture may affect gesture recognition in multimodal interfaces because of variations in gesture type, gesture frequency, and gesture occurrence".

This hypothesis brings together the subjects of gesture, multimodal interfaces, gesture segmentation and culture based theories. The three metrics stated in the hypothesis are gesture type, frequency and occurrence.

- Gesture Type. The gesture type is based on McNeill's classification. It is believed that certain types of gestures could be attributed to different cultures; therefore, it is important to analyze the type of gesture that is mostly performed.
- Frequency. The frequency is measured as the number of gestures performed by a participant divided by the period of the gesture of the same participant. This way we obtain the gestures per second which will help assess speed of gesture performance and point out what gestures are most significant for a gesture recognition system.
- Occurrence. Occurrence measures the appearance of the gestures. This once again tries to identify if certain gestures are culture-oriented or task-oriented (i.e., related to the task being performed).

B. *Experiment Guidelines*

The task to be performed consists of describing two chairs (See Figure 2). Participants were encouraged to use as many gestures as possible, just as in [21]. The analysis methodology is via video analysis using a video annotation tool called Anvil.

 ISBN: 978-1-61208-195-3

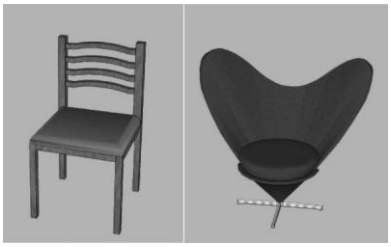


Figure 2. Classical chair (left) , Abstract chair (right)

1) Procedure

There were a total of 8 Latin American participants and 11 Anglo-Celtics videotaped, but only the ones with clearer hand gestures and comprehension of the task were chosen. A criterion for deselecting a video footage for analysis was either the lack of gestures, or the lack of iconic gestures which are the focus of this study.

The final selection was 5 participants from each sample group, totaling 10 participants. For the purpose of the experiment two samples were needed, one with English as a first language (Anglo-Celtics, and one with English as a second language (Latin Americans). For the second sample, it was important that they were sufficiently proficient and immersed in an English speaking country (Australia) for the past 6 months.

2) Gesture coding

In gesture analysis, we first analyzed the videos and segmented the video footages. For each occurrence, what was recorded was the name of the gesture type (repetition, beat, iconic, metaphoric, deictic junk) that was performed by the participant. These correspond to McNeill's classification, but the repetition gesture (which is a type of deictic gesture) was coded separately because of the difference in language. Repetition was considered to be a potential factor that reflects culture, as uncertainty in the language, or description, could be characterized this way. Junk gestures were identified as gestures without a particular meaning. This could be a gesture that the user takes the gesture back (which is a "mistake") or made some transition movements.

Gestures are separated by pauses, and a pause is defined as a temporary stop in action or speech [22]. The purpose of this pause was to eliminate the period of inactivity at the beginning of a video, when the participant explains what he or she might do, or when the participant states that he or she has ended.

TABLE 1. WORDS DERIVED FROM VIDEOS FOR CHAIR 1 AND CHAIR 2

Words	**Samples**			
	Anglo-Celtic	***Latin American***	***Both***	***Total***
Chair 1	9	13	6	28
Chair 2	13	10	5	30

3) Speech Coding

Words were not a requirement, yet the participants talked through their depictions. As a result, the "verbal descriptions more significantly used were coded. These were classified as adjectives, parts of the chair, verbs, order and shapes (See Table 1). It was found that each iconic and metaphoric gesture is related to at least one word, reflecting the participant's cognition.

Finally, we obtained close to 10 minutes of monologue object descriptions in a video footage. Seconds were used as the time measuring unit.

IV. RESULTS

Numerically, Anglo-Celtics did not display too much variation (SD) between chair descriptions, regardless of the second ones unordinary structure (See Table 2).

The Anglo-Celtic participants used more gestures on average to describe Chair 2. On the contrary, the Latin Americans used less number of gestures to describe the same chair. The reason behind this could be the degree of comfort in using a language when describing the abstract. This reflects how the language and increase of the complication of the task have an influence in cognition. Given that the features found in the abstract chair are not as common as the features found in the classic chair, this sample may have had more trouble in finding a way to explain words or shapes in the abstract chair.

The SD was again higher with the Anglo-Celtics, which made it hard to identify a pattern. On the other hand, the Latin Americans had a smaller SD and more frequent gestures, meaning shorter, concise, and common gestures by most of the participants. Their gesture frequency is higher in Chair 1, and increases in Chair 2. This could be because Latin Americans scored higher results in junk gestures in the second Chair.

Latin Americans had more frequent gestures in both chairs meaning that they performed more gestures per

TABLE 2. VIDEO ANALYSIS FOR CHAIR 1 AND CHAIR 2

Chair	**Metrics**						
	Sample	***Avg gesture duration***	***Total no of gestures***	***Avg gestures***	***SD***	***Avg gesture Time***	***Frequency***
Chair 1	Anglo-Celtic	1.84	65	12.8	5.63	22.74	0.56
Chair 1	Latin American	1.49	59	11.8	2.16	17.81	0.66
Chair 2	Anglo-Celtic	1.73	65	13	7.17	23	0.56
Chair 2	Latin American	1.67	43	8.6	2.88	14.22	0.60

 ISBN: 978-1-61208-195-3

second, even though they had fewer gestures in total. The smaller count of gestures by Latin Americans is justified by the lesser time in which they performed the gestures.

Given the distribution of gestures, we can see that in general, iconic gestures decrease with Chair 2, in contrast, junk and deictic gestures appear more.

