MMEDIA 2011

The Third International Conferences on Advances in Multimedia

April 17-22, 2011

Budapest, Hungary

MMEDIA 2011 Editors

Dumitru Dan Burdescu, University of Craiova, Romania

Philip Davies, Bournemouth and Poole College, UK

David Newell, Bournemouth University, UK

MMEDIA 2011

Foreword

The Third International Conferences on Advances in Multimedia [MMEDIA 2011], held between April 17 and 22 in Budapest, Hungary, provided an international forum by researchers, students, and professionals for presenting recent research results on advances in multimedia, mobile and ubiquitous multimedia and to bring 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 humans 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 requires substantial effort to integrate both.

The new technological achievements in terms of speed and the quality are expanding and creating a vast variety of multimedia services like voice, email, short messages, Internet access, m-commerce, to 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 2011 Technical Program Committee, as well as the numerous reviewers. The creation of such a broad and high quality conference program would not have been possible without their involvement. We also kindly thank all the authors who dedicated much of their time and efforts to contribute to MMEDIA 2011. 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 2011 organizing committee for their help in handling the logistics and for their work to make this professional meeting a success.

We hope that MMEDIA 2011 was a successful international forum for the exchange of ideas and results between academia and industry and for the promotion of progress in multimedia.

We are convinced that the participants found the event useful and communications very open. We also hope the attendees enjoyed the historic charm of Budapest, Hungary.

MMEDIA 2011 Chairs:

Petros Belimpasakis, Nokia Research Center, Finland
Yannick Benezeth, Orange Labs (France Telecom Research Center in Rennes), France
Laszlo Böszörmenyi, University Klagenfurt, Austria
Trista Chen, Gracenote Inc. / Sony Corporation of America, USA
Noël Crespi, Institut Telecom, France
Dumitru Dan Burdescu, University of Craiova, Romania
Philip Davis, Bournemouth and Poole College, UK
Antonio Liotta, Eindhoven University of Technology, The Netherlands
Jonathan Loo, Middlesex University - Hendon, UK
Alexander C. Loui, Kodak Research Labs / Eastman Kodak Company-Rochester, USA
Tao Mei, Microsoft Research Asia, China
David Newell, Bournemouth University, UK
Joel Rodrigues, Instituto de Telecomunicações / University of Beira Interior, Portugal
Patrice Rondao Alface, Alcatel-Lucent Bell Labs - Antwerp, Belgium
Sandra Sendra Compte, Polytechnic University of Valencia, Spain

MMEDIA 2011

Committee

MMEDIA Steering Committee

Laszlo Böszörmenyi, University Klagenfurt, Austria
Dumitru Dan Burdescu, University of Craiova, Romania
Philip Davies, Bournemouth and Poole College, UK
Antonio Liotta, Eindhoven University of Technology, The Netherlands
David Newell, Bournemouth University, UK

MMEDIA Advisory Chairs

Joel Rodrigues, Instituto de Telecomunicações / University of Beira Interior, Portugal
Noël Crespi, Institut Telecom, France
Jonathan Loo, Middlesex University - Hendon, UK

MMEDIA Industry/Research Chairs

Petros Belimpasakis, Nokia Research Center, Finland
Alexander C. Loui, Kodak Research Labs / Eastman Kodak Company-Rochester, USA
Tao Mei, Microsoft Research Asia, China
Yannick Benezeth, Orange Labs (France Telecom Research Center in Rennes), France
Patrice Rondao Alface, Alcatel-Lucent Bell Labs - Antwerp, Belgium
Trista Chen, Gracenote Inc. / Sony Corporation of America, USA

MMEDIA Publicity Chair

Sandra Sendra Compte, Polytechnic University of Valencia, Spain

MMEDIA 2011 Technical Program Committee

Max Agueh, LACSC - ECE Paris, France
Hakiri Akram, Université Paul Sabatier - Toulouse, France
Nancy Alonistioti, N.K. University of Athens, Greece
Giuseppe Amato, ISTI-CNR, Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo" - Pisa, Italy
Ramazan S. Aygun, University of Alabama in Huntsville, USA
Eduard Babulak, The University of the South Pacific - Suva, Fiji
Werner Bailer, Joanneum Research Forschungsgesellschaft mbH - Graz, Austria
Andrew D. Bagdanov, Universita Autonoma de Barcelona, Spain
Petros Belimpasakis, Nokia Research Center, Finland
Yannick Benezeth, Orange Labs (France Telecom Research Center in Rennes), France
Sid-Ahmed Berrani, Orange-ftgroup, France
Laszlo Böszörmenyi, University Klagenfurt, Austria
Marius Brezovan, University of Craiova, Romania
Dumitru Burdescu, University of Craiova, Romania
Helmar Burkhart, Universität Basel, Switzerland
Rodrigo Capobianco Guido, University of Sao Paulo, Brazil
Eduardo Cerqueira, Federal University of Para, Brazil
Damon Chandler, Oklahoma State University - Stillwater, USA

Vincent Charvillat, ENSEEIHT/IRIT - Toulouse, France
Kuan-Ta Chen, Academia Sinica, Taiwan
Shu-Ching Chen, Florida International University - Miami, USA
Trista Chen, Gracenote Inc. / Sony Corporation of America, USA
Nicola Corriero, University of Bari, Italy
Noël Crespi, Institut Telecom, France
Philip Davies, Bournemouth and Poole College, UK
Vincenzo De Florio, University of Antwerp & IBBT, Belgium
Thierry Declerck, DFKI GmbH - Saarbrücken, Germany
Manfred del Fabro, Klagenfurt University, Austria
David Doermann, University of Maryland - College Park, USA
Vlastislav Dohnal, Masaryk University, Brno, Czech Republic
Nick Evans, EURECOM - Sophia Antipolis, France
Fabrizio Falchi, ISTI-CNR, Pisa, Italy
Lorenzo Favalli, University of Pavia, Italy
Alexander Felfernig, Graz University of Technology, Austria
Farshad Fotouhi, Wayne State University - Detroit, USA
Tapio Frantti, VTT Technical Research Centre of Finland, Finland
Gerald Friedland, International Computer Science Institute - Berkeley, USA
Kaori Fujinami, Tokyo University of Agriculture and Technology, Japan
Eugen Ganea, University of Craiova, Romania
Lefteris G. Gortzis, University of Patras, Greece
Valerie Gouet-Brunet, Conservatoire National des Arts et Métiers - Paris, France
Christos Grecos, University of the West of Scotland, UK
Stefanos Gritzalis, University of the Aegean - Karlovassi, Greece
William I. Grosky, University of Michigan-Dearborn, USA
Marcin Grzegorzek, University of Siegen, Germany
Angela Guercio, Kent State University at Stark - North Canton, USA
Hermann Hellwagner, Klagenfurt University, Austria
Benoit Huet, Eurecom Sophia-Antipolis, France
Razib Iqbal, University of Ottawa, Canada
Juergen Jaehnert, Universität Stuttgart, Germany
Young Sub Jo, Hyundai Mobis Co. Ltd., Korea
Eleni Kaplani, TEI of Patra, Greece
Dimitrios Katsaros, University of Thessaly - Volos, Greece
Markku Kojo, University of Helsinki, Finland
Yiannis Kompatsiaris, Informatics and Telematics Institute, Greece
Harald Kosch, University of Passau, Germany
Lambros Lambrinos, Cyprus University of Technology - Limassol, Cyprus
Thi Hoàng Ngân Lê, Concordia University - Montreal, Canada
Shiguo Lian, France Telecom R&D (Orange Labs) Beijing, China
Anthony Y. H. Liao, Asia University, Taiwan
Antonio Liotta, Eindhoven University of Technology, The Netherlands
Jonathan Loo, Middlesex University - Hendon, UK
Alexander C. Loui, Kodak Research Labs / Eastman Kodak Company-Rochester, USA
Hongli Luo, Indiana University - Purdue University Fort Wayne, USA
Cristina Mairal Garcés de Marcilla, LACSC-ECE, France
Sathiamoorthy Manoharan, University of Auckland, New Zealand
Ketan Mayer-Patel, University of North Carolina - Chapel Hill, USA
Tao Mei, Microsoft Research Asia, China
Vlado Menkovski, Eindhoven University of Technology, The Netherlands
Vasileios Mezaris, Informatics and Telematics Institute Centre for Research and Technology Hellas - Thermi-Thessaloniki, Greece

Annett Mitschick, Technical University of Dresden, Germany
Ayman Moghnieh, Universitat Pompeu Fabra - Barcelona, Spain
Parag S. Mogre, Technische Universitaet Darmstadt, Germany
Jean-Claude Moissinac, TELECOM ParisTech, France
Mu Mu, Lancaster University, UK
Jogesh Muppala, The Hong Kong University of Science and Technology, Hong Kong
David Newell, Bournemouth University, UK
Petros Nicopolitidis, Aristotle University of Thessaloniki, Greece
Jordi Ortiz Murillo, University of Murcia, Spain
Eleni Patouni, University of Athens, Greece
Tom Pfeifer, Waterford Institute of Technology, Ireland
Wei Qu, Graduate University of Chinese Academy of Sciences, China
Gianluca Reali, Università degli Studi di Perugia, Italy
Bernhard Rinner, Klagenfurt University, Austria
Joel Rodrigues, Instituto de Telecomunicações / University of Beira Interior, Portugal
Piotr Romaniak, AGH University of Science and Technology - Krakow, Poland
Patrice Rondao Alface, Alcatel-Lucent Bell Labs - Antwerp, Belgium
Tarek Saadawi, City University of New York, USA
Reza Sahandi, Bournemouth University, UK
Klaus Schöffmann, Klagenfurt University, Austria
Patrick Seeling, University of Wisconsin-Stevens Point, USA
Anita Sobe, Klagenfurt University, Austria
Yuqing Song, Chinese Academy of Sciences, China
Peter L. Stanchev, Kettering University - Flint, USA
Liana Stanescu, University of Craiova, Romania
Stephan Steglich, Fraunhofer FOKUS - Research Institute for Open Communication Systems, Germany
Cosmin Stoica Spahiu, Univeristy of Craiova, Romania
Mehmet R. Tolun, Cankaya University, Turkey
Nicolas Tsapatsoulis, Cyprus University of Technology, Cyprus
Andreas Uhl, Salzburg University, Austria
Binod Vaidya, Instituto de Telecomunicações / University of Beira Interior, Portugal
Andreas Veglis, Aristotle University of Thessaloniki, Greece
Janne Vehkaperä, VTT Technical Research Centre of Finland - Oulu, Finland
Anne Verroust-Blondet, INRIA Paris - Rocquencourt, France
Qin Xin, SIMULA, Norway
Shigang Yue, University of Lincoln, UK
Sherali Zeadally, University of the District of Columbia, USA
Yu Zheng, Microsoft Research Asia, China
Yongxin Zhu, Shanghai Jiao Tong University, China

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

Context-Aware Scalable Multimedia Content Delivery Platform for Heterogeneous Mobile Devices

Kwong Huang Goh, Jo Yew Tham and Tianxi Zhang
Institute for Infocomm Research, A*STAR, Singapore
{khgoh, jytham, tzhang@i2r.a-star.edu.sg}

Timo Laakko
VTT Technical Research Centre of Finland, Finland
{Timo.Laakko@vtt.fi}

Abstract—The vision of making any multimedia content to be always accessible to every mobile user over the network is great challenge. There are two major obstacles. First, the network capacity limits the amount and the quality of multimedia content that can be made available to every user at each time instant. Secondly, it is extremely difficult for the user to browse and search for the desired content from the huge multimedia library that is available. The work in this paper aims to address these obstacles to a certain extent. In order to overcome the adverse effect of the bandwidth limitation on the quality and quantity of the delivered multimedia content, we propose using the SVC and adaptive layered streaming approach. The second problem is addressed by utilizing context-aware personal content adaptation and efficient metadata processing to reduce the burden of user navigation in the large pool of media content. A proof-of-concept prototype that integrates the two technologies was developed. Cross country test trials were conducted to demonstrate the capabilities and practical use cases of the integrated context-aware scalable multimedia content delivery system for heterogeneous mobile devices.

***Index Terms*—Context, Scalable, Multimedia, Delivery**

I. Introduction

Mobile devices have become a common and essential commodity for everyone. In recent years, multimedia features are being integrated into mobile devices. Unsurprisingly, the demand for mobile multimedia content and services has been on the rise.

A. The Desired User Experience

For an ideal user experience, any multimedia content should be readily accessible on-demand over the network at anytime. From the user perspective, it is an essential requirement to have as smooth multimedia services as possible, e.g., based on his/her personal contextual habits independently from the applied heterogeneous delivery channel. This is a difficult task with the currently popular video encoding and streaming technologies, i.e., TCP/RTP streaming of H.264 videos. This is because H.264 video stream does not allow bitstream truncation for adaptation. Therefore, in order to cater for different network bandwidth and playback device capabilities, multiple copies of a single video has to be generated. An example is the different resolution options available at YouTube and Apple Movie Trailers. However, the conditions and quality of the delivery channel can change in the duration of a video stream. When this happens, current available technology is not able to automatically upgrade or downgrade the bitstream rate for improved video quality playback.

B. Scalable Video Streaming for Heterogeneous Devices

Different mobile devices have different processing and display capabilities. Moreover, the same device model is likely to have different bandwidth constraints which depend on the user subscription and the network conditions. Given such heterogeneous conditions of the mobile devices, it is necessary to have custom encoded video streams (in terms quality and rate) to cover for different possible device and environment settings in order to achieve optimal viewing experience. However, this is near impossible with the currently available video coding and streaming technologies in the market. Current video encoding and streaming technologies, such as TCP and RTP streaming of H.264 encoded videos, would have to encode and stream these different quality video streams separately and hence the huge transmission bandwidth is required for all heterogeneous devices. Furthermore, the content management is also tedious for encoding and maintenance of different video quality streams.

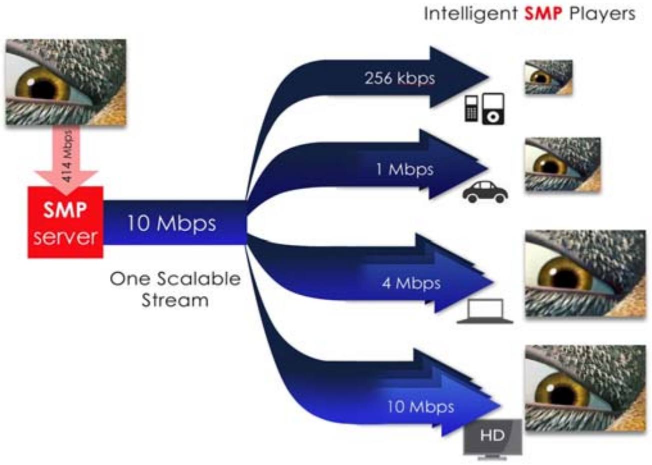


Figure 1. Advantages of scalable video coding & streaming

In this work, we use scalable video coding (SVC) which is an extension of the H.264/AVC standard [1][2][3]. Figure 1 below illustrates the advantages of SVC for heterogeneous streaming. With SVC, a scalable stream can provide adaptively different numbers of video layers to heterogeneous clients, according to the client’s processing capability and available bandwidth. In terms of content management, only one-time encoding of each of the video content is required and hence simplified the content management process. Some other related work can be found in [4].

We have integrated the context awareness aspects for positive content viewing experience and the scalable multimedia content delivery platform, to build a context-aware scalable multimedia **Co**ntent **De**livery platform for heterogeneous mobile devices (CoDe).

The next section of this paper describes the context discovery and personalization of the integrated platform and section III describes the end-to-end scalable multimedia platform from encoding, streaming to decoding and playback. The integrated platform test is described in section IV followed by the conclusions.

II. Context Discovery and Personalization

Personalization is based on the stored and semantically refined context information of the user. The user preferences are included into the delivery context [5]. Personalization aims to increase the acceptance of the set of information. It helps the user to get relevant content in the current situation. A platform describing personal preferences in each context is developed. The semantic of context information is used as a basis for adaptation and personalization.

A. Context-Aware Server

A user context may contain parts such as: spatio-temporal (place, time), environment, personal, task, social. User context information is derived from lower level context information. A low-level context is composed from different sources (sensors, network connection, user preferences, user agent profile etc), for example, measuring location, 3D acceleration, vibration, time, etc. The context information can also be given explicitly.

The location context is fetched from the GPS (outdoor), Cellular ID (indoor) or WLAN hotspots (indoor). In Cellular ID based positioning, ID of the used base station is sent to the server, where it is searched from the list of base stations and its location is returned. The accuracy of cell ID based positioning is inferior to GPS positioning, but it consumes less battery resource. Similar method is used in the WLAN hotspot detection; the phone scans for unique Basic Service Set Identifiers (BSSID) of available WLAN access points, which are then compared to the predefined list of known WLAN hotspots.

The technical context consists of device properties, such as display size, user agent, compatible formats, battery life, available space and other capabilities and limitations. Static information about the user agent could be collected from UAPROF header included in devices HTTP requests or during the registration of the device. Dynamic information, such as battery status information, should be updated periodically to the context module. Network context keeps up the information about available connection types to adapt the provided content in the most suitable format. Context module could also take advantage of external online context sources such as weather service or global calendar service.

B. Service Personalization

Personalization service retrieves user context and context history information from context management services. A user profile contains information about the user for personalization. Personalization module helps the user to get relevant content and services in the current situation. Table 1 shows the context information used for adaptation and personalization.

Context Information used for Personalization	
Context Data	Used in Movie Recommendation
Gender	Yes
Age	Yes
Language	Yes
Interest	Yes
Country	Yes
Screen	No
Time of the day	Yes
Time of the week	Yes
Network	No
Free Time	Yes
Mood	Yes
Activity	Yes
Location	Yes

Table 1. Context information used for service personalization

C. Media Content Analysis, Tagging and Retrieval

Media content needs to be analysed properly in order to get it utilized appropriately. Content analysis also includes tools for content management, and it takes into account content duration information, numbers of scalable layers encoded, scalable resolutions available, the content genre information, etc. Before the server retrieves relevant content for the specific user, content analysis is required to find out the useful personalized information, which can be user's age, gender, interest, language, etc. Based on this user information, a relationship between multimedia content and user profile can be built up and saved in server's database. Therefore, whenever and wherever the user wants to get their interested multimedia content, the server can satisfy their requirement by simply retrieving relevant information from database.

In order to facilitate the user's search for the multimedia content created by them, tagging mechanism could be used to assist the implementation of personalization. For instance, when user creates a new video, he can assign text-based tags to it which can facilitate video searching via tags. Content analysis can realize the service personalization and tagging can optimize the dynamic freedom of the system.

III. Scalable Multimedia Platform (SMP)

The scalable multimedia platform (SMP) developed in this work, as depicted in figure 2, is the integration of transcoding non-scalable media content into scalable media content; live layered streaming; server content management; and mobile client device decoders, into a next generation mobile entertainment solution. The SMP scalable content comprises of a base layer and several enhancement layers. Depending on the client's capabilities, only the appropriate audio/video layers or sub-streams will be abstracted from the same copy of the scalable media content for real-time delivery from the SMP server to the client.

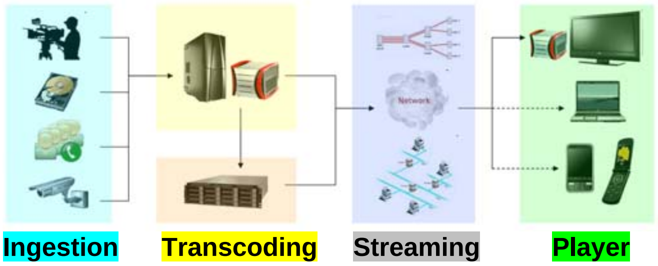


Figure 2. Scalable Multimedia Platform (SMP)

In general, the scalable video content can support scaling in spatial resolutions, bit-rate, and visual quality. The following sub-sections will further describe the details of the main modules of the SMP used in this paper.

A. The Media Content Transcoding Module

The content transcoding module converts a video from any supported compressed format to an ISO/IEC standard-compliant SVC video bitstream with AAC-LC (Advanced Audio Coding, low complexity profile) audio bitstream. Both the video and audio bitstream are multiplexed into the standard MP4 file container format (MPEG-4 Part 14 or ISO/IEC 14496-14). All media content that is stored in the video library is represented using the above scalable video format. The SVC encoder makes use of fast algorithms in [6]. The media track of the SVC compressed video in the MP4 file format is hinted accordingly to support several spatial and temporal resolutions [7]. This module provides the required application interface (API) for specifying the desired scalable encoding parameters of each media file ingested.

B. The Scalable Streaming Server

The streaming server consists of the scalable video streaming module which reads in a particular scalable media file from the video library and streams it in an instant-on-demand mode to the requesting client player. It employs the Real-Time Streaming Protocol (RTSP) with RTP over UDP for the media delivery. A different client player can simultaneously request from the streaming server either the same or a different media stream for playback. A unique feature of this module is that it allows the streaming server to automatically tailor the scalable video stream delivery to each of the requesting client player. This module is responsible for the automated selection of scalable media sub-streams for real-time delivery to the requesting client player.

C. Client Decoding and Player

The client decoding and player SDK comprises of the media buffering and decoding module, and the media streaming and adaptation module. The scalable media buffering and decoding module enables smooth media sub-streams management and decoding of the media for playback. This module also ensures robust networked media delivery and error concealment [8]. The scalable media decoder can also be deployed as a Microsoft's DirectShow filter plug-ins for the windows media player.

The scalable media streaming and adaptation module enables real-time reception of scalable media sub-streams requested from the Server System. It employs the Real-Time Streaming Protocol (RTSP) with RTP over UDP for the media delivery. The client player may request a different version of the media file depending on its own current capabilities such as the available processing power and resolution of the display device. This module automatically adjusts by requesting only the pertinent media sub-streams from the streaming server for delivery and playback on the client player.

IV. Integrated CoDe Platform for Test Trials

Figure 3 illustrates the CoDe's service-oriented architectural design between the clients and servers. It highlights the main client-side and server-side modules, together with their software implementation interfaces, communication/network protocols, operation system, and programming language environments. A desktop GUI application that was first developed using Nokia's Qt C++ language. The codes were portable on Windows, Linux, Mac OS as well as Windows Mobile operating systems. The demo application for the streaming test trial is Video-on-Demand (VoD). Subsequent applications such as News-on-Demand or live broadcast events can be added. The client retrieves the VoD metadata from the VoD server backend via Web Services. The Web Server is running Microsoft IIS and exposes the VoD application interfaces through WSDL. The VoD application server is running on Windows Server OS, and operates on Windows J2E framework. The VoD metadata is stored persistently on the MySQL Server.

Multimedia content source is compressed into SVC (Scalable Video Coding) with AAC (Advanced Audio Coding) standard formats, and stored as a hinted MP4 file container format that can be readily streamed via a RTSP/RTP streaming server (such as the Darwin Streaming Server). The compressed content is stored in a file server. The streaming server communicates with the client player via the Scalable Multimedia Platform Protocol (SMPP), which is an extended version of the standard Real-Time Streaming Protocol (RTSP). It is responsible for establishing the hand-shaking with the client player to exchange information about the media file, and for setting up a media session for packet-based streaming via the Real-Time Protocol (RTP) over User Datagram Protocol (UDP). The SMPP further supports dynamic media stream adaptation between client and server.

Client GUI for VoD application was demonstrated during the streaming test trial for Windows based laptop and mobile phones such as Windows-Mobile based smart phones and iPhone. For client-side scalable video playback, each PC/notebook client was installed with the relevant plug-ins, namely the SVC decoder plug-in and SMP protocol (SMPP) streaming plug-in. These plug-in was developed using the Microsoft DirectShow architecture. The plug-ins was integrated into the VoD GUI desktop application (via Nokia's Qt-Phonon framework) or it can also be embedded into a media player (such as the Windows Media Player) that is integrated into a web page. For mobile phone clients such as iPhone, web-based browser for the VoD was used and video streaming is performed via HTTP streaming of the SVC base layer to the iPhone's H.264 player. For Windows Mobile based phone, the PC-based GUI was ported to the Windows-Mobile OS for the VoD trial.

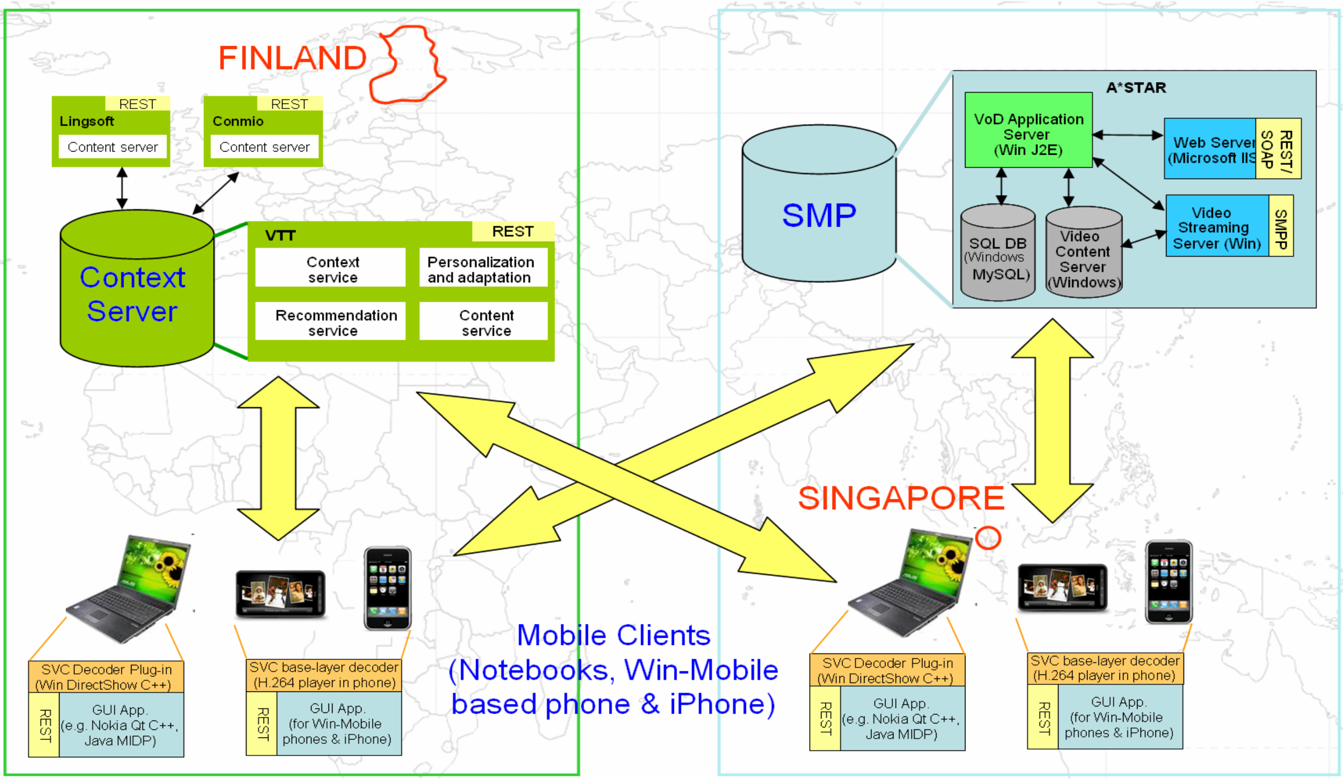
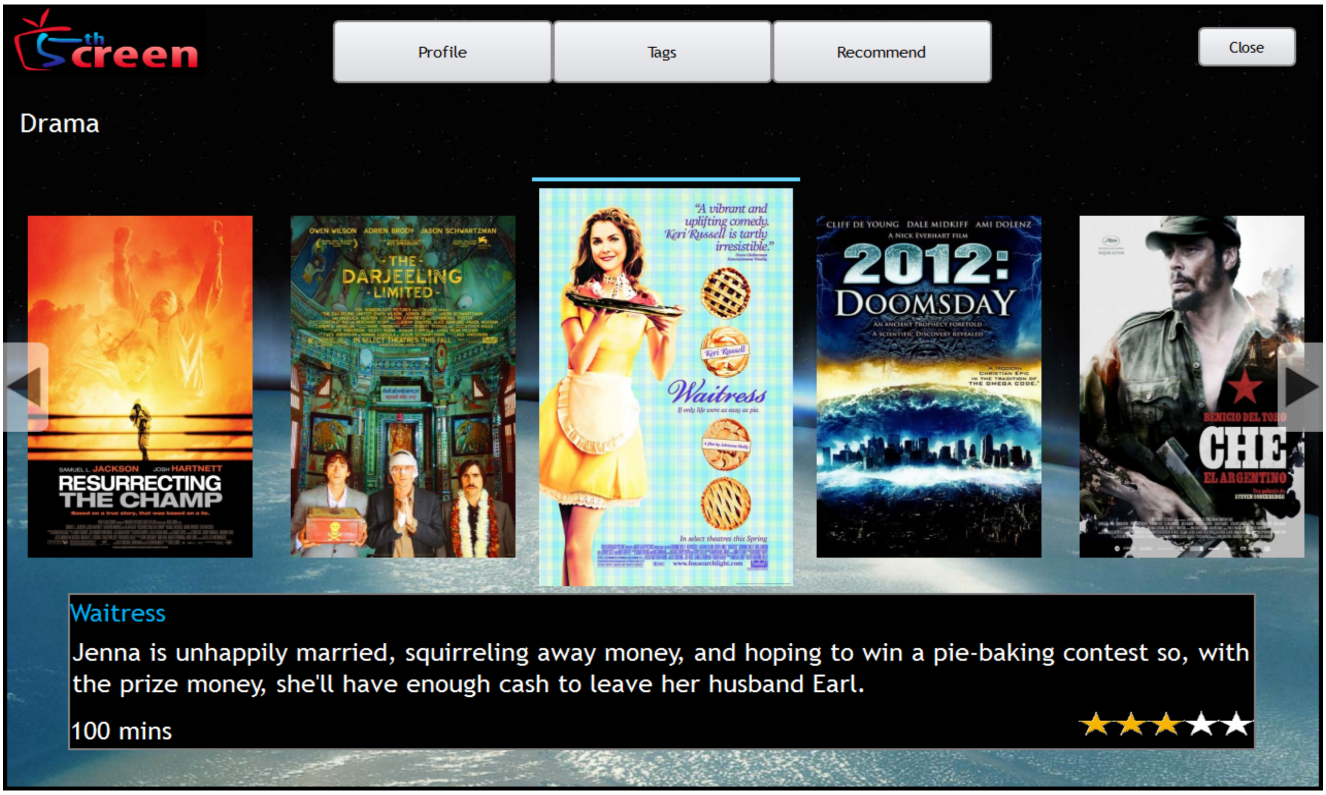

FINLAND
REST
Lingsoft
Content server
REST
Conmio
Content server
Context Server
VTT
REST
Context service
Personalization and adaptation
Recommendation service
Content service
SVC Decoder Plug-in (Win DirectShow C++)
REST
GUI App. (e.g. Nokia Qt C++, Java MIDP)
SVC base-layer decoder (H.264 player in phone)
REST
GUI App. (for Win-Mobile phones & iPhone)
Mobile Clients (Notebooks, Win-Mobile based phone & iPhone)
SMP
A*STAR
VoD Application Server (Win J2E)
Web Server (Microsoft IIS
REST/ SOAP
Video Streaming Server (Win)
SMPP
SQL DB (Windows MySQL)
Video Content Server (Windows)
SINGAPORE
SVC Decoder Plug-in (Win DirectShow C++)
REST
GUI App. (e.g. Nokia Qt C++, Java MIDP)
SVC base-layer decoder (H.264 player in phone)
REST
GUI App. (for Win-Mobile phones & iPhone)

Figure 3. CoDe's service-oriented architectural platform and geographical common trials

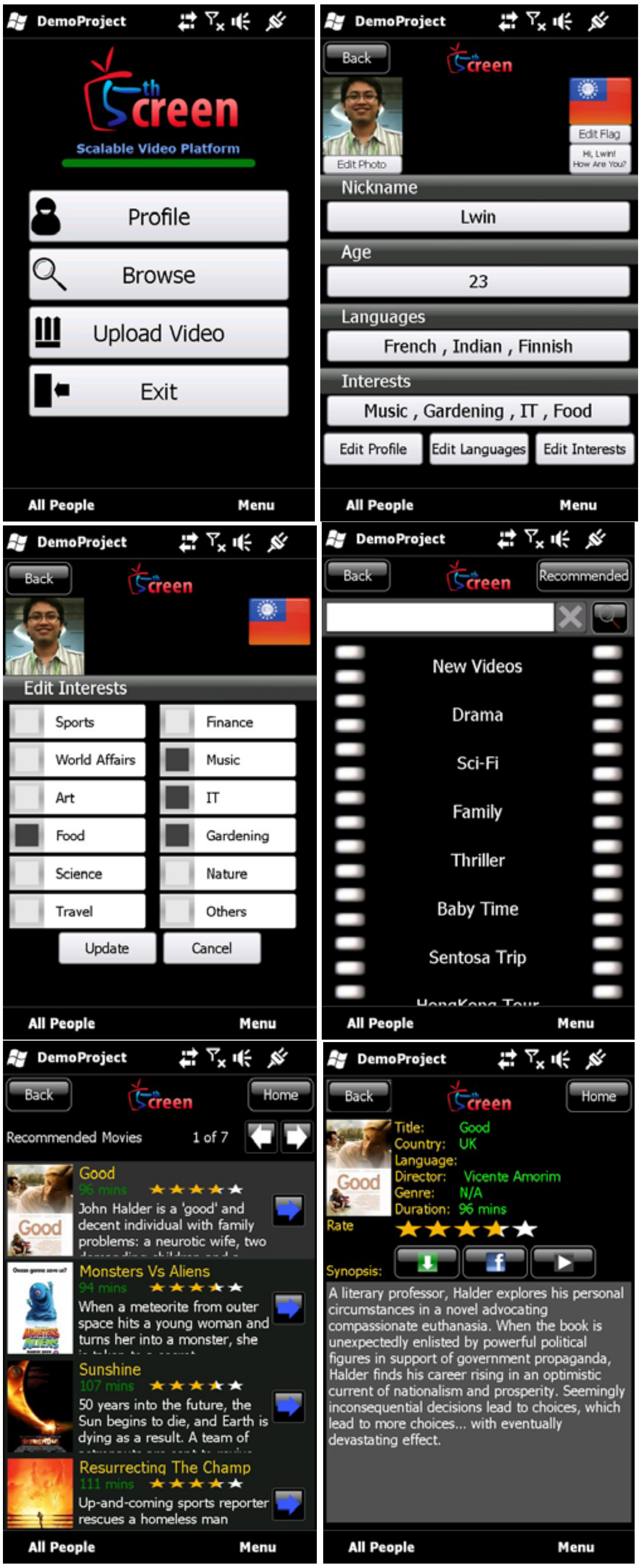


Waitress

Jenna is unhappily married, squirreling away money, and hoping to win a pie-baking contest so, with the prize money, she'll have enough cash to leave her husband Earl.

100 mins ★★★☆☆

Figure 4. PC client GUI's application named as 5thScreen, showing the context-aware recommended



Nickname: Lwin

Age: 23

Languages: French , Indian , Finnish

Interests: Music , Gardening , IT , Food

Edit Interests

□ Sports	□ Finance
□ World Affairs	■ Music
□ Art	■ IT
■ Food	■ Gardening
□ Science	□ Nature
□ Travel	□ Others

Update Cancel

Recommended Movies 1 of 7

Good
96 mins ★★★★★
John Halder is a 'good' and decent individual with family problems: a neurotic wife, two

Monsters Vs Aliens
94 mins ★★★★★
When a meteorite from outer space hits a young woman and turns her into a monster, she

Sunshine
107 mins ★★★★★
50 years into the future, the Sun begins to die, and Earth is dying as a result. A team of

Resurrecting The Champ
111 mins ★★★★★
Up-and-coming sports reporter rescues a homeless man

Title: Good
Country: UK
Language:
Director: Vicente Amorim
Genre: N/A
Duration: 96 mins
Rate ★★★★★

Synopsis:
A literary professor, Halder explores his personal circumstances in a novel advocating compassionate euthanasia. When the book is unexpectedly enlisted by powerful political figures in support of government propaganda, Halder finds his career rising in an optimistic current of nationalism and prosperity. Seemingly inconsequential decisions lead to choices, which lead to more choices... with eventually devastating effect.

Figure 5. Smart phone VoD application GUI, from top- left: main page, top-right: profile page, middle-left: interest-edit page, middle-right: video tag-list, bottom-left: video browser page and bottom-right: video info page

The context server was located in Finland whereas the scalable multimedia platform was located in Singapore. The cross-country context-aware scalable media streaming test trial has been successfully conducted in both Finland and Singapore. The context client in the laptop and the smart phones communicate with the context server located in Finland to obtain the context information. With the context information the client VoD application then communicate with the database server in Singapore to obtain the list of recommended list of movies for viewing. With information of the client devices such as the CPU, network type and screen sizes, the client can then make request to the streaming server located in Singapore for the suitable scalable video layers to be streamed to the client. Generally, with the context information, the client is able to make good recommendation; and the video streaming is smooth with the correct choice of scalable layer being streamed. Figure 4 shows the laptop client GUI's recommendation of movie list in a browser based on the context-aware information. Figure 5 shows six selected screen shots of the smart phones GUI.

Figure 6 shows the live context-aware scalable media streaming, in which the laptop is streaming the base layer plus 1 resolution enhancement layer via the internet, whereas the smart phone is streaming only the base layer via 3G connection, from the same scalable file stored at the streaming server.

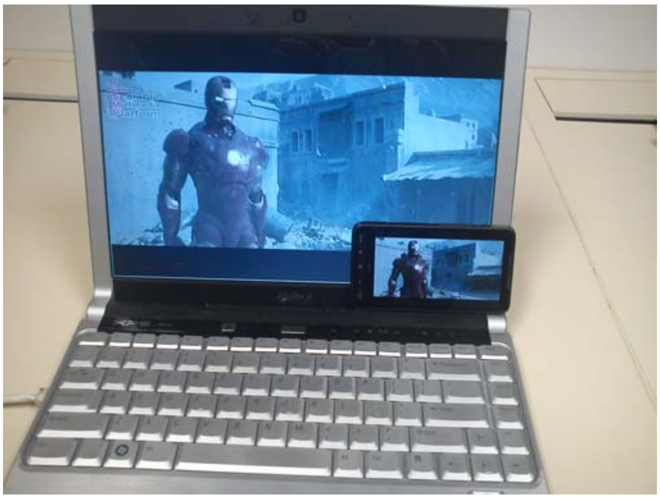


Figure 6. Demo picture of the cross-country context-aware scalable media streaming test

V. CONCLUSIONS

An integrated context-aware scalable multimedia content delivery for heterogeneous mobile systems is developed and trial-tested for cross-country content streaming. The proof-of-concept prototype of a context-aware scalable media delivery for heterogeneous devices has shown good context-aware use-cases with video streaming for best possible quality under the constraints of client device capability, network conditions and user preferences.

The current proof-of-concept platform only makes use of the context information for video recommendation service and to decide at the client side the number of scalable layers to be streamed. Full video streaming adaptation, i.e., on-the-fly adaptation to network conditions with error resilience and concealment, is yet to be integrated. The amount of context information used is also quite limited. Future work will address these limitations.

References

[1] JSVM software. Available from CVS repository. :pserver:jvtuser@garcon.ient.rwth-aachen.de:/cvs/. (Last access in Jun 2010).

[2] ITU-T Rec. H.264jISO/IEC 14496-10. Advanced video coding for generic audio-visual services, 2005.

[3] Schwarz, H., Marpe, D., and Wiegand, T.,: “Overview of the scalable video coding extension of the H.264/AVC standard”, IEEE Trans. CSVT. v17 i9. pp. 1103-1120.

[4] D. Pichon, P. Seite, and JM. Bonnin, “Context-aware delivery of video content to mobile users”, ACM Mobility Conference, ISBN. 978-1-60558-536-9, Nice, France, Sep 2009.

[5] Laakko, T.: Context-Aware Web Content Adaptation for Mobile User Agents. R. Nayak, et al. (Eds.): Evolution of the Web in Artificial Intel. Environ., SCI 130, Springer-Verlag, pp. 69–99, 2008.

[6] W. S. Lee, Y. H. Tan, J. Y. Tham, K. H. Goh, and D. J. Wu: “LACING: An improved motion estimation framework for scalable video coding”, ACM International Multimedia Conference, Vancouver, Canada, pp. 165-168, Oct 2008.

[7] H. Gao, J. Y. Tham, K. H. Goh, W. S. Lee, and K. S. Aw, “MP4 File Creator for SVC Adaptive Video Streaming”, IEEE Intl Conf Internet Technology and Applications (iTAP), pp. 1-4, Wuhan, China, Apr 2010.

[8] H. Gao, J. Y. Tham, W. S. Lee, and K. H. Goh, “Slice error concealment based on size-adaptive edge-weighted matching and motion vector outlier rejection,” Proceedings of the Second APSIPA Annual Summit and Conference (APSIPA ASC 2010), pp. 1058–1063, Biopolis, Singapore, 14-17 December 2010.