The Latin Americans used more words for Chair 1 and less in Chair 2 (See Table 1). Less gestures and words in Chair 2 could probably mean a better selection of words and gestures, or the lack of vocabulary. The higher words count for Chair 1 must mean a higher degree of confidence, or more predictable and structured ideas on behalf of the Latin Americans.

A. Findings

After analyzing the performance of both samples, in this section we reveal the results of the metrics stated in the hypothesis: gesture type, frequency and occurrence.

1) Frequency

Gesture frequency indicates that overall the Latin American sample performed more gestures per second; however, this evidence is not enough to say that a certain sample was more expressive than the other. The use of gestures involves various factors, such as the comfort of a person had in front of the camera, or the confidence with the object being described, as well as the language. Chair 1 had Iconic and repetition gestures with higher frequency in both samples. Chair 2 on the other hand had an increase in junk and metaphoric gestures. The most significant gestures for the gesture recognition were the iconic ones as well as repetitions, and subsequently they are the ones that convey the representation of the chair.

2) Occurrence

There are no junk and deictic gestures in the description of Chair 1 for the Anglo-Celtic sample, but they do appear in Chair 2. We can see that number of gestures increases in Chair 2. This means that the occurrence of gestures was related to the task, not to the culture. Since Chair 2 was more complex and there was a need for more explanation by the user.

3) Gesture Type.

The results for gesture types show that in Chair 1, the iconic gestures were close to 50% in both sample groups. In Chair 2, the iconic gestures diminish and metaphoric gestures increase for the Latin American sample group. Again, this may be related to the complexity of the chairs.

V. DISCUSSION

Now we may relate the gesture metrics to the cultural attributions made by both Hofstede and Hall (Section 2.2). As Anglo-Celtics are low context cultures, they used more words and gestures in longer time, since they took time to explain the chair in detail. On the other hand, Latin Americans, which represent the high context culture, performed fewer gestures, in shorter time and used fewer words. The element that calls for attention is the higher use of metaphoric gestures, as this is a characteristic of a society that relies on reading between the lines and letting nonverbal cues explain the meaning.

Continuing with the culture analysis, we will now state the relation of the gesture performance with Hofstede's cultural dimensions. The connection between these dimension (experiment, cultural aspects, participants, results) are displayed in Table 3. As mentioned before, the traits that are mostly reflected are IDV, UAI, and MAS.

- IDV. This trait could be related in fact that the SD between samples is higher with the Anglo-Celtic cultures reflecting the societies high individualism index (IDV, 82). On the other hand, the low SD with the Latin Americans shows the low individualism index (IDV, 20).
- UAI. This trait could be reflected in the overall impression of Chair 2. The Anglo-Celtic sample did not vary too much in gesture means and time from one chair to another, showing greater comfort with adverse situations (UAI, 43). It is possible to say that Latin Americans showed their high uncertainty avoidance (UAI, 86) since they use less time and limited gestures, possibly sticking to "what they knew" instead of managing the abstract.
- MAS. This trait could be related to the fact that the Anglo-Celtics as a low context culture are more assertive (MAS, 63), in comparison to the Latin Americans that are more human-oriented and therefore there is a higher use of metaphors (MAS, 47) in their descriptions.

The Latin Americans in this sample have more of an advantage with the language as they have been immersed in the culture and language for the past 6 months. Regardless,

TABLE 3. INTEGRATION OF EXPERIMENT AND CULTURE

Sample	**Metrics**			
	Context	***Predominant culture trait***	***Metric Evidence***	***Predominant Gesture Type***
Anglo-Celtic	Low context (assertive, rely con words)	Individualism Masculinity	High SD Constant gestures beween chairs More gestures and more time	Iconic
Latin American	High context (rely heavily on non verbal communication)	High Uncertainty Avoidance Collectivism	Low SD Fewer Gestures in the second Chair Fewer gestures in less time	Metaphoric Repetition

they still performed fewer gestures and chose different words.

VI. CONCLUSION AND FUTURE WORKS

A. How these affect multimodal interfaces?

We started this paper in order to prove if multimodal interfaces could be affected by a user's culture. After the literature review, we have seen that any interaction is a result of user, task and input. Apart from performance or stability issues, multimodal interfaces are subject to a context problem. In the international scene, depending on where participants are from, their style of communication will vary. This analysis arrived to the conclusion that as the complexity of a task increases, so does the use and type of gestures. The metrics stated in the hypothesis influence multimodal interfaces and their performance in the following ways:

- Frequency may affect the recognition rate because of the need for faster, more efficient algorithms.
- Occurrence also affects interaction due to the possibility of absence (zero occurrences) of certain gestures that may convey functionality (i.e. iconic).
- Gesture type, as well as occurrence, also affects the goal that the user wishes to attain. Identifying and classifying certain gestures due to their use during trials would permit the identification of type tendencies and will help embed differences in the development of the gesture recognition tool.

Due to the "freedom" that hand gestures provide, gesture based interfaces gain popularity. The aim of HCI is to have users strongly prefer to adopt the new technologies for interaction because of the usability opportunities they provide. Culture influences a user's openness and a more conservative or traditional culture, like the Latin American, could take more time to adapt, this was visible with the frequency rate difference between the academic and abstract chair. The performance of a multimodal interface will not only be affected by the task being performed, but by the cultural background and language skills of the user. Therefore, the design of gesture-based interfaces not only requires a multidisciplinary approach, but also a culturally sensitive one.

We acknowledge that future studies need a larger sample size. Similarly, future studies may also work on the consistency of the annotations by having more than one person in charge of coding the gestures. Also, the results could have significant variations if the experiment is carried out in Spanish, the native language of the second sample. Further research studies can also attempt to investigate the effects of gender on performance.

ACKNOWLEDGMENT

This research is supported by the Australian Research Council Discovery grant DP0988088 to Kavakli, titled "A Gesture-Based Interface for Designing in Virtual Reality". Also, the experiment on which this project was based is inspired by Jing Liu's previous work "Temporal Relation between speech and co-verbal iconic gestures in multimodal interface design". Authors are grateful to her for her support in sharing data.

REFERENCES

[1] Wachs, P., Kölsch, M., Stern H., Edan Y. (2011) Vision-Based Hand-Gesture Applications. Communications of the ACM. vol. 54. no. 2., pp. 60-71 DOI:10.1145/1897816.1897838

[2] Kita, S. (ed.) (2003). Pointing and Placing. Pointing: Where Language, Culture and Cognition Meet. Retrieved April 2012 from http://www-psych.stanford.edu/~herb/2000s/Clark.Pointing.placing.03.pdf

[3] Bavelas J., Gerwing J., Sutton Ch., and Prevost D. (2008) Gesturing on the telephone: Independent effects of dialogue and visibility Journal of Memory and Language Volume: 58, Issue: 2, pp. 495-520 ISSN: 0749596X DOI: 10.1016/j.jml.2007.02.004

[4] Visser, W., and Maher, M.L. (2011).The Role of gesture in designing. Artificial Intelligence for Engineering Design, Analysis and Manufacturing Vol 25, pp. 213-220. Cambridge University Press DOI: 10.1017/S08900604/11

[5] Gullberg, M., (2010) Methodological reflections on gesture analysis in second language acquisition and bilingualism research. Second Language Research 26,1 ; pp. 75–102. DOI: 10.1177/0267658309337639

[6] Karray, F., Alemzadeh, M., Saleh, J., and Arab, N. (2008) Human-Computer Interaction: Overview on State of the Art. International journal on smart sensing and intelligent systems, 1,1 Retrieved April 2012 from http://www.s2is.org/Issues/v1/n1/papers/paper9.pdf

[7] Lanfranco, A., Castellanos, A., Desai, J., Meyers, W. (2004) Robotic Surgery. A Current Perspective Ann Surg. 2004 January; 239(1): pp. 14–21. DOI: 10.1097/01.sla.0000103020.19595.7d

[8] Purdue University (2011, February 3). Future surgeons may use robotic nurse, 'gesture recognition'. ScienceDaily. Retrieved April 2012, from http://www.sciencedaily.com-/releases/2011/02/110203152548.htm

[9] Van den Hoven, E., and Mazealek, A. (2010). Grasping Gestures: Gesturing with physical artifacts. Artificial Intelligence for Engineering Design, Analysis and Manufacturing, 25, pp. 255-271. DOI: 10.1017/S0890060411000072

[10] Kim, M.J., and Maher, M.L. (2008). The impact of tangible user interfaces on spatial cognition during collaborative design. Design Studies 29(3), pp. 222–253 DOI: 10.1016/j.bbr.2011.03.031

[11] Berry, M. (2009) The Importance of Bodily Gesture Example 1. McNeill's classification of gesture types Retrieved April 2012 from http://www.mtosmt.org/issues/mto.09.15.5/berry_ex1.html

[12] Hofstede, G. H., Hofstede G. J. ,and Minkov, M. (2010) Cultures and organizations: software of the mind : intercultural cooperation and its importance for survival. (3rd ed) McGraw-Hill Professional ISBN 0071664181, 9780071664189

[13] Murphie, A. , and Potts, J. (2003) Culture and Technology. N.Y: Palgrave Mcmillan

[14] Hall, S. (2007) Review: Representation: Cultural Representations and Signifying Practices (Culture, Media and Identities Series) London, SAGE Publications Limited

[15] Aaron Marcus and Associates, Inc (2010) Cross-Cultural User-Interface Design for Work, Home, and On the Way: Accounting for

 ISBN: 978-1-61208-195-3

Cultural Preferences, Acceptance, and Constraints. ACM SIGGRAPH ASIA 2010 Courses. DOI 10.1145/1900520.1900525

[16] Rehm, M., Bee, N., and André, E. (2008) Wave like an Egyptian — Acceleration based gesture recognition for culture-specific interactions. In Proceedings of HCI 2008 Culture, Creativity, Interaction, pp. 13–22, 2008.

[17] Yammiyavar, P., Clemmensen, T., and Kumar, J. (2008). Influence of cultural background on non-verbal communication in a usability testing situation. International Journal of Design, 2(2), 31-40. Retrieved April 2012 from http://www.ijdesign.org/ojs/index.php/IJDesign/article/view/313/164

[18] Hofstede, G. (2011) Dimensions of national Cultures Retrieved April 2012 from http://www.geerthofstede.nl/culture/dimensions-of-national-cultures.aspx

[19] Hall, E.T. (1976) Beyond Culture. Anchor Books

[20] Changing Minds (2011) Hall's cultural factors. Retrieved April 2012 from http://changingminds.org/explanations/culture/hall_culture.htm

[21] Liu J., and Kavakli, M. (2011) Temporal Relation between speech and co-verbal iconic gestures in multimodal interface design. Retrieved April 2012 from http://coral2.spectrum.uni-bielefeld.de/gespin2011/final/Liu.pdf