Low Complexity Corner Detector Using CUDA for Multimedia Applications

Rajat Phull, Pradip Mainali, Qiong Yang
Interuniversitair Micro-Electronica Centrum vzw.
Interdisciplinary Institute for BroadBand Technology
Kapeldreef 75, Leuven B-3001, Belgium
rajatphull@gmail.com, {pradip.mainali,qiong.yang}@imec.be

Patrice Rondao Alface
Alcatel-Lucent Bell Labs
Copernicuslaan 50, Antwerp B-2018, Belgium
patrice.rondao_alface@alcatel-lucent.com

Henk Sips
Delft University of Technology
Mekelweg 4, 2628 CD Delft, The Netherlands
sips@ewi.tudelft.nl

Abstract—**High speed feature detection is a requirement for many real-time multimedia and computer vision applications. In previous work, the Harris and KLT algorithms were redesigned to increase the performance by reducing the algorithmic complexity, resulting in the Low Complexity Corner Detector algorithm. To attain further speedup, this paper proposes the implementation of this low complexity corner detector algorithm on a parallel computing architecture, namely a GPU using Compute Unified Device Architecture (CUDA). We show that the low complexity corner detector is 2-3 times faster than the Harris corner detector algorithm on the same GPU platform.**

Keywords-LoCoCo; Harris feature detector; GPU; CUDA

I. INTRODUCTION

High speed detection of feature points is a fundamental requirement of many real-time computer vision and multimedia applications such as image matching, immersive communications, augmented reality, object recognition, mosaicing etc. Among robust and real-time feature point detectors (a good survey can be found in [16]), corner detection algorithms such as Harris [3] and KLT [2] are widely used for their lower complexity compared to SIFT [14] and SURF [15]. Several implementations of corner feature detection algorithms exist on GPUs. Sinha et. al. [4] proposed an implementation of the KLT corner detector for GPUs. However, the last step of non-maximum suppression of the cornerness response was performed on the CPU; this limits the potential speedup that can be obtained by a corner detector algorithm. Teixeira et. al [5] proposed their own implementation of non-maximum suppression for GPUs. They attained a significant speedup but introduced an imprecision of one pixel in the localization of corner points. Moreover, their method used a 3x3 Prewitt filter instead of a 9x9 Gaussian filter to compute image gradients. Both the algorithms [4][5] were implemented using the traditional OpenGL GPGPU API.

For modern GPU architectures, the CUDA [7] framework is supported by Nvidia GPUs for general purpose parallel computing. Compared to traditional GPUs and APIs such as Direct 3D or OpenGL, CUDA provides much more flexibility to manage and utilize GPU resources in order to fully exploit data parallelism in an application. Moreover, CUDA provides a high level programming model and a straightforward method of writing scalable parallel programs to be executed on the GPU. To our knowledge, none of the corner detector algorithm has fully exploited the computational power of CUDA.

Pradip et al. [1] proposed a Low Complexity Corner detector algorithm, which reduces the complexity of the Harris and KLT corner detectors by using a box kernel, integral image, and efficient non-maximum suppression. It achieves a complexity reduction by a factor of 8 on a CPU platform. By exploiting the computation power of CUDA, this paper proposes an efficient mapping of this low complexity corner detector on GPU. The implementation outperforms the execution time of existing state-of-the-art corner detector algorithms on GPUs [4][5].

The remaining of the paper is organized as follows: the low complexity corner detector algorithm is described in Section 2 in order to make the work self-contained. The mapping of the low complexity corner detector on GPU is described in Section 3. Section 4 shows the experimental results and Section 5 concludes the paper.

II. LOCOCO : LOW-COMPLEXITY CORNER DETECTOR

Harris feature detector is based on the local autocorrelation function within a small window of each pixel as shown in (1) and (2), which measures the local change of intensities due to the shifts in a local window:

$$C(\mathbf{p}) = \sum_{\mathbf{x}\in W}\left\{\begin{bmatrix} g_x^2(\mathbf{x}) & g_x(\mathbf{x})g_y(\mathbf{x}) \\ g_x(\mathbf{x})g_y(\mathbf{x}) & g_y^2(\mathbf{x}) \end{bmatrix}\times v(\mathbf{x})\right\} = \begin{bmatrix} G_{xx} & G_{xy} \\ G_{xy} & G_{yy} \end{bmatrix} \qquad (1)$$

$$\mathbf{x} = (x, y)$$

$$g_i = \partial_i(\mathbf{g}\otimes I) = (\partial_i\mathbf{g})\otimes I, i \in (x, y) \qquad (2)$$

where v(**x**) is a weighting function, which is usually Gaussian or uniform, **p** is a center pixel, W is a window centered at **p**, I is the original image, **g** is Gaussian, and g_x and g_y are the convolution of the Gaussian first order partial derivative with I in x and y directions at point (x, y), respectively.

As shown in (1), the Harris corner detector algorithm computes image derivatives using the Gaussian derivative kernel, computes cornerness response and suppresses non-maximum points to obtain the corner points. LoCoCo reduces the computational complexity of the Harris algorithm in each step. First, by using the integral image and box kernel, the computational cost of gradients is reduced. The box kernel is obtained by approximating the first order Gaussian derivative kernel. Second, many repeated calculations for computation of cornerness according to (1) are reduced by the use of the integral image. Finally, the combination of sorting (to rank cornerness responses) and non-maximum suppression is replaced by the efficient non-maximum suppression [6]. The LoCoCo algorithm is summarized as follows:

1. Calculate the integral image for the original image I.
2. Compute gradients gx and gy by using the integral image and the box kernel approximation.
3. Create the integral images for g2x, g2y and gxy Then, evaluate (1) and (2) and compute the cornerness response. With the use of the integral image, each element of (2) can be evaluated in 4 memory accesses and 3 operations.
4. Efficient non-maximum suppression is performed to suppress the non-maximum point instead of sorting and performing non-maximum suppression.

By following the above mentioned steps, LoCoCo achieves comparable feature detection results and a speedup factor of 8 with respect to Harris on the CPU platform. More details and experimental results are presented in [1].

III. Mapping LoCoCo Using CUDA

This section explains the mapping of each step of LoCoCo on the GPU using CUDA. In CUDA, the kernel is visualized as a grid, which consists of multiple parallel thread blocks; each thread block can contain up to 512 parallel threads. It is the responsibility of the programmer to choose the number of blocks per grid and the number of threads per block. Once the kernel is launched, the grid blocks are distributed on the parallel multiprocessors as described in [7]. The global memory exists off-chip and is accessible by all threads. The shared memory is on-chip and the threads within a block can communicate and cooperate using the shared memory as well as the thread synchronization mechanism. As described in [12], high performance on CUDA can be achieved by allowing a massive number of active threads to exploit the large number of cores, hence hiding memory latency by computations.

A. *Integral Image*

LoCoCo makes an extensive use of the integral image. We propose an efficient method to map the computation of the integral image on the GPU. The computation of the integral image can be separated in two stages. As shown in Figure 1(a), the prefix sum is calculated for each row. After completing the processing on rows, as shown in Fig. 1(b), the prefix sum is applied to each column, thus resulting into an integral image.

The prefix sum is computed for all rows in parallel by using the efficient parallel scan algorithm designed for GPUs [8]. The key idea of this algorithm is to divide the block of data into warp-sized chunks and all scan primitives are built upon the set of primitive intra-warp scan routines. The warps execute instructions in SIMD fashion and synchronization is not needed in order to share data within a warp. Thus, the intra-warp scan routine performs scan operations over a warp of 32 threads and computes the prefix sum for 32 elements without requiring any synchronization operation.

The computation of the prefix sum on a row is performed by allocating a thread block to that row and dividing it into warp-sized chunks. All the warps are scanned in parallel using an intra-warp scan routine. Next, the partial results of each scan are accumulated and adjusted to get the scan for the complete row. The reduced number of synchronization steps and various optimizations, such as efficiently exploiting shared memory and performing an initial serial scan of multiple input elements when read from global memory, makes it one of the fastest scans yet designed for the GPU [8].

In order to evaluate the subsequent prefix sum on the columns, the prefix sum result of the rows is transposed and a new row-based scan is launched. Transpose between the two steps help to maintain coalesced access to the global memory [7]. The resultant integral image is not transposed back again to correct the orientation, since the computation of integral image in the subsequent step of computing the cornerness response leads to another transpose, yielding the restored image.

B. *Gradients*

The strategy used to parallelize this step is based upon creating many threads to exploit the large number of cores. The image is partitioned into a regular grid of blocks. The width and height of these blocks are equal to the 16th of the width and height of the image, respectively. Each thread in that block can be mapped to a pixel location and computes the gradient value corresponding to that pixel of the image. Therefore, for the box kernel, the computations performed by the threads in a block are independent of each other. The CUDA kernel is launched wherein each thread performs eight memory accesses and seven operations in parallel to calculate g_x or g_y, corresponding to each pixel. The step is complete when the gradients are computed for all the pixels in the image.

C. *Cornerness Response*

The strategy used to compute the integral image for g^2_x, g^2_y and g_{xy} is the same as described in Section 3.1. In order to obtain g^2_x, g^2_y and g_{xy}, the scan algorithm is modified such that each element of g_x and g_y is squared or multiplied with each other when fetched from global memory to shared memory.

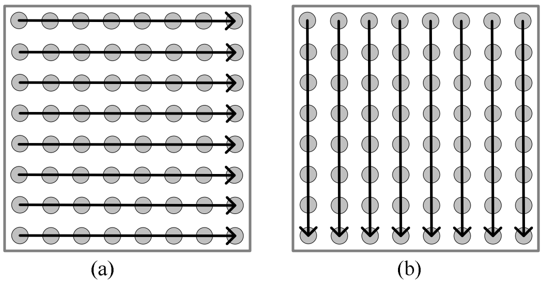


Figure 1. (a) Prefix sum on rows (b) Prefix sum on columns

The computation of the summation within a window and cornerness response corresponding to each pixel, is based on a similar strategy as described in previous section is adopted so that each thread in every block can be mapped to a pixel location in the image. The CUDA kernel is launched wherein each thread performs four memory accesses and three operations in parallel to calculate the window sum corresponding to each pixel followed by the computation of cornerness response. The step is complete when the cornerness response is calculated for all the pixels in the image.

D. *Efficient Non-Maximum Suppression*

Access to off-chip global memory is slow and requires 200 to 300 cycles per access. This latency can be hidden by launching a massive number of active threads [12]. But this technique does not give enough speedup for algorithms that have repeated calculations or are bounded by memory accesses. Another technique to attain speedup is to use low latency on-chip shared memory and reuse data among all the threads in a thread block to reduce the number of accesses to the global memory [12]. As non-maximum suppression incorporates repeated calculations on a small region of pixels therefore shared memory is exploited to reduce the number of accesses to the global memory.

The suppression algorithm is implemented for a *d* by *d* (*d*=9) neighborhood. The kernel is launched wherein threads in each thread block fetches (2*d* x 2*d*) pixels from global memory to the shared memory. At this point the contents of shared memory can be visualized as four sub-blocks as shown in Figure 2(a). The kernel is implemented in such a way that the threads in a thread block compute the maximum value of the cornerness response in each of the sub-blocks in parallel. These maximum values are termed as candidate local-maximas (max1, max2, max3 and max4). The maximum value in each of the sub-blocks is calculated using parallel reduction [10], where the add operation is replaced with a comparison operator. For each candidate maxima, the threads in a thread block fetch the local neighborhood pixels to the shared memory if the value is greater than the predefined threshold, as shown in Figure 2(b). The maximum value is computed in each of the blocks using [10]. If the candidate maxima remains maximum in the local neighborhood then it is marked as corner point else the point is suppressed.

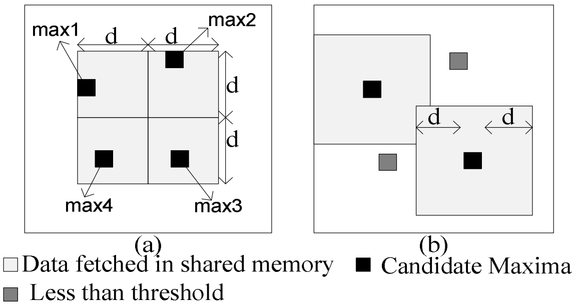


Figure 2. Efficient Non-Maximum Suppression

IV. Results and Discussion

The execution time of LoCoCo is evaluated on a CPU and on a GPU. To measure the effectiveness of LoCoCo on GPUs, the execution times are compared with our CUDA based implementation of the Harris corner detector on the same GPU.

For Harris, the Gaussian derivative is implemented by using the separable Gaussian convolution kernel [11], which requires less computations compared to the 2D convolution. In order to measure the cornerness response, the gradients are squared and multiplied with each other. Furthermore, the summation within the window is implemented by utilizing the separable convolution. The separable filter can be used to sum the pixels within the window by setting the coefficients of separable filters to 1. This method of implementation runs much faster than computing the naive sum of all pixels within a window. As described in [4], the sorting for cornerness response is performed on the CPU and this involves transferring the complete cornerness image back to the CPU. Instead of adopting this approach, the sorting is performed on the GPU by using an efficient sorting algorithm as presented in [9]. After this step, non-maximum suppression is performed on the GPU. Thus, the Harris algorithm completely runs on the GPU and this implementation is utilized to have a fair comparison with LoCoCo implementation on the GPU.

As shown in Figure 3, the GPU implementation of LoCoCo is around 14 times faster than the corresponding CPU implementation. The speedup is mainly due to the fact that computation of the integral image and efficient non-maximum suppression is efficiently parallelized using CUDA.

The comparison of execution time of both LoCoCo and Harris on GPU is shown in Figure 4 for different image and kernel sizes. As shown in Figure 4(a), for various image sizes and a fixed kernel size of 9x9, the LoCoCo implementation on GPU is around 2 times faster than the Harris corner detector on GPU. The original Harris algorithm uses Gaussian convolution instead of Integral image computation and box kernel approximations; contrary to CPU programming the execution time of the GPU implementation of convolution for small kernel size (9x9) is comparable to the time taken by the computation of the integral image and the box kernel approximation.

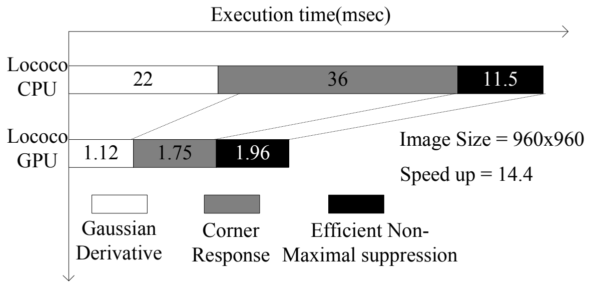


Figure 3. Execution time of LoCoCo on CPU and GPU (CPU: Intel Core 2 Duo E6750, 2.66 GHz and 2 GB RAM, GPU: Nvidia GeForce GTX 280, 1.296 GHz, 1 GB Global memory, 16 KB shared memory per core)

With the box kernel approximation, the speedup for a kernel size of 9x9 is mainly due to the fact that LoCoCo replaces the combination of feature sorting and non-maximum suppression in Harris by efficient non-maximum suppression. As shown in Figure 4(b), for a kernel size of 31x31, LoCoCo is 3 times faster than the Harris. As the kernel size increases, the computation of the integral image and box kernel approximation remains unaffected but the execution time for the convolution increases significantly. For applications that require multi-scale estimation, the convolution must be computed for each scale, while only one execution of the integral image allows for the computation of all the scales. In that case, the LoCoCo algorithm turns out to be much more efficient than the Harris algorithm.

Table 1 presents the comparison of LoCoCo using CUDA with other state-of-the-art implementations of corner detectors. Accelerated corner detector [5] provides a full implementation of the Harris corner detector on GPU by proposing its own version of non-maximum suppression. The proposed non-maximum suppression has two different variants, one in which the cornerness response image is compressed and the other in which the cornerness response image is not compressed. The lossy compression of the cornerness response image introduces a precision error of one pixel in localization of the corner points whereas our method does not introduce any precision error or compression of the cornerness response image. A comparison is made with the version of the accelerated corner detector in which the cornerness response image is not compressed.

The time taken by this implementation is reported in [13]. The timings presented in Table 1 include the time to transfer data between the GPU and the CPU. Notice that execution times reported in [4][5] are related to a GeForce 8800 GTX while the execution times of our contribution have been measured on a GeForce GTX 280. Even though reference [17] indicates that GTX280 delivers twice the performance of GeForce 8800, it can still be inferred that our method is the fastest compared to state-of-the-art implementations of corner detectors reported till now.

TABLE I. COMPARISON WITH OTHER METHODS

Algorithm	Time (ms)	Image Size	Platform
L.Teixeira [5]	7.3	640x480	GeForce 8800 GTX
Sinha [4]	61.7	720x576	GeForce 8800 GTX + AMD Athlon 64 X2 Dual Core 4400 (one core used)
Our Method	2.4	640x480	GeForce 280 GTX

V. CONCLUSION AND FUTURE WORK

In this paper, we proposed an efficient implementation of low complexity corner detector using CUDA. Corner detection using CUDA has not been reported so far. Our method greatly exploits the data parallelism and achieves a speedup factor of 14 with respect to CPU. Experimental result shows that low complexity corner detector is around 2-3 times faster than Harris on a GPU. With the increase of kernel size, the execution time of our method remains close to constant while the execution time of the Harris increases, thus achieving further speedups.

REFERENCES

[1] P. Mainali, Q. Yang, G. Lafruit, R. Lauwereins and L. Van Gool, "LoCoCo: Low Complexity Corner Detector", ICASSP 2010, pp. 810-813.

[2] C. Tomasi and T. Kanade, "Detection and tracking of point features", Technical Report CMU, April 1991

[3] C. Harris and M. Stephen, "A Combined corner and edge detector" In Proce. of Alvey Vision Conf., pp. 147-151, 1988

[4] S. Sinha, J. Frahm and M. Pollefeys, "GPU-based video feature tracking and Matching", in EDGE 2006, workshop on Edge Computing Using New Commodity Architectures, 2006

[5] L. Teixeira, W. Celes and M. Gattass, "Accelerated Corner Detector Algorithms", in BMVC, 2008

[6] A. Neubeck and L. V. Gool, "Efficient non-maximum suppression", in ICPR 2006, Vol. 3, pp. 850-855.

[7] http://developer.nvidia.com/object/cuda.html.

[8] S. Sengupta, M. Harris, and M. Garland. "Efficient parallel scan algorithms for GPUs". NVIDIA Technical Report NVR-2008-003, December 2008

[9] N. Satish, M. Harris, and M. Garland. "Designing efficient sorting algorithms for manycore GPUs", Proc. 23rd IEEE IPDPS2009, May 2009

[10] M. Harris, "Optimizing Parallel Reduction in CUDA", NVIDIA Developer Technology

[11] V. Podlozhnyuk, "Image Convolution with CUDA", Nvidia CUDA 2.0 SDK convolution separable document

[12] S. Ryoo, C. Rodrigues, S. Baghsorkhi, S. Stone, D. Kirk and W. Hwu. "Optimization principles and application performance evaluation of a multithreaded GPU using CUDA". In Proc. 13th ACM SIGPLAN, 2008.

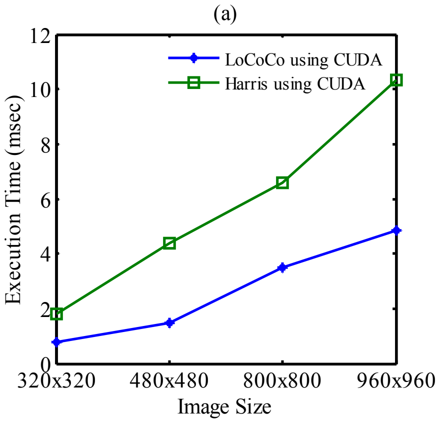
[13] J. F. Ohmer and N. J. Redding, "GPU-Accelerated KLT Tracking with Monte-Carlo-Based Feature Reselection", DICTA 2008.

[14] D. Lowe "Distinctive image features from scale-invariant keypoints" International Journal of Computer Vision, 60, 2 (2004), pp. 91-110

[15] H. Bay, A. Ess, T. Tuytelaars, L. Van Gool, "SURF: Speeded Up Robust Features", Computer Vision and Image Understanding (CVIU), Vol. 110, No. 3, pp. 346--359, 2008

[16] T. Tuytelaars, K. Mikolajczyk "Local Invariant Feature Detectors: A Survey", Foundations and Trends in Computer Graphics and Vision, Vol. 3, nb 3, pp 177-280, 2008.

[17] http://www.nvidia.com/content/PDF/fermi_white_papers/N.Brookwood_NVIDIA_Solves_the_GPU_Computing_Puzzle1.pdf



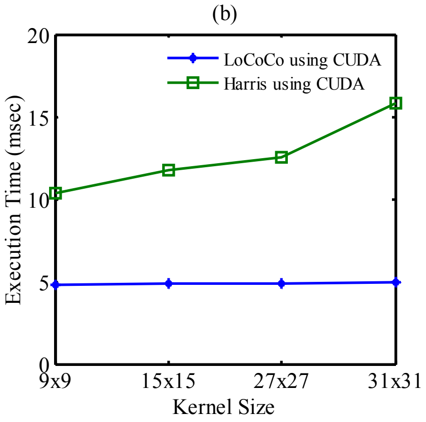


Figure 4. Execution time for LoCoCo and Harris on Nvidia GeForce GTX 280 GPU (a) Comparison for different image sizes (b) Comparison for different kernel sizes

Real-Time Temporal Control of Musical Processes

Raphaël Marczak
INSERM, U642, Rennes,
F-35000, France
Université de Rennes 1,
LTSI, Rennes, F-35000, France
raphael.marczak@gmail.com

Myriam Desainte-Catherine
Université de Bordeaux
IPB, UMR5800,
SCRIME, ENSEIRB-Matmeca
Talence, France
myriam.desainte-catherine@labri.fr

Antoine Allombert
LIPN
Université Paris 13
Villetaneuse, France
antoine.allombert@lipn.univ-paris13.fr

***Abstract*—The temporal control of the execution of multimedia processes is a crucial point for a number of application fields. We propose a formalism for authoring multimedia scripts that involves dynamic triggerings. In addition, we propose an abstract machine able to execute these scripts, by adapting the temporal organization of the processes according to the dynamic triggerings. This machine must implement some temporal features such as fast forward or GOTO functionalities. We present some algorithms performing these functionalities. A last, we mention some evaluation by users and possible future works.**

***Keywords*-Musical Processes; Petri network; Real-Time; Flexible time; Iscore**

I. Introduction

Controlling the temporal unfolding of multimedia processes is a useful but challenging task for several kinds of applications. Software systems like *Flash* are quite useful for creating multimedia contents for animation but they fail at precisely specifying synchronization between several processes. From live-performance to composition of interactive music or even oral presentations, the conception of a complex script implies controlling the temporal sequence of several processes as well as their coordination.

In most applications from the musical domain, time control is often performed by introducing continuous control giving the tempo of a score. In such systems, durations of musical events are defined according to a time unit that is stretched in order to fit real-time purposes. For example, a score-follower [1] or a real-time accompaniment system [2] is listening to a performer playing a score, which is known. By analyzing the sound played by the musician, the system is able to extract the notes and to follow the fluctuations of the tempo played by the musician. In consequence, the system can use those fluctuations of tempo to control the time unit of the score that he must play to accompany the musician. But such systems often fail at synchronizing precisely processes beginning or ending.

In the VIRAGE project [3], we addressed the problem of executing a script of multimedia processes for controlling light, sound or video, for live-performance while adapting the launching of the processes according to the play of actors on the stage. We introduced and implemented the *Iscore* system that uses discrete controls to precisely synchronize multimedia processes according to a script [4].

In this paper, we present added features to the *Iscore* system that permit real-time temporal control. The innovation of this system lies in the melding of discrete and continuous controls over temporal unfolding of the processes. With this system, the user can firstly place processes on a time-line and specify temporal constraints between starting or ending dates of processes. Secondly, he can specify interactive points that will come during execution flow of the system. The link between those two paradigms, that is time-line and time-flow, holds thanks to temporal relations that are verified by the system both at writing and execution time. At last, during execution, the user can modify the speed of the processes execution in a continuous way and he is also able to use a GOTO feature enabling him to skip or repeat some parts of the script.

In Section II, we briefly introduce the *Iscore* system and the implementation of the management of discrete events thanks to an abstract machine constituted of a Petri Net. In Section III, we present the new features, that is modification of the execution speed and GOTO, and how they are implemented in the context of the abstract machine.

II. Iscore system

The *Iscore* system has been fully described in [5], let us remember the bases of its conception and implementation.

The main question addressed by the *Iscore* system is the authoring and interpretation of musical scores of electro-acoustic music. This problem has been enlarged to a more general model for interactive multimedia scripts. The *Iscore* has typically two sides : the authoring side allows an author to design a multimedia script that can be modified during its execution, while the performance side executes the multimedia scripts and allows a performer to take benefit from the interaction possibilities. We studied the case in which the interaction possibilities consists in modifying the date of some steps of the multimedia processes involved in a script, as well as the speed of execution of these processes.

ISBN: 978-1-61208-129-8

In order to prevent excessive modifications of a script during its execution, an author can define some boundaries that a performer must always respect. Since we only consider temporal modifications as interaction possibilities, the limits upon the interaction possibilities consist in temporal relations that must be respected during each performance.

A. *Authoring side*

One can find an example of a script on Figure 1. In this representation, the time-line is horizontal and left-to-right oriented. The author can temporally organized some *temporal objects*, which are represented as boxes. A *temporal object* can be *simple*, i.e., it represents the execution of a multimedia process such as the objects *sound* and *red* on the example ; or it can be *complex*, i.e. it represents the execution of a group of *temporal objects* such as the *lights* object. Each complex temporal objects holds its own time-line with a specific time speed. Therefore, a script can be performed with heterogeneous time speeds in its complex *temporal objects*.

A *temporal object* presents some *control points*, represented by circles on the top and bottom borders of the objects. *Control points* represent some particular moments of the execution of the *temporal object*. The beginning and the end of an object are naturally considered as particular moments. Other intermediate moments can also be considered. The author can temporally organize a script by adding some temporal relations between the control points of the objects involved in this script.

Since these relations can be defined between points, they can be of two types : *precedence* and *posteriority*. Formally, a temporal relation tr can be defined by a 6-uple :

$$tr = < t, p_1, p_2, \Delta_{min}, \Delta_{max} >$$

- t is a type (*Pre* or *Post*)
- p_1 and p_2 are control points
- Δ_{min} and Δ_{max} are real values in $[0, \infty]$

If tr is a precedence relation, then it imposes the inequality :

$$\Delta_{min} \leq d(p_2) - d(p_1) \leq \Delta_{max}$$

where $d(p_i)$ is the date of p_i.

The inequalities imposed by the temporal relations must be respected during each execution.

The possibilities of interaction are expressed through *interaction points*, represented by red circles. An *interaction point* turns a control point into a dynamic one. A dynamic control point must be explicitly triggered by the performer during the execution. On the contrary, the other control points (the static ones) are triggered by the system.

In order to execute interactive scripts, we had to design an abstract machine that can trigger the static control points, accept the triggering of the dynamic control points and respect the temporal relations.

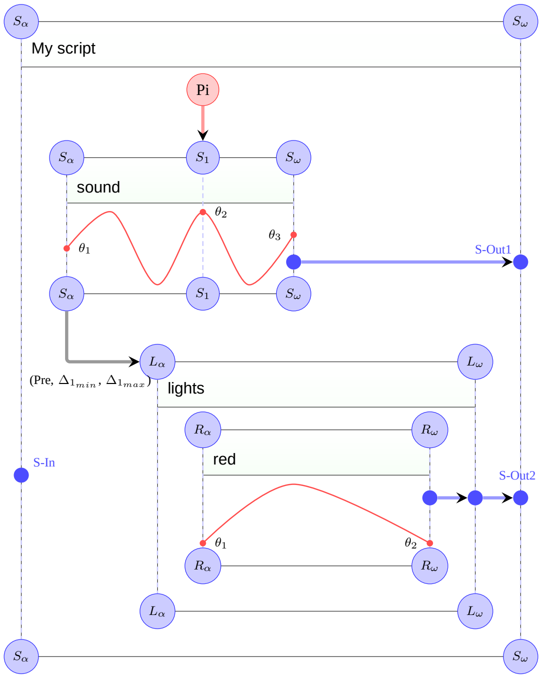


Figure 1. An example of an interactive script

B. *ECO machine*

We call the abstract machine designed for executing the interactive scripts the *ECOMachine*, for *Environment*, *Controls* and *Outputs*. One can find a representation of this machine on Figure 2. The term *Environment* must be understood as *temporal environment*. It carries all temporal information specified in the script. This information is represented in a time-stream Petri net [6]. One can see an example of such a net on Figure 3. To generate the environment associated to a script, we transform the script into a Petri net according to the following method. Each control point is turned into a transition. If a temporal constraint imposes the simultaneity of different control points, their transitions are merged. If a precedence relation is specified between a control point p_1 and a control point p_2, a sequence arc/place/arc is added between the transition of p_1 and the transition of p_2. The type of Petri net that we use accepts a time range on each arc. This time range allows us to represent the possible values for a time interval between control points. In addition, the crossing of a transition that represents a dynamic control event is conditioned by receiving an external control message.

Then, the term *Controls* represents the flow of control messages that trigger the dynamic control points. When a control point of a temporal object is triggered, the associated step of the process represented by the object is triggered. The part of the machine, which runs the multimedia process, is called the *processor*. It receives messages from the Petri net and sends the data produced by processes through the

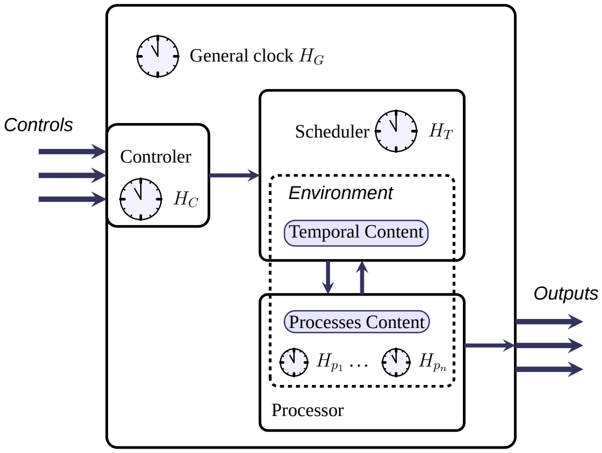


Figure 2. The ECOmachine

Outputs flow.

The structure of this machine allows us to temporally control the triggering of the control points, as well as the speed of execution.

III. Real-time temporal control

We present the *Iscore* system without speed modification, we provide some algorithms to manage this system in an efficient way, and then we introduce what modifications are needed to allow speed modification as well as the GOTO feature.

A. Discrete Temporal Control System

For the discrete temporal control, we apply the petri network crossing rules.

1) Static And Dynamic Transitions: A script without trigger points, i.e., without interaction, only contains static transitions. These transitions are triggered by the system. When an interaction is needed, the compositor can decide to add a trigger point on a processes start or end, the transition becomes dynamic [3]. He can also decide a time range in which this event can be triggered.

In the Petri network, this mechanism is implemented by a transition, which is waiting for an event. Each of its ingoing arcs has minimum and maximum values that match the time range decided by the compositor. The minimum value corresponds to the minimum time after which we can cross the transition. The maximum value corresponds to the time when the transition will be crossed, whether the event was received or not.

2) Priority FIFO Algorithms: During the editing process, the time taken by each user action is not really important. However during real time execution, we cannot allow a procedure to be time consuming. The most many-time called procedure is the one, which decides if a transition should be crossed or not. It is impossible to check all the transitions all time because it would be a long process. So we decided to use a priority FIFO that can be filled with priorityActions.

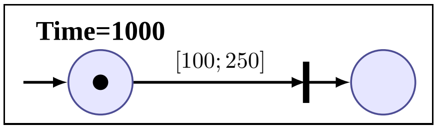


Figure 3. Here is a part of a Petri net produced by the transformation of an interactive score. During the execution, a token is produced in the left place at absolute time 1000. This leads to the creation of a START action labeled with the date 1100, as well as an END action labeled with date 1250. These two actions are put in the priority FIFO.

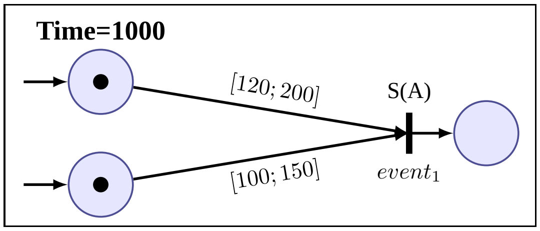


Figure 4. An example of a dynamic control point

PriorityAction

A PriorityAction is defined by a transition, a date, a type (START or END) and a boolean stating if the action is still enabled. The type is START when a transition could be crossed (for example when we just wait for the event to come), and END when a transition must be crossed whether the event was received or not.

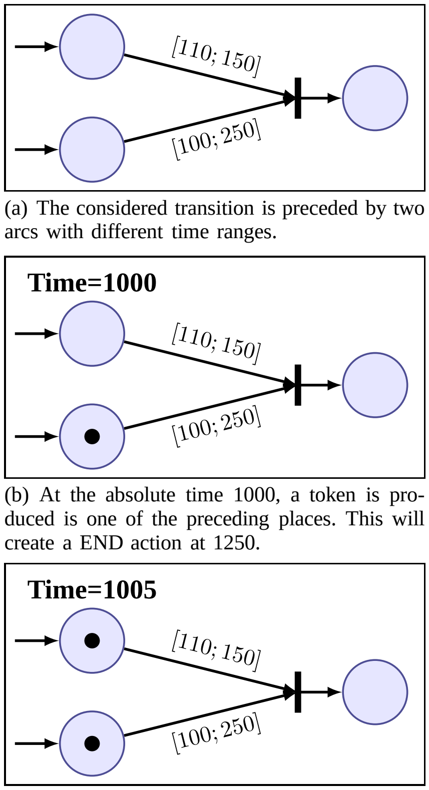
Filling and updating the FIFO

These actions will be the elements of a chronologically priority FIFO.

Filling the FIFO is not an easy procedure. If fact, START and END actions should be computed in real time, but they depend on how the tokens arrived in the places before the transition. The Figure 4 shows a simple example. In this example, the considered transition represents a dynamic control point. This means that during the execution, the system will wait for receiving a external trigger message called $event_1$. When this message is received, the system crosses the transition, which leads to start a process called *A*. The system must respect the time ranges introduced by the user. Therefore, if the two tokens are produced simultaneously at absolute time 1000, the transition can be crossed between time 1120 and time 1150. Then the system will ignore a message $event_1$ that would arrive before time 1120 and it will automatically crossed the transition if no message has been received at time 1150.

One can find a more complex example on Figure 5.

A transition knows at every time, which ones of its ingoing arcs are active, i.e., when a token is present in the



(a) The considered transition is preceded by two arcs with different time ranges.

(b) At the absolute time 1000, a token is produced is one of the preceding places. This will create a END action at 1250.

(c) At absolute time 1005, a token is produced in the second place. This will create a START action labeled with date 1115 (the maximum of 1000 + 100 and 1005 + 110) and a update the END action labeled with date 1155 (minimum of 1005 + 150 and 1000 + 250)

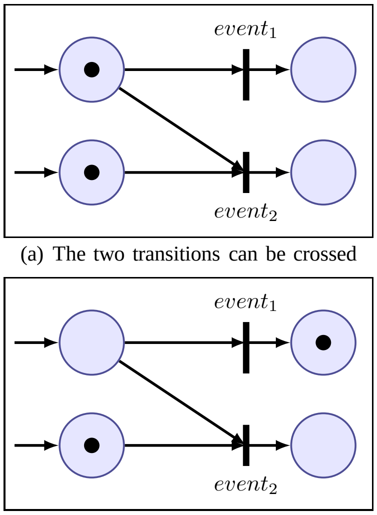
Figure 5. An example in which, two tokens arrive at different times. This implies that the system needs to update the date of an action

place linked by this arc. When the incoming arcs are not all active, we can only calculate and update the END action. When all arcs are actives, we can add the START action. When we update an action, we just disable the previous one and add a new one, for efficiency purpose.

When a token arrives in an empty place, all the outgoing arcs are stated as active. At this moment, the END and START actions are computed. (See the Algorithm 1).

Computing the FIFO elements

At each ECO Machine cycle, the makeOneStep procedure (Algorithm 2) is called. It handles all the actions, which dates are lower or equal to the current date. If an action is disabled, it simply removes it. If an action is END, it forces the transition to be crossed (if it is impossible, an exception should be thrown). If an action is START, it fills a list of sensitized transition, i.e., transition waiting for an event. After that, it handles all the sensitized transitions. A transition that can finally not be triggered will be removed from the list (for example if a place is involved in two or more transitions, and another transition was previously crossed, consuming the token).



(a) The two transitions can be crossed

(b) The message $event_1$ is received, the corresponding transition is crossed. The second transition cannot be crossed anymore.

Figure 6. In this example, the crossing of a transition leads to the impossibility to cross another transition.

Updating transition state

The last algorithm (Algorithm 3) handles a transition crossing. It consumes and produces token in the corresponding places, executes the external actions (i.e., the start or end of a processes), and resets the transition state. In fact, when a token is consumed, it could destabilize the system (for example when a token is involved in two or more transitions, as seen in the Figure 6).

B. *Continuous And Discrete Temporal Control System*

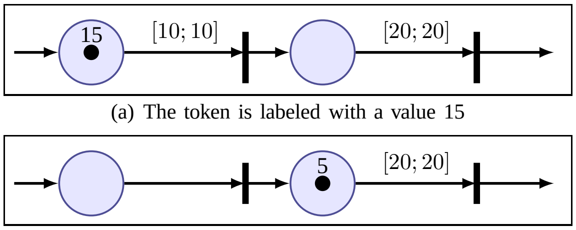
It is very useful to have the possibility to accelerate or decelerate the script. For example if we need to accelerate the fade out for music and light. But these features could not be possible without the concept of numbered tokens.

1) Deceleration: Decelerating is not a complex part of the speed modification. In fact, if all the time values are set in a millisecond precision, the precision of the Petri network time is in microsecond. So for decelerating, a simple multiplication of the next computed delta time by a factor between 0 and 1 is sufficient.

2) Acceleration: Accelerating is much more challenging. A first idea could be to call makeOneStep more often, but this would too much computation time, and it would be unsafe in real-time processing. Our solution is to use stamped tokens, but it is not straight forward solution. Each number on a token represents the remaining time to be handled.

Algorithm 1 setArcAsActive

```
Require: $arc\ transition\ petriNet$
$transition.labelInGoingArcAsActive(arc)$
$currentDate \leftarrow petriNet.getCurrentDate()$
$startDate \leftarrow arc.getMinDate() + currentDate$
$endDate \leftarrow arc.getMaxDate() + currentDate$
if $startDate > transition.getStartDate()$ then
    $transition.setStartDate(startDate)$
end if
if $endDate < transition.getEndDate()$ then
    $transition.setEndDate(EndDate)$
    $petriNet.addPriorityTransitionAction(transition, END, endDate)$
end if{/* if an END action already exist for this transition, it will be disabled */}
if $transition.allInGoingArcAreActive()$ then
    $petriNet.addPriorityTransitionAction(transition, START, startDate)$
end if{/* if an START action already exist for this transition, it will be disabled */}
```



(a) The token is labeled with a value 15

(b) After the crossing of the first transition, the token is label-led with the value 5

Figure 7. An example of the spreading of a label-led token

- The makeOneStep algorithm will precise a token value when a transition is crossed, matching the remaining time after the crossing (currentTime - actionTime).
- When an arc is stated as active, the token value must be subtracted from the minimum and maximum arc values. See Figure 7.
- If this subtraction makes the transition crossable (for example if the value on the second arc on the Figure 7 is $[15; 20]$), this transition should be added to a transition list, handled at the end of the makeOneStep algorithm. And this should be repeated as often as the numbered token are disseminated in the Petri network.

3) Processes: The new speed must also be given to all currently running processes for them to adapt their computation, and must be provided to each processes launched next.

4) GOTO: The GOTO currently implemented can be seen as a very fast acceleration, where all dynamic transitions are turned into static transitions. It was an acceptable solution for a first version of the VIRAGE sequencer, but it has several limits.The artists using the VIRAGE sequencer made a lot of feedbacks about it.

Algorithm 2 makeOneStep

```
Require: $petriNet$
$transitionActionFIFO \leftarrow petriNet.getTransitionActionFIFO()$
$currentDate \leftarrow petriNet.getCurrentDate()$
while $(transitionActionFIFO.size() \neq 0) \wedge (transitionActionFIFO.top().date() \leq currentDate)$ do
    $topAction \leftarrow transitionActionFIFO.top()$
    $topTransition \leftarrow topAction.getTransition()$
    $transitionActionFIFO.pop()$
    if $topAction.isEnable()$ then
        if $topAction.getType() = START$ then
            $topTransition.declareAsSensitized()$
        else if $topAction.getType() = END$ then
            if $topTransition.allInGoingArcsAreActive()$ then
                $topTransition.crossTransition()$
            else
                $throw\ incoherentStateException$
            end if
        end if
    end if
end while
$sensitizedTransitionList \leftarrow petriNet.getSensitizedTransitionList()$
for each $currentSensitizedTransition$ in $sensitizedTransitionList$ do
    if $!currentSensitizedTransition.allInGoingArcsAreActive()$ then
        $sensitizedTransitionList.remove(currentSensitizedTransition)$
    else if $(petriNet.hasReceivedEvent(sentizedTransition.getEvent())) \vee (sensitizedTranstion.isStatic())$ then
        $sentizedTranstion.crossTransition()$
        $sensitizedTransitionList.remove(currentSensitizedTransition)$
    end if
end for
petriNet.resetEvents()
```

The first one is that if all processes are played rapidly, they are also played integrally. When a GOTO is performed, the artists usually do not want all the intermediate values, but only the last state of each processes. For example, if a processes computes a sound fade-in in a normal execution

Algorithm 3 crossTransition

```
Require: $transition\ petriNet$
$inGoingArcs \leftarrow transition.getInGoingArcs()$
for each $inArc$ in $inGoingArcs$ do
    $inArc.consumeToken()$
    if inArc.nbToken = 0 then
        $transitionList = inArc.getPlace().getSuccessorsTransition()$
        for each $transitionToReset$ in $transitionList$ do
            $transitionToRest.resetArcState()$
        end for{/* Resetting a transition arc state means looking for all predecessors places (after disabling the END and START action), and activate corresponding arc (with the previous values) if there is still a token in the place */}
    end if
end for
for each $externAction$ in $transition.getExternAction()$ do
    $externAction.execute()$
end for{/* an action could be a process start or end */}
$outGoingArcs \leftarrow transition.getOutGoingArcs()$
for each $outArc$ in $outGoingArcs$ do
    $outArc.produceToken()$
end for
$transition.resetArcState()$
```

; in a GOTO situation, only the last value is useful. A good solution can be to execute all processes without sending the results, and only broadcast the last results of each processes. Another solution can be to regularly save the ECOMachine state, and its processes, and perform the GOTO from the closest saved state.

The second one is that some processes can not be accelerated, for example a light, which need 5 seconds to be correctly initialized. A solution can be to precise some processes as GOTO-rigid, and execute them completely even in a GOTO situation.

Finally, when some processes are in the correct state, for example the light is correctly initialized by a previous execution, artists do not want to have a complete reinitialization. A solution can be to have an interaction with this processes by asking is current state, and skip the GOTO-rigid part if the initialization is already performed.

C. Validation

These features were tested and validated by the artists involved in the VIRAGE project. An Agile method (SCRUM) was set up to improve communication between developers and artists. A bug tracker allowed fast corrections and performances were made to test and present these features during frequent meetings [3].

IV. Conclusion

In this paper we presented a novel system for controlling in real-time the temporal unfolding of multimedia processes. For that purpose, we mix several temporal paradigms. Firstly, we use a time-line model to place processes start and end dates as well as temporal relations between them. Secondly, the execution uses a time-flow model in which the processes are executed while holding the temporal relations stated on the time-line between dates. In this last model, temporal synchronization is performed thanks to discrete controls associated to processes dates, while continuous controls can be performed to control the speed of the processes. The implementation of such continuous controls in a discrete model as a Petri net was not straight-forward. We proposed solutions that have to be enhanced. In particular, the GOTO feature should not execute all processes but only those, which have a persistent effect on the future. As a matter of fact, a fade-in, which is followed by a fade-out of the light can be completely skipped. However, the move of a camera should be performed before the recording process execution. Logical relations have to take place between processes in order to skip processes properly. In the future, we want to open the scripts in order to give choices to the user, which can depend on what is going on in real-time. Such a system should be useful also in the context of museography and improvisation.

References

[1] A. Cont, “Antescofo : Anticipatory synchronization and control of interactive parameters in computer music,” in *Proc. of the International Computer Music Conference (ICMC08), Belfast, North Irland*, 2008.

[2] R. Dannenberg, “A language for interactive audio applications,” in *Proc. of the International Computer Music Conference (ICMC02), San Francisco, USA*, 2002.

[3] A. Allombert, R. Marczak, M. Desainte-Catherine, P. Baltazar, and L. Garnier, “Virage : Designing an interactive intermedia sequencer from users requirements and the background,” in *Proc. of the International Computer Music Conference (ICMC10), New-York, USA*, 2010.

[4] A. Allombert, M. Desainte-Catherine, J. Larralde, and G. Assayag, “A system of interactive scores based on qualitative and quantitative temporal constraints,” in *Pr. of the 4rd International Conference on Digital Arts (ARTECH 2008), Porto, Portugal*, November 2008.

[5] A. Allombert, “Aspects temporels d’un système de partitions musicales interactives pour la composition et l’interprétation,” 2009.

[6] M. Diaz, *Petri Nets: Fundamental Models, Verification and Applications*. Wiley-Blackwell, 2008.

Complexity Scalable Video Decoding Scheme for H.264/AVC

Hoyoung Lee[1], Jaehwan Kim[2], Luong Pham Van[3], Bongsoo Jung[4], Kwangpyo Choi[5], Younghun Joo[6], and Byeungwoo Jeon[7]
[1,2,3,7] Departments of Electronic and Electrical Engineering, Sungkyunkwan University, Suwon, KOREA
[1]hoiing@skku.edu, [2]kjh4759@skku.edu, [3]pvluong@skku.edu,[7]bjeon@skku.edu
[4,5,6] Samsung Electronics Co., LTD., Suwon, KOREA
[4]bongsoo.jung@samung.com, [5]kp5.choi@samsung.com, [6]mrzoo@samsung.com

***Abstract*— Recent proliferation of portable devices requires video contents playable on virtually any portable devices. However, their limited computing resources pose significant restriction to accomplish real-time decoding of high resolution or high quality video contents. To solve the problem, we propose a complexity scalable video decoding scheme for real-time playback on portable devices. In this paper, we analyze the complexity of H.264/AVC video decoding elements and develop a complexity scalable decoding scheme by simplifying motion compensation and deblocking filtering. Experimental results with the H.264/AVC main profile coded bitstream show that its decoding complexity can be reduced up to 26% without significant loss in subjective quality as compared to the conventional H.264/AVC decoder.**

Keywords-H.264/AVC; video decoder; complexity scalable decoding

I. Introduction

The development of portable devices and prevalence of wireless communication infrastructures make the multimedia services very popular. The video services such as IPTV, video conferencing, or video telephony are the most popular multimedia services. However, the limited battery capacity and the computational performance of the portable devices are major restriction to implement real-time playback of video contents on portable devices. Moreover, increasing demand of playback of higher resolution or quality video contents on portable platforms makes it quite difficult to attain real-time video.

To attain the real-time video decoding, we have to decrease its workload by reducing computational complexity of decoding process. The significant problem in such reduction is the possible huge degradation of objective quality due to processing mismatch between encoder and decoder. Since distorted pictures caused by reduced complexity process are used again for reference picture for the following pictures to be decoded, propagation of distortion error will become larger so that it causes even more degradation of objective quality as time goes on. However, due to some characteristics of human visual system, certain degradation in video quality might be tolerable to a certain extent.

Therefore, many complexity scalable video decoding algorithms have been developed to make a good compromise between complexity reduction and subjective quality drop [1]-[4]. Peng [1] proposed a discrete-cosine transform (DCT)-based complexity scalable video decoder via pruning the DCT data. Chen *et al.* [2] expended the IDCT pruning approach [1] by using a simpler interpolation filtering method according to frame types in motion compensation. Lei *et al.* [3] proposed a complexity scalable algorithms in AVS video codec (Audio and Video coding Standard in China) using a loop filter and luminance interpolation in motion compensation scaling method. Its encoder sends some information about the loop filter and the luminance interpolation to a decoder for complexity control of a decoder. H. Nam *el at* [4] proposed complexity scalable H.264 decoder with downsized decoding. W. Ji *el at* [5] proposed energy-scalable video decoding algorithms.

The H.264/AVC standard achieves high coding efficiency with many advanced coding tools such as variable block size motion compensation, multiple reference frames, quarter-pel motion vector accuracy, context adaptive entropy coding, etc. [6]. To implement a complexity scalable H.264/AVC decoder, we analyze its decoding complexity by decoding tool by tool and develop some complexity control parameters which our proposed complexity scalable decoding scheme utilizes in its decoding complexity control.

The rest of the paper is organized as follows. In Section II, we analyze H.264/AVC video decoding functions from the viewpoint of complexity control. In Section III, we describe the proposed method for complexity control of decoder. Experimental results are given in Section IV. Finally, we make some conclusions in Section V.

II. Video decoding elements for complexity control

H.264/AVC decoder performs variable length decoding (VLD) of incoming bitstream and then reconstructs various syntax elements such as motion vector, reference index, quantization parameter, and residual data of slices. The residual data are obtained through inverse quantization (IQ) and inverse transform (IT). Following, they are combined with a predictor which is generated either by motion compensation or by intra prediction. Subsequently, reconstructed picture is generated through deblocking filtering process. To evaluate the video decoding elements from the view point of complexity control, we analyze the complexity of them. Table I depicts the complexity profiling

ISBN: 978-1-61208-129-8

TABLE I. COMPLEXITY PROFILE OF DECODING ELEMENTS

Decoding elements	Complexity rate(%)
Motion compensation	27.51
Variable length decoding(VLD)	25.19
Deblocking filter	16.65
IQ/IT	10.65
Reconstruction	3.08
Intra prediction	0.57
Others	16.34

result of H.264/AVC decoder in terms of decoding time. Its analysis is based on bitstreams conforming to the H.264/AVC main profile with IBBPBBP structure where every 60th frame is coded as I picture.

As shown in Table I, motion compensation and VLD are the major complex elements in a video decoder. This means, other than the VLD, motion compensation is the most complex process in decoder. Second major complex element is the deblocking filter, which is applied to reconstructed picture after finished the decoding process. It is an important element to improve a subjective quality of reconstructed video, especially, in low bit rates bitstreams. By the way, IQ/IT takes only about 10% of complexity in which IT occupies more computational complexity than IQ. In the previous investigation [1], it is found out that complexity control of IT process brings a significant quality loss. Therefore, in this paper, we decide two decoding functions of motion compensation and deblocking filter for complexity control.

A. Motion compensation

In H.264/AVC motion compensation for luma component, pixel value at fractional quarter-pel positions generated according to motion vectors. Fig. 1 depicts their positions. Predicted values at half-pel positions are generated by an one-dimensional 6-tap FIR interpolation filtering horizontally or vertically. A sample value at quarter-pel position is generated by averaging values at two nearest half-pel and integer positions. Computational complexity to generate the fractional samples is different depending on sample positions as depicted in Table II. Samples at quarter-pel positions labeled as *f*, *i*, *k*, *q* are the most complex positions. On the other hands, half-pel samples labeled *b*, *h* are the least complex ones. To reduce the complexity of interpolation filtering, we simplify the interpolation filtering of each sample position using adjacent integer-pel samples. For example, a simplified sample value at quarter sample position labeled as *a* is derived,

$$\begin{aligned} a &= (G+b+1)/2 \\ &= \{G+(E-5\cdot F+20\cdot G+20\cdot H-5\cdot I+J+16)/32+1\}/2 \\ &= (E-5\cdot F+52\cdot G+20\cdot H-5\cdot I+J+48)/64 \\ &\simeq (48\cdot G+16\cdot H+32)/64 \end{aligned} \tag{1}$$

where *G* is at an integer position as depicted in Fig. 1, and *b* is at half-pel position derived by 6-tap FIR interpolation filtering. To use a shift operation, we adjust a rounding value appropriately. In Table II, we propose a simplified luma interpolation filtering of each fractional sample. Motion compensation for chroma components are generated by a bilinear interpolation of four neighboring integer samples as,

$$\begin{aligned} a = ((8-xFrac_c)\times(8-yFrac_c)\times A+xFrac_c\times(8-yFrac_c)\times B+ \\ (8-xFrac_c)\times yFrac_c\times C+xFrac_c\times yFrac_c\times D+32)/64 \end{aligned} \tag{3}$$

where *a* is a predicted chroma sample value and *A*,*B* *C*,*D* are the integer-pel position samples. $xFrac_c$ and $yFrac_c$ are the fractional offsets of fractional samples. To reduce the complexity for chroma interpolation filtering, a predicted chroma sample is copied from nearest neighboring integer-pel sample.

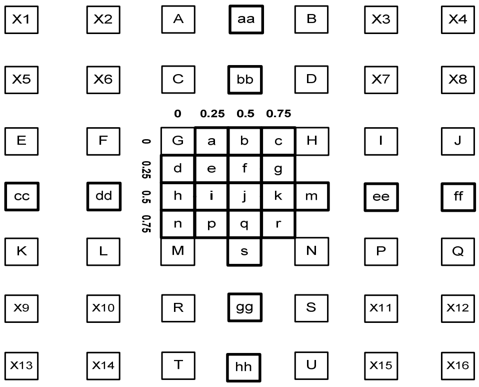


Figure 1. Fractional sample positions for quarter sample luma interpolation

TABLE II. COMPLEXITY COMPARISON OF INTERPOLATION FILTERING FOR H.264/AVC AND THE PROPOSED METHOD

	Sample position	Interpolation by H.264/AVC	Simplified interpolation by the proposed
I	b(0.5,0)	6-tap	*(G+H+1)/2*
	h(0.5,0)	6-tap	*(G+M+1)/2*
II	a(0.25,0)	6-tap+2-tap	*(48G + 16H+32)/64*
	c(0.75,0)	6-tap+2-tap	*(16G + 48H+32)/64*
	d(0,0.25)	6-tap+2-tap	*(48G + 16M+32)/64*
	n(0,0.75)	6-tap+2-tap	*(16G + 48M+32)/64*
III	e(0.25,0.25)	(6-tap)×2+2-tap	*(2G + H+M+2)/4*
	g(0.25,0.75)	(6-tap)×2+2-tap	*(G + 2M+N+2)/4*
	p(0.75,0.25)	(6-tap)×2+2-tap	*(2H + N+G+2)/4*
	r(0.75,0.75)	(6-tap)×2+2-tap	*(H + 2N+M+2)/4*
IV	j(0.5,0.5)	(6-tap)×6+6-tap	*(G + M + H + N+2)/4*
V	f(0.5,0.25)	(6-tap)×6+6-tap+2-tap	*(3G + 3H + M + N+4)/8*
	i(0.5,0.75)	(6-tap)×6+6-tap+2-tap	*(G + H + 3M + 3N+4)/8*
	k(0.25,0.5)	(6-tap)×6+6-tap+2-tap	*(3G + H + 3M + N+4)/8*
	q(0.75,0.5)	(6-tap)×6+6-tap+2-tap	*(G + 3H + M + 3N+4)/8*

B. Deblocking filter

In H.264/AVC video coding, its deblocking filter consists of three processing phases: boundary strength decision, filtering decision, and actual pixel filtering.

In the boundary strength decision of H.264/AVC, boundary strength parameter (BS) is determined by the rules in [6] for each boundary of 4×4 block. BS can be 0 ~ 4 according to the rules. In our experiment, we found out that most often selected BS value is 0 or 2. Furthermore blocking artifact is more noticeable in flat and simple regions than in complex textured regions [7]. In H.264/AVC, flat and simple regions are often predicted in large partitions such as 16×16 or 16×8, 8×16. On the other hand, complex regions are mostly coded under 8×8 sub-block partitions. Using these observations, we re-design a boundary strength decision process as shown in Fig. 2.

Filtering process for sample sets (p_0, q_0) only takes place when following condition is satisfied [6]:

$$\begin{gathered} BS > 0 \ and \\ |p_0 - q_0| < \alpha \,\&\&\, |p_1 - p_0| < \beta \,\&\&\, |q_1 - q_0| \le \beta \end{gathered} \tag{4}$$

where α, β are thresholds dependent on quantization parameter QP. Since filtering decision process has lots of comparison operation and it is performed for each edge, complexity would be much increased. In our simplified deblocking filter, filtering decision is performed just one time per each 4×4 block using an average value of samples such as:

$$\begin{gathered} BS' > 0 \ and \\ |p_{0Avg} - q_{0Avg}| < \alpha \,\&\&\, |p_{1Avg} - p_{0Avg}| < \beta \,\&\&\, |q_{1Avg} - q_{0Avg}| \le \beta \end{gathered} \tag{5}$$

where p_{Avg} and q_{Avg} is an average value of samples in P and Q 4 4 blocks.

After the boundary strength decision and filtering decision, actual filtering process is applied to each block boundary. In H.264/ AVC, filtering strength is different depending on BS value. If BS < 4, a 4-tap FIR filter is applied with input samples p_0, p_1, q_0, q_1, and producing outputs p_0' and q_0'. In case of BS being 4, a 5-tap or 4-tap filter is applied according to conditions [6]. However, the proposed simplified deblocking filter applies a 2-tap FIR weak filter when BS' of a current block boundary is 1 according to the proposed boundary strength decision rules. When BS' value is larger than 1, we apply same filtering method like to that of H.264/AVC.

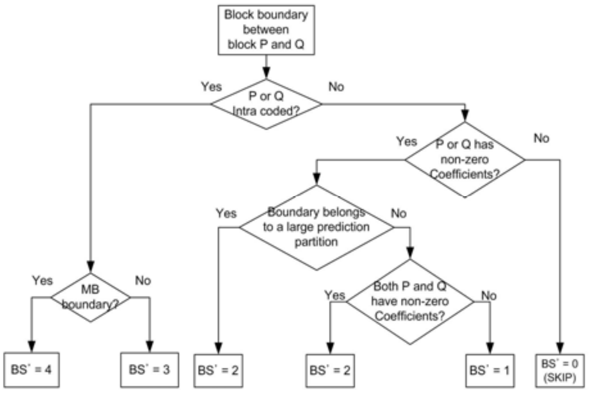

Block boundary between block P and Q
Yes
P or Q Intra coded?
No
Yes
P or Q has non-zero Coefficients?
No
Yes
Boundary belongs to a large prediction partition
No
Yes
MB boundary?
No
Yes
Both P and Q have non-zero Coefficients?
No
BS' = 4
BS' = 3
BS' = 2
BS' = 2
BS' = 1
BS' = 0 (SKIP)

Figure 2. Simplified boundary strength decision

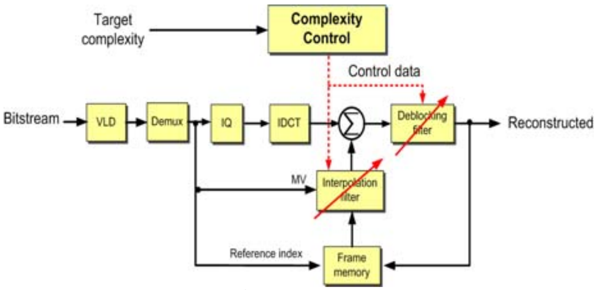


Figure 3. Block diagram of proposed complexity scalable decoder

TABLE III. PROPOSED MOTION COMPENSATION COMPLEXITY REDUCTION LEVEL (luma)

MCR_{Level}	Complexity reduction sample position
0	No reduction
1	*f,i,q,k*
2	*j*+ (MCR_{Level}=1)
3	*e, p, r, g*+ (MCR_{Level}=2)
4	*a, c, d, n*+ (MCR_{Level}=3)
5	*b, h*+ (MCR_{Level}=4)

TABLE IV. PROPOSED COMPLEXITY REDUCTION LEVEL FOR DEBLOCKING FILTER

DFR_{Level}	I Slice	P Slice	B Slice
0	No reduction		
1	a	a	b
2	a	b	b
3	b	b	b
4	b	b	c
5	c	c	c

(a: conventional deblocking filter, b: simplified deblocking filter, c: forced deblocking filter off)

III. COMPLEXITY SCALABLE VIDEO DECODING SCHEME

Fig. 3 depicts a block diagram of the proposed complexity scalable decoding scheme. The complexity scalable video decoder has a minimum quality loss within the maximum complexity reduction by controlling the control parameters. Therefore, we have to find an optimal complexity control level of parameters which satisfy the minimum quality loss. To find an optimum control level of parameters, we evaluate complexity-distortion (C-D) performance according to various complexity control parameters.

A. Motion compensation complexity control

In this paper, we apply a simplified interpolation filtering method according to fractional sample positions as proposed in Table II. The complexity scalability for motion compensation can be attained by controlling the number of luma samples which are involved in the simplified interpolation filtering. Table III depicts a proposed motion compensation complexity reduction level (MCR_{Level}). As depicted in Table III, we reduce the complexity of interpolation filtering for luma samples from the most complex sample position labeled as *f, i, k, q* to the least complex position labeled as *b, h* according to MCR_{Level}. Furthermore in order to control the motion compensation for chroma samples, we apply the proposed simplified chroma interpolation method when the MCR_{Level} is larger than 0.

ISBN: 978-1-61208-129-8

B. Deblocking filter complexity control

To control the complexity of deblocking filter, we apply a simplified deblocking filter and selectively make the deblocking filter off according to slice type as depicted in Table IV. We define a complexity control level for deblocking filter, namely deblocking filter reduction level (DFR_{Level}). When DFR_{Level} is 0, we use the conventional H.264/AVC deblocking filter without any complexity reduction. For increasing DFR_{Level}, we apply the proposed simplified deblocking filter from B slice to I slice. To attain the maximum complexity reduction of deblocking filter, we switch off the deblocking filtering process for all slice types.

C. Complexity control scheme

We define two complexity control parameters as above; those are DFR_{Level}, MCR_{Level} respectively. To find an optimum control level which satisfies a minimum distortion loss, we evaluate complexity-distortion(C-D) performance [8]. The complexity reduction and distortion is measured by following equations:

$$AST[\%]=\frac{DecodingTime(\text{reference})-DecodingTime(\text{proposed})}{DecodingTime(\text{reference})}\times 100 \qquad (11)$$

$$\Delta PSNR = PSNR(\text{proposed}) - PSNR(\text{reference})$$

where AST is an average saving time and ΔPSNR is a difference in objective quality between reference and the proposed method and used for quality distortion.

Fig. 4 depicts complexity-distortion performance in terms of (DFR_{Level}, MCR_{Level}). In Fig. 4, each point represents MCR_{Level}s according to $DFR_{Level.}$ And the optimal complexity control points are drawn with a line. Table V shows the complexity control parameter level *g* determined by (DFR_{Level}, MCR_{Level}) which minimize distortion and maximize complexity reduction as shown in Fig. 4. We have set up 6 different complexity levels and expected distortion and complexity reduction according to complexity level *g* as in Table V.

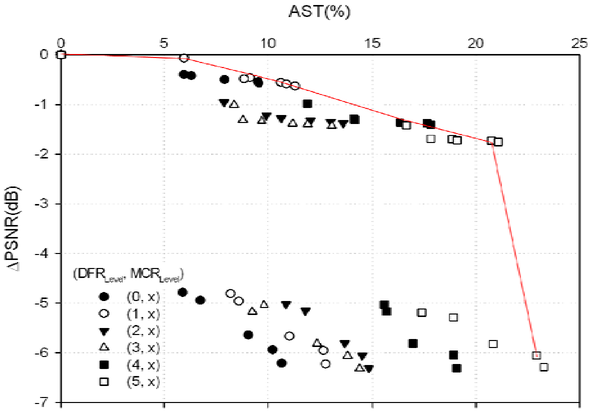


Figure 4. Joint complexity-distortion curve

TABLE V. JOINT COMPLEXITY REDUCTION LEVEL

Complexity level(*g*)	DFR_{Level}	MCR_{Level}	Relative Distortion	AST(%)
0	0	0	0	0
1	1	0	-0.072	5.9533
2	1	3	-0.5625	10.6242
3	4	3	-1.3635	16.325
4	5	4	-1.735	20.7446
5	5	5	-6.2835	23.2606

TABLE VI. EXPERIMENTAL RESULTS OF THE PROPOSED METHOD

Seq.	*g*	ΔPSNRY	ΔPSNRCb	ΔPSNRCr	AST(%)
city	0	0	0	0	0
	1	-0.06	-0.01	0.00	5.41
	2	-0.70	-3.97	-4.39	12.89
	3	-1.37	-4.03	-4.48	18.61
	4	-1.33	-4.76	-5.16	22.84
	5	-7.04	-4.76	-5.16	26.06
harbour	0	0	0	0	0
	1	-0.11	-0.03	-0.03	5.89
	2	-0.91	-3.99	-2.84	10.47
	3	-1.77	-4.43	-3.26	17.95
	4	-2.24	-5.64	-4.40	22.68
	5	-7.93	-5.64	-4.40	25.08

IV. EXPERIMENTAL RESULTS

To evaluate performance of the proposed scheme, we implemented it on JM17.0 H.264/AVC reference software. Bitstreams for experiments are coded as H.264/AVC main profile with GOP size = 60 under IBBPBBP structures. The number of reference frames is 5, and one picture is coded as one slice. QP is set to 22, 27, 32, and 37. The number of total encoded pictures is set to 300. We used two standard definition sequences for experiments: city and harbor.

The performance of the proposed scheme is measured by AST(%) and ΔPSNR(dB). Table VI shows experimental results of the proposed scheme. We can see that the proposed scheme attains similar complexity scalability and complexity reduction as estimated. However, we can find a huge distortion when complexity control level *g* is 5. Since P slice is used to as a reference slice to another slices, the error according to complexity control propagates to another slices. When *g* is 5, we reduce the complexity of motion compensation in P slices to attain the maximum complexity reduction. The errors in motion compensation in P slices are propagated to other slices. Fig. 5 shows the subjective qualities according to complexity control level *g*. As depicted in Fig. 5, we can identify that there is no significant subjective quality loss until *g*=4. However, the complexity control level *g* is 5, we can see some subjective quality loss, but it also maintains acceptable visual quality.

V. CONCLUSIONS

In this paper, a complexity scalable H.264/AVC decoding scheme is proposed using two control parameters. The proposed scheme can control complexity of a decoder with variable complexity control levels with the minimum quality loss. However, the proposed scheme has a restriction according to slice type. Our future work will find another complexity control variables which can control the complexity regardless of slice types and reduce more complexity as well. We will also develop complexity estimation and control method.

ACKNOWLEDGMENT

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (No. 20100008030).

REFERENCES

[1] S. Peng, “Complexity scalable video decoding via IDCT data pruning,” in Proc. of IEEE conference on consumer electronics 2001. Los Angeles, CA, June 2001.

[2] Y. Chen, Z. Zhong, T. Lan, S. Peng, and K. van Zon, "Regulated complexity scalable MPEG-2 video decoding for media processors," IEEE Transactions on Circuits and Systems for Video Technology, vol.12, no.8, pp. 678- 687, Aug 2002.

[3] C. Lei, Y. Chen, and W. Ji., “A complexity scalable decoder in an AVS video codec,” in Proc. of the 6th International Conference on Advances in Mobile Computing and Multimedia (MoMM '08), ACM, New York, USA.

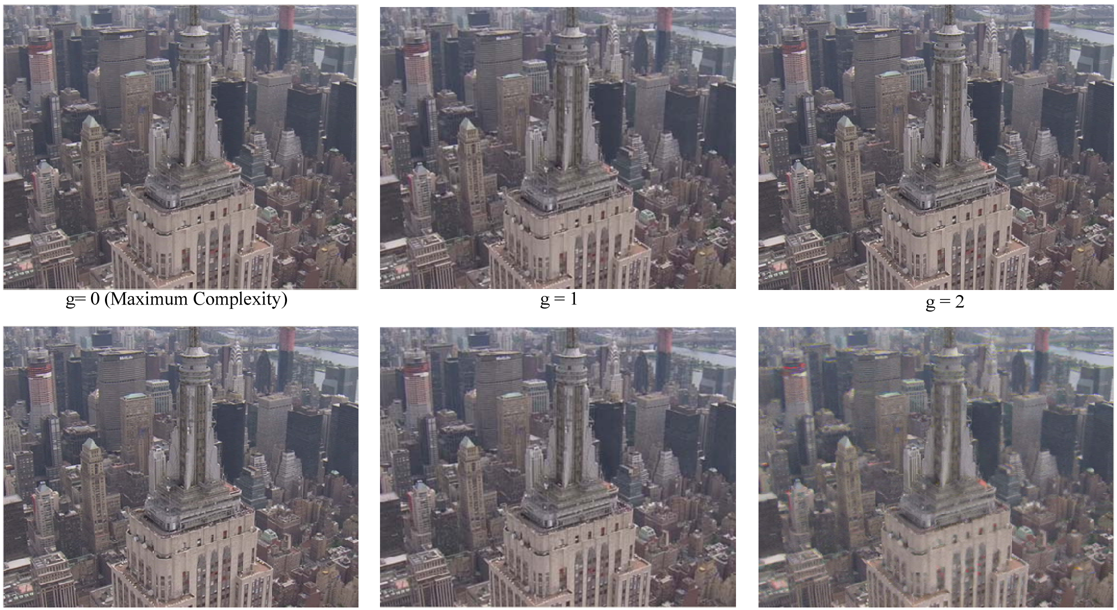
[4] H. Nam, J. Jeong, K. Byun, J. Kim, and S. Ko , "A complexity scalable H.264 decoder with downsizing capability for mobile devices," Consumer Electronics, IEEE Transactions on , vol. 56, no. 2, pp. 1025-1033, May 2010.

[5] W. Ji, M. Chen, X. Ge, P. Li, and Y. Chen, “ESVD: An Integrated Energy Scalable Framework for Low-Power Video Decoding Systems,” EURASIP Journal on Wireless Communications and Networking, vol. 2010, Article ID 234131, 14 pages, Jun. 2010.

[6] T. Wiegand, G. J. Sullivan, G. Bjøntegaard, and A. Luthra, “Overview of the H.264/AVC video coding standard,” IEEE Trans. Circuits Syst. VideoTechnol., vol. 13, no. 7, pp. 560–576, Jul. 2003.

[7] S. D. Kim, J. Yi, H. M. Kim, and J. B. Ra, “A deblocking filter withtwo separate modes in block-based video coding,” IEEE Trans. Circuits Syst. Video Technol., vol. 9, pp. 156–160, Feb. 1999.

[8] H. Lee, B. Jung, J. Jung, and B. Jeon, “Computational complexity scalable scheme for power-aware H.264/AVC encoding,” in Proc. of IEEE International Workshop on Multimedia Signal Process (MMSP '09), Rio de Janeiro, Brazil, pp. 1–6, Oct. 2009.



g= 0 (Maximum Complexity)

g = 1

g = 2

g = 3

g = 4

g= 5 (Minimum Complexity)

Figure 5. Subjective quality of proposed method (City sequence, QP 22 299th slice)

A Formal Model for the Specification and Analysis of HLA-based Distributed Multimedia Interactive Simulation using Hierarchical Time Stream Petri Nets

Akram Hakiri[1,2], Michel Diaz[1,2]
[1]*CNRS ; LAAS ; 7 avenue du colonel Roche, F-31077 Toulouse, France*
[2] *Universit de Toulouse ; UPS, INSA, INP, ISAE ; LAAS ; F-31077 Toulouse, France*
Email: {Hakiri, Diaz}@laas.fr

***Abstract*—This paper proposes a formal model for the specification and analysis of distributed multimedia simulation. This model is based on Hierarchical Timed Stream Petri Nets (HTSPN), which has been proposed for specifying temporal and logical constraints in high level multimedia description and simulation. It takes into account a powerful synchronization definition between different flows issued from distributed multimedia systems. A simulation was done using a special Java-based framework to assess the methodology and analyze the expression and interpretation power of HTSPNs. For instance, such an interpreted model permits powerful analysis techniques for validating the quality of service in computer networks before protocol implementation. Consequently, it allows the specification of both the temporal non-determinism of weakly distributed applications and the temporal variability of the multimedia processing. An example is used to demonstrate the capabilities of this scheme to specify the QoS requirements of simulated applications.**

Keywords-Formal Model; Distributed Multimedia Simulation; HLA; HTSPN.

I. Introduction

The specification and the verification of temporal and logical properties of distributed multimedia interactive simulation is a fundamental step to be conducted before implementation. Therefore, on one hand, synchronization schemes [6] bring important contributions to the emerging concepts of distributed simulation systems, especially when these systems must maintain temporal relations between various streams. On the other hand, HLA-based applications [1]) need structural approaches to specify the synchronization scenarios between intra-flow, inter-flows and inter-objects to allow an adequate management of the system resources. This paper suggests to use a formal model based on Hierarchical Stream Timed Petri Nets to specify and analyze synchronization constrains between synchronized units in intra-flow and inter-flow cases for the specification and the verification of the next generation of distributed interactive multimedia simulation.

The proposed model is applied to an HLA based simulation which includes audio, video and interactive streams issued from a selected application. Using the power modeling of Petri Nets suggests the specification of a requested quality of services in distributed asynchronous multimedia application. The synchronization scheme developed here discus applications that involve HLA and are built on HLA-RTI APIs. Its aim is to facilitate the editing phase and the development time required to deliver high fidelity simulation that will respect all structural, temporal and logical application related constraints.

This paper is organized as follow: after a brief introduction, Section II introduces the motivation of multimedia formal specification. In Section III Petri Nets have been selected for specifying distributed multimedia applications. Section IV presents a set of QoS requirements to be used in distributed multimedia simulation. Section V introduces the formal model and shows analysis results. Transport architecture is presented in Section VI and conclusion is given in Section VII.

II. Motivation

In the general case, flows need to satisfy natural synchronization constraints and synthetic synchronization constraints between applications. The natural synchronization constraints are intrinsic to the flow itself and need to be respected when presented to the remote hosts to ensure the comprehension of the associated information. For example, in SECAM systems, a video frame is be displayed 25 times per second. These constraints are given by codecs. Synthetic constraints are imposed by the application itself and results from the abstract global synchronization specified by developers. For instance, an audio stream must start when a given event occurs. To handle the granularity of these constraints, synchronization units have to process the information units, providing a way to modulate a synchronization scheme for each flow and to provide an optimal control of each flow with respect to the system resources.

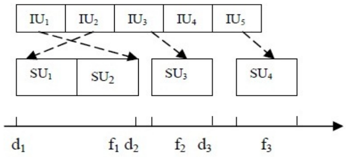


Figure 1. Correspondence between Information Units and Synchronization Units

Figure 1 shows the correspondence between the information units (UI) and the synchronization units (SU). Each synchronization unit, noted SU_i, is associated to a sequence of information units (IU_i), together with each starting date d_i and finishing date f_i. The granularity of the information induces the performance of the scheduling protocol. In order to process streams using a set of available resources, the synchronization scheme has to use synchronization units adapted to the specific flow and to the synchronization of the media acquisition.

The specification of the synchronization scheme should be specified using a formal model to conduct an editing and a verification phase before any implementation. The analysis of the possible expressive power required for distributed interactive systems lead to select a formalism based on Petri nets (PN).

Other research contributions have been proposed to provide a formal approach for the specification of the distributed multimedia communication. Authors in [2] and [7] used Petri Nets to simulate a complex military simulation system with HLA, to manage its concurrent properties and to specify the synchronization problems for the simulated commands. Author in [10] explores the impact of Time Service Management in HLA (HLA Time Management Service) to specify an engine based on Stochastic Petri Nets to run the distributed simulation, and proposes the use of HLA as a platform of reference to compare different approaches for partitioning and distributing application executions. [8] presents an approach based on Colored Petri Nets to reduce the bandwidth usage for distributed simulations using HLA. In [9], the authors propose a colored temporal PN model for simulating the federation execution. The proposed model aims to assist developers of HLA simulations to design high-level simulation and to specify the constraints of the simulation.

However, as we outlined in our previous work [3] also with other related works, these approaches do not provide a structured model, and do not provide a comprehensive qualitative analysis of the simulation. Furthermore, no quantitative analysis was presented, particularly when specifying the temporal constraints and the performance analysis of information exchanged during the simulation. As a consequence, in this paper we use the same Hierarchical Time Stream Petri Nets formalism to extend the power of the previous models to express the spatial, temporal, logical and semantic structures that appear in the distributed interactive multimedia simulation.

Our contribution, is an extension this previous works, but with another validation tool, uses primarily a temporal model because it induces a required flexible management of system resources and allows expressing of the non determinism that may occur when a time de-synchronization occurs between different distributed streams, especially when these flows are very heterogeneous, such as the union of streaming media (audio, video, images) and streaming interaction flows coming from the actors of the virtual environments. That is, we can find a tradeoff between two targets: synchronization of stream to reduce the end-to-end latency and eliminating delay jitter. Hence, we aim to improve those QoS parameters and we add real-time scheduling approach for stream synchronization.

III. Hierarchical Time Stream Petri Nets

The HTSPN [4] (see also the HTSPN formalism in our previous work cited in Section II) model is an extended Petri Net model that used timed arcs for the modeling of multimedia processing (communication, presentation...). The temporal jitter appearing inside weakly synchronous multimedia systems is modeled by the arc Temporal Validity Interval (TVI). These arcs TVI are tuple [x, n, y], where x, n and y are respectively the minimum, the nominal and the maximum admissible durations of the related processing. Such way of multimedia systems modeling allows the expression of both the temporal non-determinism of weakly synchronization in distributed multimedia applications and the admissible temporal variability of multimedia objects.

Temporal drifts between multimedia streams can be fully and accurately specified with the help of 9 different synchronization semantics that can be selectively associated with transitions. As a consequence, HTSPNs appear to be a powerful tool for the formal modeling, analysis, verification and simulation of distributed multimedia simulation systems. HTSPN models allow three fundamental concepts to be formally described with powerful temporal extensions: the atomic, the composite and the link components.

Atomic Component: an atomic component is modeled in HTSPNs by an arc with a TVI and a place associated with one atomic resources type, for example video data with [8, 10, 12] as TVA. Atomic synchronization layers aim to describe synchronization constraints inside atomic components by specifying intra-stream synchronization.

Link Component: a link is modeled in HTSPN by a timed arc (L, t), where L is the link (to be layered) place. The TVA associated with the link introduces the timed link concepts. Using the HTSPN firing rules [5], timed links allow the modeling and the formal specification of the transversal semantics of the application layer.

Composite Component: the composite component provides a hierarchical structuring mechanism based on the recursive composition of atomic and composite component through the use of sub-nets. The HTSPN use these composite type places that are not only structurally, but also temporally, equivalent to a (sub) net. A composite layer is able to describe inter-stream synchronization constraints.

IV. QOS REQUIREMENTS IN DISTRIBUTED MULTIMEDIA SIMULATION

The quality of the mono-media presentation describes the quality of the discontinuity of a single stream. This discontinuity occurs for instance when data are lost; it can cause a significant loss of synchronization, and it becomes very important to optimize the quality of the presentation at the receiver side to present the application. The end to end latency defines the maximum allowable transfer delay between two remote entities. This period corresponds, for example, to the delay when a sender pronounces a word and when the receiver receives the sound. This delay should not exceed a given limit since it affects the interactive communication between the remote users. The intra-stream synchronization ensures the compliance with the time constraints of the timing units for each stream. The synchronization level is given for each flow by the temporal validity intervals (nominal delay, allowable jitter) of each synchronization unit. The intra-stream jitter is given by (1) and illustrated in Figure 2. $\tau(n)$ is the arrival time of object *n* and the maximum allowable jitter intra-flow (equation (2)) is then $2\times\epsilon$'.

$$\epsilon'_{min} \le \tau(n-1) - \tau(n) \le \epsilon'_{max} \tag{1}$$

τ(n) is an intra-flow object presented at time n, 2*ϵ' is the intra-flow allowable jitter.

$$-\epsilon'_{min} = T' - \epsilon' \tag{2}$$
$$\epsilon'_{max} = T' + \epsilon' \tag{3}$$

To ensure the receipt of *n* objects within a time interval, one has to guarantee the constraints of the quality of service of the intra-flow, i.e. the maximum value of the global jitter of the synchronization of n objects is given as the sum of all jitters that exist between all consecutive units (2*ϵ per period).

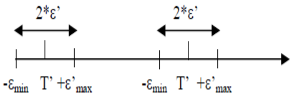


Figure 2. Jitter in intra-stream

For the intra-stream synchronization, the QoS requirements, that should be satisfied for instance when an audio and a video streams need to be synchronized, depend on the communication variability. For instance, at the receiver site, if two units of two different flows arrive at 2 different times t1 and t2, the correctness of their synchronization has to be deduced from the specification and the presentation constraints: the synchronization scheme should provide the acceptable interval for synchronized units of the flows, and should define some actions to eliminate the streams discontinuities. As an example, if a flow is behind the other(s) (is late) de-synchronization will occur and the discontinuity may become visible (when sound is no more synchronized with video, this problem is called ”Lip-Synchronization). Relation (3) and Figure 3 specifies a periodic traffic, with period T, and an inter-flow jitter equal to 2*ϵ for one period.

$$\epsilon_{min} \le \tau(x1, x2) \le \epsilon'_{max} \tag{4}$$
$$\epsilon_{min} = T - \epsilon \tag{5}$$
$$\epsilon'_{max} = T - \epsilon \tag{6}$$

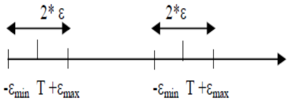


Figure 3. Inter-stream Jitter

It is clear from the above equations that the flows must be sent periodically. In particular, the packet size is a very important criteria that has to be carefully chosen for the QoS constraints to be fulfilled. Indeed, for example for audio data, the length of the packet affects the time required to produce it. Table I gives the packet size with respect to the data that has to be transmitted through a network. it shows how audio data should be prepared and sent over Networks. Column 1 presents the sampling frequency that produces the audio data. Column 2 gives the time necessary to produce an IP packet (1518 Bytes). For example, with a 8 KHz sample frequency, the time required to produce this packet is 189 ms. In order to fulfill the QoS requirements for distributed media application, this delay need to be short enough because it delays the packets and implies the quality of the interactivity and of the presentation at the receiver side. Then, Columns 3, 4 and 5 show the size of the frame, given an interval of time, to satisfy the quality of the presentation. It seems that a delay of 20 ms is for sure a correct value because it sends a high audio delay quality with the respect to the frame size. The requirement of low latency means that it is better for the senders to send small packets frequently rather than large packets seldomly. Let us assume that the acceptance purpose is to provide a 150 ms end-to-end latency: 50 ms can be taken as the maximum time allowed for preparing and sending a packet, also for processing and presenting it in the receiving application, and

Table I
RELATIONSHIP BETWEEN PACKET SIZE AND PROCESSING TIME

Frequency (Khz)	Time (ms) IP Packet	Frame Size in 50ms	Frame Size in 30ms	Frame Size in 20ms
8	189	400	240	160
11	69	1101	660	440
22	34	2201	1321	881
44	17	4403	2642	1761
96	8	9606	5764	3843

also can be the propagation delay in the network. Distributed multimedia applications are not only presentation driven; they are also data-driven.