[22] Pause (2011) In Oxford Dictionaries. Retrieved April 2012 from http://oxforddictionaries.com/definition/pause

 ISBN: 978-1-61208-195-3

Automatic Discovery and Composition of Multimedia Adaptation Services

Jean-Claude Moissinac
Telecom ParisTech, LTCI UMR 4141
46 rue Barrault
F-75 634 Paris cedex
jean-claude.moissinac @ telecom-paristech.fr

Abstract—**In this paper, after presenting the context, which indicates a considerable increase in the need for the adaptation of multimedia documents, we show that these results can be obtained by the composition of basic services. Nevertheless, this requires the availability of semantic descriptions of services, for which a shared vocabulary and good practices still need to be defined. We identify a series of works that can contribute to this process and offer basic guidelines to establish these descriptions. This article especially highlights the importance of the development of semantic descriptions of several important families of multimedia processing and proposes a structure that is used to build and organize such descriptions.**

Keywords-multimedia; semantic web services; adaptation; service composition.

I. INTRODUCTION

Our environment is enriched every day by a greater number of communicating devices and multimedia document providers. From a user point-of-view, each of us today takes advantage of a finite number of devices, usually personal: a telephone, television, laptop, tablet, portable media player. The great variety in the features offered by each of these devices requires services returned to the terminals that are adapted to them. Tomorrow we'll probably be able to benefit from the functions offered by equipment found in the different places we move to [29].

From a provider of multimedia content point-of-view, this growing complexity is a problem. A provider is often obliged to offer the same multimedia content in several formats and presentations. The current methods of adaptation are not sufficient to cope with the variability of situations that must be taken into account: preferences or needs of users, equipment available, and context of use.

In this paper, we show why this situation makes it necessary to implement adaptation processes that are widely configurable and propose a methodology to do this.

Given the variety of multimedia documents that users are exchanging, it is difficult to require a producer of multimedia content to provide as many versions of a document as possible contexts of use. It is necessary to identify ways to adapt a variety of documents to different contexts, either known at the time the content is put online or unknown until the time of the consultation.

We consider it desirable to offer to Internet players the ability to provide processing resources for the adaptation of multimedia documents. We must define the methodology and establish the prerequisites to allow such operations.

Section 2 presents two usage scenarios that illustrate the need for the dynamic processing of service compositions for multimedia. Section 3 presents a set of technologies and works which can contribute, or have contributed to, the proposal of this article. Section 4 describes a general architecture for adaptation of multimedia documents. Section 5 provides guidance for the descriptions of processing services which focus on our work. Finally, Section 6 presents the next steps as we see them.

II. EXAMPLE SCENARIOS

To light the way, we present two usage scenarios, one inspired by [30] and [14] as an extension of work published in 2004 which is centered on the user, the other responding to the needs of multimedia providers.

A. Campus scenario

We assume that we are on the campus of an international university. Some courses are available as multimedia documents.

There are different situations in which the content is used: during a classroom course, to follow and annotate the current presentation; at home to learn; or, later, when using the knowledge acquired during the course.

Users access to that content in various ways as well. For example, a user preferring English might be using a terminal with a small screen (5cm x 5cm) and a good resolution (800x 600) with Wi-Fi access while another will be on a wired network with a large screen, and prefers Spanish. One user might be in a location where he can activate the sound, another not. Disabled users can be taken into account; for example, the text will be displayed larger for the visually impaired or will be converted by a Text-To-Speech utility if the context permits.

Finally, the emergence of new devices, tablets, media players with new features for restitution of the media, but also the ability to interact, requires taking into account these new modes of access.

In this scenario, it appears necessary to have a system that dynamically configures itself to provide the best adaptation of a multimedia document in a context only known at the time of the request. The system cannot be

 ISBN: 978-1-61208-195-3

limited to a fixed set of adaptations. It must be able to be configured dynamically and be extensible.

B. Broadcast ecosystem scenario

Another scenario can be found in the broadcast production industry, which needs to adapt a lot of content to many different user contexts.

The media industry has many new opportunities to exploit its productions and archives: mobile multimedia, on-demand content, new products built on archives, etc. To do that, the media industry must do a lot of various processing, dependent on the target.

To achieve this aim, the media industry must be able to provide different sort of processing, depending on the targeted user context; such modern media production facilities must to function enable to compose processes from a rich list of elementary processes such as transfer and storage, capture, transform, etc. This scenario is illustrated in the current effort of standardization of FIMS [8].

Our contribution focuses on the development of semantic descriptions of basic adaptation services, based on ontologies. These descriptions help to meet the need mentioned above, but may have many other uses in applications of Semantic Web Services. In the next section, we will discuss a series of works that contribute to, and complement our approach.

III. RELATED WORKS

In this section, we will discuss a set of works that may contribute to the approach presented in this document.

First, we assume that basic services will be accessed via the Internet. We include them in the generic class of Web services, either REST [7] or SOAP services [20] or other technologies to make services available on the Internet.

In order to achieve automated operation of these services, they must have a description formally usable by IT processes. A minimum concerns the description of each service interface; for this, the most common technology is WSDL [1]. We will see below that this is not enough.

We want to use a set of basic services that will work together to create a more complex service. For this, many works concern the composition of services. They generally focus on the fact that a developer creates a process by calling a set of Web Services. Major efforts are focused on this type of software production process applied to 'business' in business. A language emerged to describe the workflow created to oversee the services called: WS-BPEL[13]. 'runtimes' are able to supervise the execution of a process defined with WS-BPEL (e.g. we are working with ODE [23]).