Therefore, a formal model must describe both the data and their presentation. In addition, the model must provide means for representing the logical and temporal compositions of their interactions. In order to specify the best choice for an audio packet size, Figure 4 gives the functional point able to satisfy our interactive requirements. It displays the variability of the packet size with respect to the time needed to produce the packet. These curves provide the Temporal Validity Interval (TVI) which will be used in our formal model. It follows that the best value of TVI is given by [15, 20, 25], where 20 ms is the nominal time to produce a packet, and there is a maximal drift per period of 2*5=10 ms.

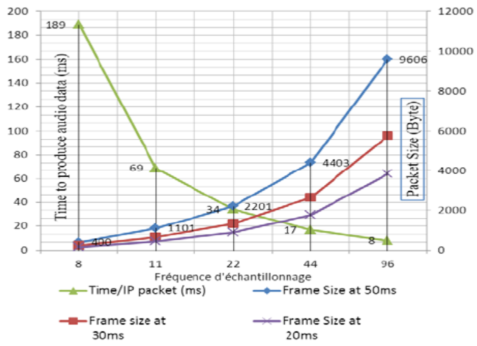


Figure 4. Variability of packet size with the processing time

The first curve in Figure 5 indicates the time required to produce a 1518 Byte audio data frame. From the left, this curve shows for example that a voice corresponding to a sequence of samples of 8 bits at a frequency of 8 kHz leads to produce a sample every 125 ms; for a packet size of 1518 Bytes, the sender must wait 189 ms to produce and start to send only this packet. This value does not fulfill the QoS requirements needed to transmit the packets of interactive applications. A 22 kHz sampling allows a processing and production time equal to 34ms. However, it does not lead after these 34 ms to a packet that exceed the proposed maximum size of 1518 Bytes, it is about 2201 Bytes. As said before, we selected 50 ms as the maximum time allowed to produce a packet at the sender side, and as a consequence it is clear in Figure 4 that it is not possible to send a full IP packet. If a 22 KHz sampling frequency would have been selected, it would have fulfilled all temporal and length QoS requirements: the delay to produce the audio data frame is 20ms for a packet size of 881 Bytes, but the packet size is rather short. Using the 50 ms values lead to start the specification of the formal model.

Notice that if some problems come from the network, and if then the different flows are not received at the same time, some application incoherence could results and the corresponding flows need to be re-synchronized, if possible, at the receiver side. For example, as applications of distributed simulations incorporate multimedia flows, together with flows resulting from the interactive system control, they may become incoherent after crossing a (wide area or other) disrupting network. To ensure consistency between these flows, an adequate synchronization scheme between these flows is necessary and has to be specified.

Such synchronization between the flows can be defined by successive steps, for example first by ensuring the synchronization in each streams, second between the different multimedia streams, and, third by ensuring the synchronization between these multimedia flows and the control flows of the distributed interactive simulation.

V. FORMAL MODEL OVER HLA-RTI

Basically, the application (Figure 5) is a platform for distributed interactive simulation, and it allows end users to interact by voice, video and distributed simulation events sent in real time. Such an application consists of at least three streams: the audio and video streams captured by a camera and the flow coming from the modification of the virtual environment of the distributed simulation. The synchronization scheme considered involves three types of flow synchronization:

- Intra-stream synchronization between the objects of each flow
- Inter-stream synchronization between the audio and video streams to meet the timing constraints often called Lip- Synchronization.
- Inter-stream synchronization between the two (audio / video) streams and the control stream of the distributed interactive simulation.

The intra-stream synchronization considers one flow, the inter-stream synchronization considers all flows, and specifies the acceptable inter-stream drift. The constraints of intra-stream synchronization which must be verified for each flow are:

- Units have an audio synchronization nominal duration of 20 ms by assuming a jitter of 5 ms. That is to say, the temporal validity interval of each unit of the sync audio is [15, 20, 25].
- The video synchronization unit has 40 ms as a nominal duration and a jitter of 10ms. The synchronization interval validity of the video is then [30, 40, 50].
- The synchronization units of the distributed interactive simulation flow have a nominal duration of 20 ms wit a 5 ms jitter. The temporal validity interval of this flow is then [15, 20, 25].

The corresponding HTSPN synchronization model is defined by a three levels representation: the link level considers

Copyright (c) IARIA, 2011. ISBN: 978-1-61208-129-8

the application level, and depends on the developer choices (the application reference is given in Figure 6. The temporal validity interval, [60, 80, 100], at this layer corresponds to the inter-stream synchronization and will be explained later on. It means that the transition will be fired in the interval min, max=[60, 100], the time being started when the transition is enabled, i.e. when the places have all one token.

Thus, knowing that the sound has to be produced and sent in less than 20 ms (Figure 4), and that the image in less than 40 ms (given by the application), we measured the processing of the interactive event: it has been found to be 16. The delay of the interactive flow must be driven by the audio stream because the audio media is the most time sensitive one, and the audio stream will be then selected master stream: it control the time schedule for the firings of the transitions. Therefore, the number of places of this stream must be a multiple integer of the number of the video and simulation units. As a consequence, the synchronization transition will be defined at the rendezvous which occurs at a period equal to the LCM (Least Common Multiple) of the nominal durations of the three streams, i.e. at time equal to 80 ms, the LCM of (16, 20, 40). The granularity of the synchronization is determined by the maximum acceptable inter-stream drift. As the audio stream has a possible drift of 5 ms, the advance of the interactive flow results only from the cumulative effect of the drifts of this flow.

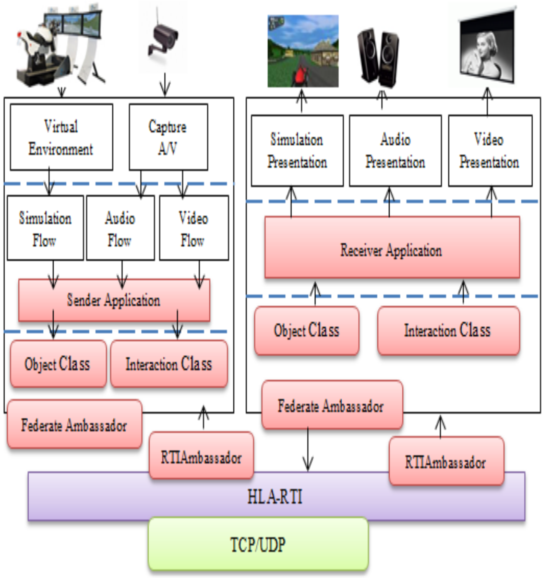


Figure 5. Platform used for the formal specification

The allowable drift of the video is 10ms: this drift is achieved by the treatment of 2 units of synchronization, which is the treatment of 4 units of sync audio and 5 units of flow synchronization Interactive. The formal modeling of this approach is given by three hierarchical levels and 5 HTSPN nets. Figure 6.a shows the highest level. This highest level specifies the full constraints of the inter-stream synchronization between the audio stream, modeled by Aud, the flow of the interactive simulation, modeled by the Sim, and the video stream, modeled by VID.

The atomic or composite components and materials are managed at the HLA-RTI level. The link layer is independent of the middleware, and it represents the application level. Figure 5 describes the synchronization architecture of the distributed interactive simulation governing the HLA-RTI middleware. Within the composite layer, Places, from Sim1 to Sim5, represent the objects of the distributed simulation, AUD1 to AUD4 represent the audio objects and places VID1 and VID2 describe the video objects. Each circle represents a data packet: to ensure the synchronization between the streams, at the model defines a synchronization for each set of 5 packets of distributed simulation (i.e. 544 bytes per packet), of 4 audio packets (of size 881 byte packet) and of 2 video packets. The composite layer fulfills this inter-stream synchronization and prepares the link synchronization Layer.

This net specifies in particular the control that must be implemented to ensure the adequate synchronization between these three flows, e.g. to ensure that the video stream is no later than 30ms compared to the other flows. This control must be applied with a maximum granularity of 20 ms, corresponding to two units of video synchronization, 4 units of sync audio and 5 units of sync interactive flow. The purpose of this architecture is to express all the specified timing requirements. During the simulation, HLA-RTI supports the transmission of audio, video and interactive streaming from the sender application to the remote hosts. It allows both the transport layer and control layer. HLA defines two types of information exchange: the objects and the interactions. Objects are inherently persistent during the simulation, represented by atomic component; they implement the flow control. The intra-stream synchronization is managed by the objects that control the constraints of quality of service required for the flows. The interactions are persistent and will be able to natively transport the flows between the Federates. Finally, the places Sim1 to Sim5 represent the objects of the distributed simulation.

As described in Figure 6, the first point of inter-stream synchronization is of type ”MASTER”, with the audio stream as ”MASTER” is placed at the point go after a nominal duration equal to 80ms (100ms maximum). This synchronization is likely to induce the acceleration (respectively deceleration) of the video and audio flows after 5 units of synchronization and can also cause a delay or the loose of the video stream. The abstract place *Sim* is specified by the subnet shown at the top of Figure 5. This HTSPN model controls explicitly the advance of the interactive simulation flow with respect to the other flows. Given the jitter units of 10 ms for the video and of 5 ms the audio stream, then after 5 intra-synchronized objects of the interactive flow, this stream can be up to 20ms ahead of the other flows. The control of the jitter of this stream should be done by the HLA-RTI middleware to ensure that all constraints of synchronization with the other streams are enforced. The HLA Objects should control independently each stream using native HLA APIs *UpdateAttributeValues()* and *reflectAttributeValue()*. These functions are able not only to control the advance of a flow compared to the others, but also to ensure the intra-stream synchronization.

The HLA-RTI APIs *sendInteraction()* and *receiveInteraction()* could be used to send data.
Because audio and video objects do not need in many cases to be exchanged between federates, their data packet should be send using the HLA interactions. HLA provides many other APIs that can use in the implementation. As the synchronization is implemented at the receiving side, to schedule the data reception, the API tick(T1,T2) should be used with two arguments that are the minimum and the maximum values used in the temporal validity interval; for example tick(12,20) has to be used.

VI. MULTIMEDIA TRANSPORT ARCHITECTURE

Sender and receiver are involved in the stream transmission. As a requirement of the HLA-RTI middleware, both participants are federates and should follow the HLA rules in order to be compliant with the specification. Hence, RTI supports both ”Reliable” and ”Best Effort” communication mode. Since Multimedia stream need to be send continuously, it is necessary to optimize the throughput and the reduce the end-to-end latency. This solution need UDP-based ”Best Effort” transport protocol.

As we outlined in Section IV, multimedia packets need low latency to meet the QoS requirements, therefore it is mandatory to schedule a stream transmission task in order to share the system resources with other tasks. The synchronization interval validity are used to meet the requirements of the schedule system interval timer provided by the underling operating system. The interval timer allows the application to schedule periodic timer events. Thus, the application receives and requests timer messages at the Temporal Validity Interval (TVI) given in each arc of the HTSPN model- that is, the TVI allows the application to schedule the timer events within the TVI resolution, that is the timer interval of the *upadeintercation()* and *sendInteraction()* function is caller in this regular time resolution. In fact, real-time stream transmission over large scale networks adds latency and jitter due to the router scheduling and admission control within the router queues. Using the the value admissible in the TVI is twofold:(1) the re-synchronization of the media frames in the presentation layer at the receiver side without using reliable stream control (TCP protocol), the end-to-end latency can be carefully controller before the stream being displayed, and (2) allows the receiver buffer handling the received stream with minimum frame lost and eliminates jitter issues. Likewise, The longer the reconstruction buffer is, the larger the jitter can be reduced.

VII. Conclusion

We have presented a formal model based on Hierarchical Temporal Stream Petri Nets for the synchronization of distributed interactive multimedia systems. This model is able to describe applications implemented using an HLA distributed simulation. It offers a good modeling power for at the same time the expression and the analysis of temporal constraints in such systems. It also allowed us to specify precisely, completely and in a unified way the multi-level logical, temporal and semantics timing constraints that are fundamental for synchronized distributed applications.

Taking into account all these constraints early in the design process leads to a rather efficient development of distributed applications and reduces the cost of this development. Our future work is to design and implement by this model a full distributed application that has been developed to remotely teach car drivers.

References

[1] SISO-STD-004.1-2004 - Dynamic Link Compatible HLA API Standard for the HLA Interface Specification (IEEE 1516.1)

[2] P. Senac et al., *Modeling logical and temporal synchronization in hypermedia systems*, IEEE Journal on Selected Areas in Communications, Vol.14, N1, pp. 84-103, January 1996

[3] A. Hakiri and al., *Multi-level Model for Synchronizing Temporal Streams on HLA-Based Distributed Multimedia Applications Using HTSPN*, Second International Conferences on Advances in Multimedia (MMEDIA), 2010, pp. 140-147

[4] M. Diaz and P. Senac, *Time Stream Petri Nets a model for timed multimedia information*, in *Petri Nets: Fundamental models, Verification amd Applications*, ISBN 978-1-84821-079-0, 2009

[5] R. Willrich et al., *Multimedia Authoring with Hierarchical Timed Stream Petri Nets and Java, Multimedia Tools and Applications*, Journal of Multimedia Tools and Applications, Volume 16 Issue 1-2, January-February 2002

[6] G. Blakowski and R. Steinmetz, *A media synchronization survey: reference model, specification, and case studies*, IEEE Journal on Selected Areas in Communications.IBM Eur. Networking Center, Heidelberg;

[7] L. Xie et al., *Complex System Simulation Based on Petri Net Combined with HLA*, 1st International Workshop on Education Technology and Computer Science, 2009. vol. 3, pp .205-208.

[8] M. B. Kpatcha et al., *Exploring impact of time management services on HLA-based Petri Nets Simulation Engine. Simulation Methods and Applications: Simulation Practice and Theory*. Volume 9, Issues 3-5, 15 April 2002, pp. 143-166

[9] S. Combettes and A. Nketsa, *Interoperability Compliance Validation Of HLA Federations Based On Colored Petri Nets*, 2003 EURO SIW 03E-SIW-084.

[10] R. Guha, M. Bassiouni, and G. Schow, *A Framework For Modeling High Level Architecture (Hla) Using Petri Nets*, University of Central Florida. Department of Computer Science. Orlando, FL 32816

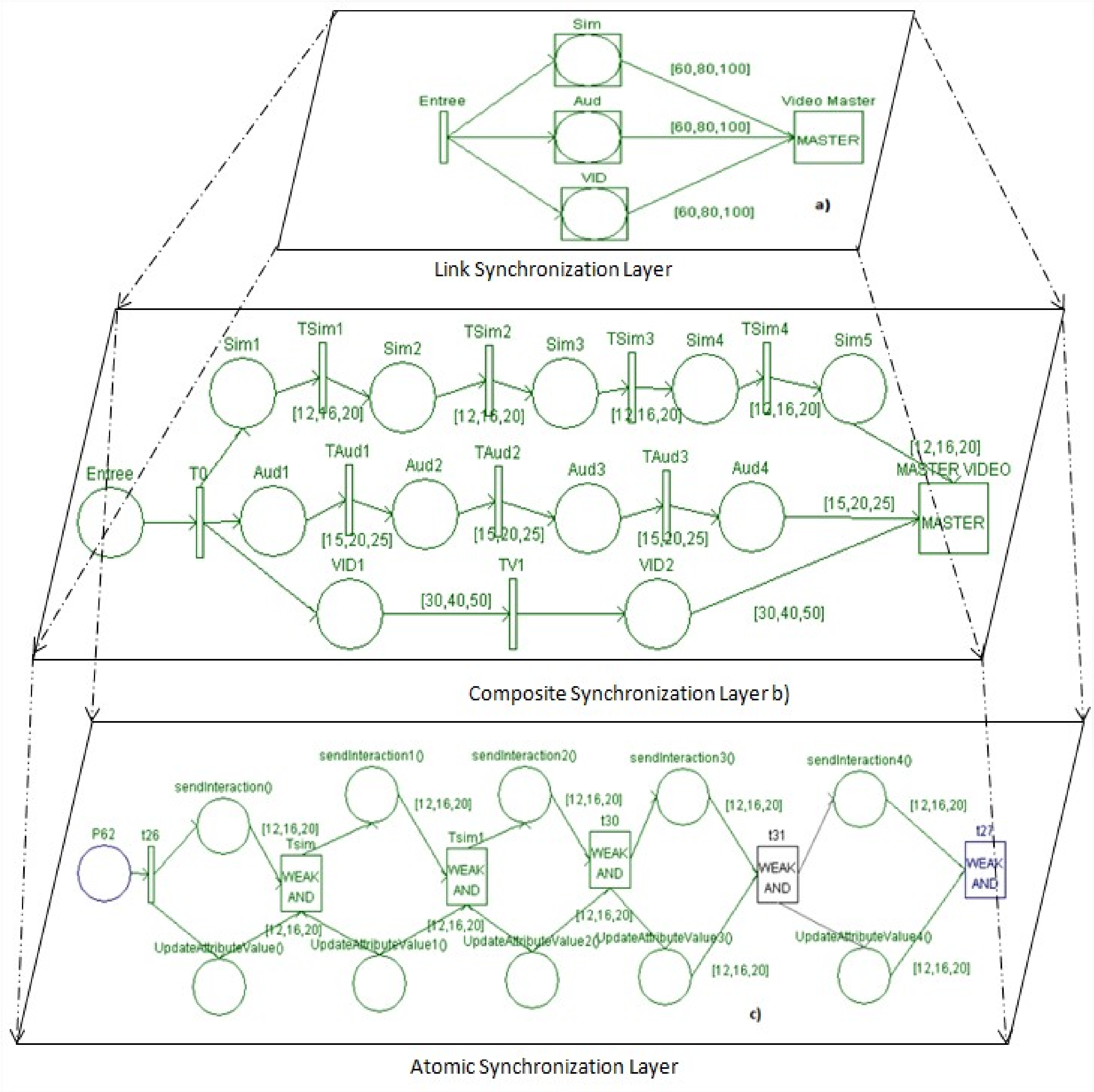


Figure 6. Distributed Multimedia Synchronization scheme over HLA-RTI. a) The link synchronization Layer, directly connected to the application, b) the composite synchronization layer for the inter-stream synchronization Layer and c) the atomic synchronization layer which of the intra-flow synchronization

The Anatomy of an Adaptive Multimedia Presentation System (AMPS)

Nick Rowe
Faculty of Technology
Bournemouth and Poole College
Bournemouth, UK
nrowe@bpc.ac.uk

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

***Abstract*—The use of multimedia presentations within learning environments is described and guidelines for the design of good E-Learning systems are identified. It is argued that a linear sequential presentation of knowledge segments is effective, but that the user is provided with optional links to relevant segments during the presentation. The synchronisation of multiple media is considered and the design of a prototype E-Learning system is discussed. The segmentation of material is then discussed and how the information can be stored in a data repository consider with respect to the requirement of accessing linked segments. Finally, the nature of adaptivity is discussed leading to a discussion of the salient parts of an adaptive multimedia presentation system.**

Keywords – multimedia, hypermedia, E-Learning, learning objects, adaptive, education.

I. MULTIMEDIA FOR LEARNING

Over the last fifteen years or so, there have many studies on using multimedia presentations to assist the learning process. Many applications have been designed to utilize the potential afforded by the use of computer-based learning systems. However, the early promise of these systems has not resulted in the widespread use of strong computer-based multimedia mechanisms within the learning environment. Instead, weak forms of multimedia have been favoured elevating form over content. It is perhaps hardly surprising that Craig, [6], shows that its use is not associated with a significant improvement in student grades.

This flexible 'one size fits all' approach to multimedia presentation makes it popular, but, Burke and James, [4], show within a business education environment, teaching abstract, conceptual and theoretical content with multimedia are more likely to be effective. However, for quantitative material requiring problem solving it may not be so effective. In these situations, they go on to say, the use of step-by-step instruction that allows students to see problems worked out in real time were more effective. This does not mean that multimedia applications cannot perform the latter tasks, it simply means that applications popularly used by teachers and lecturers generally do not do it.

So the dilemma here may be that in order to produce rich multimedia presentations which are inherently more complex, the authoring process will also need to be complex and therefore time consuming. But where is the starting point for the design of such systems? Gagne et al, [11], offers clear, if obvious, guidelines for the design of good E-Learning environments:

1. Gain the learner's attention (reception).
2. Inform the learner of the objectives (expectancy).
3. Stimulate recall of prior learning (retrieval).
4. Present the learning stimulus (selective perception).
5. Provide learning guidance (semantic encoding).
6. Elicit appropriate performance (responding).
7. Provide feedback (reinforcement).
8. Assess the learner's performance (retrieval).
9. Enhance retention and transfer (generalisation).

Also, if time and money is to be invested in the production of such materials the effect on learning outcomes needs to be clear. Krippel et al, [13], recently argued that this information is not readily available and that the true effect of multimedia technologies on learning outcomes remains unclear. More research is needed to examine educational environments where these new technologies are used to indentify improvements or underperformance over conventional pedagogies. It also needs to identify successful characteristics within certain contexts. Krippel argues that only with this evidence will educators be able to use multimedia technologies efficiently and effectively.

II. LESSON LAYOUT

If a multimedia presentation is to be designed to emulate a lesson or lecture, a good starting place would be to analyze the structure of a typical lesson and identify elements that will transfer well to these presentations. The difficulty here is that there no such thing as a typical lesson and very often delivery is adapted based on the content, teaching style and many other parameters.

One element that can be considered is the layout of a lesson and that it is usually planned. In other words, the content of the lesson has been identified by the teacher. This means that at its inception the lesson is rigid and linear. This is not to say the lesson itself is rigid, it will be adapted by the teacher based on an interaction with the learners. Deviation from the plan is acceptable; however, usually the main objectives learning outcomes will remain intact. Beasley and Smyth, [1], noted that despite multimedia learning environment giving an opportunity to explore their material in a more active, non-linear fashion, students exclusively studied the material linearly. They go on to say that this was possibly due to not being given any specific information on how to study in this way. Extending this slightly further it could be said that we are not taught to learn in this way.

ISBN: 978-1-61208-129-8

Interestingly in this study, two features used in a non-linear manner were the hyperlinked glossary and the search facility.

In essence, a learning environment needs to bring together learning units and construct them into a linear form based on the learning objectives. Then at the delivery stage it needs to provide the learner with optional mechanisms to deviate from the planned path. These mechanisms can be extended to include elements seen in the classroom such as asking questions and requesting topics to be explained in more detail and providing optional links to allow the user to view related topics.

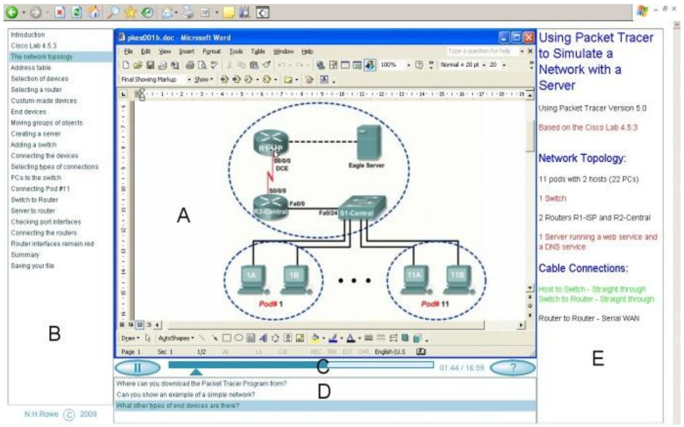
III. Multiple media

Using multimedia for learning is not new and does not need to be computer-based. Teachers have used it for hundreds of years. Using more than one medium to relay information improves the efficiency of the communication. Ellis, [10], notes that the importance of multiple channels for the delivery of educational content can be found in the theory of multi-channel communication. This confirms that when information is presented by more than one channel, there will be additional reinforcement, resulting in greater retention and improved learning.

With computer-based systems the problem is not now having the computing power to present rich multimedia content as it was in the past. There may still be issues with network bandwidth and heavily hit servers, but the problems are now usually centered on the synchronization of the different media. Languages like SMIL, [5], 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.

IV. The Development of a Prototype e-Learning System

Using the principles of lesson delivery and synchronised multi-focus multimedia elements, a prototype was developed using Adobe Flash, [8].



Using Packet Tracer to Simulate a Network with a Server

Using Packet Tracer Version 5.0

Based on the Cisco Lab 4.5.3

Network Topology:

11 pods with 2 hosts (22 PCs)

1 Switch

2 Routers R1-ISP and R2-Central

1 Server running a web service and a DNS service

Cable Connections:

Host to Switch - Straight through
Switch to Router - Straight through

Router to Router - Serial WAN

Where can you download the Packet Tracer Program from?
Can you show an example of a simple network?
What other types of end devices are there?

Figure 1: Screen Layout of a Multi-focus E-Learning System.

Figure 1 shows the screen layout of such a system. Here, five elements are synchronized to act from the same timeline. Element A is a traditional audio-visual presentation, B provides a table of contents that can be clicked to move within the presentation. C is a normal temporal control, D is a frequently asked questions section and E is incrementally loading HTML, (iHTML). Here the content, text and images, is displayed in real-time. Each segment of the HTML code is given a time-stamp and is not displayed until that time is reached in the presentation.

Authoring the table of contents and iHTML code is relatively easy and is carried out as a post-processing activity. The author watches the audio visual presentation through the system in the role of lecturer and is given access to additional functions that allow table of content titles and the segments of iHTML to be added to the system. These are then automatically entered into an XML configuration file and displayed during playback by users of the system accessing the system in the student role.

V. Asking Questions within the Presentation

Panel C, in Figure 1, as well as allowing temporal control, contains a button that allows the user to ask the system a question. When the button is pressed it activates a question dialogue that allows the student user to enter a text-based question to be read by the author of the content: the lecturer. This question is marked with the time it was asked in the presentation. This question and time stamp are appended to a file on the server running the E-Learning system.

These questions that have been asked by any student user of the system are available to users of the system entering in a lecturer role. In this role, all questions that have been asked can be viewed and when selected the lecturer is taken to the part in the presentation where the question was asked. The opportunity is then given to the lecturer to answer the question with a short additional video. Once published, this video is available to all student users of the system and the question is displayed in the same manner as the table of contents being highlighted as it is relevant in questions panel, (D). However, the answer presentation is only played if the student selects the question. This allows the presentation to continue uninterrupted unless the student specifically wants to see the answer to that particular question. If a question is played the main presentation is paused while the answer video is played and resumed from the paused position when the answer video has ended. Thus, the student is given the option to view previously asked questions.

With the publishing of answers to asked questions, during the life of the presentation more questions are likely to be asked and therefore the presentation matures over time and provides more supplementary information useful to a learner viewing the presentation for the first time.

VI. Adaptive E-Learning

Allowing learners to ask questions within the presentation and optionally view answers to previously asked questions is, in some measure, adapting the presentation to learner requirements. Generally, acknowledging the important relation between individual learners and education has along history. Shute and Towle, [15], 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 itemize four components of 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. 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 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, [3], defines such systems as giving a presentation that is adapted specifically to the user’s knowledge of the subject and suggest 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 some adaptive engine. The second implication leads on to another area. This is that the presentation needs to be reduced to fragments to allow these elements to be manipulated. These fragments are generally termed *learning objects* and much research has been done around their use.

A good example of adaptive navigational support offered by an application is AHA! an open source adaptive hypermedia platform, [9]. 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.

VII. 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*, [12]. 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, [2]. However, this interoperability and reusability may have been overstated in the past. McGreal, [14], 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.

Boyle, [2], 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 our 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, [2], 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.

VIII. 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., [9], 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. 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 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.

To assist this process each segment is associated with a set of data relating to it. This data can contain simple information like name and description and also link to data used during its presentation like the iHTML text. Since this text is tightly bound with the original presentation it provides useful information to base decisions on linking one segment with another.

IX. THE STORAGE OF INFORMATION

The segmentation of individual learning objects has ultimately to be reference to the ontology of that subject area. The storage of information needs to be indexed in order for it to be retrievable. Each node is provided with a unique address which defines its location on the ordered tree. The addressing system is chosen in such a way that it corresponds a knowledge hierarchy that is specified by sections, sub-sections, sub-sub-sections etc. see Figure 2. The ordered tree also provides the ability to define segmentation. Consider a video clip divided into 8 segments A to H. Each segment corresponds to a knowledge division or a set of knowledge divisions in the subject ontology. One typical association is seen in Figure 2.

1	1.1	1.1.1		A	0
		1.1.2			
	1.2	1.2.1	1.2.1.1	B	20
			1.2.1.2		
		1.2.2	1.2.2.1		
			1.2.2.2		
	1.3	1.3.1		C	60
		1.3.2			
	1.4	1.4.1		D	80
	1.5	1.5.1		E	90
		1.5.2			
		1.5.3			
	1.6	1.6.1		F	110
	1.7	1.7.1		G	120
		1.7.2			
		1.7.3			
		1.7.4			
		1.7.5			
		1.4.1	1.4.1.1	H	200
			1.4.1.2		
			1.4.1.3		
			1.4.1.4		
					250

Figure 2: Association of ontology divisions with video segments

X. ONTOLOGIES

According to Gruber, in a computing context, an ontology is "*an explicit specification of a conceptualisation"* [17]. This has been refined by Struder as "*a formal, explicit specification of a shared conceptualization"* where 'formal' indicates that the language of ontologies should be readable by machines as well as humans and where 'a shared conceptualization' indicates that this specification constitutes a community reference which allows the sharing of a consistent understanding of what information means and further makes possible interoperability between systems.

Usually ontologies are represented as knowledge hierarchies with the most general concepts at the top and more detailed and specific concepts at lower levels [16]. The structure of these knowledge hierarchies is naturally representable as networks, where each node on the network represents a unit of knowledge. Although many different network topologies are possible in theory such a linear, circular, hub/spoke, tree etc., the ontological model that we will be using here will be a simple ordered tree.

The ordered tree network is distinguished by 1. there is only one route from any node to any other node and 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.

Node addressing

The first step in building an operational structure is to reference the components of the ontology which we do by providing each node with a unique address. We adopt a positional system to delineate each sub-section within a knowledge hierarchy where each section, sub-section, sub-sub-section etc. is represented by series of numbers separated by points. This has the advantage of being scalable and universal in application. See Figure 3

Each node is represented by a unique vector. Thus

|X> = |1,2,1,1>
|Y> = |1,4,1,3>
|Z> = |1,3,2,0>

The knowledge tree network can alternatively be fully represented by the adjacency matrix A_{ij} where

$$|X_i\rangle \;= \sum_{j=1}^{n} A_{ij}$$

1 1.1 1.1.1
1.1.2
1.2
1.2.1
1.2.1.1 |X>
1.2.1.2
1.2.2
1.2.2.1
1.2.2.2
1.3
1.3.1
1.3.2 |Z>
1.4
1.4.1
1.4.1.1
1.4.1.2
1.4.1.3 |Y>
1.4.1.4
1.5
1.5.1
1.5.2
1.5.3
1.6
1.6.1
1.7
1.7.1
1.7.2
1.7.3
1.7.4
1.7.5
1.4.1

Figure 3: Example of unique address system for knowledge hierarchy

In the case of our example presented in Figure 3 it can be expressed in the adjacency matrix in Figure 4. This matrix is symmetric.

	1	1	1.1.1	1.1.2	1	1.2.1	1.2.1.1	1.2.1.2	1.2.2	1.2.2.1	1.2.2.2	1	1.3.1	1.3.2	1	1.4.1	1.4.1.1	1.4.1.2	1.4.1.3	1.4.1.4	2	1.5.1	1.5.2	1.5.3	2	1.6.1	2	1.7.1	1.7.2	1.7.3	1.7.4	1.7.5
1	0	1			1							1			1						1				1		1					
1.1	1	0	1	1																												
1.1.1		1	0																													
1.1.2		1		0																												
1.2	1				0	1			1																							
1.2.1					1	0																										
1.2.1.1							0																									
1.2.1.2								0																								
1.2.2					1				0	1	1																					
1.2.2.1									1	0																						
1.2.2.2									1		0																					
1.3	1											0	1	1																		
1.3.1												1	0																			
1.3.2												1		0																		
1.4	1														0	1																
1.4.1															1	0																
1.4.1.1																	0															
1.4.1.2																		0														
1.4.1.3																			0													
1.4.1.4																				0												
1.5	1																				0	1	1	1								
1.5.1																					1	0										
1.5.2																					1		0									
1.5.3																					1			0								
1.6	1																								0	1						
1.6.1																									1	0						
1.7	1																										0	1	1	1	1	1
1.7.1																											1	0				
1.7.2																											1		0			
1.7.3																											1			0		
1.7.4																											1				0	
1.7.5																											1					0

Figure 4: Adjacency matrix

Once a nodel address system is specified it then becomes possible to give quantitative values to terms such as ‘level of detail’, ‘difficulty’, ‘proximity’, ‘strength of links’ etc.

We define the following terms based on this nodel address system.

Difficulty: we define the difficulty of a knowledge node to be equal to the degree of centrality of the node – 1. In other words it is equal to the number of sub-nodes that are connected to a given node. Although it might be argued that this is a crude measure of ‘difficulty’ it has the advantage of being directly related to the complexity of the knowledge node and by association can be used as a measure of the difficulty.

Level: the level of a knowledge node as the same as the tree level of the node which is equal to the dimension of the representative vector of the node. Thus the level of node |X> = |1, 2, 1, 1> is 4 while the level of node |Z> = |1, 3, 2> is 2. We say that the **level** of a knowledge node is equal to its **importance** and represents the level of detail that a knowledge node contains.

Distance: this is a measure of how close two nodes are on the ontology. The degree of separation of knowledge segments is dependent upon the level of the nodes. Nodes at level 3 are an order of magnitude closer than nodes at level 2 and those at level 2 an order of magnitude closer than at level 1. We therefore define distance between nodes as the number of nodes traversed divided by the order of magnitude of their level. Thus two neighbouring nodes at level 1 will have a separation of 1, while two nodes at level 2 will have a separation of 0.1 and those at level 3 a separation of 0.01 Distance is therefore a measure of how

ISBN: 978-1-61208-129-8

close two knowledge segments are related to the subject ontology. For a tree network this is a unique value that indicates the **strength of connection** between two knowledge segments.

XI. Ontological Calculus

In order to determine the quantitative value of each of these terms it is required to define the algorithms or operations on the node addresses that will provide the appropriate values determined by the definitions. This set of operations will form a calculus enabling the manipulation of ontology.

A high level segment such as |1.1> contains less detail than a lower level segment such as |1.2.1.1> The level of a knowledge vector is given by multiplying the normalized vector by the unit covector. We define the unit covector of n dimensions $<U_n| = <1\ldots\ 1,1,1|$ where there are n elements.

The normalization of a knowledge vector |X> we represent as N|X> where N is the normalization operator. Hence the level of the knowledge vector |X> is given by:

Level = $<U_n|N|X>$

Thus for the case of |X> = |1,2,1,1> we have

Level |X> = <U|N|1,2,1,1>
= <1,1,1,1|1,1,1,1>
= 4

Similarly

Level |Z> = <U|N|1,3,2,0>
= <1,1,1,1|1,1,1,0>
= 3

Distance algorithm
We define the n-dimensional Level covector
$<L_n| = <n,\ldots\ 3,2,1|$

The distance of two nodes is given by the modulus of the difference of their node addresses multiplied by the Level Order of Magnitude covector <LOM| where
$<LOM_4| = <1, 0.1, 0.01, 0.001|$

Thus for the two vector addresses |X> and |Y> their proximity is given by:

Proximity|Y>|X> = $<LOM_4|(|Y> - |X>)$
= $<LOM_4|(|1,4,1,3> - |1,2,1,1>)$
= $<LOM_4|0,2,0,2>$
= <1, 0.1, 0.01, 0.001 |0,2,0,2>
= <1x0 + 0.1x2 + 0.01x0 + 0.001x2>
= 0.202

Similarly the proximity of |Z> to |Y> is

Proximity|Y>|Z> = $<L_4^2|(|Y> - |Z>)$
= <1, 0.1, 0.01, 0.001 |0,1,1,3>
= 0.113

And similarly
Proximity|Z>|X> = $<L_4^2|(|Y> - |X>)$
= <1, 0.1, 0.01, 0.001 |0,1,1,1>
= 0.111

It should be clear from these examples that proximity is not associative.

Proximity|Y>|X> $\neq$ Proximity|Z>|X> + Proximity|Y>|Z>

Difficulty
The difficulty of a segment is defined to be equal to the degree of centrality of the node minus one. The degree of centrality is determined by the Adjacency matrix of the ontology A_{ij}

The degree of a node is the number of connections to it. We will denote the degree of knowledge vector $|X_i>$ as

Difficulty = $<D|X_i> = \sum_{j=1}^{n} A_{ij}$

These sets of algorithms form a calculus which enable clear metrics to be determined that can be calculated and fed into the AMPS system to facilitate adaption.

XII. 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., [7], 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. Figure 4 shows both the assembly language and the CAF.

```
(a)     if GM.Concept.weight > 10
        then ( PM.GM.Concept.show = true )

(b)     <CAF>
        <domainmodel>
        <concept>
        <name>Adaptive</name>
<concept>
<name>Adaptive HyperMedia</name>
<attribute>
        <name>title</name>
        <contents>Adaptive HyperMedia</contents>
</attribute>
…
</concept>
…
</domainmodel>
</CAF>
```

Figure 5: (a) A typical fragment of assembly language, (b) A fragment of the CAF file in XML format

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. The declaration and use of this intermediate language has another advantage. Each system can be developed and refined independently: one system generates the CAF, the other uses it. CAFs, specifically designed for testing, can be used by the presentation system.

XIII. Adapting Materials in an E-Learning System

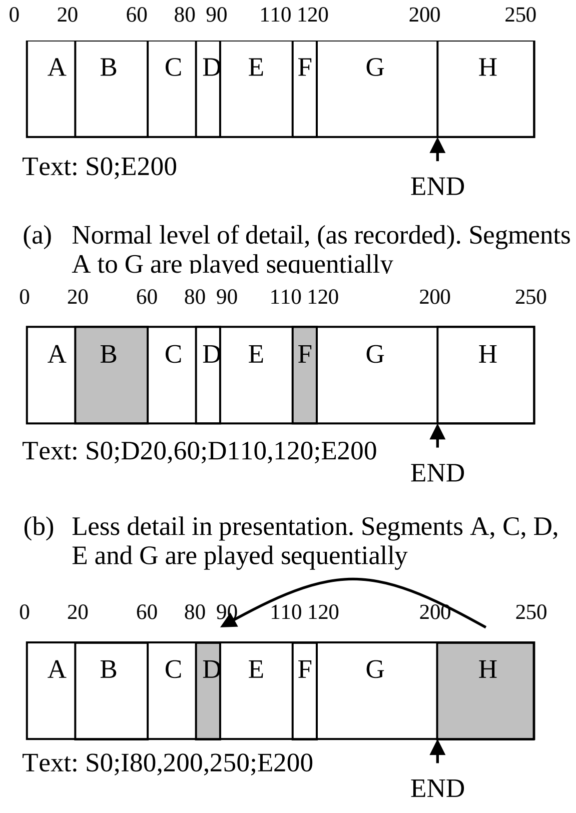
Once the decision to establishing the segment as the heart of an E-Learning system has been made, the rest of the system can be designed around it. Entities including the user and materials to test the user knowledge can be included in the E-Learning database.

In the development of the materials the educational concepts must be isolated from a unit of a course and developed into learning objects. The syllabus of a unit consists of an ordered set of concepts and a course is an ordered set of units. Each concept is formed into a segment. Initially a segment contains audio-visual resources required for its presentation. The authorship sequence continues by adding addition data to the segment including references to the AV file and the iHTML file used during presentation.

To make the 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.

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. 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 system.

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 C 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 H in the place of the smaller fragment D. Here, more detail can be added to specific parts of the file and therefore particular concepts are elaborated within the segment. These additional fragments are added to the end of the media file and are 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 at different levels.



(a) Normal level of detail, (as recorded). Segments A to G are played sequentially

(b) Less detail in presentation. Segments A, C, D, E and G are played sequentially

(c) More detail in presentation. Segments A, B, C, H, E, F and G are played sequentially

Figure 6: Three levels of detail from a single audio-visual fragment.

XIV. CONCLUSION

The E-Learning presentation system 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. It also allows fragments to be added or removed from the segment as required adapting to the user requirements. At any stage in the presentation the detail can be manually increased or decreased. Questions can be asked, the answers published onto the system as a linked segment. Other segments within the data repository are displayed that may be relevant to the current segment. The algorithm to do this is contained in a separate system that has access to the same E-Learning database and acts independently from the presentation system. As this system discovers links between the segments in the repository they are added to the database by adding links to each segment. When the segment is presented to the user as part of a lesson these link are displayed giving the user the optional ability to display these linked segments. A strength variable keeps track of the relevance of the links and this can be displayed to the user.

The presentation side of the system runs from meta-data provided from an XML configuration file created at the time the presentation is requested by the user. Information on the user's progress is obtained from the database to pick the level of detail required for each segment. This information is obtained from the results of previously attempted tests and from changes made by the user if the segment has been previously viewed by the user.

The XML configuration file will consist of a number of essential elements for the presentation of the lesson:

- An ordered list of the segments contained in the lesson
- For each segment a list of allowed detail levels along with an adaptive descriptor for each detailing the way the content will be manipulated for that particular level and the synchronization information to present additional material, (for example iHTML blocks)
- For each segment, a list of other linked segments that are considered relevant, together with a metric indicating the strength of that relevance. The answers to previous questions asked by viewers of that segment can also form linked segments with a high value of relevance.

REFERENCES

[1] Beasley, N., Smyth, K., 2004. Expected and Actual Student Use of an Online Learning Environment: A Critical Analysis. *Electronic Journal on e-Learning.* 2(1). 43-50.

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

[3] Brusilovsky, P., 2001, Adaptive Hypermedia. *User Modeling and User-Adapted Interaction* 11. 87-110.

[4] Burke L.A., James, K.E., 2008. PowerPoint-Based lectures in business education of student-perceived novelty and effectiveness. *Business Communication Quarterly*, 71(3), 277-296.

[5] 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.

[6] Craig, R. J., Amernic, J.H., 2006. PowerPoint presentation technology and the dynamics of teaching. *Innovation in Higher Education*. 31(3), 147 - 168

[7] 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). Available from: citeseerx.ist.psu.edu/ viewdoc/download. [Accessed 10 March 2010]

[8] 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.

[9] 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 10 March 2010].

[10] Ellis, T. 2004. Animating to build higher cognitive understanding: A model for studying multimedia effectiveness in education. *Journal of Engineering Education*.

[11] Gagne, R. M., Briggs, L.J., Wager, W.W. 1992. *Principles of Instructional Design*. Wadsworth Publishing Co.

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

[13] Krippel, G., KcKee,A.J., Moody, J., 2010. Multimedia use in higher education: promises and pitfalls**.** *Journal of Instructional Pedagogies*, Vol 3. Available from: http://www.aabri.com/jip.html [Accessed 10 March 2010]

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

[15] Shute, V., Towle, B., 2003. Adaptive E-Learning. *Educational Psychologist*. 38(2), 105–114

[16] Novak, J.D., and Cañas, A.J. 2006. *The theory underlying concept maps and how to construct them*. Technical Report IHMC CmapTools 2006-01, Institute for Human and Machine Cognition.

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

A User Interface for Spatio-Temporal 'Eventually' Queries using Gamepad

Vineetha Bettaiah
University of Alabama in Huntsville
Department of Computer Science
Huntsville, USA
vb0003@cs.uah.edu

Ramazan S. Aygün
University of Alabama in Huntsville
Department of Compute Science
Huntsville, USA
raygun@cs.uah.edu

***Abstract*— Video databases have both spatial and temporal components. Querying and retrieval of spatio-temporal content is a challenging task due to lack of simple user interfaces. In this paper, we propose a system to allow the user to interactively build "eventually" queries in video databases. In eventually queries, the user just needs to provide the starting state (or information) and the ending state without providing the intermediate states. This helps the user setup queries without knowing all details. Our system uses a methodology similar to the one in gaming. Queries are built by displaying natural videos based on gamepad commands rather than on a graphical interface using a mouse or a keyboard. The system uses a semantic sequence state graph (S^{3}G) to search the database. The system is applied on a tennis video database. This paper, proposes a novel spatio-temporal query and retrieval system with user friendly interface for developing "eventually" type spatio-temporal queries using gamepad.**

Keywords-Video querying and retrieval; interactive query interface; eventually queries.

I. Introduction

The video databases have both spatial and temporal dimension. The process of retrieving spatio-temporal objects and events that span space and time domains is known as spatio-temporal query. The design of a good spatio-temporal query system should consider representation of the spatio-temporal information, query building, and simplicity. The *representation* to model the spatio-temporal system must be sophisticated enough to capture the semantic contents. Such a system should be able to represent objects, events, and changes in the data. It may be hard to build a spatio-temporal query instantaneously. Therefore, the system should allow the user to build the spatio-temporal queries incrementally to retrieve one or a sequence of many events which, cause the specified action. The spatio-temporal query system must also be simple enough to be used by a general-purpose user and should not require them to know the internal representation of the database.

Significant effort has been made on querying spatio-temporal databases and many of the approaches are based on developing new languages or extending the existing query languages such as SQL [11] or developing interfaces for the user to build a spatio-temporal query. STQL (Spatio-Temporal Query Language) [7] demonstrates how SQL can be extended to query spatial objects that change over time. It extends SQL by adding features like a set of spatio-temporal predicates such as *disjoint, meet, overlap, coveredBy, covers, inside, contains*, and *equal*. 2198 predicates are identified between two evolving regions. Such a large number of predicates make it practically impossible to name the predicates as well as their utilization by the user. Jain *et al.* [1] uses pattern matching properties of SQL to express spatio-temporal queries. Since the data is represented as strings based on a grammar, it is possible to apply pattern matching techniques. In [3], "conceptual-neighborhood-graph" (or "closest-topological-relationship-graph") is developed based on spatio-temporal relationships like overlap, meet. This graph is used to retrieve spatial objects that changes over time [4].

Besides query languages based on SQL, visual query languages have also been proposed to query spatio-temporal data since the data has at least spatial component. Icons are usually used to represent objects. Lvis supports querying moving objects [5][6]. Query-By-Trace [8], Visual Interactive Query Interface [10] and Visual Query system S-TVQL [9] are other examples of visual querying interfaces for spatio-temporal content. All the above approaches present difficulty in analyzing the query for the novice user. Naik [2] provides a user interface for querying tennis video databases. The user chooses (or click) the locations of players and the ball on the available interface for each instance in the query. However, point-and-click approach using a graphical court view is tedious and does not provide an intuitive method of building queries.

In this paper, we focus on the *eventually* operator in temporal logic. If the user is interested in the next available state from a current state, we basically call it as a *'next'* query. If the user is interested in whether a state is reachable from a current state, we call it as an *'eventually'* query. Eventually type query result allows the user to visualize all intermediate steps to reach the given state. These types of query allow the user to specify two states and view all intermediate events and states between them and also relieve the user from trying to recollect every possible next event to query in case of "next" query. In other words, the user does not need to specify all intermediate steps. It is possible that the user may not know or not interested in intermediate steps.

Since video databases may require spatio-temporal queries that include three dimensions, it is hard to build such queries without a proper user interface. Especially, incorporating temporal dimension is difficult. We observe that one of the common environments in which users provide spatio-temporal inputs to the system is the environment of video games. In these eniveronments, a player (or a user) provides spatio-temporal inputs of objects using a gamepad. In our system, the queries are built by displaying natural videos based on gamepad commands rather than on a graphical interface. There are three components in the system: building the query, searching and retrieval of clips, and displaying query result. Semantic sequence state graph (S^3G) is used to search the database. A query is built incrementally as a sequence of queries. Though this paper describes the query building process using "Eventually" type queries, the process is applicable to build other types of spatio-temporal queries. We illustrate the system on tennis videos.

Our paper is organized as follows. The following section provides background about the database and indexing. Section III describes how a gamepad is used for "eventually" queries. Our examples and illustrations are provided in Section IV. The last section concludes our paper.

II. Background

In this section, we provide information about our semantic modeling and retrieval system (SMART) and our semantic sequence state graph (S^3G) for indexing and retrieval of videos from a tennis video database.

A. SMART

The semantic content of a video corresponds to high-level information in the video. SMART [1] models objects, events, sequence of events and the resulting spatio-temporal interactions among objects in the video. A sample application on tennis videos that utilizes SMART is developed for modeling and retrieval of semantic contents in a tennis video. The semantic contents of a tennis video are modeled using a set of objects, a set of events, a set of locations on the court besides a set of camera views and a set of production rules (grammar) which, are given in [1].

Objects: The set of objects Σ_O contains three objects: the ball b, the first player U and the second player V:

$$\Sigma_O = \{U, V, b\}.$$

Events: The set of events Σ_E contains two distinct events: the forehand shot F, and the backhand shot B:

$$\Sigma_E = \{F, B\}.$$

Locations: The tennis court is divided into 13 non overlapping regions including the net N as shown in Figure 1. The set of locations Σ_L includes all these 13 regions:

$$\Sigma_L = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, N\}$$

The production rules are used to encode the semantic contents of the videos as a set of strings. Each video clip is represented with one string.

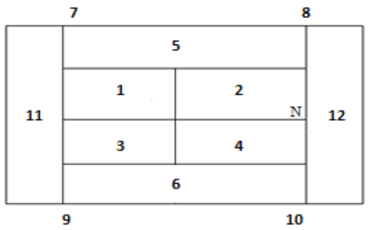


Figure 1. Regions of the tennis court.

B. S^3G - Semantic sequence state graph.

While SMART [1] represents the semantic information in a video as a set of strings, the semantic sequence state graph (S^3G) [2] represents the same information in the form of a graph. In tennis video, each object (ball, $player_1$, $player_2$) can be in any of the 13 possible locations. Therefore, theoretically, there are a maximum of 13^3 patterns of assigning 3 objects to 13 locations. Each assignment pattern defines a unique state in S^3G, and the maximum number of states in S^3G is less than 13^3 due to game constraints. S^3G also reduces the number of states by maintaining only states that are present in the video database. An event from the set of all possible events Σ = {F_1 ($player_1$ hits forehand), F_2 ($player_2$ hits forehand), B_1 ($player_1$ hits backhand), B_2 ($player_2$ hits backhand)} makes the objects move causing state-to-state transitions. Note that S^3G may have cycles as a state may be visited many times during the game. Thus, in S^3G, the semantic information of a clip is represented by a sequence of states and transitions, starting from one of the 8 possible states (four serve locations and two players). The semantic information of all clips, together, represents the semantic information of the video.

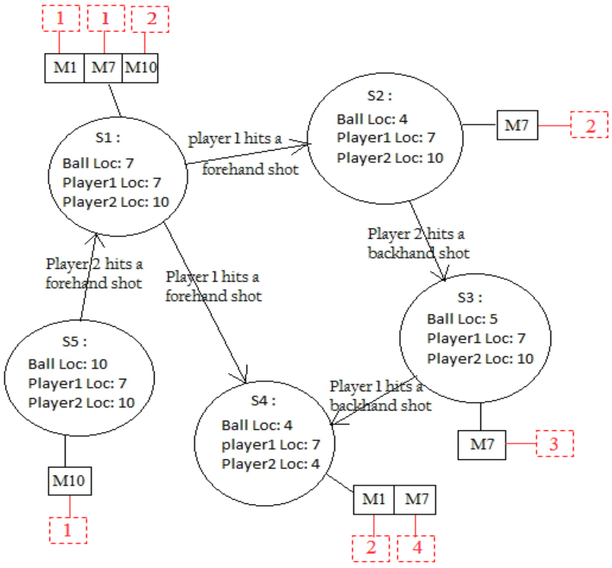


Figure 2. Construction of S^3G from SMART string data.

Example: S^3G in Figure 2 is built for three clips of a video: M1 = A[U] C[U7b7V10 b4 V4], M7 = A[U] C[U7b7V10 b4 BV10 b5 BU7 b4 V4], and M10 = A[V] C[U7b10V10 b7

]. The letter *A* represents a close-view of a player, while letter *C* represents court-view. The sub-string for court-view is indexed by S^3G. In M1, $Player_1$ (U) serves an ace from location 7 as Player2 moves to location 4 to receive, but $Player_2$ fails. In M7, $Player_1$ serves from position 7 again; Player2 responds from location 5 with a backhand shot and the ball hits location 5; and $Player_1$ hits with a backhand shot at location 7 and the ball goes to location 4. In M10, $Player_2$ serves the ball from location 10 and the ball hits location 7. For example, nodes S_1, S_2, S_3, S_4 and the state transition from $S_1 \rightarrow S_2$, $S_2 \rightarrow S_3$, $S_3 \rightarrow S_4$, represent M7.

The semantic sequence state graph, as described above, has a limitation. From the string representation of M1, it is clear that the temporal order of states in M1 is (S_1, S_4) indicating that S_1 is the first state and S_4 is the second state. Similarly, temporal orders of states in M7 and in M10 are (S_1, S_2, S_3, S_4) and (S_5, S_1), respectively. The initial S^3G did not contain temporal orders of states in various clips. The lack of temporal order information could lead to the retrieval of clips that do not satisfy the criteria specified by the query. For example, if a query specifies a direct transition from S_1 to S_4 through a forehand shot from $player_1$, the system will retrieve two clips M1 and M7 because both clips are attached to S_1 as well as S_4. However, note that in M7 there is no direct transition from S_1 to S_4. Retrieval of incorrect clips also occurs when a state is visited multiple times during graph traversal. This is possible due to the occurrence of several instances of a same state in a single clip. This limitation can easily be resolved by attaching to each clip a list of temporal orders (ranks) of the state as shown in Figure 2 by dotted red squares. With the enhanced S^3G, the retrieval of clips is a two step process. In step 1, clips common to all states involved in the query are selected. In step 2, clips in which states do not satisfy the temporal order constraints are deleted. In addition, timings of these ranks are stored in the database. For example, time of S_1 in clip M1 may be at 127th second, time of S_4 in clip M1 may be at 129th second. Hence the event, $player_1$ hits a forehand shot from S_1 to S_4 starts from 127th second and ends at 129th second. These are represented as *StartTime* and *EndTime* respectively for each event

III. INTERACTIVE RETRIEVAL OF VIDEO CLIPS USING GAMEPAD.

A user-friendly interface is needed to get spatio-temporal input from the user. There were some approaches in the past to handle spatio-temporal queries. For example, Jain *et al.* [1] developed a user interface with drop-down menus to get input from the user. These inputs are first used to develop a SQL string pattern that can be mapped to spatio-temporal expressions. Then a SQL query is built. On the other hand, Naik *et al.* [2] develop a graphical user interface with a mouse point-and-click approach using court-view for tennis. Basically, the user needs to choose an object and then click where the object should be on the court. After a state is built, the next or future states are built in a similar fashion. In this paper, we propose a better user interface than the previous approaches: a) we provide a court-view (from a tennis video) to the user for interactions and b) the inputs are obtained using a gamepad. The gamepad provides multiple buttons that enable switching objects and locating them on the court-view.

A. *Features of Gamepad*

The gamepad provides a set of interactions as shown in Figure 3 to build spatio-temporal queries: an 8-way switch (A) for directional controls, a set of 10 buttons, B={b_1, b_2, …, b_{10}}, a record (R) button to store the current information, and a record/search (R/S) button.

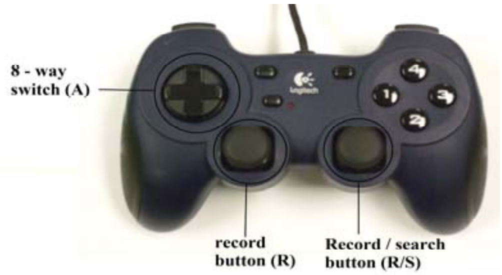


Figure 3. Gamepad features

B. *Mapping Gamepad Input to Semantic Information in S^3G*

The critical part of querying is to map the input from the gamepad to semantic information for tennis video to be used for retrieval. Each feature or button (e.g., button b_1) of gamepad produces a numerical input when pressed. These numerical inputs are mapped to semantic information. For example, the 8-way switch has eight active positions corresponding to eight directions. The switch generates unique code for each direction. Thus it is used to input the direction of movement of the object on the tennis court while building the query. Based on the input from switch A and other features of the gamepad, our system builds the query and provides as input to the retrieval system based. Every node in S^3G, represents particular semantic information about ball-player location. The search process matches the semantic information specified by the query with the semantic information of the video represented by S^3G. If a match is found, clips attached to the matching nodes are retrieved from the database and displayed to the user.

C. *Algorithm for Eventually Query*

The system retrieves the desired video clips from the database using an approach that has three main steps. First, spatio-temporal queries are built interactively using the gamepad. Secondly, a search process is initiated where clips that include the states of interest are identified by applying queries to our graph based indexing structure

(S^{3}G) and fetched from the database. Finally, the system provides a visual display of the query results by displaying all relevant events in real time using original video in the database. An implementation of “Eventually” query is given by following algorithm.

```
void Eventually_Query ( )
begin
 // $I_{Lb}$, $I_{L1}$, $I_{L2}$ represents location of ball, player$_1$, player$_2$
respectively of the InitialState $I_S$
// $F_{Lb}$, $F_{L1}$, $F_{L2}$ represents location of ball, player$_1$, player$_2$
respectively of the FinalState $F_S$
// S.clipList  denotes the list of clips associated with state S
  $I_{Lb} \leftarrow$ Initial_ball_loc
  $I_{L1} \leftarrow$ Initial_player$_1$_loc
  $I_{L2} \leftarrow$ Initial_player$_2$_loc

  $I_S \leftarrow (I_{Lb}, I_{L1}, I_{L2})$

  $F_{Lb} \leftarrow$ Final_ball_loc
  $F_{L1} \leftarrow$ Final_player$_1$_loc
  $F_{L2} \leftarrow$ Final_player$_2$_loc
  $F_S \leftarrow (F_{Lb}, F_{L1}, F_{L2})$

  QueriedClipsList ← Ø   // List of clips having subset of the
intermediate events between $I_S$ and $F_S$
  OutputClipList ← Ø   // List of clips having all the events
between $I_S$ and $F_S$

  $I_S$_Present = SearchState ($I_S$)
  $F_S$_Present = SearchState ($F_S$)
  if  $I_S$_Present = = false  OR  $F_S$_Present = = false  then
          Print “Query state does not exist”
  else
     $Im_S$ = SelectConnectedStates( $I_S$, $F_S$)
                                                // $Im_S$ - IntermediateState
if  $Im_S$ = = NULL
         Print “Cannot reach the FinalState from InitialState“
    else
      for i in range( 0 to $Im_S$.size) :
         QueriedClipsList=QueriedClipsList ∪ $Im_S$[i].clipList
          for k in range ( 0 to $Im_S$[i].clipList.size)
               for l in range ( 0 to $Im_S$[i+1].clipList.size)
                    if $Im_S$[i].clipList[k] == $Im_S$[i+1].clipList[l]
                       if $Im_S$[i].clipList[k].order ==
                           $Im_S$[i+1].clipList[l]+1
                            OutputClipList = OutputClipList ∩
                                                    $Im_S$[i].clipList[k]
                       endif
                    endif
               end
            end
       end
    endif
  endif
end
```

“Eventually Query” button on the UI in Figure 4. The system displays icons for all three objects in locations corresponding to the current state. An “Eventually” query requires the specification of an *InitialState* and a *FinalState*. The user may select current state as the InitialState and record (I_{Lb}, I_{L1}, I_{L2}) by pressing *R/S* button or may specify an arbitrary InitialState by using the gamepad. Switch *A* is used to move and position objects on the tennis court in a pre-determined order (ball, player$_1$ and player$_2$) by moving their icons. The position of each icon is constantly displayed on the UI window. The record button *R* is used to record the locations of the ball and player$_1$ (I_{Lb}, I_{L1}) for the *InitialState*. After positioning player$_2$, *R/S* button is used to record its location (I_{L2}) to complete the information needed to fully specify the *InitialState*. Similarly, the *FinalState* (F_{Lb}, F_{L1}, F_{L2}) is also specified using switch *A*, buttons *R* and *R/S*. However, this time, when *R/S* button is pressed, it not only records the location of player$_2$, but also initiates the search process.

In the second step, the query built is executed to determine if there is a sequence of consecutive events that takes the game from *InitialState(S_i)* to *FinalState(S_j)* in S^{3}G. Note thst S_i and S_j are used to denote initial and final state instead of I_S and F_S for convenience. If S_i or S_j is not present in S^{3}G, the algorithm terminates saying that queried events are not present in the tennis video. If both the states are present, the system finds all possible paths from S_i to S_j using a graph-traversal algorithm. One clip may completely include a path or a path may be spanned by a sequence of successive clips. Let the set of clips associated with S_k be C_k for i<= k <= j. The clips present in the list $\{C_i \cap C_{i+1} \cap C_{i+2} . . . . \cap C_j \}$ are identified, and each clip in which, the states satisfy the order constraint is placed in OutputClipList as a clip that includes the path completely. The clips present in the list $\{C_i \cap C_{i+1} \mathrm{U}\ C_{i+1} \cap C_{i+2}\ \mathrm{U}\ C_{i+2} \cap C_{i+3} . . .\ \mathrm{U}\ C_{j-1} \cap C_j\}$ are identified and this list is called QuriedClipList. A virtual clip which takes the game from S_i to S_j is created by reordering these clips in increasing order.

As the query is built, the clips that satisfy the conditions specified by the user are made available to the user. Each clip has *StartTime* and *EndTime* that determines the timings of the beginning and ending of a clip, respectively. The user query may not involve retrieving back to back clips from the same video. Therefore, the clip is played for the user, and it is paused automatically at *FinalState* to let the user define the next query. Also the current QueriedClipList and the OutputClipList are displayed in the UI. OutputClipList contains the set of clips that satisfy the user query. However, it may be possible that when the user reaches a step in query building process, there might not be any clip that satisfies all the conditions specified by the user so far. Our system also maintains QueriedClipList that maintains all clips that satisfied all sub-queries. The user may check either list to see the relevant clips.

IV. ILLUSTRATION

Figure 4 displays the user interface for building a query. It has three components: tennis video display, the court view, and buttons for functionalities and drop boxes for query results. The tennis video display is the major component for building a query. As the user builds a query, a corresponding clip is shown to the user. The tennis court view helps the user to associate objects with locations. There are three buttons available: “New Query” to start a new query, “Eventually Query” to skip some states during query building process, and “Query History” to visualize the query built so far. The drop boxes are used to see the list of clips that satisfy the conditions during a query building process.

The three steps involved in the query process is supported by the UI. The first step of building the query is done in the query window, “searching step” is done in the background and search results are appropriately displayed into two lists mentioned. And the last step is done by displaying the queried event in the query window. The black ring icons represent each of the player’s location (L_1 and L_2) and yellow icon represents the ball location (L_b).

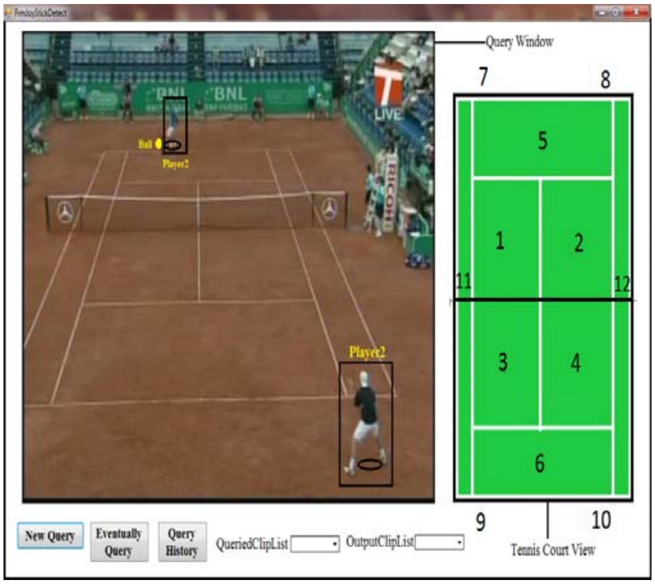


Figure 4. Snapshot of the User Interface

When the “Eventually Query” button is pressed any time during the query process, the icons for the players and the ball appear in the query window as shown in Figure 4. This allows the user to provide the location L_1, L_2 and L_b for the Player1, Player2 and Ball using the features of gamepad. These locations are recorded as InitialState I_S (I_{L1}, I_{L2}, I_{Lb}). As shown in Figure 5 I_{L1}= location 8, I_{L2} = location 6 and I_{Lb} = location 3. Similarly, FinalState F_S is provided by moving the corresponding icons using features of gamepad as shown in Figure 6. Thus F_S (F_{L1}, F_{L2}, $F_{Lb)}$ = (5, 9, 2). Figure 6 also shows the snapshot after providing the eventually query have been executed. QueriedClipList contains four clips that match any intermediate event between I_S and F_S. OutputClipList shows one clip that has the entire events between I_S and F_S.

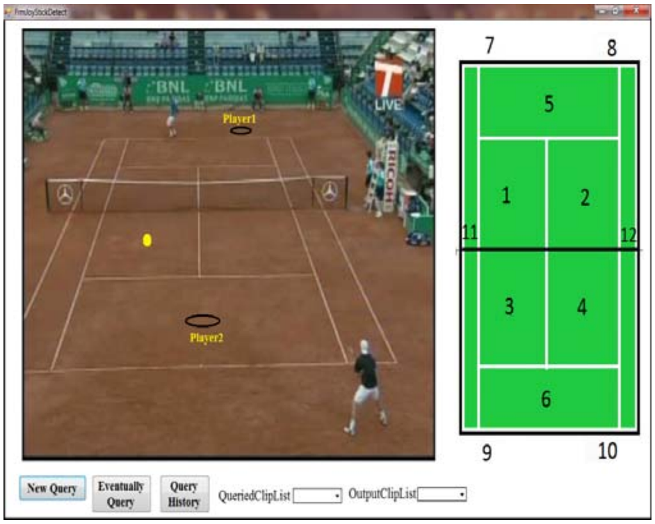


Figure 5. Snapshot after providing then InitialState I_S

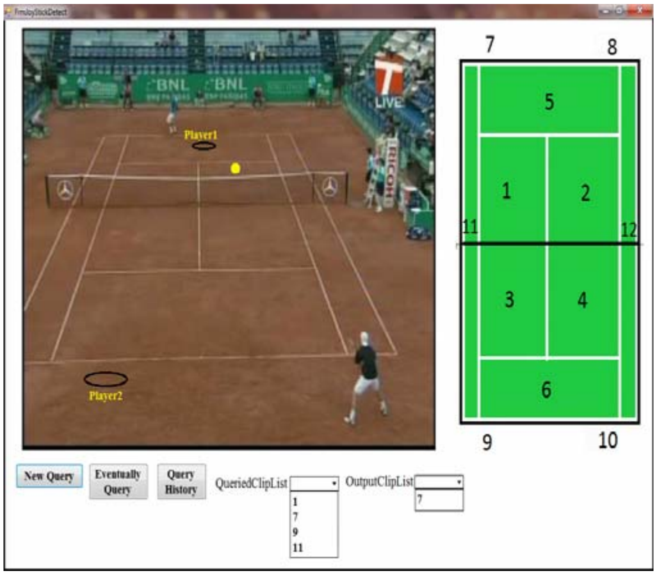


Figure 6. Snapshot showing the queried FinalState F_S and results after executing the “Eventually” query

V. USABILITY STUDY

A usability study was conducted to compare our gamepad user interface (GI) with the mouse interface (MI) that uses point-and-click approach developed by Naik [2] satisfaction as metrics (ISO, 1998). Ten users who were almost randomly selected to participate in the study were trained to use both interfaces and then were asked to build five test queries of varying complexity to take measurements to assess the three metrics. The environment was designed to ensure that the study was fair and unbiased.

The user satisfaction was measured using preference and ease-of-use. Preference is a measure that indicates the likelihood of using one interface over the other. After completing all queries each user was asked to indicate his or her preference on a scale from 1 to 5 (1 – I definitely choose MI, 2 – I prefer MI over GI, 3 – I have no preference, 4 – I prefer GI over MI, 5 – I definitely choose GI). The metric *ease-of-use* was also ranked on a scale of 1 to 5 (1 – very low, 2 – low, 3 – average, 4 – high, 5 – very high) for both user interfaces. Based on the user data, it was concluded that the users overwhelmingly (9 out of 10) preferred GI over MI with preference receiving an average score of 3.7/5.0. For the metric ease-of-use, all users ranked GI high (score 4) and MI average (score 3), respectively. This clearly indicates the gamepad interface causes less user discomfort than the mouse interface. The overall opinion of users also favored GI over MI.

VI. CONCLUSION

This paper presented an innovative user friendly system for retrieving the desired clips from tennis game video using a gamepad. The system allows user to build spatio-temporal 'eventually' queries. Eventually query is an important type of query since it is usually difficult to have a proper user interface but it is very important since the user does not need to provide all the details about a query. We have used S^3G to build eventually queries. We have developed an interface that gives the feeling of playing a game as the inputs are received through a gamepad. As future work, we look into other types of temporal queries. We also plan to specify the type of shot (forehand and backhand) while the user builds a query.

Though the indexing capability of S^3G is described for tennis game videos it can easily be used for indexing general videos like other games and news events. In all videos objects interact because of events caused by objects or natural phenomena in a limited space over time. The only differences among different types of videos are number of objects, events, spatial layout, and the associated semantics. Therefore, S^3G can be used to index general video. However, the number of states and the number of arcs (transitions) may become very large if the video involves too many objects and events. If the designer takes sufficient care to minimize the number of states and transitions based on the number of active objects and relevant events then S^3G can be used effectively. For example, in a basketball game, there are ten players on the court and one ball. Passing the ball, dribbling from one location to another, shooting are examples of events (state transitions).

At present, S^3G is built from the string representation of the video manually generated by SMART [1]. Future research should focus on automating the generation of the string data using video analysis techniques. Also work for automatic feature extraction for building content based image retrieval is going on in parallel in our group. It is suggested that experiments be conducted in an environment of a collection of videos. Minor modifications are needed to S^3G to accommodate retrieval of selected clips from multiple videos. However, the optimization for searching S^3G is limited since we are interested in all paths for an eventually query to retrieve all relevant clips. In the future, the probabilistic relationship between states through transitions can also be studied.

ACKNOWLEDGMENT

This material is based upon work supported by National Science Foundation under Grant No. 0812307.

REFERENCES

[1] Jain, V. and Aygun, R. S., "SMART: A grammar -based semantic video modeling and representation," IEEE Southeastcon, 2008, pp. 247-251.

[2] Naik, M., Jain, V., and Aygun, R. S., "S3G: A Semantic Sequence State Graph for Indexing Spatio-temporal Data - A Tennis Video Database Application," IEEE International Conference on Semantic Computing, 2008, pp. 66-73.

[3] Erwig, M. and Schneider, M., "Spatio-Temporal Predicates," IEEE Trans. on Knowledge and Data Eng., vol. 14, no. 4, July 2002, pp. 881-901.

[4] Erwig, M. and Schneider, M., "Developments in spatio-temporal query languages," Tenth International Workshop, Database and Expert Systems Applications, 1999, pp. 441-449.

[5] Bonhomme, C., Trépied, C., Aufaure, M., and Laurini, R., "A visual language for querying spatio-temporal databases," Proceedings of the 7th ACM international Symposium on Advances in Geographic information Systems, 1999, pp. 34-39.

[6] Sourina O., "Visual 3D Querying of Spatio-Temporal Data," International Conference on Cyberworlds, 2006, pp. 147-156.

[7] Erwig, M. and Schneider, M., "Spatio-Temporal Predicates," Technical Report, FernUniversit at Hagen, 1999.

[8] Erwig, M. and Schneider M., "Query-by-Trace. Visual Predicate Specification in Spatio-Temporal Databases," Proceedings of the 5th IFIP Conf. on Visual Databases, 2000, pp. 199-218.

[9] Cavalcanti, V. M., Schiel, U., and de Souza Baptista, C., "Querying spatio-temporal databases using a visual environment," Proceedings of the Working Conference on Advanced Visual interfaces, 2006, pp. 412-419.

[10] Li, X. and Chang, S. K., "An Interactive Visual Query Interface on Spatial/temporal Data," Proceedings of the Tenth International Conference on Distributed Multimedia Systems, 2004, pp. 257-262.

[11] Silberschatz, A., Korth, H. F., and Sudarshan, S., Database System Concepts, 3rd Ed., McGraw Hill, 1997.

Development of a Data Model for an Adaptive Multimedia Presentation System

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
scutts@bournemouth.ac.uk

Nick Rowe
Faculty of Technology
Bournemouth and Poole College
Bournemouth, UK
nrowe@bpc.ac.uk

Abstract

We investigate the requirements and nature of data models for a multimedia learning system that presents adaptable learning objects based on a range of stimuli provided by the student and tutor. A conceptual model is explored together with a proposal for an implementation using the well-known relational data model. We also investigate how to describe the learning objects in the form of hierarchical subject ontology. An ontological calculus is created to allow knowledge metrics to be constructed for evaluation within data models. We further consider the limitations of the relational abstract data model to accurately represent the meaning and understanding of learning objects and contrast this with less structured data models implicit in ontological hierarchies. Our findings indicate that more consideration is needed into how to match traditional data models with ontological structures, especially in the area of database integrity constraints.

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

I. INTRODUCTION

In previous work [1], we proposed an Adaptive Multimedia Presentation (AMP) System to provide a semi-automated tool for learning that adapts to students' needs. A prototype was constructed and evaluated in a real class environment in the Cisco Academy at Bournemouth University [2]. This showed that undergraduate students liked using the AMPS, but would prefer more 'adaptability' in the presentation of materials. The results led the writers to conclude that more investigation was needed to find alternative, flexible methods of multimedia learning object creation, storage and retrieval. The principal aim of this paper it to look further at the conceptual, semantic, and ontological data modelling issues involved in the making a more rigorous AMP system implementation.

In section II, we set out our understanding of the learning object concept and its role in our AMP system. In section III, we look at the role of adaption and the staging of its implementation. In section IV, we present a conceptual model of AMPS and relate it to subject ontologies. In section V, we create the necessary ontology calculus to enable us to produce knowledge metrics that feed into our AMP system and use the structure of the ontology itself as a reference point for the construction of learning objects. Section VI indicates how all of this might be implemented in a relational data model, while section VII reflects on the appropriateness of using relational models for hierarchical structures. The paper concludes with Section VIII indicating future directions.

II. 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, [3]. 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 supposed benefits is that these objects can be reused [4]. However, interoperability and reusability may have been overstated. McGreal, [5], points out the difficulties in reusing a learning object in a different environment. This is principally because it is difficult to create learning objects independent of context. The likelihood is that the object bears the imprint of the ideology and culture in which it was produced. Links between objects in different contexts may still be useful to students, because it provides another way to see a concept, and may well provide alternative applications and examples. Boyle, [4], describes the learning object as a wrapper around content. The wrapper describes the structure of the object and includes the metadata about the object. The learning object is packaged in a standard container format which can be stored in a database. The included metadata permits fast effective searches to retrieve learning objects suitable for a particular purpose.

The Linking of Learning Objects

Breaking up knowledge into learning objects based on the content structure highlights the importance of two aspects of the presentation of materials. Firstly, a lesson can be considered to be a selected sequential set of segments and secondly, any segment presented may be connected to another segment in the database.

Authoring a lesson becomes a process of

- choosing related segments in the database
- creating new segments
- attaching metadata to the new segments to allow them to be linked, once published.

III. ADAPTING CONTENT

Adaptation can take many forms but it is important to realise that adaption, as in nature – so in computing, is always in response to a particular stimulus.

Stage	Stimulus	Adaption	Method
1	Student emails	production of new video segments	Manual
2	Student prior knowledge	selection of video segments	pre-lesson test
3	Student ability	selection of video segments	Real-time response

Figure 1: Staging of Adaptive Methods

The AMP system is at present only adaptive at stage 1 in responding to manually produced additional video segments to the stimulus of student emails. This is considered a low level of adaption and is not automatic. The adaption is performed by the tutor rather than the AMP system and thus requires a huge manual effort to respond to requests for further information. We plan to increase the number of stimuli which produce automatic responses. Possible stimuli will include student prior knowledge and student ability, which we call the "student signature" and will be developed further in another paper.

In order to introduce adaptation into AMPS, segments are presented with different levels of detail for each student according to the

1. level set by the original author of the segment in deciding a preferred presentation level.
2. tutor model in the AMPS can override the author level by using test information about the student's level of knowledge.
3. student is allowed to alter the level of detail presented.
4. selections of level can be made persistent.

A typical lesson segment will be 2-5 minutes long. The presentation system plays an AV file in real-time leaving the original segment intact. Metadata carried with the segment is used to cue synchronized events such as the display of an incremental HTML file. Here the file is formatted as a normal HTML file presented in paragraphs by adding it to a display box at a time specified by an adaptive descriptor.

IV. Conceptual Model for an Adaptive Presentation System

AMPS Data Schema

The aim of the AMP system is to link together learning objects as segments so students can explore by following links, regardless of the lesson or course on which the student begins their journey. The AMP system needs to respond to the meaning of segments to enable the automation of data link creation.

The aim of this section is to carefully define terms and concepts used in the AMP system model. Textual definitions are given and then a representation is derived of the system as an ordered graph.

AMP System -Textual Description of Terms

Administrator

The administrator is a role which completes any task not related to the courses or their content. The role of adding, editing or removing students would be considered an administration task. The role of adding, deleting and editing courses, lessons and resources may be completed by the Tutors or Authors. This role may be given to human effort or automated.

Answer

An answer is the answer to a single question available on a test for the student. Questions and answers are to be determined and designed by the author. This may become another task the system can automatically undertake.

Author

The author creates the segments, lessons and/or courses by means of implementation and editing. The author may be the same person as the tutor.

Tutor

The tutor determines the intended content of courses, lessons and resources and may instruct the author on the construction of materials for delivery.
This role may be partly replaced by automation in the future.

Student

The student is the course subscriber, or person learning the course content and committed to completing a course. Once all courses to which the student has subscribed are complete the student ceases to be a student. The student may be subscribing to many courses at any one time. Subscription may be limited or prevented by the delivering institution or the tutor.

Course

The content is delivered as a set of lessons related by the sequence in which they are to be presented to the student. The content can be referred to by a single attribute known as the course title. The set of content related learning segments the student is committed to complete, or is given access to, by completing an enrolment or subscription process.

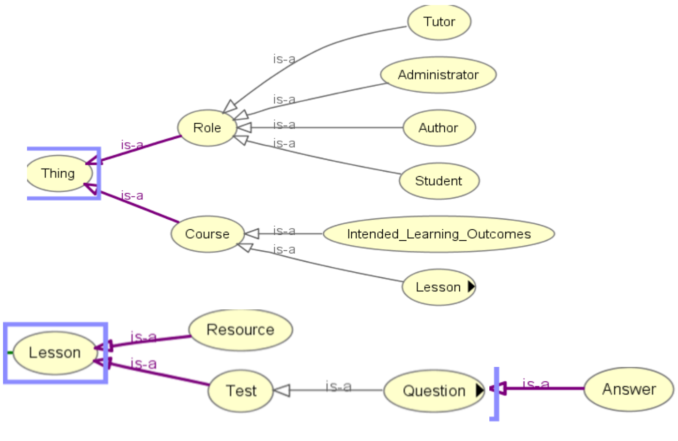


Figure 2: Inherited attributes

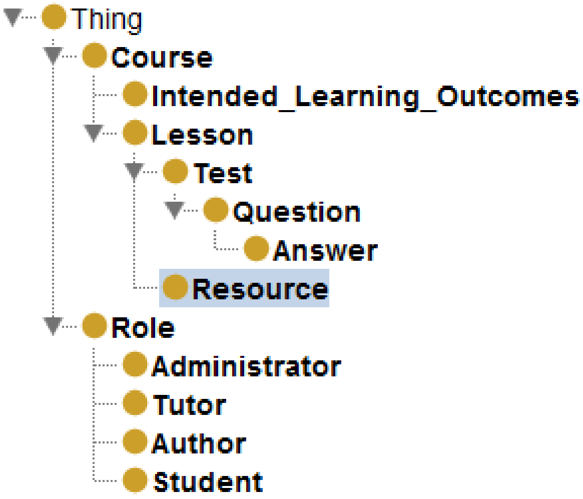
Lesson
A lesson is a set of learning segments related by the sequence in which they are to be delivered. As a course is normally made of several lessons so a lesson is normally constructed from several segments of media and the sequence related to those resources. A single lesson can be referred to by its title. It forms part of a course or a number of courses at any one time.

Segment
A segment is the description of the timeline of a single piece of media, or part thereof. Each segment conveys to the student a single point of learning. The granularity of what constitutes a single point of learning is to be determined by the tutor and constructed into a resource by the author. A segment is part of a lesson and a set of segments can be identified by the lesson title and the sequence identifier within the lesson. The delivery of the entire sequence of segments may vary if student knowledge has been proven and tested to deem a particular resource need not be presented to the student. This is part of the personalisation process.

Test
A test tests the students' knowledge of the content of a section of the curriculum. That section may be based on the course or lesson level.

Question
This is a question available on a test. There may be many answers for each question where the MCQ format is used. A set of questions is formed to become a test for a lesson or course. Questions and answers are to be determined and designed by the author. This may become another task the system can automatically undertake in future.

Class Hierarchy of Terms in Protégé Software
This can be expressed more compactly in ontological form in terms of classes and entities. The relationships between terms in the class hierarchy are shown in Figure 3 modelled in Protégé [6].



- Thing
  - Course
    - Intended_Learning_Outcomes
    - Lesson
      - Test
        - Question
          - Answer
      - Resource
  - Role
    - Administrator
    - Tutor
    - Author
    - Student

Figure 3: Ontological Structure represented in Protégé

V. Ontological Calculus

Since the storage of information needs to be indexed in order for it to be retrievable, the segmentation of individual learning objects will need to reference the ontology of the subject knowledge area in order for it to be retrieved and structured into lessons. It will be essential therefore to construct full subject ontology [7] to which all the learning segments are related.