Developers can read a service specification written in plain text to understand its role or do a search in a warehouse of services such as UDDI [23]- to find a service that meets their expectations.

To create an automatic dialing service, the WSDL description is not enough to describe the fact that it takes as input an image given by an URL to access it in the Internet; or to understand what transformation is applied to the image –the transformation is only known by its name. We need to have the role and effects of each service described: which is the role of a semantic description of services. Several techniques for semantic description of services have been proposed, including: SAWSDL [6], OWL-S [19], WSML [3]. The use of OWL-S to describe media adapters, for example, has been proposed as part of MPEG-21 [22]. The need arose to describe some effects of a service using rules. In common parlance, such a rule can define a part of the effect of an operation to resize an image 'if the object has a media width and that the service is applied, then the width of the media object will be changed'. The SWRL language [10] was proposed to represent such rules.

Planning an automated composition of services has so far resulted in only a few works. As for multimedia, the proposed solutions are, for example, heuristics [15], a systematic exploration of possibilities [16] or more complex methods based on rules describing a form of expertise [35][12][28]. An interesting solution was proposed by [8]. And it concluded, however, with the idea that ontology for multimedia adaptation services could help to solve the problems left open by the proposed solution. The search for such ontology and how to use it is at the heart of our proposal. More recently, [23] proposes a way to describe services with the goal to automatically build mashups. This work focuses on problems of automatically composing services with heterogeneous descriptions in heterogeneous domains and gives ideas on how to solve that important problem. Our work focuses on getting good enough descriptions in one domain, multimedia, to establish either widely used standard descriptions or to easily make a match from our description to another. In [34], we find an in-depth analysis of a composition process in the aim of performing various semantic analyses on multimedia content.

Very significant work was carried out around these concepts in the context of the European Initiative ESSI: WSML [3] is a language defined to formalize the modeling of web services offered by WSMO [35]; WSMX [36] defines a runtime environment and set of services.

Numerous studies have focused over the last decade on the adaptation of multimedia documents. For examples and significant elements of state of the art in this work; see [23] [32]. Pellan [23] proposed a method to directly choose an appropriate service, knowing an initial application and context. The proposal focuses on choosing a service tailored to a context; it has a real contribution in the way to obtain the appropriate service itself in a space of predefined variant of the service.

 ISBN: 978-1-61208-195-3

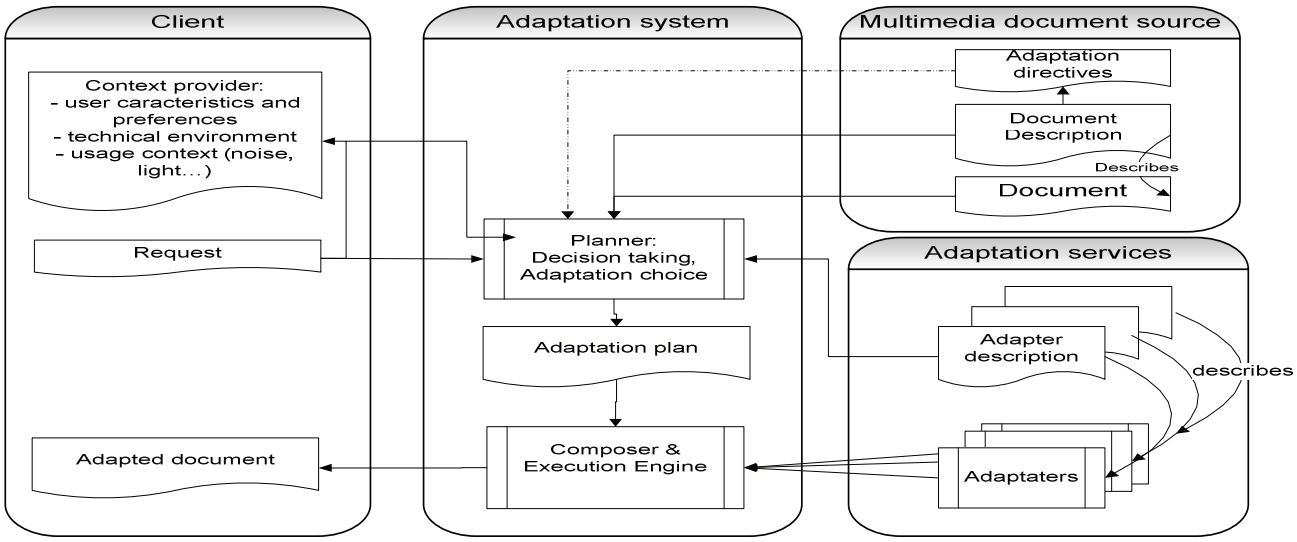


Figure 1. General architecture for distributed adaptation

MPEG-21 DIA [22] defines the desired adaptations to the media composing a document. This approach is insufficient because there is no possibility of describing dependencies between the different media adaptations. For example, if I describe that below a certain screen size, I no longer display a certain image, I have freed the space that can be used for text. But this dependence on the adaptation of the text based on that of the image cannot be described by DIA.

Even today, in many cases, adaptation is performed either at the server level by responding to a request with a different answer depending on what is known about the context of issuance of the request, or at the client -for example, the script by exploiting what is known locally in the terminal and about its user. The notion of proxy adaptation was introduced and is used by the network industry [14][23].

A very thorough study of the semantic description of multimedia services has been led by Bernhard Reiterer; the results are mainly available in German [28]. Very little research focuses on automatic composition applied to multimedia services; Derdour et al. [5] proposes a methodology for assembling basic services that provides services via mediation in order to make the entries of some services compatible with the output of other services.

IV. GENERAL ARCHITECTURE FOR DISTRIBUTED ADAPTATION

Fig. 1 shows the general principle of a distributed adaptation system as envisioned in this article.

A user requests a multimedia document or service. His request passes through an adaptation system. It is accompanied by an explicit or implicit context of use. We have presented in [21] an extended description of the following concepts. The main parts of an adaptation system are:

- the planner: it takes as input a description of the multimedia document and a description of a context and produces an adaptation graph, which describes the composition of a set of elementary steps, possibly subject to conditions, performed in parallel or in sequence, resulting in the appropriate document;
- context provider: as many works deal with the collection and provision of context, we leave this question out of our field of research and consider that a 'black box' is available and provides a context; there is a dependency between the context provider and the planner: the planner must be able to understand and use the context model of the context send by the context provider the context provider gives a context on demand or send an event each time a change occurs in the context;
- the source of multimedia documents: a component must manage the access to the multimedia document and its metadata description (source, nature...) which is to be adapted
- the composer searches for the needed services, while the runtime executes the plan and supervises the execution of selected services; at the end, it provides the result or a link to the result,
- elementary adapters: these components provide a specific adaptation for a part of the document.

The general principle is as follows. A consumer initiates an adaptation cycle. He/she sends a request to the planner; part of this request consists of a reference to a multimedia document and part of a reference to a context. The planner

 ISBN: 978-1-61208-195-3

uses this information to apply its adaptation algorithms and find a plan. It decides what will be the sequence of adaptation operations. The adaptation plan is sent to the composer who seeks elementary adapters to compose the ready to execute representation of the plan. The execution of the sequence is supervised by the runtime and returns a reference to access the resulting document.

V. DESCRIPTION OF ADAPTERS

A. Adaptaters for basic media

Most of the adapters we want to consider perform an elementary operation on a media category.

We first need to list the media types which must be taken into account. Beyond an obvious list (text, image, audio, video), we believe it is necessary to consider other media. Two examples are:

- A musical score, which is not a text or an image or sound, although it may be transformed into these three forms,
- A map, which is neither text nor an image, but an object which is much more complex.

In the present state of our descriptions, we have also introduced: 2D graphics, animated 2D graphics, 3D graphics and animated 3D graphics.

It is necessary to establish a methodology that allows some extensions, to define a new media that is useful, for example, in a specific activity, or some specializations, e.g. to distinguish speech from music as two kinds of audio documents.

We also see that some media have evolve in time; we are not yet sure of the best classification to be adopted. (Is a 2D graphic a degenerate case of an animated 2D graphics? Is an animated 2D graphic a 2D graphic extended by a description of a temporal process?)

Each media type must be associated with a list of characteristics that define it. Many ontologies have attempted to define the most common media and their representative characteristics. This situation is due to the fact that each ontology has its relevance in a given application. The W3C has taken steps to ensure correspondence among the models whenever possible [33]. In a preliminary study in 2007 [10], about two dozen models for describing media types, at least some media types, were identified. We found more during our work. In the work on WSMX, the concept of 'Data Mediation' [1] was introduced and can be a way to cope in an automated way with this problem. We find a similar concept in [5] in the work on the adaptation of input/output in a UML diagram representing the processing of a multimedia document.

B. Adapters

In [21], we described the top-level categories - transmoder, transformer, transcoder, extractor, composer- that we use as the basis for the definition of service classes.

- transmoder: changes a media from one modality to another –like creating an image of a text,
- transcoder: changes the format used to code a media without changing any other parameter –like transcoding an image from Jpeg to PNG)
- transformer: changes one or several intrinsic parameters of a media –like changing the size of an image,
- extractor: extracts each media and rules of composition from a composed multimedia document,
- composer: creates a composed multimedia document from a set of media and some rules to compose them.

These categories are then refined according to the media they take as input, the output they provide and the changes they perform on the media. We undertook a systematic description of adapters and have already identified about forty relevant types of adapters.

For example:

- text to speech is a text to audio transmoder whose input is mainly a text and output is an audio sequence,
- scaling of an image is a transformer that goes from one image to another image by changing certain characteristics.

We can see the existing services as being instanciated from the semantic description of some classes of services:

- class ‘transformer/scaling’, applicable to several media types: image, video, 2D graphics, animated 2D graphics,
- class ‘transformer/summary’ applicable to text, video, audio...

Whenever possible an adapter will be in one of the main classes. One last class has been defined to contain all adapters that are not clearly an instance of one of the previous classes. This last class is to be avoided because it conveys the poorest semantic. This class will include such additional adapters specific to a particular treatment on a type or a specific document format, for example, an adapter for PowerPoint documents or any document type specific to a specific activity domain.

Each adapter must have a basic WSDL description to conform to the call mechanisms of SOA services. But, as is now well identified, WSDL only provides technical information on how the call is made and no information on the meaning of the parameters, the nature of the result, the preconditions for the call or the effect of service execution on the surrounding world. To some extent, this information will be inherited from the ontology. However, each service may need a specific description not defined by the ontology, for example:

- the type of a parameter does not have an exact correspondence in the ontology and we need to define the mapping between the types and provide

 ISBN: 978-1-61208-195-3

the type expected by the service (e.g., string versus URL),

- constraints on a parameter, for example, the width and height to resize an image can be limited to be homothetic,
- technical constraints imposed by the instance of the process, for example, the size of data transmitted is limited,
- pre-conditions of a nature that have nothing to do with the functions of the service (access control, security, etc.) can be attached to a service; these conditions involve concepts that are beyond the functional area under consideration -multimedia, and other concepts are necessary: other descriptors, other ontologies.