An ontology can be represented as a tree network where there is one and only one path between two nodes. While an ontology specifies the structure and relationships within a body of knowledge it is also necessary to determine metrics in the structure which can be used to provide measures of attributes such as complexity, level of detail or closeness of subject areas. The first step to defining these metrics is to provide each node with a unique address which defines its location on the ordered tree.

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. [8] This structure is then used to label an ontology where fragments of knowledge have an order determined by their pre-requisites. Thus a body of knowledge is divided into section, sub-section, sub-sub-section etc. and so we adopt an addressing system which corresponds to this knowledge hierarchy where each address is correspondingly specified by sections, sub-sections, sub-sub-sections etc.

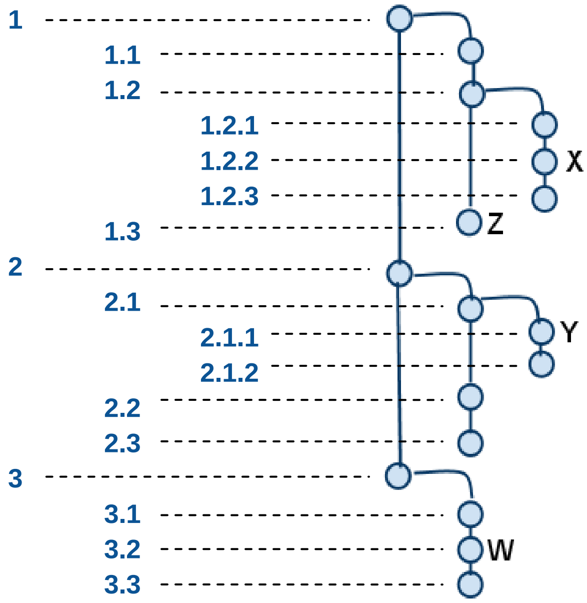


Figure 4: Knowledge hierarchy corresponding to an ordered tree

Node Address Notation
Our unique addressing system for each node is in the form of an array which has entries providing positional representation for each node level. For simplicity we

use the matrix notation of Bras and Kets borrowed from quantum formalism where <X| represents the left ideal (row) and |X> represents the right ideal (column) of the matrix array.

Thus from Figure 4 we find node |X> = [1, 2, 2] and node |Y> = [2,1,1]

The first stage in producing a calculus which can be used for the determination of knowledge metrics is the mathematical representation of unique addresses for nodes within the tree network. We also define the following representations for specific elements:

The unity ideal <U| = [1, 1, 1…] and correspondingly |U> = the equivalent column vector extendable to n dimensions

The level ideal <L| = [1, 2, 3,...] and correspondingly |L> = the equivalent column vector extendable to n dimensions

We will also have cause to make use of the Level Order of Magnitude ideal <LOM| = [1, 0.1, 0.01, …]. In addition we make the following definitions:

Segment: a segment is defined as a node together with all its sub-nodes. The total number of nodes in a segment is a measure of the amount of detail contained within a segment of knowledge and can be

Complexity: we define complexity of a knowledge node to be equal to the degree centrality minus 1which is the measure of the number of sub-nodes that are connected to a given node. Thus a knowledge node composed of many sub-nodes or subdivisions is deemed to be more difficult than one with fewer subdivisions and is a measure of **difficulty** of the knowledge node.

Level: We designate the term level applied to each node by the position it occupies in the representation.
Thus the level of node |X> is 3 while the level of node |Z> is 2. We say that the **level** of a knowledge node is equal to its **importance** and represents the level of detail that a knowledge node contains.

Distance: the distance or separation of one node from another is a measure of how close two knowledge segments are related to the subject ontology. For a tree network this is a unique value determined by the number of steps between the nodes. However the separation of knowledge segments is dependent also upon the level of the nodes traversed (i.e. nodes at level 3 are an order of magnitude closer than nodes at level 2 and those at level 2 an order of magnitude closer than at level 1). We therefore define distance between nodes as the number of nodes traversed divided by the order of magnitude of their level. Thus two neighbouring nodes at level 1 will have a separation of 1, while two nodes at level 2 will have a separation of 0.1 and those at level 3 a separation of 0.01 etc. Distance is a measure of the **strength of connection** between two nodes.

Thus to obtain the distance between two node we use an algorithm which takes the modulus of the difference between the nodes and multiplies it by the level order of magnitude vector.

Distance |X>|Y> = <LOM|[|X> - |Y>]

Thus for the nodes in Figure 4 we have the following assigned addresses

|W> = [3, 2, 0]
|X> = [1, 2, 2]
|Y> = [2, 1, 1]
|Z> = [1, 3, 0]

Hence the distance D between various nodes is

D[|X> - |W>] = [1, 0.1, 0.01] (|[1, 2, 2] - [3, 2, 0]|)
= [1, 0.1, 0.01]|[2, 0, 2]
= 2.02
While
D[|X> - |Y>] = 1.11
D[|Y> - |W>] = 1.11
D[|Y> - |Z>] = 1.21
D[|X> - |Z>] = 0.12

If we have a general node |A> = |a.b.c> then distance of |X> = |i.j.k> from |A> is given by the algorithm:

D |A> - |X> = (|a-i| * 1) + (|b-j| * 0.1) + (|c-k| * 0.01)

This set of algorithms form a calculus which enable clear metrics to be determined that can be calculated and fed into the AMPS system to facilitate adaption.

VI. IMPLEMENTING THE DATA MODEL

Our AMP system has been structured on a relational data model, where the user data together with the knowledge content data (in the form of learning object segments) is held in a relational database

Figure 5 depicts the rudimentary data model of the AMP system which has been derived using Chen's ERA method. We expect the segment entity to hold such attributes as **Level** (a measure of the importance of the segment) and **Complexity** (a measure of the difficulty of a knowledge node) as well as **Strength** of nodel links (a measure of the ontological proximity of the knowledge areas). Each of these three metrics are determined through the ontology calculus.

Adaption is performed by additional metrics attributed to the student entity and the tutor entity. A student signature will contain a measure of the prior knowledge of the student to enable adaption at level 2, and student ability to enable adaption at level 3 in real time as indicated in Figure 1.

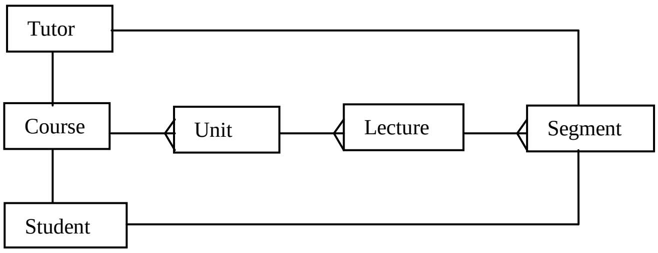


Figure 5: Schematic representation of basic ERD

However it should be noted that this data model requires that the knowledge tree (or subject ontology) is contained within the relational structure along the content-backbone of

COURSE-UNIT-LECTURE-SEGMENT

However the appropriateness of using a relational model to represent the semantics of the learning system is potentially problematic and we turn now to a consideration of the issues involved in using a relational data model to hold an essentially hierarchical ontological structure.

VII. Evaluation of Data Model Limitations

The object of this paper has been to produce a robust architecture and design independent of any given implementation model. A key issue has emerged that concerns the AMPS architecture. Specifically, the suitability is in question, of a relational database to store and retrieve learning objects in real time to dynamically assemble learning objects in an multi-media presentation that adapts objects to the user's learning abilities and needs.

Ted Codd introduced the relational data model with 12 rules (actually 13) of relational database management in 1969 [9]. The 'Relational Model' altered data management systems at the time because it imposes strict rules of formal logic. Previously ad hoc methods were used to stored and retrieve data items held in network or hierarchical data models [10]. These abstract data models had arisen informally from contemporary data storage structures, such as storing pointers to files that connect records. Some designers had realised that rigorous approach was needed, and were using forms of relational algebra before Codd formalised these into the abstract relational data model. The new rules ensured that stored data conformed to integrity constraints, so that a well-structured data bank only stored 'true' data and could be relied upon to only allow correct data that conformed to the integrity constraints to be stored. Other data was rejected as false.

Recently, extensions to the relational model have been suggested that create novel data models and some have been used in commercial products such as the object database called ObjectStore [11] [12]. More or less, these data models allow semi-structured or unstructured data to be stored in 'relational' databases. However, such extensions do not adhere strictly to the relational model and are considered to be ad hoc.

Whilst modelling the AMP system, the writers have found it necessary to strictly define the use of database concept by rigorously defining terms. For example, it has become of vital importance to distinguish the terms 'abstract data model' from the usage of 'data model' as implicit in database design methods.

Furthermore, we need to carefully consider structures that describe learning objects, or segments, in a subject hierarchy– which we have identified as an ordered tree structure or ontology – and contrast it with storing of that data in a relational database. The first of these is essentially hierarchical, while the abstract data model used for storing the data is relational.

While designing the AMP system data model, an attempt was made to model learning objects consisting of lesson 'segments' in a relational database. The writers have encountered e-learning application data that is structured in different ways. This is typically blocks of text, for example HTML, or multimedia data types such as animations, that are linked in network hierarchies such as tree structures, rather than simple data types normally stored using the abstract relational model. This needs further investigation.

'Semantic Modelling 'is the attempted representation of 'meaning' to allow systems to interact 'intelligently' with users [13]. However, relational databases understand little about the data they store, and what it actually means to humans. Relational database management systems 'understand' only simple data types and certain integrity constraints. Understanding or meaning is left to the user of a database when using the relational model. For the writers, semantic modelling is about the structure of meaning rather than the structure of data. The relaxation of integrity constraints in still controlled ways is required to maintain a rigorous logic to data stored, and this where most change to data management methods is occurred. The modifications result in changes to the abstract data model, in addition to the ways data itself is viewed.

Recent developments are essentially partial reversals to less rigorous, more ad hoc abstract data models, similar to the pre-relational ones, and their use is unhelpful in the context of the pure relational model.

Semantic Modelling is often referred to as a form of 'Data Modelling' (e.g. applying Chen's ERA modelling to a problem domain[14] [15]) to capture persistent data. This is useful as an aid to database design, but is distinct from the writers' interpretation of semantic modelling used in this paper.

Semantic modelling in a rigorous sense ought to relate to capturing the 'meaning' or 'intelligent interpretation' of data. The other commonly used meaning of semantic modelling relates to data modelling to design an implementation on a DBMS which is an implementation of the (abstract) relational model. Further research is needed into the nature of rigorous models, tools and techniques for semantic modelling, tools. It is a growing research area for multimedia systems in general, and the complexities of interpretation of meaning, semantics and data are compounded by the adaptive features of the AMP system.

VIII. Conclusion

Investigations into semantic models and semantic modelling should be strictly logical explorations into how data models and integrity constraints can be modified without rendering the database contents (facts, meanings, and intelligent interpretations) uncertain or meaningless.

Meta-learning by the AMP system requires awareness that it is participating in a learning process and therefore needs an explicit, built in ‘tutor model’. The current AMP system implicitly assumes there is a real-life tutor who will perform the role of the tutor model, which involves intelligent and experienced selection of learning objects appropriate to the student.

In future, we need to construct a full, robust tutor model to automate the segmentation process, which needs detailed investigation of the nature of meta-learning []. 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 ‘tutor model’.

Finally, although work discussed in this paper answered research questions posed in previous papers, it has indicated further questions with a different emphasis:

1. What is the usability level of the user interface and how can this be further improved?
2. What further adaptation features are required and how are they to be evaluated?
3. What model is best employed to define the interaction between the interface and the adaptation engine?
4. What is the full specification of the ontology required and how is it captured?
5. How should database schemas be constructed for the AMPS for real-time extension at data and meta-levels?
6. How should the ontology engine structure be modelled and evaluated? Can fuzzy logic or data mining techniques be candidates for a useful algorithm?
7. How do we determine the appropriate definition of an API, possibly by means of an IDL, between the ontology, the adaptation engine and the AMP system’s user interface?

We leave these questions to further papers.

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 28 January 2010]

[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, Charles, Landis, Gordon, Orenstein, Jack, Weinreb, Dan., (1991). ‘The Objectstore Database System’, *Communications of the ACM* 34 (10): 50–63.

[13] Date, C., Darwen, H. & Mcgoveran, D. (1998). ‘Relational Database Writings 1994-1997’, Addison Wesley.

[14] Chen, P. ‘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.

A Semantic Approach for the Repurposing of Audiovisual Objects

Benjamin Diemert, Marie-Hélène Abel, Claude Moulin
Université de Technologie de Compiègne, France
Email: (benjamin.diemert, marie-helene.abel, claude.moulin)@utc.fr

***Abstract*—In this paper, we address the issue of audiovisual document modeling to promote its repurposing. It requires to identify the various components of the document that can be reused during professional production. We describe a use case which incorporates User Generated Content and considers various exploitation forms and medium. We define a conceptual model and present real applications using it. The main contribution of this paper is the conceptual model which distinguishes between editorial, encoding and sequencing aspect of an audiovisual document.**

***Keywords*-audiovisual content; audiovisual document; repurposing; metaproduction**

I. Introduction

The main issue at stake in the audiovisual field is the lack of standard formats to support content creation and repurposing. The modeling of the audiovisual document is still an open debate which should not be restricted to the choice of a media wrapper. Indeed, audiovisual documents are not just the digital files delivered to the viewer at the end of the production chain. They are a montage of content picked up in various files and ordered given editorial guidelines to convey a message. A fine modeling of the audiovisual objects is needed to promote a common view on the audiovisual document components and thus foster their repurposing.

A closer look on audiovisual repurposing – also called metaproduction by [1] – reveals various practices ranging from sampling of content fragments, to complete reuse of document's parts via reencoding of video material to fit technical restrictions. Each practice enables the exploitation of a different level of the audiovisual document, i.e. the content assemblage, the editorial structure or the material realisation. Repurposing is thus a desired practice because it reduces production costs and opens up new exploitation forms.

Another significant shift to promote audiovisual repurposing is the ability to integrate User-Generated Contents (UGC) into professional production. Since the widespread of quality camera in mainstream electronic devices, amateurs are now able to create audiovisual content in a glimpse. They can now compete with professionals for the capture of unpredictable events such as natural disasters. However, their lack of training in audiovisual production requires additional efforts to leverage the quality of UGC to professional criteria. This trend enables professionals to diversify their sources of content as well as produce content in a more participative way.

In any case, repurposing opportunities require a shared conceptual model of the audiovisual document components to be seized and exploited. Ontologies as formal representations of a conceptualization can provide such a model. The intent is to provide public conceptual models that can be mapped together and extended to fit more specific use [2]. Hence, an ontology of audiovisual document could be related to content description ontologies but also specialized to handle specific kinds of documents. Information about an audiovisual document created from such an ontology could then be linked to other data on the Semantic Web and thus integrate both media- and domain-dependent descriptions.

In this paper, we present a use case implying UGC integrated into a professional production and show modeling aspects required by it (section II). Then, we explain how our ontology copes with these requirements (section III) and compare it to other related standards and models (section IV). Finally, we show how applications are using our model to provide a solution for UGC repurposing in a similar situation to our use case (section V).

II. Use Case and Requirements

A. Use Case

Let's consider a public channel interested in the capture of a cultural event such as an opera, a play or a concert. The channel's producers are interested in three kinds of content; *material from the event* itself, *interviews* of the director and performer and *comments from members of the audience*. Producer's concern is to minimize the number of people employed and the equipment used while ensuring the material quality and maximizing its exploitation. Thus, they envisage to split the work and call for amateurs' contributions:

- the capture of the event requires a professional staff with sufficient technical skills and various recording equipments for a long period of time (equipment set up, recording test etc.).
- interviews can be performed by an another team composed of only a journalist and a cameraman.
- comments from the audience members can be shot by the spectators themselves before and after the performance. In order to ensure its quality, they should be assisted and follow guidelines.

Copyright (c) IARIA, 2011. ISBN: 978-1-61208-129-8

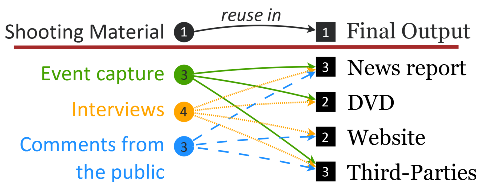


Figure 1. Overview of the material created in the use-case and their exploitation opportunities

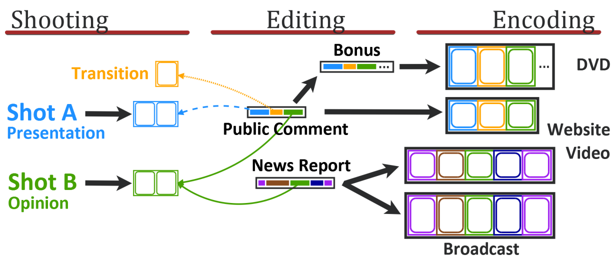


Figure 2. Detailed view on the spectator's comments processing for each exploitation case

Concerning the material's exploitation, Fig.1 depicts 4 possible exploitation cases (in *italic*):

- a *news report* for the channel is created from an assemblage of every kinds of content
- materials from the event are kept for a *DVD* with the interview and best comments from the public as bonus content
- some comments from the public and the interview are reused on the auditorium's *website* for promotion
- any part of the material can be sold to a *third-parties*

All exploitations reuse some shooting material but require to be produced by distinct processes in order to fit the specificities of each distribution medium/channel and meet the expectations of the intended audience. However, the shooting is organised independently from the rest of the production chain. While every process share the resulting material, it is managed differently to produce the final content. This leads to variations in the material encoding quality and the content editorial formatting. For instance, a news report has usually a more intense pace than the bonus content of a DVD and thus different editing of the same material. Moreover, bandwidth constraints are quite different between broadcast and web distribution as between dvd and vhs tape recording.

In our case, the multiplicity of production teams and the diversity of repurposing situations does not admit a simple solution like manual file annotation and sharing. With each team dealing with other's material, there is a need to clear up what result of the process will be reused for each exploitation case. From now on, we focus on the processing of the spectators comments shown Fig. 2:

- Each comment *shooting* is divided in two shots, one for the spectator's presentation and another to express his/her opinion on the performance. These shots are the raw material shared by DVD bonus, promotion website and the news report exploitation case.
- The two shots are *edited* differently according to the editorial structure defined by the professional producers of each exploitation cases.
  - The DVD and the promotion website share also parts of the editing line. Both assemble the shots together with a transition in between. A selection is directly published on the website while another editing with a compilation of the best ones is created for the DVD bonus content.
  - For the news report, only the opinion shot is integrated into the report structure along with jingles and comments from a journalist.
- Each final edit is then *encoded* specifically to fit the distribution constraints, the highest quality for DVD and channel broadcast then a lower quality for the website distribution. Note that the news report is distributed both on the public channel and on the channel website, resulting in two different encoding of the same edit.

B. *Modeling requirements*

The details of this use case specifies more than a simple reuse of material. It specifies the kind of processing that support repurposing and which defines thus the modeling requirements for the audiovisual document:

- the *reencoding* of edited material to fit the technical parameters proper to each distribution medium/channel (news report distributed by channel broadcasting and internet).
- the reuse of shooting materials in two distinct editorial structure (*resequencing* of the opinion shot in the website and news report editing).
- the reuse of a part of an editorial structure into another editorial structure (*repurposing* of the public comment editing into the DVD bonus editing).

In the next section, we define a conceptual structure which represent the audiovisual document. The principle of our modeling is to identify distinct components of the audiovisual document in order to handle them separetely and thus enables these three kinds of reuse.

III. AUDIOVISUAL OBJECT MODEL

The purpose of this section is to detail the modeling of an audiovisual document. Firstly, we clarify the distinction between video material and content sequence (subsection III-A), then present an example extracted from the use case which illustrates reencoding and resequencing (subsection III-B). Secondly, we distinguish between content montage and editorial structure (subsection III-C) then show an example which depicts repurposing (subsection III-D).

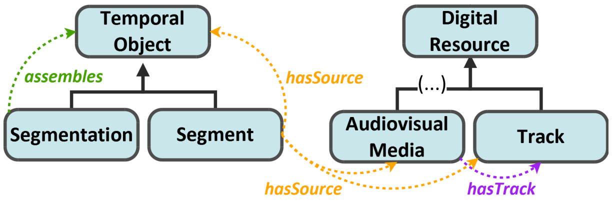


Figure 3. Concepts representing digital file, audiovisual content and their relations

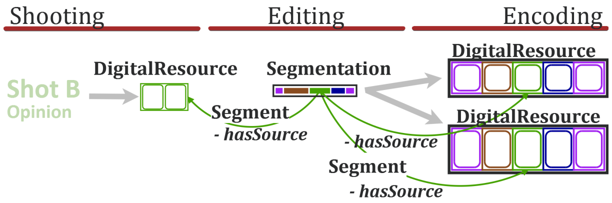


Figure 4. Modeling of the opinion shot used for the creation of a News Report

A. Resource versus Content

The challenge of audiovisual object modeling is to distinguish its various components and relate them. The audiovisual object is all together a recording, a content and a document. The recording is a digital file stored on a medium. The content refers to the viewing or the playing of this recording, i.e. what can be actually perceived by a human or sensed by a machine. The document is a communicative object created and structured in order to convey a message to others.

In this first part, we will focus on the recording and playing part of the audiovisual object. The recording is created from the encoding of a video flow with a compression algorihtm such as MPEG-2 or H.264 and encaspulated in a digital file according to a wrapper format such as AVI or MKV. The audiovisual content flow is reconstructed from the decoding of a series of bits. One important feature of this relation is that content can be recreated from various recording – with different encodings for instance.

From this first description of the audiovisual object, we have defined a distinction already highlighted by [3] between its recording form (*DigitalResource*) and its playing form (*TemporalContent*). Fig. 3 depicts the main relations between these concepts and their specializations:

- A **DigitalResource** is a sequence of bits with its own address (*accessPath* attribute), like a digital file or a portion of a digital file. We distinguish between two kinds of DigitalResource:
  - a **MediaWrapper** is a DigitalResource which encapsulates any kind of media resources (audiovisual, picture, text etc.) according to a given format (*wrapperFormat* attribute). A MediaWrapper details the name of the encoding algorithm used to create it (*encodingMethod* attribute which in the case of video use the FourCC identifier. FourCC provides a four character code identifier as well as a short description for 331 video codecs. A MediaWrapper also specifies the main encoding parameters such as *samplingRate* (frequence of value picking) and *bitResolution* (value encoding precision). We specialize this concept for resource holding audiovisual content with the concept of *AudiovisualMedia*.
  - a **Track** is a portion of a digital file holding a series of content which may be unwounded to create temporal content, like an audiovisual, audio or subtitle track. Track shares the same attribute than MediaWrapper to define its encoding method and parameters.
- A **TemporalObject** is an object which content is unwinded and revealed progressively during a fixed period of time (*objectDuration* attribute). We distinguish two kinds of TemporalObject:
  - a **Segment** is a temporal selection on a flow. It uses the timecode system from the source as a reference to define its starting position in the flow (*timeCodeIn* attribute) and its duration to define its ending position. The source of a Segment (relation *hasSource*) designates the object on which the timecode selection is made. It can be one or more TemporalObject, MediaWrapper or Track.
  - a **Segmentation** is an ordered collection of TemporalObject. Each TemporalObject is related to the Segmentation by an *assembles* relation. Segmentations can be used to represent the temporal content montage or a particular content indexing.

The goal of this modeling is to keep record of the relation between content and the ressources which provides the video material. We explain now how to use these concepts to model the reencoding and resequencing of audiovisual material.

B. Reencoding and Resequencing

With the concepts of DigitalResource and TemporalObject, we can represent the three stages of our use case like shown Fig.4:

- The shooting stage creates DigitalResources, such as the audience's member opinion shot.
- The editing stage creates a Segmentation which represent a simpler version of the actual editing line. The Segmentation is composed of Segments pointing to the DigitalResource used as raw material – thanks to the hasSource relation. In our scheme, the hasSource

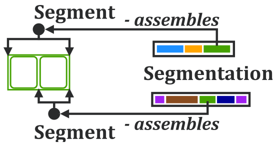


Figure 5. Resequencing of the same DigitalResource implies two different Segments here included in two Segmentations

attribute linking the opinion shot DigitalResource to its Segment is represented by a thin green arrow line.

- The Encoding stage creates two DigitalResources as outputs. One Segment is enough to represent the content sequence corresponding to the opinion shot and encoded differently in each DigitalResource. Indeed, the Segment has the same beginning and duration but points out to distinct DigitalResource through distinct hasSource relations.

This example shows that the *hasSource* relation between a Segment and a DigitalResource allows a multiplicity of technical variations (encoding method, wrapper format) and copies to be linked to the same content. This linking enables thus to deal with reencoding of the same content.

In the same example, if some part of the opinion shot needed to be cut away for the news report, then this selection would be represented by another Segment with another beginning and duration. Fig.5 shows an overview of this case where the news report Segmentation would rather integrate this Segment. In addition, the selection could be directly made on the spectator's comment Segmentation as it is also a temporal object with a duration. Thus, the Segment concept allows us to deal with the resequencing of content into various Segmentations.

C. Content versus Editorial Structure

An audiovisual object is not only a content recorded on a medium, but also a document created by humans to convey a message to other humans. Thereby, the audiovisual object is related to an intended message defined during pre-production, an actual content realized during production and a perceived message interpreted during the viewing. Usually, a document refers to a genre which prescribes a pattern to structure the message. The genre pattern aims at easing the message transmission by providing a common reference to the document's creators and viewers. For instance, everybody expects an interview to be composed by a series of questions and answers.

We distinguish between the document which refers to a genre and the editorial objects which form its actual structure. A common view is to consider the document as a production result while the editorial objects represent the creation units. The document is created one editorial object at a time, but it is ordered, sold and distributed as a whole. At pre-production stage, the editorial objects hold intentions on the content to be produced. At production stage, materials are created or acquired to express at best these intentions. At the editing stage, contents montage can reveal several particular points of view on the way to realize the original intentions. At post-production, the editing is finalized and the document can be packaged for distribution.

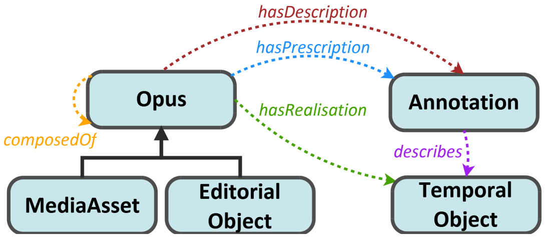


Figure 6. Concepts representing content, editorial structure, annotation and their relations

From these details on the fabrication of the audiovisual object, we have defined a distinction between an editorial perspective and the assemblage of video material. This distinction enables us to separate the editorial structure (*EditorialObject*) of audiovisual document (*MediaAsset*) from its potential realisations (TemporalObject defined previously). In addition, we can describe each of these elements with specific *Annotation* such as script extract, dialogue, signal analysis etc. Our Annotation concept is generic and extensible as defined in [4]. We do not defined it further in order to focus on audiovisual objects modeling, even if we are well aware that annotations are needed to describe the audiovisual object components. We just point out that different kinds of annotation can be used and explain where they should be attached. Fig.6 depicts the main relations between these concepts:

- a **MediaAsset** is a document intended to be published. Its structure varies for each specialization of media asset and defines thus a new genre or format. For instance, an interview is composed of questions and answers.
- an **EditorialObject** is a document's fragment which composes the editorial structure of the MediaAsset. EditorialObject's composition can also be structured by editorial rules. For instance, a shot can be divided in subshots to detail complex camera movements like travelling followed by a panning. For audiovisual documents, we specify generic and basic kinds of editorial objects:
  - **Shot** is a series of uncut frames, which means a group of pictures recorded between two pushes on the record button (on and off). Shot can be composed of SubShot or FilmedElement.
  - **SubShot** is a part of a Shot with a continuous camera movement or a constant framing.
  - **FilmedElement** is something that appears in the

frame, usually a main part like an actor, an object etc. rather than the background.

MediaAsset and EditorialObject can be merged into the more generic concept of **Opus**. An Opus is a repurposable piece of work with an editorial consistency. Its purpose is to enrich the modeling of the piece of work by binding it to other concepts:

- a prescriptive annotation which reflects the editorial intentions to realize (relation *hasPrescription* with Annotation as range). Such an annotation can integrate script extract or dialogue.
- a portion of video material which has either been the result of an original creation or a selection of existing content (relation *hasRealisation* with TemporalObject as range).
- other Opus which contribute to define an editorial structure (*composedOf* relation).
- descriptive annotations which reflect the editorial choices made during production (relation *hasDescription* with Annotation as range). Such annotations can be considered as an updated and extended version of the prescriptive annotations. They can integrate extracts from the final script, actual dialogue but also conceptual indexing from a controlled vocabulary or an ontology.

Note that there is another kind of Annotation which is directly connected to a content sequence (relation *describes* with TemporalObject as range). These Annotation can be composed of video material analysis which are specific to the content and not to the editorial choices made.

D. Repurposing

The goal of repurposing is to support the reuse of editorial objects inside various documents. As an example, we explain how the spectator's comment can be repurposed. The whole comment is modeled as a MediaAsset, while the presentation and opinion shots are EditorialObjects. After the shooting, two similar selections (Segment) are made on the best opinion shot (DigitalResource). One very short for the news report, another much larger for the promotion website and the DVD bonus content. In this case, the *hasRealisation* relation between EditorialObject and Segment allows the two selections to be related to the same editorial object. As a consequence, each content sequences is easier to find from the other and both benefit from the annotation of the editorial object.

After the montage of the spectator's comments, the editor of the DVD bonus content wants to reuse the montage of the best comments. In Fig.7, we depict our modeling of this example. The bonus content is modeled as another document (MediaAsset) which reuses some existing comments (MediaAsset). Here, the repurposing is made without changes in the editing of the reused documents. The *composedOf* relation between two MediaAssets represents the hierarchical integration of one in the other while the montage is modeled by a Segmentation. In this case, the bonus montage (Segmentation) is made by an ordered assemblage of the comment's montage (Segmentation) through the *assembles* relation.

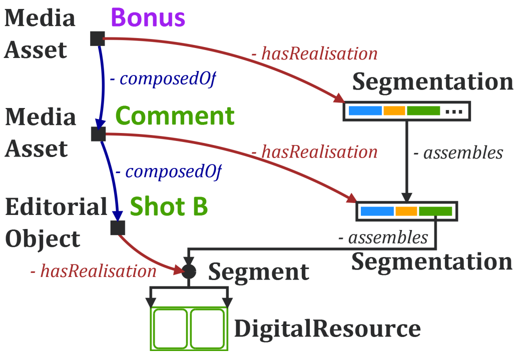


Figure 7. Repurposing of public comments in DVD bonus content

IV. Related Works

Despite a full conversion of the prominent MPEG-7 standard into a OWL Full ontology achieved by [5], the syntactic and semantic ambiguity of MPEG-7 demonstrated by [6] remains a real concern for data integration. The COMM ontology has clarified how formal semantics could be added to the MPEG-7 using patterns from the DOLCE foundational ontology [2]. COMM also contributed to highlight the need for separation and interrelation between low-level signal features, content sequence and annotation. Compared to COMM, our model defines an additional representation level to cope with editorial composition. With this additional level, our model enables to manage video material, content sequence and editorial composition indenpendently.

V. Applications

The audiovisual object model presented in the previous section has been developped in the course of the MediaMap project [7]. MediaMap is a Celtic project which aims at innovating in the area of audiovisual content production, in particular in the niche of UGC. In this project, we have formalized our model into an OWL-DL ontology so it is used in the project's applications: a shooting assistant for UGC production developped by SkemA and a search interface dedicated to audiovisual professionals developped by Exalead. SkemA is specialized in the developpment of Web and Mobile video platforms and Exalead a solution provider for entreprise and web search. Both applications use our ontology for representing an actual situation similar to the use case described in section II. That situation concerned the Tannhaüser opera, composed by Richard Wagner and directed by Jan Fabre.

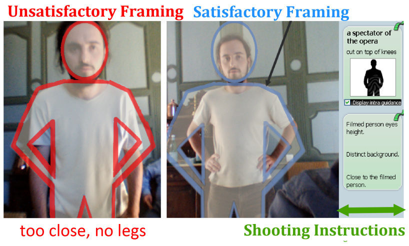


Figure 8. Guidance during shooting provides a summary of the instructions and a real-time framing indicator

A. UGC Shooting Assistant

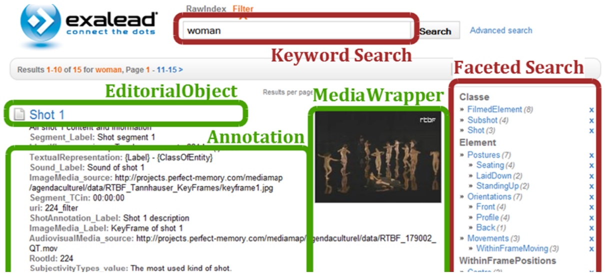
The ”CameraMate” shooting assistant is a software embedded into an ad-hoc prototype camera developped by Vitec Multimedia. Vitec Multimedia is a hardware provider specialized in digital video equipments. It is designed with a form-like interface to manage the shooting workflow for the amateur cameraman. Guidance are given before and during shooting according to predefined editorial recommendations defined using an audiovisual scripting vocabulary. This high-level vocabulary is related to low-level signal analysis algorithms which provide real-time indicators of the relevancy of the shooting according to the recommendations – see Fig.8.

The output of the CameraMate is conform to the examples detailed in the previous section. For instance, each shot is modeled as an EditorialObject with the shooting recommendations as prescriptive Annotations. The video material captured by the camera is a MediaWrapper. Finally, when the mission is done the cameraman can send the result to the professional producer which ordered the shooting.

B. Audiovisual Search for Producer

Once the UGC is retrieved by the professional producers it is stored in a semantic repository developped by Memnon. Memnon is a service and solution provider for media digitization and archiving. Exalead is in charge of indexing the audiovisual objects and provides then a dedicated-version of its search solution, as shown Fig.9. Professional can use the interface to retrieve any kind of audiovisual objects described before and thus enables repurposing. The results are presented as attached to an editorial object (MediaAsset or EditorialObject). Annotations provide general and audiovisual descriptions of the editorial object. MediaWrappers provide video and encoding parameters. The interface has two distinct search methods:

- from keywords indexing full-text annotation of audiovisual objects, like general description.
- from existing attributes and their values (faceted search) which provide a mechanism to filter the result set. This feature offers progressive filtering possibilities enabling a conceptual navigation in the result set thanks to our Annotation model.



Results 1-10 of 15 for woman, Page 1 - 11-15 >

Shot 1

All shot 1 content and information

Segment_Label: Shot segment 1

TextualRepresentation: {Label} - {ClassOfEntity}
Sound_Label: Sound of shot 1
ImageMedia_source: http://projects.perfect-memory.com/mediamap/agendaculturel/data/RTBF_Tannhauser_KeyFrames/keyframe1.jpg
Segment_TCin: 00:00:00
uri: 224_filter
ShotAnnotation_Label: Shot 1 description
ImageMedia_Label: KeyFrame of shot 1
AudiovisualMedia_source: http://projects.perfect-memory.com/mediamap/agendaculturel/data/RTBF_179002_QT.mov
RootId: 224
SubjectivityTypes_value: The most used kind of shot.

Classe
- FilmedElement (8) x
- Subshot (4) x
- Shot (3) x

Element
- Postures (7) x
  - Seating (4) x
  - LaidDown (2) x
  - StandingUp (2) x
- Orientations (7) x
  - Front (4) x
  - Profile (4) x
  - Back (1) x
- Movements (3) x
  - WithinFrameMoving (3) x

WithinFramePositions

Figure 9. Search interface with keyword and faceted search on audiovisual objects

VI. Conclusion

In this article we have defined a conceptual model of the audiovisual document. Our work contributes to identify and clarify the definition of all the objects composing an audiovisual document. We present a use case involving professional and UGC productions to demonstrate its expressiveness. Finally, we present applications using our model to provide shooting assistance and enhanced audiovisual search. We are currently working on the evaluation of our model in the more general context of collaborative audiovisual production including web 2.0 technologies. For that purpose, we consider essential the use of an organizational memory. This can be seen as a platform with different services fostering the exchange of knowledge, information and audiovisual resources [8] inside a company.

References

[1] J. V. Ossenbruggen, F. Nack, and L. Hardman, “That obscure object of desire: multimedia metadata on the Web, Part-1,” *Multimedia, IEEE*, pp. 38–48, 2004.

[2] R. Arndt, R. Troncy, S. Staab, L. Hardman, and M. Vacura, “COMM: designing a well-founded multimedia ontology for the web,” *The Semantic Web*, pp. 30–43, 2007.

[3] P. Morizet-Mahoudeaux and B. Bachimont, “Indexing and Mining Audiovisual Data,” *Lecture Notes in Computer Science*, vol. 3430, no. 5, pp. 34–58, 2005.

[4] L. Hardman, v. Obrenovi, F. Nack, B. Kerhervé, and K. Piersol, “Canonical processes of semantically annotated media production,” *Multimedia Systems*, pp. 327–340, 2008.

[5] R. Garcia and O. Celma, “Semantic Integration and Retrieval of Multimedia Metadata,” in *SemAnnot 2005*, Ireland, 2005.

[6] F. Nack, J. V. Ossenbruggen, and L. Hardman, “That obscure object of desire: multimedia metadata on the Web, part 2,” *Multimedia, IEEE*, vol. 12, no. 1, pp. 54–63, Jan. 2005.

[7] “Mediamap project,” Web site: http://www.mediamapproject.org/ [Last accessed: February 2011].

[8] M.-H. Abel and A. Leblanc, “Knowledge Sharing via the E-MEMORAe2.0 Platform,” in *ICICKM 2009*, Montreal, Canada, 2009, pp. 10–19.

Indexing Support Vector Machines for Efficient top-k Classification

Giuseppe Amato, Paolo Bolettieri, Fabrizio Falchi, Fausto Rabitti, Pasquale Savino
ISTI-CNR
Pisa, Italy
Email: g.amato@isti.cnr.it, p.bolettieri@isti.cnr.it, f.falchi@isti.cnr.it, f.rabitti@isti.cnr.it, p.savino@isti.cnr.it

***Abstract*—This paper proposes an approach to efficiently execute approximate *top-k classification* (that is, identifying the best k elements of a class) using Support Vector Machines, in web-scale datasets, without significant loss of effectiveness. The novelty of the proposed approach, with respect to other approaches in literature, is that it allows speeding-up several classifiers, each one defined with different kernels and kernel parameters, by using one single index.**

***Keywords*-Image Classification; Support Vector Machines; Similarity Searching.**

I. Introduction

The classification problem is typically defined as follows. Given a set of classes $c_1, \ldots, c_n$, and an object o, the classification problem is to decide which classes o belongs to. In this paper, on the other hand, we address the classification problem from an *Information Retrieval* perspective. Let $c_1, \ldots, c_n$ be n classes, and DS a very large dataset (for instance the World Wide Web) of unclassified objects. Given a class c_i, we want to retrieve the k objects of DS having the highest likelihood to belong to c_i. In this case, the class c_i can be considered as a query, and the best k objects that belong to c_i as the answer to the query. We call *top-k classification* this formulation of the classification problem.

Support Vector Machines (SVMs) [1], are widely used to perform automatic supervised classification. The aim of this paper is to propose a strategy that, given a set of SVM classifiers defined for a set of classes $c_1, \ldots, c_n$, and a dataset DS, executes top-k classification very efficiently. More specifically, we do not address the problem of learning classifiers, for which several solutions already exist as mentioned in Section III. Rather, given an SVM classifier for any class c_i, and a dataset DS, our aim is to efficiently search in DS for the best k objects that belong to c_i.

As we will see, we propose an approximated approach. That is, our approach returns an imprecise result, compared to the result that would have been obtained by performing a sequential scan of the entire dataset DS and applying the available classifiers to every object. However, the experiments will show a small imprecision, compared to an improvement of efficiency of orders of magnitude.

The rest of the paper is organized as follows. First we discuss related work. Next we briefly introduce the SVM. In Sections IV and V, we present the top-k classification, while in Section VI, the index structure used in the experiments. Finally in Sections VII and VIII, we describe the settings of the experiments and the analysis of the results.

II. Related Work

Efficient top-k classification techniques were proposed in [2], [3], by leveraging on the property that instances in the feature space lie on a hypersphere. This approach is able to use one index for various classes obtained using Support Vector Machines and built using the same kernel. In [4], the authors propose a method for efficient top-k classification based on boosting. In [5], the authors propose an efficient method for retrieving the instances closest to the separating hyperplane (the most ambiguous instances) to support active relevance feedback.

The novelty of the proposed approach, with respect to other methods existing in literature, is that it allows supporting and speeding-up the use of several classifiers, each one defined with different kernels and/or kernel parameters, by using one single index in the *input space* [1] of the dataset.

III. Introduction to SVM

An SVM [1] builds classifiers by learning from a training set that is composed of both positive and negative examples.

In many cases, in order to be able to separate element that belong to the class from those that do not belong to the class, it is convenient to map vectors, representing elements, in an higher dimensional vector space using a mapping function $\Phi(\cdot)$. Omitting several theoretical details (see [1] for more information), the learning phase determines a vector ω such that the decision function

$$f(o) = < \omega, \Phi(o) > +b \tag{1}$$

is able to optimally classify most of the training set examples ($< \omega, \Phi(o) >$ is the dot product between vectors ω and $\Phi(o)$). When the decision function is positive it indicates that an object belongs to the class. A popular learning algorithm is the kernel-based version of the adatron algorithm.

The SVM literature often call *input space* the space where objects are defined, and *feature space* the space where

[1] the space in which objects are originally represented (see Section III).

objects are mapped by $\Phi(\cdot)$. We will use this terminology in the reminder of the paper.

SVM methods do not define the mapping function explicitly, but use the properties of the kernel functions to perform learning and classification. A kernel function K, defined as $K(o_i, o_j) =< \Phi(o_i)^T, \Phi(o_j) >$, computes the dot-product of o_i and o_j in the feature space. There are simple kernel functions that easily compute the dot-product of objects mapped in very high or even infinite dimensional spaces without even knowing the actual mapping functions.

It can be proven [1] that the kernel-based decision function defined above, can be also represented in the dual form

$$f(o) = \sum_{(o_i, y_i) \in T} y_i \alpha_i K(o, o_i) + b \qquad (2)$$

where o_i are the element of the training set and y_i is 1 or to -1 according to the fact that the training object o_i is a positive or a negative sample of the class to learn. In this formulation, the learning phase consists in finding the parameters α_i, which basically determine the contribution of each example o_i of the training set to the solution of the learning problem, rather than the vector ω. Most of the α_i, obtained in the training phase, will be equal to 0. So, in order to compute the decision function f, we only need to maintain the training objects for which the α_i are greater than 0. These objects are the *support vectors*.

IV. Top-k classification

Let f_c be a decision function defined according to Equation 2 for the class c. The value $f_c(o)$ indicates the degree of membership of the object o to the class c. Large positive values indicate high membership of o to c; large negative values indicate that o does not belong to c; values close to zero indicate uncertainty.

The top-k classification problem can also be formulated as follows. Given a decision function f_c and a dataset DS, retrieve the k objects $o_1, \ldots, o_k$ in DS for which $f_c(o_i), i = 1 \ldots k$, is larger than when applied to any other object in DS. More formally:

Definition: 1: Let DS be a dataset of objects, c a class, and f_c the decision function for c. We define

$$\begin{aligned} \textit{top–k}(DS, c) =&\{o_1, \ldots, o_k \in DS \mid \\ &\forall o \in (DS \backslash \{o_1, \ldots, o_k\}), \\ &f_c(o) \leq f_c(o_i), i = 1, \ldots, k\} \end{aligned}$$

A. Approximate top-k classification

Clearly, $top-k(DS, c)$ can be computed with a sequential scan of the whole dataset. However, this is very inefficient when DS is very large. Suppose we have a set of candidates $CS \subseteq DS$ for class c. Then, top-$k(CS, c)$ is an approximation of top-$k(DS, c)$. However, consider that if CS is chosen carefully, top-$k(CS, c)$ will not necessarily differ very much from top-$k(DS, c)$. For instance, if $CS =$ top-$k(DS, c)$, then top-$k(CS, c)$ =top-$k(DS, c)$. According to this, given CS, approximate top-k classification can be performed by applying the decision function to the objects of CS rather than all objects in DS. Provided that CS is much smaller that DS ($\#CS \ll \#DS^2$), this process will be much more efficient than exhaustively classifying all objects in DS.

In the next section, we will propose a strategy to obtain CS, by using techniques of nearest neighbors searching, in such a way that approximation will be highly accurate and CS is much smaller than DS.

V. top-k classification by means of Nearest Neighbors Search

The training set T_c for a class c consists of positive and negative examples. Let us denote as PT_c and NT_c respectively the positive and negative training objects ($T_c = PT_c \bigcup NT_c$). As discussed in Section III, the learning phase identifies the αs parameters for the decision function, and implicitly the support vectors (the training vectors whose α_i are strictly greater than 0). Let us denote as $PSV_c \subseteq PT_c$ and $NSV_c \subseteq NT_c$ respectively the positive and negative support vectors identified after the training of the classifier for a class c. The decision function given by Equation 2 can be rewritten as

$$f_c(o) = \sum_{p \in PSV_c} \alpha_p K(o, p) - \sum_{n \in NSV_c} \alpha_n K(o, n) + b$$

The formula above just uses the support vectors and disregards the elements of the training set whose αs are 0, since they do not provide any contribution to f_c.

From the definition of f_c given above, it is easy to see that the objects o of the dataset that are very similar, according to K, to several positive support vectors and are dissimilar to several negative support vectors, have higher chances to return an high value when f_c is applied to them. In fact, the kernel K can be seen as a similarity function. That is, K returns large values when the two compared objects are similar and small values when the two objects are not similar. This suggests a strategy to select from DS a subset of promising candidates for c. In short, we can first search the objects of the dataset that are closer, according to K to each positive support vector. Then, we apply the decision function f_c only to the selected candidates to find the best k matches.

More formally, the candidate set $CS_c \subseteq DS$ for class c can be obtained as

$$CS_c = \bigcup_{p \in PSV_c} NN_K(p, s, DS) \qquad (3)$$

where $NN_K(p, s, DS)$, is a nearest neighbors query, which returns the s objects of DS most similar to p,

[2] we use $\#$ to indicate the cardinality of a set

according to kernel K, for some s much smaller than the size of DS.

The selection of the k best objects matching the class c can now be obtained by applying f_c only to objects in CS_s, which is significantly smaller than DS.

VI. Using an index structure in the input space

Several scalable techniques can be found in literature to efficiently process nearest neighbors search queries [6], [7]. However, even if, leveraging on these techniques, Equation 3 can be computed very efficiently, there are two reasons why it is not practical to build an index using K directly (that is in the feature space):

1) Access methods for similarity search typically rely on the fact that the similarity (the kernel K in our case) can be expressed in terms of a distance (dissimilarity) function, which should satisfy some specific properties, like for instance the metric postulates. A distance function can always be derived from a kernel K. However, such function, in many cases is not suitable to be used to build an efficient index structure. In fact, given that the kernel K compares elements in an high dimensional feature space, the underlying distribution of distances might not be convenient and we might incur in the curse of dimensionality [8].
2) We would like to support several classifiers for the same dataset, each recognizing a different class. Different classifiers might require different kernels and different kernel parameters (that is, different similarity functions). If we succeed to find a suitable distance function for a certain kernel K and we create an index with this distance, we are bound to the specific kernel K and its kernel parameters (for instance different σs in the case of the RBF kernel), so we are bound to a specific classifier. To support several classifiers, we would need several indexes, each for a specific kernel and kernel parameters.

Next section shows how to solve the above problems by building *one single index in the input space* to serve various kernels, provided that they satisfy some conditions.

A. Kernels that allow using a single index in the input space

Instead of deriving a distance function from a kernel, in many cases it is possible define a kernel in terms of a convenient distance function as follows:

$$K(o_1, o_2) = g(d(o_1, o_2)) \qquad (4)$$

where $d : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$ is a distance function between objects of the input space $\mathcal{D}$.

Many widely used kernels can be defined using Equation 4. For instance, if d is the Euclidean distance (L_2), with $g(x) = e^{-\frac{x^2}{2\sigma^2}}$ we obtain the RBF kernel; with $g = -x^\beta$, for $\beta > 0$, we obtain the power kernel; when d is the L_1 distance, with $g(x) = e^{-\gamma x}$ we obtain the Laplacian kernel. Other examples exist.

Let $\overline{\mathbb{R}} = d(\mathcal{D}, \mathcal{D}), \overline{\mathbb{R}} \subseteq \mathbb{R}$, that is $\overline{\mathbb{R}}$ is the set of possible values that d can give for any arbitrary pair $o_1, o_2 \in \mathcal{D}$. If g is monotonously decreasing over $\overline{\mathbb{R}}$, then the k objects closest to o, with respect to d, are the k objects most similar to o in the feature space, with respect to K.

In other words, *given any kernel K defined according to Equation 4, with a monotonous decreasing g and the same distance d,we have that*[3] $NN_K(p, s, DS) = NN_d(p, s, DS)$.

This implies that the most similar objects to a support vector in the feature space induced by K, are exactly the objects closest to the same support vector, in the input space. Therefore, we can use just one single access method defined using d and built in the input space, to search for the nearest neighbors in the feature space induced by a large class of kernels defined in terms of d. This, as discussed in Section V, also gives us the possibility of identifying the subset of promising candidates just working in the input space, for the same class of kernels[4].

Note also that the gs, that produce the RBF, Laplacian, and power kernels, are all monotonously decreasing in $\mathbb{R}^+$. Therefore, in this case, any d, which always returns positive values, allows this technique to be used.

VII. Experiment settings

Tests of the proposed techniques were executed on a single 2.4GHz Core 2 Quad CPU, using the CoPhIR dataset [11]. CoPhIR consists of 106 millions images, taken from Flickr, described by MPEG-7 visual descriptors. In the tests we used the first set of one millions images taken from CoPhIR. The access method used to efficiently search for objects close to the support vectors, in the input space, is the MI-File [12] (Metric Inverted File). The MI-File is a disk maintained index, based on inverted files, which supports efficient and scalable approximate similarity search on data represented in metric spaces. To define the kernel for the support vector machine, according to Equation 4, we used $g(x) = e^{-\frac{x^2}{2\sigma^2}}$, and as d we used a combination of MPEG-7 distance functions [13]. We trained the support vector machine to recognize 5 different classes: churches, pyramids, seascapes, paintings, and temples. We used a standard kernel-based adatron with cross-validation, to learn

[3] Here we abuse with the notation, NN_K gives the most similar, while NN_d gives the closest ones.

[4] Note that kernels used with SVM must be positive definite [1] or conditionally positive definite [9]. Therefore when a kernel K is obtained from Equation 4, we must first prove this. However, in many common cases this is true. Consider that, [10], in Theorem 12, shows that when $g = e^{-tx}$, for all $t > 0$, K is positive definite iff d is negative definite and symmetric. Note that the Eucledian distance is negative definite and symmetric. In fact, given that the RBF Kernel is positive definite, then d^2, when d is the Eucledian distance, is negative definite and symmetric and, as consequence of Theorem 11 in [10], also d (the Eucledian distance) is negative definite and symmetric.

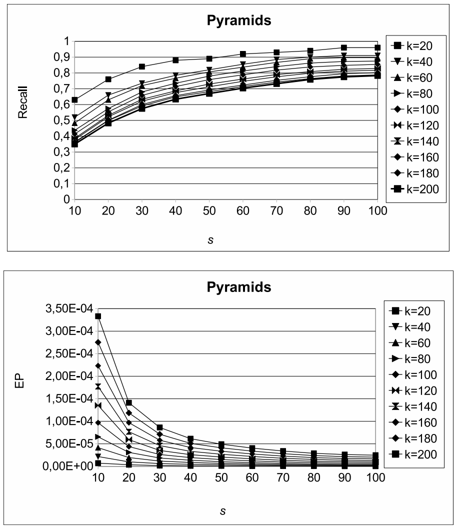


Figure 1. Quality of the of the approximate Top-k classification varying the number of nearest neighbors (s) retrieved for each support vector, according to Equation 3, for various values of k. Here, for brevity, we just show the results when the query is class Pyramids. Similar results were obtained for the other classes.

these classes. The set of candidates CS_c for a class c was obtained according to Equation 3. $NN_d(p, s, DS)$ searches were executed efficiently using the MI-File populated with all objects in the dataset according to distance d. The number s of nearest neighbors to each support vector ranged from 10 to 100, that is respectively 100.000 and 10.000 times smaller than the size of the dataset. Approximate Top-$k(DS, c)$ was executed computing Top-$k(CS_c, c)$ for various values of k ranging from 10 up to 200.

VIII. Analysis of the experiments

It is important to stress that this paper does not propose a new classification technique. Rather, given an SVM classifier built using standard tools, we propose a technique to perform top-k classification much faster than exhaustively classify all objects of a huge dataset. In this respect, given a classifier, our experiments aim at comparing the techniques that we propose, for efficient approximate top-k classification, against the exhaustive solution for top-k classification, which solve the top-k problem by sequentially and systematically classifying all objects of the dataset.

The evaluation of the quality of the top-k classification results consists of two parts: 1) an objective and quantitative evaluation of the error introduced by the use of the approximate classification and 2) a subjective evaluation based on real user feedback.

Both evaluations required to perform an *exhaustive classification* (sequential classification of the entire dataset), that was compared with the proposed approximate technique.

The objective evaluation was carried out by computing the measures of recall and error on the position [6]. More precisely, given a class c, the recall at k, R_k, is the percentage of the best k objects retrieved by the approximate method that also appear in the best k identified by the exhaustive classification as belonging to c.

The error in the position at k (EP_k) measures the quality of the ranking obtained by the approximate method with respect to the exhaustive one. It gives the average shifting of elements in the rank in percentage with respect to the size of the dataset.

More formally, recall at k is

$$R_k = \frac{\#(S_k \cap S_k^A)}{\#S_k} \tag{5}$$

and the error on position at k is

$$EP_k = \frac{\sum_{o \in S_k^A} |OX(o) - S_k^A(o)|}{\#S_k^A \cdot \#X}, \tag{6}$$

where S_k and S_k^A are the k best matches to c found respectively by the exhaustive classification of the entire dataset and by the our approximate method. OX is the ordering of the entire dataset X with respect to the decision function f_c for class c. For example, if o_1 is the most appropriate object that belongs to the class c, o_2 is the second and o_3 is the third, $OX(o_1) = 1$, $OX(o_2) = 2$ and $OX(o_3) = 3$. $S_k^A(o)$ is the position of o in the rank of k best matches found by the approximate classification.

The subjective evaluation, based on user feedbacks was performed by asking 5 students to blindly judge the results obtained with the exhaustive and approximate classification. To compare the two results we used the precision at k measure defined as follows:

$$P_k = \frac{\#(S_k \cap S^c)}{\#S_k} \tag{7}$$

where S_k is the result obtained by either the approximate or the exhaustive classification method, and S^c is the set of images correctly classified for c. $S_k \cap S^c$ was obtained by asking the users to select the correct results in S_k (blindly for exhaustive and approximate classification). Precision at k tells us the percentage of the k retrieved elements that belong to the class c according to the user judgement.

A. Approximate vs exhaustive classification

We first performed experiments to see how, according to Equation 3, the choice of the number s of retrieved nearest neighbor dataset objects to a support vector affects the quality of the approximation. Results are reported in Figure 1. We varied s from 10 to 100. For brevity, in the Figure we report results just for Pyramids. However, similar considerations can be made for the other classes. We can see that, in the chosen range of s, recall increases with s and saturates when s is around 90. On the other hand, the error

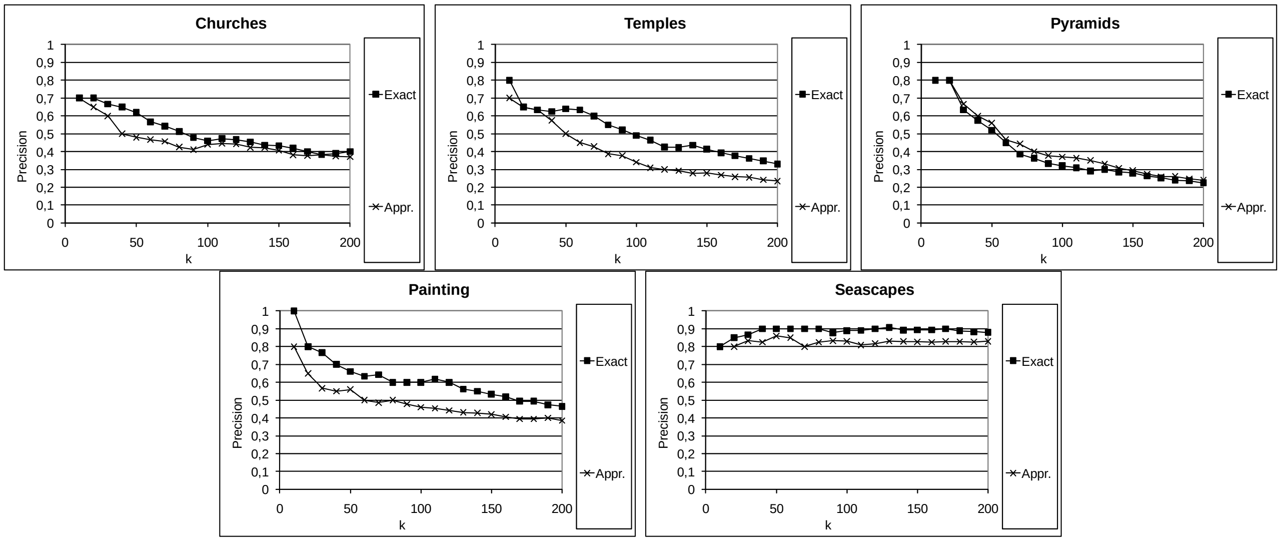


Figure 2. Precision of the approximate Top-k classification and the exact Top-k classification, for various values of k, as judged by users.

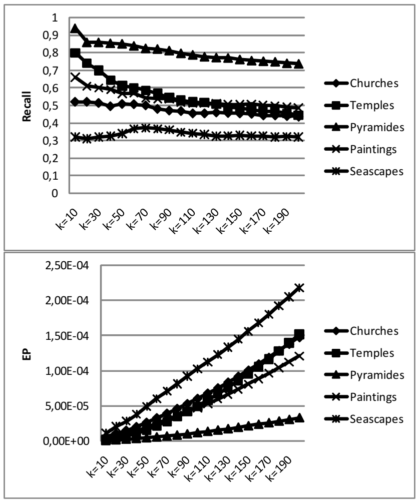


Figure 3. Recall and EP of the approximate top-k classification for several classes as queries, when s is fixed to 100, for various values of k, considering the exact top-k classification the ground truth.

on position decreases rapidly when s increases. Very small error values are already obtained when s is 60. This means that, recall increases and missed objects are also substituted by better objects when s increases.

Let us now discuss the quality of the approximate classification when s is fixed to 100, and we vary the number k of objects retrieved for a class. Results of these experiments are shown in Figure 3. For each class considered (i.e. churches, temples, pyramids, paintings, and seascapes) we plot the recall and the EP vs the number of best matches k to a class. We note that both recall and EP are influenced by k: worse results are typically obtained when k increases. However, we can see that the reduction of the recall is in many cases minimal with respect to the increase of k, while the increase of the error on the position is more evident. This means that on average the approximate classification strategy is always able to find the same percentage of correct objects (almost stable recall), even if missed correct objects are replaced by worse objects (worsening EP). For instance, let us consider the class *pyramid*. Recall is around 0.9 for $k = 10$ and it goes to 0.75 when $k = 200$. That is, 1 out of 10 images is missed when we retrieve 10 objects, while a bit less than 3 out of 10 images are missed when we retrieve 200. The error in position is almost 0 when $k = 10$ and it is also negligible until $k = 100$. Thus, the ordering is practically maintained in the approximate result. When k increases more, the quality of ranking degrades. For instance, in case of $k = 200$, the error in position is about 0.00003. That is with a dataset of 1 million objects the average shift was of 30 positions, with respect to the exact rank.

We should also consider that the time required to perform exhaustive classification of the entire dataset, for a given class was 39 minutes, on average. Good approximate classification of the same dataset can be obtained on average in 1.5 minutes,thus the approximate classification is *more than one order of magnitude faster than exhaustive classification.*

B. User evaluation

Results discussed above were obtained comparing approximate classification against exhaustive classification algorithms, using the exhaustive classification as a ground truth. However, generally even the exhaustive classification presents some imprecisions, which can be evaluated when users are called to judge the result, or by using real ground truths. As we will see in the following, surprisingly, users do not see much difference between exact and approximate results. This means that the degradation from exact to approximate classification is purely mathematical, and it is not significantly perceived by users.

The test discussed in this section evaluates the difference of quality between the exhaustive top-k execution and the approximate top-k execution, as perceived by real users. To obtain this, we asked 5 students to blindly judge the results, obtained by the exhaustive and approximate classification, by selecting the good and the wrong images. Based on this, we computed the average precision using Equation 7. Results are shown in Figure 2 for various choices of the number of best matches k.

It can be seen that generally there is no significant difference between the precision of the exhaustive and the approximate classification, even thought the approximate classification is much faster. In fact, precision measured for the exhaustive and approximate classification has practically the same trend when k varies. In addition, generally the difference in precision, between the exhaustive and the approximate classification, is smaller than 10%.

A separate discussion is needed for the Seascape classifier. In the experiments discussed in previous section, results for the Seascape class were worse than all the other classes (see Figure 3). In fact, recall was always below 0.4. On the other hand, the user perceived precision of the approximate classification is very high and always above 0.8. It can be seen, also, that the exhaustive classification has also a precision above 0.9 in most cases. When the approximate classification is used, missed images are always substituted by other images that are deemed to be still good by human evaluators, offering an high precision. Therefore, approximation makes sense also in this case.

It is also worth mentioning that in one case of the tested classes, the approximate classification performed even better than the exhaustive one. In fact, it can be seen that for the Pyramids class, the curve of the approximate classification is always higher than that of the exhaustive one. This, we believe, is a further proof that no significant information is actually lost during the approximation: the lost information is mainly noisy information.

IX. Conclusions

Science is becoming data-dominated. New data-intensive computing paradigms are emerging that differ from the traditional techniques, where Big Data was not a fundamental issue [14]. We have presented an approximate technique for efficiently executing top-k classification tasks on very large datasets. The proposed technique is some orders of magnitude faster with respect to exhaustive classification and the accuracy of approximate results is very high.

The peculiarity of the proposed technique is that it is able to use one single index built in the input space to support top-k classification tasks on several classes defined using various kernels and kernel parameters. We discussed the properties that the kernel has to satisfy so that they can be used with the proposed technique and we have seen that many widely used kernels are in fact included.

Acknowledgements

This work was partially supported by the VISITO Tuscany POR CREO FESR 2007-2013 project, funded by the Regione Toscana.

References

[1] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*, Cambridge University Press, March 2000.

[2] A. Qamra and E. Y. Chang, “Using pivots to index for support vector machine queries,” in *CVDB ’05*, New York, NY, USA, 2005, pp. 59–64, ACM.

[3] N-Panda and E. Y. Chang, “Exploiting geometry for support vector machine indexing,” in *Proceedings of SIAM International Data Mining Conference, SDM*, 2005, pp. 322–333.

[4] S. Litayem, A. Joly, and N. Boujemaa, “Interactive objects retrieval with efficient boosting,” in *Proceedings of ACM Multimedia*, 2009, pp. 545–548.

[5] M. Crucianu, D. Estevez, V. Oria, and J.P. Tarel, “Speeding up active relevance feedback with approximate knn retrieval for hyperplane queries,” *Int. J. Imaging Syst. Technol.*, vol. 18, no. 2-3, pp. 150–159, 2008.

[6] P. Zezula, G. Amato, V. Dohnal, and M. Batko, *Similarity Search - The Metric Space Approach*, vol. 32 of *Advances in Database Systems*, Springer, 2006.

[7] H. Samet, *Foundations of Multidimensional and Metric Data Structures*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2005.

[8] K. S. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, “When is ”nearest neighbor” meaningful?,” in *ICDT ’99, Proceedings*. 1999, vol. 1540 of *LNCS*, pp. 217–235, Springer.

[9] S. Boughorbel, J.P. Tarel, and N. Boujemaa, “Conditionally positive definite kernels for svm based image recognition,” in *IEEE ICME 2005*. July 2005, pp. 113–116, IEEE.

[10] C. Cortes, P. Haffner, and M. Mohri, “Rational kernels: Theory and algorithms,” *J. Mach. Learn. Res.*, vol. 5, pp. 1035–1062, 2004.

[11] P. Bolettieri, A. Esuli, F. Falchi, et al., “Enabling content-based image retrieval in very large digital libraries,” in *Second Workshop on VLDL*, 2009, pp. 43–50.

[12] G. Aamato and P. Savino, “Approximate similarity search in metric spaces using inverted files,” in *InfoScale ’08*. 2008, pp. 1–10, ICST.

[13] M. Batko, F. Falchi, C. Lucchese, et al., “Building a web-scale image similarity search system,” *Multimedia Tools and Applications*, vol. 47, no. 3, pp. 599–629, 2009.

[14] T. Hey, S. Tansley, and K. Tolle, Eds., *The Fourth Paradigm - Data Intensive Scientific Discovery*, Microsoft Res., 2009.

3D Object Retrieval and Pose Estimation for a Single-view Query Image in a Mobile Environment

Yoon-Sik Tak
Department of Electrical Engineering
Korea University
Seoul, Korea
Life993@korea.ac.kr

Eenjun Hwang
Department of Electrical Engineering
Korea University
Seoul, Korea
ehwang04@korea.ac.kr

***Abstract*—3D object retrieval and its pose estimation for a single view query image are essential operations in many specialized applications. With the recent deployment of various mobile devices, such operations require near real-time performance. However, most of the existing methods are not appropriate for mobile devices, due to their massive resource requirements. In this paper, we propose new 3D object retrieval and pose estimation schemes that can be used on a client-server platform. In order to accomplish this, we first construct both a sparse and a full index on the shape feature of the objects for the client and the server, respectively. Then, the client (the mobile device) retrieves the candidate camera view images that are similar to the query image by using the sparse index. The server refines the results by using the full index and then computes the exact pose by using the SIFT (Scale Invariant Feature Transform) features. In the experiment, we show that our prototype system based on the proposed scheme can achieve an excellent performance.**

Keywords- 3D object retrieval, pose estimation, shape-based retrieval, distance curve, SIFT.

I. INTRODUCTION

3D object retrieval and pose estimation are popular operations in various applications, such as robotic vision, medical image analysis, unmanned aerial vehicles (UAVs), and manufacturing automation. For instance, such operations can be used by robots to recognize diverse objects and change them to some specific pose for further actions such as assembly [1].

Depending on the hardware requirements, existing studies on the problem can be classified into three groups: The first group uses specialized equipment, such as the CMU high speed VLSI range sensor found in [2] and the CCD camera and laser scanner found in [3]. The second group uses multiple cameras. For example, in [4], the pose estimation was done using a pair of ground and onboard cameras for an autonomous unmanned aerial vehicle. In [5], a linear algorithm for computing the 3D points and the camera positions from multiple perspective views was proposed. The third group uses a single camera. For instance, in [6], a fully automatic solution using the Contracting Curve Density algorithm with speedup factors and SIFT features was proposed for a 3D object pose tracking. In [7], a pose tracking scheme based on the SIFT features and the Ferns was proposed for the classification of objects on mobile phones. The SIFT and Ferns were simplified for mobile phones at the cost of accuracy, due to their resource requirements. In our previous works [8][9], we introduced a simple object type classification scheme based on the shape symmetry and presented a time-consuming but accurate client-server collaboration scheme for 3D object retrieval in a mobile environment.

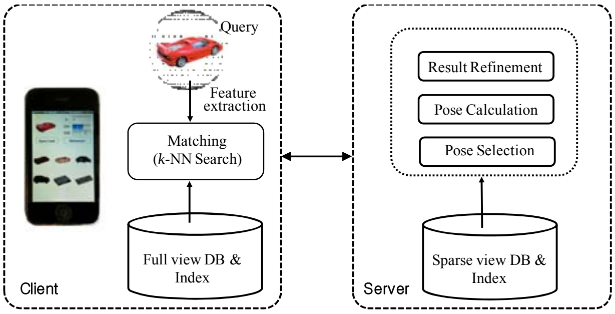


Figure 1. An illustration of system flow

In this study, we ameliorate our previous work in the following directions: (i) We present a new object type classification scheme, which can determine the type of an object automatically using the shape difference curve. ii) We present another client-server collaboration scheme which takes an heuristic approach to determine the object pose faster. We compare the performance of those two collaboration schemes through extensive experiments

In order to achieve good object recognition, we construct two indexes with different granularities based on the shape of the objects: The sparse index is constructed for the client in order to perform an approximate matching using large angle view images. Therefore its size is small compared to a full index. Conversely, the full index is constructed for the server using the small angle view images; these can be used for a more accurate matching. We propose two different client-server collaboration schemes in order to achieve load balancing. Basically, the client performs an approximate matching using the sparse index. The server refines it by using the full index. Since different view images of an object could give the same shape, we use their SIFT features for an accurate pose estimation. Figure 1 shows the overall architecture of our scheme.

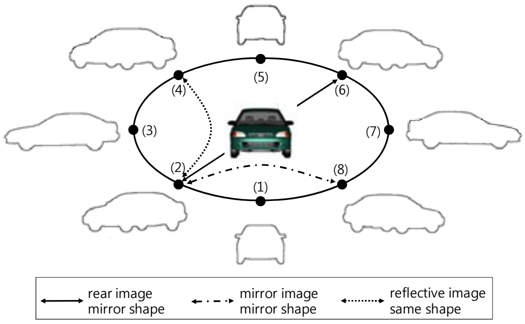


Figure 2. An illustration of shape patterns

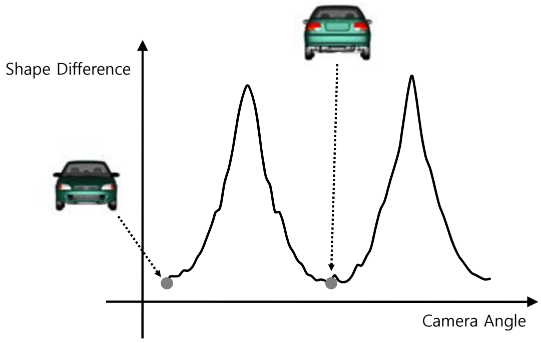


Figure 3. The shape difference curve of a car object

II. The View Index Organization

A. The Shape Represenstation

One straightforward method for shape based 3D object recognition is to consider all of the distinct camera views of the objects. For each camera view image, we first extract its shape contour and then calculate its distance curve by connecting the distances between the center point and all of the points along the shape contour. Considering the distance curve as sequence data, we can use well-known sequence matching techniques for retrieval purposes. In addition, we can construct a multidimensional index based on their DFT (Discrete Fourier Transform) values.

B. Camera View Skimming

Most real-world objects have bisymmetry in the front and/or on the side. Depending on the object symmetry, different camera view images of an object can have a same or a mirrored shape. Formally, for an image at an arbitrary angle, we can define three related images according to the camera viewpoint, as seen in Figure 2: the rear image, the mirror image, and the reflective image. Based on these images, the following properties are exhibited depending on the object symmetry: (i) for a bisymmetrical object, the mirror image has the mirror shape of the object; (ii) for any object, the rear image also has the mirror shape of the object; and (iii) the reflective image has the same shape of the given image. Based on these facts, we can remove the redundant camera view images that have the same or mirrored shapes from the index without sacrificing any matching accuracy. This facilitates the reduction of the index size and improves the matching speed. More specifically, our camera view skimming scheme consists of two parts: i) Mirror image pairing, and ii) Camera view pruning.

Mirror image pairing: For any type of object, an image and its rear image have mirrored shapes. Distance curves of mirrored shapes are simply the reverse of each other, and their DFT values are the same. Therefore, we can pair these mirror shaped view images via a set of DFT values. By pairing the mirror-shaped views during indexing and restoring the reversed curve during matching, the index size can be reduced by 50%.

Camera view pruning: Since all of the viewpoints of 3D objects can be generated by a combination of horizontal and vertical camera movements, we can consider the object symmetry in two planes: the horizontal and the vertical. Depending on the front and side symmetries, we can define four object types per plane. For instance, the horizon plane has four object types: H1- H4:

H1: Represents objects that have the same shape with respect to all horizontal camera views (e.g., a sphere).

H2: Represents objects that have front symmetry. Their distance curves repeat every 90 degrees (e.g., cars).

H3: Represents objects that have front symmetry and the same front and side views. Their distance curve repeats more frequently than that of H2. (e.g., dice).

H4: Represents objects that have no repeating shape pattern. We can define the vertical object types in the same manner.

We define vertical object types in the same manner, except that we consider the front and top views of the objects in the vertical plane. By combining the horizontal and vertical types, we can define 16 different object types.

Object type classification: Depending on the object type, the index entries for the redundant camera view images can be removed from the index without sacrificing any matching accuracy. Even though most real-world objects have certain level of symmetry, it is not easy to determine the exact object type automatically. In order to perform this efficiently, we developed a new object type classification method based on the shape difference curve. Informally speaking, the shape difference curve indicates how similar each camera view image is to the base view image. The shape difference of two different view objects can be defined by the Euclidean distance of their distance curves. For any symmetric object, its base view represents the camera view where the object is exactly bisymmetric. Based on the base view image, we can calculate its shape difference curve using all the camera view images. Depending on the object type, its shape difference curve has different repeating patterns. By analyzing these repeating patterns, we can determine the object type. For instance, Figure 3 shows the shape difference curve of a sample car object. Since the car has bisymmetry, the repeating pattern appears twice in the curve. The detailed

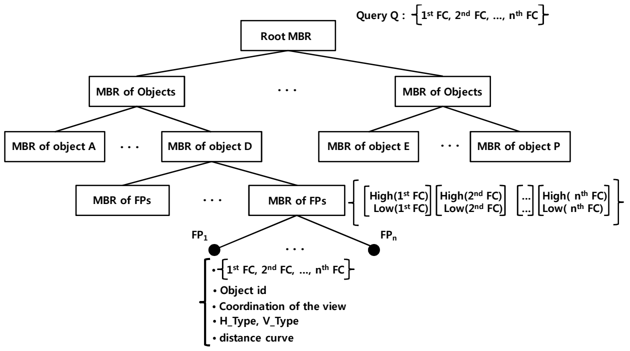


Figure 4. The overall index structure

steps for automatic object type classification are found in Algorithm 1. By considering both the horizontal & the vertical view sequences, we can automatically define object types.

Algorithm 1: *Object Classification (image sequence IS)*

```
Variable queue : Queue;
Period P = {0, 0} // {a, b} denotes view image period from a° to b°

find a base view BV from the images in IS
if there is no BV, exit the algorithm // Type 4.
else
  calculate shape distance distances of BV image for IS
  insert every image I whose shape distance is less than
    some threshold σ into the queue
  repeat until queue is empty do
    retrieve one from the queue into next
    set EndP as ⌜next.position/2⌝
    initialize P to {0, EndP} and D to 0
    for i is 2 to (180/ EndP)
      for j is 0 to EndP
        if i is odd
          D += shape distance of $I_j$ and $I_{(i-1)* EndP+j}$
        else
          D += shape distance of $I_j$ and $I_{i*EndP-j}$
    if D ≤ σ
      return P
```

C. The Index Construction

Even though the number of camera view images that need to be indexed is considerably reduced via our camera view skimming method, we still have to consider a large number of view images of the 3D objects. Therefore, an effective index structure is essential to achieve fast searching. For this purpose, we have constructed a R-tree based indexing structure based on the set of DFT coefficients obtained from the distance curves. The detailed steps for index construction are: (1) For the distance curve of each view image, we calculate a set of Fourier Coefficients (FCs) and define its Fourier Point (FP) which includes the *object id*, the *coordinate*, the *H_type*, the *V_type*, and the distance curve of the origin view. (2) For each object, we construct a subtree which contains all of the camera views of the object done by grouping the adjacent FPs using a minimum bounding rectangle (MBR). For each MBR, the lower and upper endpoints of the FCs are defined by its lower and upper bounds. (3) All of the MBRs representing the subtree in (2) are grouped again into larger MBRs until all of the MBRs of the objects are contained in the Root MBR. Figure 4 illustrates the overall structure of our index structure.

III. MATCHING

In this section, we describe how the server and client work together to retrieve similar 3D objects and estimate their exact poses using a single view image. We consider two different collaboration schemes: CS1 and CS2. The former guarantees no false dismissal by considering all of the camera view images at the server. The latter uses some heuristics to speed up the matching process at the cost of accuracy.

A. The K-NN Search in the Mobile Client

Since we represent the shape of the objects by using the distance curve, any of the existing matching frameworks proposed for the sequence data can be used. In this paper, we revise the priority queue based k-NN search algorithm [10] to find the k most similar to the objects based on our index structure. In order to prevent an unnecessary and time-consuming matching process, we hierarchically used several low bound functions, such as *Fourier_Dist*, *MINDIST* and *LB_Keogh* [10]. The revised k-NN search algorithm is described in Algorithm 2.

Algorithm 2: *k-NN Search (Q, k)*

```
Variable queue : MinPriorityQueue;
queue.push(root);
result = {};
while not queue.IsEmpty() do
  top = queue.Pop();
  if top.id is in the result
      continue;
  else
      if top is a sequence with DTW Dist.
          add top to result;
          if | result | = k
              return result;
      else if top is a leaf node
          for each Fourier Point P in top do
              queue.push(P, Fourier_Dist(Q,P));
      else if top is a Fourier Point P
          retrieve its full sequence S from DB;
          queue.push(S, LB_Keogh(Q,S));
          calculate reverse sequence S' of S //mirror shape
          queue.push(S', LB_Keogh(Q,S'));
      else if top is a sequence S with LB_Keogh Dist
          queue.push(S, DTW(Q, S));
      else
          for each child node C in top
              queue.push(C, MINDIST(Q,C));
```

B. Collaboration Scheme I (CS1)

After retrieving *k* similar objects for the given query Q, the client sends the result R to the server for refinement. The server refines it by using its full view index. In order to speed up the refinement, we use the maximum distance of R as an upper bound for the search. In order to guarantee no false dismissals, the CS1 considers all of the views of the objects during the refinement. Figure 5 shows the flow of the CS1; a brief sketch for the CS1 is shown in Algorithm 3.

Algorithm 3: *CS1 (Q, R)*

```
Variable queue : MinPriorityQueue;
1. **retrieve** MBR of Object Os whose distance from Q ≤ maximum distance of R.
2. **for** each O
3.     **if** O.id is in R
4.         **set** the UB of O as the dist. in R.
5.     **else**
6.         **set** the UB of O as maximum distance in R.
7.     **insert** O into the minimum priority queue.
8. k-NN Search_CS1(Q, k)
```

The k-NN Search_CS1(Q, k) at line 6 is a modified version of Algorithm 2. The difference is that we don't need to push the root of the index (line 1) into the queue and the following code segments need to be inserted before every push operation.

```
1. **if** top.UB ≤ D(Q, top)
2.     **continue**;
3. **set** top to one of the nodes {P, S, S', C}
4. **set** D() one of the distances {Fourier_Dist(),LB_Keogh(), DTW(), MINDIST()}
```

C. Collaboration Scheme II (CS2)

In some applications, a quick response time could be more important than the guarantee of no false dismissal. Moreover, with the huge database of 3D objects, supporting a fast search can be prioritized at the cost of accuracy. CS2 speeds up the matching process by using a heuristic algorithm at the cost of matching accuracy. This scheme is based on the assumption that if, for two camera views V1 and V2 of an object and query Q, if *angle*(V1, Q) < *angle*(V2, Q), then *dist* (V1, Q) < *dist*(V2, Q). Even though there could be some exceptions, this assumption is still valid in most cases. Based on this assumption, CS2 can refine R very quickly. Figure 5 shows the CS2 flow; the major steps are shown in Algorithm 4. Before adding the top into the result in line 5 in Algorithm 2, Algorithm 4 is called in order to refine the results of the mobile device.

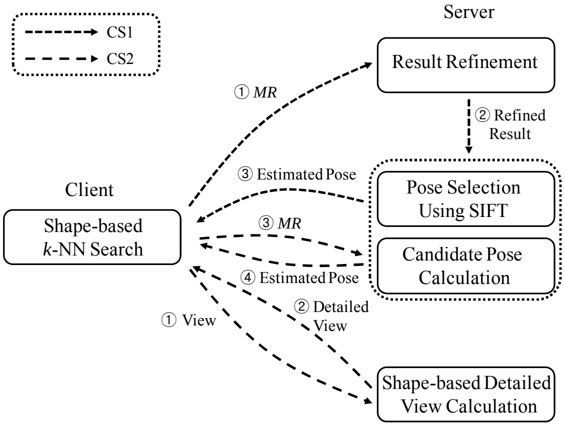


Figure 5. An overall flow for collaboration schemes

Algorithm 4: *CS2 (Q, image I, client view gap Gc)*

```
Variable queue : MinPriorityQueue;
1. **set** x as x coordination of I, y as y coordination, p as Gc.
2. **extract** 8 neighbor views N of I with the combination of $V_{x±p, y±p}$
3. **insert** every N into the queue with LB_Keogh dist. and p
4. **repeat** until |Result| = k do
5.   **if** top contains DTW distance
6.     **if** server view gap Gs ≤ top.p, set top.p to top.p /2.
7.       **extract** neighbor views N of top.p
8.       **calculate** LB_Keogh(N, Q)
9.       **insert** N into the queue with the distance and p.
10.    **else**
11.      **insert** top into Result
12.  **else**
13.    queue.push(top, DTW(Q, top));
```

D. The Candidate Pose Extraction

Since different views of an object might give the same shapes, we have to consider all of the same shaped views to give an accurate pose estimation. However, we have removed the redundant view images with the same shape obtained during the index construction. Therefore, at the pose estimation stage, we need to retrieve these images from the database or generate them dynamically from the 3D object using software tools such as CAD. Equations (1) and (2) explain the way to calculate the coordination of candidate pose views when the coordination of the base view is (*k*, *l*). HR_{period} and VR_{period} denote the period of horizontal and vertical shape pattern, respectively. For instance, HR_{period} of typical cars is 90. By combining all of the x and y coordination, we can get all of the candidate poses.

$$x = \begin{cases} HR_{period} * (i-1) + k & \text{if } i \text{ is odd} \\ HR_{period} * i - k & \text{if } i \text{ is even} \end{cases} \quad (1)$$

$$y = \begin{cases} \mathrm{VR}_{\mathrm{period}} * (j-1) + l & \text{if } i \text{ is odd} \\ \mathrm{VR}_{\mathrm{period}} * j - l & \text{if } i \text{ is even} \end{cases} \quad (2)$$

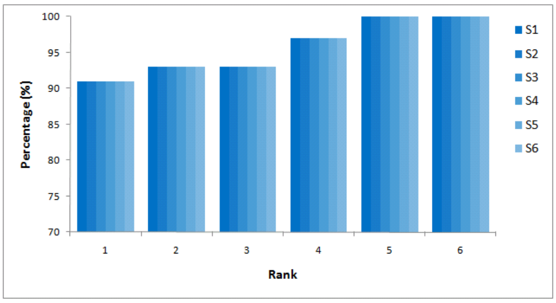


Figure 6. An accuracy comparison

where, $1 \le i \le 180/HR_{period}$ and $1 \le j \le 180/VR_{period}$

E. *The Pose Selection using SIFT Features*

Among the candidate poses, the best matching pose can be determined based on the actual visual features. This can be done by using a well-known image matching method such as SIFT [11] or SURF[12].