Several technical solutions have been proposed to complete the WSDL description for the semantic enrichment. We have begun experimenting with various solutions (OWL-S, WSML, SAWSDL and proprietary descriptions by extending WSDL). Apart from these experiments, we are trying to define the necessary descriptors, independently from a description language.

We believe that conceptually it is not the media that is provided to the adapter, but the access method of the media. In practice, the adapters receive as input a URI to access the media.

All descriptions deemed necessary in the context of Semantic Web Services (SWS) are often referred to by the acronym IOPE, Inputs Outputs-Preconditions Effects.

In the case of media processing, the minimum is to determine which characteristics of the media were changed and which descriptors are useful for the result of the transformation.

Following the work of [28], we consider different versions of the same multimedia document as variants of this document. We can describe the result of an adaptation, not exhaustively, which would not be possible, but only through changes made to formally described characteristics of the original.

As a general principle, we will consider that all the attributes of a transformed media remain the same, with the exception of those whose transformation is described. This has an advantage if a descriptor is added to a future version of a transformation: the adapters that were based on the current ontology work without that descriptor, by default, as if it were granted that they do not change the descriptor; this hypothesis seems relevant because, if it were not the case, it would mean that when we had done the initial description of the adapter, we forgot an important part of the description.

Consider two services to reduce the size of a picture, cropping and scaling. In both cases, the result is an image, which is a variant of the original image. In both cases, the width and height characteristics of the images are changed by the transformation. What differentiates the two transformations is that in the case of a crop, the image portion resulting from the processing is extracted from the original image while in the case of a change of scale, the resulting image is the result of the processing of all the image data. Most of the other features remain unchanged and can be skipped from the description. The amount of data used to represent the image is –generally- changed; we must mention that fact in the description.

Through this example, we see how difficult it can be to describe the adapters, but also the richness of the approach to build a large catalog of such descriptions. We undertook this work, which is being refined gradually; we are aware of similar work, for example at the University of Klagenfurt, but it seems that all the research projects we have identified are currently stopped.

The scientific community on multimedia adaptation and media processing and the one on Semantic Web Services will benefit from progress on these types of descriptions. Collective work will be necessary to achieve the goal of establishing shared concepts and vocabulary, to design a formal representation and to create the tools to facilitate the specification of new services based on the proposed model.

C. Adaptation of a multimedia service

In the work of Pellan [23], three dimensions of an adaptation process of a multimedia service are to be taken into account: a spatial adaptation, a temporal adaptation, and an adaptation of interactions.

We have begun to take into account the depth of all three dimensions for all media types and all categories of adaptation, but this work must still be completed.

We retain the assumption of Pellan: useful results can be obtained by considering that these three dimensions can be treated independently and that a composition of adaptations selected along each axis can be chosen.

On one hand, works such as those of [15] propose a method to adapt the layout of a document (spatial adaptation). On the other hand, we are exploring the possibilities of abstract representations of interactions [3][33][35], which could then allow concrete instantiations adapted to each situation.

VI. CONCLUSION AND FUTURE WORKS

We believe that the proposed approach will, in the future, be followed by other research. Indeed, it responds to the need to move from distributed storage on the Web to distributed processing. This approach benefits from unused software resources and from available bandwidth and processing capabilities on the Internet, usable in a decentralized manner and dynamically reconfigurable. These features could be a major asset for the spread of pervasive computing.

We think that describing all the known categories of multimedia services is possible; we have identified more than 50 categories and, probably, the categories added in the future will remain under a total of 100. Our main results are in the structuration of the categories, the principles of the description of each category and a first description of a group of categories.

 ISBN: 978-1-61208-195-3

Our future work is to complete the list of categories, clearly describing them all and, most importantly, to publish and share these descriptions to encourage their adoption.

ACKNOWLEDGMENT

This work was supported by the SOA2M project in the UBIMEDIA LAB, common with Alcatel/Lucent Bell Labs.

REFERENCES

[1] Christensen E., Curbera F., Meredith G., and Weerawarana S., "Web Services Description Language (WSDL) 1.1" http://www.w3.org/TR/wsdl [retrieved: April, 2012].

[2] Cimpian E., "Process Mediation in Semantic Web Services" 4th European Semantic Web Conference, June 2007, pp. 16-20, Innsbruck, Austria.

[3] Coutaz J., "PAC: an object oriented model for implementing userinterfaces" SIGCHI Bulletin Oct. 1987 Vol. 19 Num. 2

[4] Bruijn J. de, et al. "Web Service Modeling Language" http://www.w3.org/Submission/WSML [retrieved: April, 2012].

[5] Derdour M., Zine Ghoualmi N., Roose P., and Dalmau M., "UML Profile for Multimedia Software Architectures" Multimedia International Journal of Intelligence and Security, pp. 209-231, 2010

[6] Farrell J., et al. "Semantic Annotations for WSDL and XML Schema" http://www.w3.org/TR/sawsdl/, [retrieved: April, 2012].

[7] Fielding R., "Architectural Styles and the Design of Network-based SoftwareArchitectures" PhD Dissertation, University of California, Irvine

[8] "FIMS Media SOA Framework 1.0", September 2011, AMWA-EBU Specification

[9] Girma B, Brunie L, and Pierson J. "Content Adaptation in Distributed Multimedia Systems" Journal of Digital Information Management, special issue on Distributed Data Management, Vol. 3 No. 2, June 2005

[10] Hausenblas M. et al., "Multimedia Vocabularies on the SemanticWeb" www.w3.org/2005/Incubator/mmsem/XGR-vocabularies-20070724/ [retrieved: April, 2012].

[11] Horrocks I., Patel-Schneider P.F., Boley H., Tabet S., Grosof B., and Dean M., "SWRL: A Semantic Web Rule Language", http://www.w3.org/Submission/SWRL/ [retrieved: April, 2012].

[12] Jannach D., Leopold K., Timmerer C., and Hellwagner H., "A knowledge-based framework for multimedia adaptation", Applied Intelligence, 24 (2), pp. 109-125.

[13] Jordan D., et al. "Web Services Business Process Execution Language Version 2.0" http://docs.oasis-open.org/wsbpel/2.0/OS/wsbpel-v2.0-OS.html [retrieved: April, 2012].

[14] Kazi-Aoul Z., Demeure I. and Moissinac J.C., "PAAM: a Web Services Oriented Architecture for the Adaptation of Multimedia Composed Documents", "Parallel and Distributed Computing and Networks (NCSP)", Innsbruck, Austria.

[15] Kimiaei-Asadi M. and Dufourd J.C., "Context-Aware Semantic Adaptation of Multimedia Presentations", Proc.IEEE International Conference on Multimedia & Expo (ICME 2005), Amsterdam, pp. 362-365, July 2005

[16] Lardon J., Gravier C., and Favolle J. (2010), "DOM tree estimation and computation: overview of a new web content adaptation system". In Proceedings of the 2nd ACM SIGCHI symposium on Engineering interactive computing *systems* (EICS '10). ACM, New York, NY, USA, pp. 357-360.

[17] Lee W. "Ontology for Media Resource", http://www.w3.org/TR/mediaont-10/ (retrieved 02/28/2011)

[18] Lopez-Velasco C., Villanoca-Oliver M. Gensel J., and Martin H. « Adaptabilité à l'utilisateur dans le contexte de services Web, Extraction des connaissances : état et perspectives » Revue Nouvelle des Technologies de l'Information, RNTI-E, ed. Cépaduès, 2005, pp. 153-158

[19] Martin D., et al. "OWL-S: Semantic Markup for Web Services" http://www.w3.org/Submission/OWL-S/ [retrieved: April, 2012].

[20] Mitra N. et al., « SOAP Version 1.2 Part 0: Primer » (link check: 03.03.2011) http://www.w3.org/TR/2003/REC-soap12-part0-20030624/ [retrieved: April, 2012].

[21] Moissinac J.C., Demeure I., and Kazi-Aoul Z., « Services d'adaptation de contenus multimédia, composition de services et pair-à-pair », CRIMES 09, St-Denis de la Réunion

[22] "MPEG-21 ISO / IEC JTC1/SC29/WG11/N6168, MPEG-21 Part7: Digital Item Adaptation", March 2007

[23] OASIS, http://www.oasis-open.org/specs/index.php#uddiv [retrieved: April, 2012].

[24] ODE, http://ode.apache.org/ [retrieved: April, 2012].

[25] Pellan B. and Concolato C., "Scalable Multimedia Authoring of Documents", Multimedia Tools and Applications, vol. 43, No. 3, pp. 225-252.

[26] Pietschmann S., Radeck C., and Meißner K. "Semantics-Based Discovery, Selection and Mediation for Presentation-Oriented Mashups", in Proc. Of Mashups 2011, September 2011; Lugano, Switzerland

[27] Rabin, J. "Guidelines for 1.0 Web Content Transformation Proxies" http://www.w3.org/TR/ct-guidelines/ [retrieved: April, 2012].

[28] Reiterer B., "Beschreibung von Multimedia-Adaptierungsoperationen als Semantic Web Services" PhD dissertation, Klagenfurt University

[29] Rodriguez B.H., Moissinac J.C., and Demeure I., "Multimodal pervasive services for the semantic web", UBIMOB 2010, Lyon, France

[30] Saathoff C. and Scherp A. "Unlocking the Semantics of Multimedia Presentations in the Web with the Multimedia Metadata Ontology "

[31] Salomoni P., Mirri S., Ferretti S., and Roccette M. "E-Learning Galore!Providing Quality Educational Experiences Across a Universe of Individuals with Special Needs Through Distributed Content Adaptation", LILW 2007

[32] Scherp A. "A Component Framework for Personalized Multimedia Applications" PhD Dissertation, University of Oldenburg

[33] Vella G. "XPL, a Presentation Language based on User Interface Design Pattern" in Proc. of 6th International Conf. on Computer and Information Science, 2007, pp. 285-290.

[34] Verborgh R., Van Deursen D., Mannens E., Poppe C. and Van de Walle R. "Enabling context-aware multimedia annotation by a novel generic semantic problem-solving platform" Multimed Tools and Applications, 2011

[35] WSMO, http://www.wsmo.org [retrieved: April, 2012] .

[36] WSMX, http://www.w3.org/Submission/WSMX [retrieved: April, 2012] .

[37] Yanagida T., Nonaka H., and Kurihara M. "User-Preferred Interface Design with Abstract Interaction Description Language" in IEEE International Conference on Systems, Man, and Cybernetics, Taipei, vol. 3, pp. 2458-2463.

[38] Zeng L., Benatallah B., Dumas M., Kalagnanam J., and Sheng Q.Z., "Quality Driven Web Services Composition", Proc.International WWW Conference, 2003, Budapest., pp. 411-421

 ISBN: 978-1-61208-195-3