SURF is known to take relatively shorter time in matching than SIFT. On the other hand, SIFT shows better accuracy than SURF. In this work, we just need to consider a small number of images for pose estimation. Hence, we use the standard SIFT [11] method for matching for better accuracy even though it will take slightly longer time than the SURF method.

IV. THE EXPERIMENTS

A. *The Systems and Datasets*

In order to evaluate the performance of our proposed scheme, we implemented a prototype system. The server was equipped with an Intel Core2Duo CPU with 4 GB of RAM. iPhone 3Gs was used as the mobile client. Most of the applications at the client and server were implemented using C#. For the dataset in the experiments, we generated 259,200 views from 200 objects collected via the Internet [8]. The dataset contains diverse types of objects such as vehicles, kitchen appliances and furniture, to name a few.

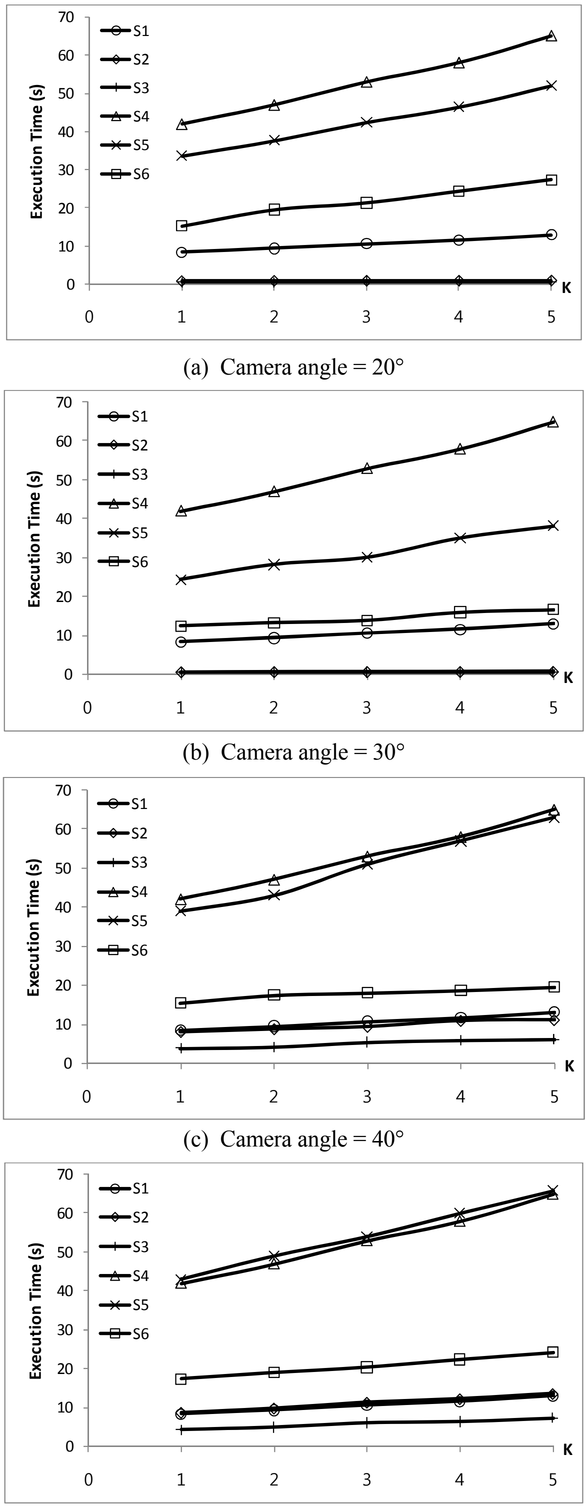
For the comparison, we considered six different system configurations that depended on the platform and the use of view skimming, as shown in Table 1.

Table 1 The System Configuration

Type	Description
S1	Server alone with camera view skimming
S2	CS1 with camera view skimming
S3	CS2 with camera view skimming
S4	Server alone without camera view skimming
S5	CS1 without camera view skimming
S6	CS2 without camera view skimming

B. *The Accuracy Comparison*

In this experiment, we show that our camera view skimming scheme does not impair the retrieval accuracy under any platform. The query input was a randomly selected view image stored in a database as a 3D model. Figure 6 shows the cumulative match curves (CMC) of the six different system configurations. For the test, we



(a) Camera angle = 20°

(b) Camera angle = 30°

(c) Camera angle = 40°

(d) Camera angle = 50°

Figure 7. The effect of camera angle on execution time

constructed indexes for 10° view images at the server and 30° view images at the client. From the graph, we observe the following facts:

1) For any platform, the camera view skimming does not have any effect on the accuracy.
2) Regardless of the view skimming, the three different platforms show the same accuracy. Theoretically, the server-alone and the CS1 guarantees the same accuracy. However, unlike CS1, CS2 cannot guarantee the same accuracy because CS2 refines the results using a heuristic approach. Therefore, CS2 shows a lower accuracy than CS1.

C. The Execution Time Comparison

In this experiment, we compare the total execution time, which includes the approximate estimation at the client and the result refinement at the server. In order to see the effect of the camera view gap size on the execution time, we considered four different camera angles for the client ranging from 20° to 50°, inclusively. In any case, the server used 10° of the camera view gap for the index construction. Since our scheme basically searches for similar objects based on the K -NN search, we measured the execution time by varying the size of the K as 1 to 5. Figure 7 shows the results. From the figure, we can observe the following facts:

1) Our camera view skimming scheme dramatically reduced the execution time.
2) A wider camera angle for view images with CS1 at the client helped to reduce the execution time since the wider camera view angle results in a smaller index at the client. However, an excessive camera angle gap can increase the execution time due to the overhead at the server for the refinement of the client result.
3) CS2 could reduce the execution time compared to CS1. From the experiment, we observe that setting the view extraction gap at the mobile client at 30° achieves a minimal searching time.

V. Conclusion

In this paper, we proposed a new shape-based client-server collaboration scheme for 3D object retrieval and pose estimation in a mobile environment. In particular, we proposed a camera view skimming scheme that reduces the index size and improves the search time using the bisymmetric property of most objects. For the pose estimation, we used the SIFT method to compare the same-shaped view images. Via various experiments on the prototype system, we demonstrated the effectiveness of our scheme. In addition, we proposed two collaboration schemes and compared their performance. Conclusively, larger camera angles used for the index at the client can reduce the index size and improve the search time. However, excessive camera angles might increase the search time at the server.

Acknowledgement

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(2010-0025395)

References

[1] A. Collet and S.S Srinivasa, "Efficient multi-view object recognition and full pose estimation," IEEE Conf. on Robotics and Automation, pp.2050-2055, 2010.

[2] D.A. Simon, M. Hebert and T Kanade, "Real-time 3-D Pose Estimation Using a High-Speed Range Sensor," IEEE Conf. on Robotics and Automation, pp.2235-2241, 1994.

[3] L. Haoxiang, W. Ying and C.W. de Silva, "Mobile Robot Localization and Object Pose Estimation Using Optical Encoder, Vision and Laser Sensors," IEEE Conf. on Automation and Logistics, pp.617-622, 2008.

[4] E. Altug and C. Taylor, "Vision-based pose estimation and control of a model helicopter," IEEE Conf. on Mechatronics, pp.316 - 321, 2004.

[5] C. Chen, D. Schonfeld and M. Mohamed, "Robust Pose Estimation Based on Sylverster’s Equation: Single and Multiple Collaborative Cameras," IEEE Conf. on Acoustics, Speech and Signal Processing, 1085-1088, 2008.

[6] G. Panin and A. Knoll, "Fully Automatic Real-Time 3D Object Tracking using Active Contour and Appearance Models," Journal of Multimedia, vol. 1 no. 7, 2006.

[7] D. Wagner, G. Reitmayr, A. Mulloni, T. Drummond and D. Schmalstieg, "Pose Tracking from Natural Features on Mobile Phones," IEEE/ACM Symposium on Mixed and Augmented Reality, pp.125-134, 2008.

[8] H. Kim, Y. Tak and E. Hwang, "Shape-based indexing scheme for camera view invariant 3-D object retrieval," Multimedia Tools and Applications, Vol. 47, No. 1, pp.7-29, 2010.

[9] Y. Tak and E. Hwang, "Indexing and Matching Scheme for Recognizing 3D Objects from Single 2D Image," Internet and Multimedia Systems and Applications, 2009.

[10] E. Keogh and C. Ratanamahatana, "Exact indexing of dynamic time warping," Knowledge and Information Systems, Vol.7, pp. 358-386, 2005.

[11] D.G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," International Journal of Computer Vision, Vol. 60, no. 2, pp. 91 - 110, 2004.

[12] H. Bay, A. Ess, T. Tuytelaars, L. V. Gool, "SURF: Speeded Up Robust Features", Computer Vision and Image Understanding (CVIU), Vol. 110, No. 3, pp. 346- 359, 2008.

Content-based Image Retrieval System for Medical Domain Using Spatial Color and Texture Histograms

Cristian Gabriel Mihai, Alina Doringa, Liana Stanescu, Dumitru Dan Burdescu, Cosmin Stoica Spahiu
Faculty of Automation, Computers and Electronics
University of Craiova
Craiova, Romania
{mihai_gabriel, stanescu, burdescu, stoica.cosmin}@software.ucv.ro
alinadoringa@hotmail.com

***Abstract*—Indexing and retrieving are two important tasks for color image databases. These tasks are possible when using the histogram as an efficient technique for content-based image retrieval domain. In this paper, a content-based image retrieval system that can be used in the medical domain is presented. The system is using an original combination of two types of histograms: a spatial color histogram and a texture histogram based on the Local Binary Pattern descriptor. Both histograms are computed in the HVC color space. The computation of the binary code associated to a Local Binary Pattern descriptor is made in an original manner using the NBS distance instead of using a simple difference between colors' components.**

Keywords- spatial color histogram; annular color histogram; content based image retrieval.

I. INTRODUCTION

Color histograms have two main advantages: easy computation and a broad applicability for a wide range of images. The main drawback is that histograms capture only global color distributions of the images and there is a lack of information about the spatial relationship among image's colors. It is highly possible that two images with similar color histograms to have a very different spatial appearance causing false positives. Because the content of images is indexed in a limited way using only color histograms it was decided that it would be an advantage to take into account the spatial information [10]. This combination has led to effective techniques for content-based image retrieval tasks based on the new spatial color histogram. It was used for our system an efficient spatial color histogram called the annular histogram [10] which is based on a density map. The experimental results have proven that this histogram outperforms the traditional color histogram and the color coherent vector [11].

Because color measurements are sensitive to varying illumination conditions texture measures can be used in many real-world applications, including for example, outdoor scene image analysis. Texture characteristics gives additional information compared to color or shape measurements of the objects. It is considered to be important in many image analysis and computer vision tasks. We have used the local binary pattern (LBP) [13] to obtain a texture histogram of patterns. LBP is one of the most used texture descriptors in medical image analysis and it has recently proven useful in describing medical images [17][18][19]. It has a low computational complexity and a low sensitivity to changes in illumination.

This descriptor is used for texture classification [13][21], face recognition [20][22][24], fingerprint identification [23], etc.

The HVC (Hue-Value-Chroma) color space [5] has been used because it represents colors along human perceptual dimensions [6]. HVC is a representation of the IE 1976 (L*a*b*) under the cylinder coordinate system. The components of a color in the HVC space are defined as:

$$H = \arctan\left(\frac{b^*}{a^*}\right), V = L^* ,\ C = \sqrt{a^{*2} + b^{*2}}\ . \quad (1)$$

The distance between two pixels in this space is computed using the NBS distance [4]. Colors with the NBS color distance below 3.0 are perceived to be almost the same color by human beings. Given a pair of colors A = (H1; V1; C1) and B = (H2; V2; C2), the NBS color distance is defined as follows [4]:

$$E_{NBS}(A, B) = 1{,}2 * \sqrt{2C_1C_2\left\{1 - \cos\left(\frac{2\pi}{100}\Delta H\right)\right\} + (\Delta C)^2 + (4\Delta V)^2}\ . \quad (2)$$

where:

$$\Delta H = |H_1 - H_2|, \Delta V = |V_1 - V_2|, \Delta C = |C_1 - C_2|. \quad (3)$$

The correspondence between the human color perception and the NBS color distance is shown in the following table:

TABLE I. CORRESPONDENCE BETWEEN HUMAN COLOR PERCEPTION AND NBS COLOR DISTANCE

NBS Value	Human Perception of Color
0 ~ 1.5	almost the same
1.5 ~ 3.0	slightly different
3.0 ~ 6.0	remarkably different
6.0 ~ 12.0	very different
12 ~	different

The paper is organized as follows: related work is discussed in Section 2. In Section 3, some details about the spatial color histograms and local binary patterns are presented. Section 4 provides a description of the modules included in the system architecture. In Section 5, the experimental results are discussed. Section 6 presents the conclusion of the paper.

II. RELATED WORK

Many research efforts have been made in the last decades to overcome the problems associated with color histograms. For content-based image retrieval systems it is more important the result of an approximate matching of two histograms than exact matching of images. Approximate matching is more useful since the interest is to retrieve similar images, rather than images identical with the sample image.

The solution proposed by Stricker and Dimai [2] divided an image into five fixed overlapping blocks. From each block the first three color moments were extracted and these were used to form a features vector of the image. Huang, et al. [3] proposed the correlogram to take into account the local color spatial correlation as well as the global distribution of this spatial correlation. In [10][12], the spatial color histograms are described for the content based image retrieval task.

A typical example of histogram-based image retrieval system is the FINDIT system developed by M.J. Swain and his colleagues [1]. The HVC color space was adopted for this system and for each image it was created a two-dimensional H-C histogram. It was used the histogram intersection to measure the similarity between a pair of images. In [4], a histogram-based image retrieval method was implemented. This method is similar to the one used in FINDIT system. The H and C-axes of the HVC color space were used, and the two axes were equally subdivided into 8 intervals, resulting in histograms of 64 bins.

Another example of a texture-based image retrieval system is the UCSB system developed by Manjunath et al. [16]. The system adopted the Gabor wavelet model to compute feature vectors of texture patterns and used the weighted L1 distance between a pair of feature vectors to measure the image similarity.

Ojala et al. [13] proposed an effective local binary pattern (LBP) method for texture analysis. They have developed powerful extensions to their approach including rotation invariance and multi resolution analysis [14]. The approach is theoretically very simple and binds together the properties of statistical and structural texture analysis. LBP and its extensions have performed very well in various comparative studies and have been applied successfully in several real-world texture analysis problems [15]. In [25], the HVC color space is used for pixels classification. The described method is based on statistical characteristics and it is used for the recognition of the airline coupons. In [27], the LBP are used to describe images of brain magnetic resonance (MR) volumes. When a query image is given the system retrieves relevant slices. LBP are used in [28] for representing salient micro-patterns in mammographic mass detection and to train a support vector machine (SVM) with the aim of distinguishing between the true recognized masses and the ones which actually are normal parenchyma.

In [30], an indexing and retrieval scheme is presented that uses the spatial color distribution. The indexing technique is based on the Gaussian Mixture modeling of the histogram of weights provided by the bilateral filtering scheme. In this way the proposed technique considers not only the global distribution of the color pixels comprising the image but also takes into account their spatial arrangement. In [8], the spatial histograms are used for region based tracking. In this context, these histograms are named *spatiograms*, which are histograms augmented with spatial means and covariances to capture a richer description of the target. In [9], it is used a medical image retrieval system that is very similar with our approach. This system is based on multiple features: color features - exploited by cumulative histograms, texture features - extracted by using gray level co-occurrence matrix (GLCM) and shape features - represented by a histogram of edge’s directions. This system extract the primitive feature of a query image and then compares it with existing features of the images from the database using a similarity measure. This similarity measure is evaluated using Euclidean distance. The experiments are made on a set containing 1000 images, covering MRI images, X-Ray images, Patology images. Retrieval Accuracy and Precision are used as performance measures. For evaluating the retrieval operation this system is using four retrieval modes: a retrieval mode based on a color histogram, a retrieval mode based on GLCM, a retrieval mode based on shape and a retrieval mode based on a combination of a color histogram and GLCM. The experimental results have shown the following results for the four retrieval modes (in the same order as before): (% Recall Accuracy : 66.1 ; % Precision: 55.3), (% Recall Accuracy : 68.4 ; % Precision: 62.6), (% Recall Accuracy : 65.12 ; % Precision: 58.3), (% Recall Accuracy : 72.3 ; % Precision: 65.4). It can be seen that using a combination of color and texture it was obtained the best retrieval result.

III. SPATIAL COLOR HISTOGRAM AND LOCAL BINARY PATTERN

A spatial color histogram is based on two main concepts: the distribution density and the density map. The distribution density can have three types: annular, angular and hybrid. The density map is obtained after performing the following algorithm:

- Calculate the centroid and the radius of each subset of pixels having the same color - considered as a geometric subset of the 2-D plan.
- Partition the enveloping disk either in annular, angular or sector (combination of annular and angular) regions.
- Count the number of pixels in each region and form a vector called the density map of a color.
- Arrange the density maps of all colors in a matrix where the density map of a color represents a matrix row; the matrix obtained is called either annular,

Copyright (c) IARIA, 2011. ISBN: 978-1-61208-129-8

angular or hybrid depending on which partition was adopted.

Figure 1 presents an example of annular distribution density vector (4, 11, 9, 5) computed by counting the number of points (starting with the center region) in each region:

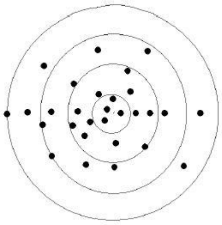
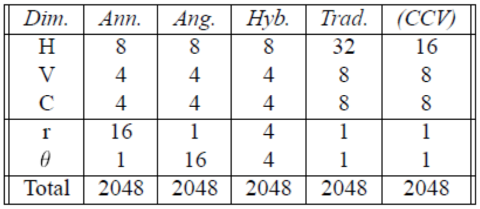


Figure 1. Annular distribution density.

In [10], five techniques for computing color histograms in the HVC color space were evaluated: the traditional histogram, color coherent vector (CCV) [11], the three types of histograms mentioned above: annular, angular, and hybrid. Features vectors were generated for each of the histograms mentioned above having the dimension equal to 2048. The color space was then quantized by making a uniform partition in each color dimension. Histogram types and the parameters used for quantification are listed in Figure 2. The experimental results have shown the following performance (in this order): annular, angular, hybrid, color coherent vector, and traditional histogram. The improvement of the spatial color histograms over the traditional one are: 36.49 % (annular), 30.41 % (angular), 26.71 % (hybrid). The improvement of the spatial color histograms over CCV are: 22.39 % (annular), 16.93 % (angular) and 13.61 % (hybrid).



Dim.	*Ann.*	*Ang.*	*Hyb.*	*Trad.*	*(CCV)*
H	8	8	8	32	16
V	4	4	4	8	8
C	4	4	4	8	8
r	16	1	4	1	1
θ	1	16	4	1	1
Total	2048	2048	2048	2048	2048

Figure 2. Histogram types and the parameters used for quantification where r, θ are polar coordinates.

LBP is one of the most used texture descriptors in medical image analysis being very useful in describing medical images. In Figure 3 it is shown how to calculate the LBP and the contrast for a pixel having 8 neighbor pixels.

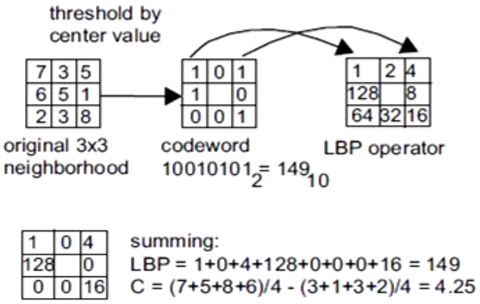


Figure 3. Computation of LBP and local contrast features.

A binary code is produced for each pixel in an image, by thresholding its neighborhood (8 pixels) with the value of the center pixel. The average of the gray levels below the value of the center pixel is subtracted from that of the gray levels above (or equal to) the center pixel. A histogram is constructed to collect up the occurrences of different binary patterns representing different types of curved edges, spots, flat areas, etc.

The original 8-bit version of the LBP operator considers only the eight nearest neighbors of each pixel. For this version there are 256 local patterns, 36 of them being rotation invariant.

The definition of the LBP has been extended to arbitrary circular neighborhoods of the pixel to achieve multi-scale analysis and rotation invariance. In Figure 4 it is presented the circularity idea behind the multi resolution approach.

The circular neighborhood definition allows obtaining a rotation invariant descriptor, but in some problems the anisotropic structural information is an important information source. To exploit this anisotropic structural information, an elliptical neighborhood definition has been used [28] for a face recognition system. This variant to the standard LBP has been named elliptical binary pattern (EBP).

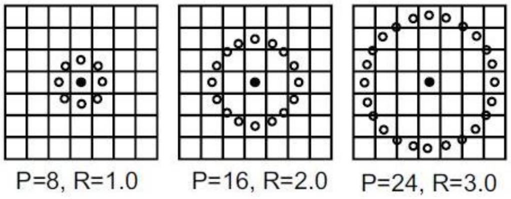


Figure 4. Circular neighborhood of pixel in multi resolution LBP.

Another variant has been proposed by [29] to solve the problem of the sensitivity to noise in near-uniform image regions. This method, called local ternary patterns (LTP), proposed a 3-valued coding that includes a threshold around zero for the evaluation of the local gray-scale difference.

The distance between two LBP histograms (histograms of patterns) can be evaluated using:

Chi square distance:

$$\chi^2(S,M) = \sum_{b=1}^{B} \frac{(S_b - M_b)^2}{S_b + M_b}. \qquad (4)$$

Histograms intersection:

$$H(S,M) = \sum_{b=1}^{B} \min(S_b, M_b). \qquad (5)$$

with a significantly smaller computational overhead.

Another extension to the original operator is the definition of so called uniform patterns. This extension was inspired by the fact that some binary patterns occur more commonly in texture images than others. A local binary pattern is called uniform if the binary pattern contains at most two bitwise transitions from 0 to 1 or vice versa when the bit pattern is traversed circularly. For example, the patterns 00000000 (0 transitions), 01110000 (2 transitions) and 11001111 (2 transitions) are uniform whereas the patterns 11001001 (4 transitions) and 01010010 (6 transitions) are not. For (8, R) neighborhood there are 58 uniform patters from the total of 256 patterns.

IV. THE SYSTEM ARCHITECTURE

The system architecture is presented in Figure 5 and contains four main modules:

- Annular histogram Module: is used to compute the annular histogram and the distribution density vector. As described above for this type of histogram the HSV space dimensions were split: H- 8 bins, V- 4 bins and C - 4 bins obtaining in this way a histogram of 8x4x4 = 128 bins. In the first step this module computes the histogram using the information extracted from the specified image. After this process completes for each histogram bin, it is computed the distribution density vector having a dimension of 16. These vectors are concatenated obtaining a single density vector having a dimension of 128×16 = 2048 that is stored in the database.

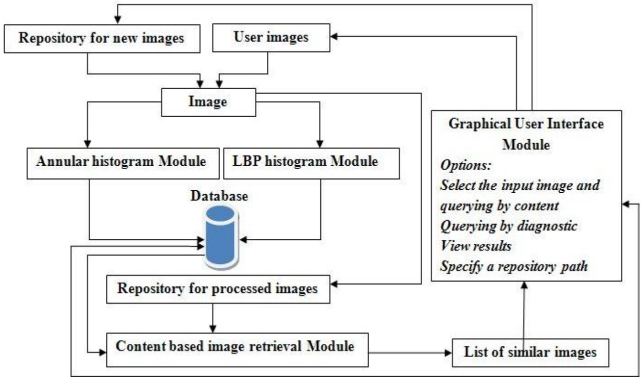

Repository for new images
User images
Image
Annular histogram Module
LBP histogram Module
Database
Graphical User Interface Module
Options:
Select the input image and querying by content
Querying by diagnostic
View results
Specify a repository path
Repository for processed images
Content based image retrieval Module
List of similar images

Figure 5. System architecture.

- LBP histogram Module – this module computes a histogram of rotational invariant local binary patterns. For our system we have used a histogram of patterns having 37 bins (36 bins for the rotation invariants and one bin for the rest of the patterns) and an original method for computing the pattern code: the NBS distance is calculated between the color components of the center pixel and the color components of a neighbor pixel; if the distance is greater that 3 (remarkably different colors) then we have 1, otherwise we have 0. In this way it is obtained the binary representation of the pattern and later the number associated with the pattern using a transformation from base 2 in base 10. The histogram of patterns is normalized and after that its content is stored in the database.
- Content based image retrieval Module: this module computes a distance D having two components :
  - D1 – the Euclidian distance between the density vector of the analyzed image and a density vector corresponding to an image already processed
  - D2 - a distance equal with 1 – HI, where *HI* is the histogram intersection between the histogram of patterns of the analyzed image and the histogram of patterns corresponding to an image already processed. The value of D is obtained as:

$$D = \sqrt{{D_1}^2 + {D_2}^2} \quad (6) \quad \text{where } D2 = 1 - HI \qquad (6)$$

- Graphical User Interface Module: is used by the user to retrieve the images that are similar with the specified input image, to specify a repository path, to retrieve the images from the database that have a diagnostic similar with the specified one – the diagnostic is specified as a text and after pressing a button a select statement is made on the database in tables *Images* and *Diagnostics* and all images having that diagnostic are returned; this option can be used to detect the total number of relevant images from the database when knowing the diagnostic of an input image.

For each input image it is returned a list of similar images having the value of the distance D smaller than a threshold value which is configurable.

V. EXPERIMENTAL RESULTS

The system has been tested using a set of 2000 images belonging to the digestive tract and they were obtained during the patients' diagnosis process. The training set contained 1800 images and the testing set 200 images. Each image from the training set has in the database information about the associated diagnostic. This information is used to detect the number of relevant images from the database having the same diagnostic. The performance and the efficiency of the information retrieval operation are measured with two parameters: recall and precision. The recall parameter measures the ability of the system to find relevant information in the database and it is defined as: the number of retrieved images that are also relevant / the total number of relevant images from the database. The precision parameter measures the accuracy of the retrieval operation and it is defined as: the number of retrieved images that are also relevant / the total number of retrieved images. For each image from the testing set it is calculated a (precision, recall) pair of values. The precision is computed easily by identifying the relevant images from the returned list and computing the value as described above. For recall we need also to detect the total number of relevant images existing in the database. This number is calculated by making the following assumption: the diagnostic associated with the input image is known. As the diagnostic is known, the option "querying the database by diagnostic" can be used. This option is in the Graphical User Interface module. In this way it is obtained the list of all images having the specified diagnostic. After this value is obtained the precision value is computed. At the end of this process 200 pairs (precision, recall) of values were obtained and used to calculate a mean precision and a mean recall. Similar with the approach described in [9] we have made an evaluation of three retrieval modes: one retrieval mode based on an annular histogram, one retrieval mode based on a LBP histogram and a retrieval mode based on a combination between an annular histogram and an LBP histogram. Using the approach described above we have obtained 3 pairs of mean values for precision and recall corresponding to the three retrieval modes (in the same order as above): (% Recall: 61.3; % Precision: 54.7), (% Recall : 64.5 ; % Precision: 59.6), (% Recall : 76.1 ; % Precision: 70.1). It can be seen that using a combination of an annular histogram and a LBP histogram it was obtained the best retrieval result.

Below, the results obtained using the following query image having the ulcer diagnostic are presented.

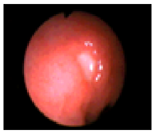


Figure 6. Query image.

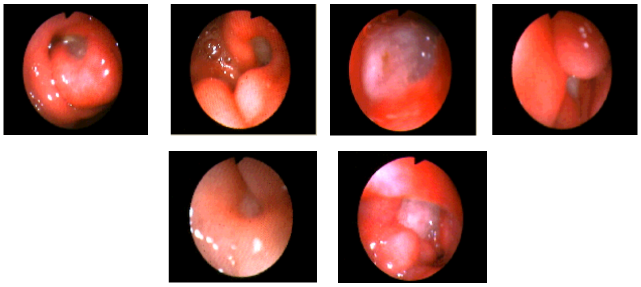


Figure 7. The similar images belonging to stomach and duodenum ulcers that were retrieved by the system.

VI. CONCLUSION

In this paper, a system for content based image retrieval that can be used in the medical domain was described.

An element of originality of this system is the usage of the combination of two histograms: annular histogram and LBP histogram for content based image retrieval.

The binary code associated to each LBP descriptor is computed using an original method based on NBS distance. This method is much better than computing simple difference between colors' components.

The limitations of the color histogram were improved by taking into account the spatial relationship between pixels.

The system was tested only on a limited dataset containing 2000 medical images, but in the future a larger dataset will be used and more experiments will be made. The experiments refer both for retrieval quality and speed.

Further extensions of the system will include shape information and the extraction of other texture-related features (Gabor and Tamura based). It should be also made comparative performance studies between this system and other existing systems (like the system described in [9]) and to include relevance feedback which is commonly used in image retrieval.

REFERENCES

[1] M. J. Swain, "Interactive indexing into image databases," in SPIE Proceedings, vol. 1908, 1993.

[2] M. Stricker and A. Dimai, "Color indexing with weak spatial constraints," in SPIE Proceedings, vol. 2670, 1996, pp. 29-40.

[3] J. Huang, S. Kumar, et al., "Image indexing using color correlograms", in Proceedings of CVPR'97, 1997, pp. 762-768.

[4] Y. Gong, "Advancing Content-based Image Retrieval By Exploiting Image Color and Region Features", Multimedia Systems, vol. 7, 1999, pp. 449-457

[5] The Japanese Society of Chromatology, The Handbook of Chromatology. University of Tokyo, 1980.

[6] M. Miyahara and Y. Yoshida, "Mathematical transform of (r,g,b) color data to munsell (h,v,c) color data," in SPIE Proceedings in Visual Communication and Image Processing, vol. 1001, 1988.

[7] M. J. Swain and D. H. Ballard. "Color indexing", International Journal of Computer Vision, vol. 7(1), 1991, pp.11–32.

[8] S.T. Birchfield and S. Rangarajan, "Spatial histograms for region-based tracking", ETRI Journal, vol.29(5), 2007, pp.697-699.

[9] B. Jyothi, Y. Madhavee Latha, and V.S.K. Reddy, "Medical Image Retrieval using Multiple Features", Advances in Computational Sciences and Technology, vol. 3 (3), 2010, pp. 387–396.

[10] A. Rao, R. K. Srihari, and Z. Zhang, "Spatial Color Histograms for Content-Based Image Retrieval", Proceedings of the Eleventh IEEE International Conference on Tools with Artificial Intelligence, 1999, pp.183-186

[11] G. Pass and R. Zabih., "Histogram refinement for content-based image retrieval", IEEE Workshop on Applications of Computer Vision, 1996, pp. 96–102.

[12] A. Rao, R.K. Srihari, and Z. Zhang, "Geometric Histogram: A Distribution of Geometric Configurations of Color Subsets", Internet Imaging, vol. 3964, 2000, pp. 91-101.

[13] T. Ojala, M. Pietikainen, and D. Harwood, "A comparative study of texture measures with classification based on feature distributions. Pattern Recognition", vol. 29(1), 1996, pp. 51-59.

[14] T. Ojala, M. Pietikainen, and T. Maenpaa T, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24(7) , 2002, pp. 971–987.

[15] T. Maaenpaa and M. Pietikainen, "Texture analysis with local binary patterns". Proceedigs Handbook of Pattern Recognition and Vision, World Scientific, 3rd Edition, 2005, pp. 197–216.

[16] B. S. Manjunath and W. Y. Ma, "Texture features for browsing and retrieval of image data", IEEE Transactions on PAMI Special Issue on Digital Libraries, 1996.

[17] L. Nanni, A. Lumini, and S. Brahnam, "Local binary patterns variants as texture descriptors for medical image analysis", Artificial Intelligence in Medicine, vol. 49(2), 2010, pp.117-125.

[18] D. Unay and A. Ekin, "Intensity versus texture for medical image search and retrieval", Proceedings of the 5th IEEE international symposium on biomedical imaging: from nano to macro (ISBI 2008), 2008, pp. 241–244.

[19] O. X. Lladó, J. Freixenet, and J. Martí, "False positive reduction in mammographic mass detection using local binary patterns", Proceedings of the medical image computing and computer-assisted intervention (MICCAI 2007*)* Brisbane, 2007, pp. 286–293.

[20] S. Marcel, Y. Rodriguez, and G. Heusch, "On the recent use of local binary patterns for face authentication", International Journal of Image and Video Processing – Special Issue on Facial Image Processing, 2007.

[21] M. Pietikäinen, T. Ojala, and & Silvén O, "Approaches to texture-based classification, segmentation and surface inspection"

C.H. Chen, L.F. Pau, and P.S.P. Wang, (eds.), Handbook of Pattern Recognition & Computer Vision, 2nd ed., World Scientific, Singapore, 1999, pp.711-736.

[22] L. Nanni and A. Lumini, “RegionBoost learning for 2D + 3D based face recognition”, Pattern Recognition Letters, vol. 28(15), 2007, pp.2063-2070.

[23] L. Nanni and A. Lumini, “Local binary patterns for a hybrid fingerprint matcher”, Pattern Recognition vol. 11, 2008, pp. 346-3466.

[24] G. Zhang, X. Huang, S. Li, and Y. Wang, “Boosting local binary pattern-based face recognition”, Lecture Notes in Computer Science, vol. 3338, Springer, 2004, pp. 180–187.

[25] Y. Heping, Z. Wang, and S. Guo, “String Extraction Based on Statistical Analysis Method in Color Space”, Proceedings of GREC, 2005, pp. 173-181.

[26] D. Unay and A. Ekin, “Intensity versus texture for medical image search and retrieval”, Proceedings of the 5th IEEE international symposium on biomedical imaging: from nano to macro (ISBI 2008), 2008, pp. 241–244.

[27] A. Oliver, X. Llado´, J. Freixenet, and J. Martı´, “False positive reduction in mammographic mass detection using local binary patterns”, Proceedings of the medical image computing and computer-assisted intervention (MICCAI 2007), 2007, pp. 286–293.

[28] S. Liao and ACS. Chung, “Face recognition by using elongated local binary patterns with average maximum distance gradient magnitude”, Asian conference on computer vision,2007,pp.672-679.

[29] X. Tan and B. Triggs, “Enhanced local texture feature sets for face recognition under difficult lighting conditions”, Analysis and modelling of faces and gestures; 2007, pp.168-182.

[30] L. Maria and S. Bogdan, “Spatial Color Distribution Based Indexing andRetrieval Scheme,” Advances in Soft Computing, vol.59, 2009, pp. 419-427.

A Clustering-based Approach to Web Image Context Extraction

Sadet Alcic and Stefan Conrad
Institute for Computer Science, Databases and Information Systems
Heinrich-Heine University, D-40225 Duesseldorf, Germany
{alcic,conrad}@cs.uni-duesseldorf.de

***Abstract*—Images on the Web come along with textual descriptions that are valuable for different applications, such as image annotation, clustering of images, image categorization, etc. But usually Web pages are poorly structured and cluttered with contents of different topics, which hinder the accurate detection of the image context. Existing approaches are based on heuristic rules and thus cannot handle the variety of documents on the Web. In this paper, we introduce a novel approach to image context extraction, building on a Web content distance measure. Utilizing this distance measure, the addressed problem can be reduced to a content clustering problem where an image is associated with the textual contents of the cluster it belongs to. Our evaluation studies confirm the validity and quality of the proposed method and demonstrate its applicability to the Web.**

***Keywords*-Image Context Extraction, Web Content Mining.**

I. Introduction

After text, image is the most basic and commonly used content on the Web. But while text semantics can be extracted from text directly, the automatic detection of image semantics is still an open issue. Considering images on the Web, we recognize a valuable advantage over isolated images: these images come in hand with other textual information on hosting Web pages that can be exploited to describe the images.

However, Web documents are usually cluttered with multi-topic contents, while at same time they do not separate these contents by explicit structure. As a consequence, the problem of estimating the image context as a (hidden) subset of the complete Web page arises. As *image context*, we understand the textual contents of a Web page that share the semantics with an image on this page.

Different parts of a Web page can be considered as possible sources for image context, namely, image url, page title, alternative text (ALT attribute), passages of surrounding text, etc. The first three have been utilized in many approaches [1], [2] due to their easy extraction and promising accuracy to describe the embedded image. However, different researchers inferred in independent empirical studies [3], [4] that these context sources do not describe images satisfactorily. The main reasons are: filenames are often generated (such as img1.jpg); the page title is mostly to general (e.g., New York Times - News); alternative text is hardly available. In comparison, the descriptions derived from passage of text surrounding an image were more reliable.

In recent years, three general approaches have been proposed to extract these passages, (i) a fixed-size window of terms [5], [4], [3], (ii) DOM tree wrappers [6], [7], (iii) content blocks derived by Web page segmentation [8], [9]. While efficient in time, the fixed-window approach is prone to precision as well as recall errors, since the extracted passage can include irrelevant content, or respectively, exclude relevant content. Wrappers are based on heuristic rules that only can cover a small subset of the possible design patterns for Web pages. Each time these patterns change, wrappers have to be adapted manually. Web page segmentation is a more principal approach to estimate the image context by associating images with the textual contents of common segments. However, most of the existing page segmentation algorithms are not designed for image context extraction and deliver to broad or to narrow segments, which affect directly the quality of context information. Due to their complexity, it is difficult to adapt existing page segmentation methods to meet the requirements of image context extraction [10], [7].

Recognizing the shortcomings of existing approaches, we present a more general solution to extract web image context by mining the Web contents of a page based on the underlying DOM tree structure. To make our approach applicable to the Web, we abstain from using visual features of contents, which are very time consuming since they need a Web page to be rendered.

Contribution. The main contribution of this work is threefold. First, we introduce a novel distance metric for Web contents based on the hierarchical structure of the DOM tree. Using this metric allows to map the contents of a Web page in a one-dimensional (1D) space. As a result, the image context extraction problem is reduced to finding context separators in 1D space. Secondly, we introduce a solution for the reduced problem by proposing a generic threshold-based clustering algorithm, which exploits the distance of adjacent contents. And finally, we evaluate the proposed method and compare its effectiveness to common approaches from the literature.

Organization. This paper is organized as follows. Section II gives a brief overview to related work. Section III introduces by example the different structural clues that can be obtained from HTML and afterwards presents our DOM-based distance metric, which is based on the preliminary thoughts. In Section IV, the clustering-based method to image content extraction is presented. Finally, the approach is evaluated in Section V and results are discussed.

II. Related Work

Many researchers were attracted by the benefits of Web image context in the past. As a result, a variety of context extraction methods, ranging from simple heuristics-based approaches to complex DOM and vision-based extractors have been proposed. Approaches like [5], [4], [3] extract a paragraphs of n-terms (n is chosen different, e.g., 10, 20, 32) surrounding the image as context. While this approach is fast and simple, it is prone to errors, i.e., when the image context is placed only under the image.

Tian et. al [6] propose a DOM-based method where the image context is selected by extracting the textual contents of the sibling nodes. Starting at the image node, the DOM tree is traversed upward until a parent node has text nodes. These are then assigned as image context. Fauzi et al. [7] distinguish three different use cases for images in HTML documents: listed images, semi-listed images, and unlisted images. To each case, context extraction rules are defined based on DOM tree.

Cai et al. [9], [8] use Vision based Page Segmentation (VIPS) [11] to partition Web documents into visual blocks. Images are assigned the text of the common visual block. VIPS is an hierarchical top-down approach, which starts with the whole page as initial block. For each block, a Degree of Coherence (DoC) is computed using heuristic rules based on the DOM Tree structure and visual cues obtained from the browser representation. The DoC value determines to what degree the contents within a block correlate to each other. It ranges from 1 to 10, while 10 represents the highest correlation. At the beginning a Permitted Degree of Coherence (PDoC) value is specified (set to 5 in [8]), which controls the segmentation granularity. If a particular block has a DoC value smaller than PDoC, this block has to be subdivided and this rule is repeated until all blocks on the bottom fulfill the mentioned condition. Hattori et al. [12] define a distance function that computes a distance between contents based on structural depth of HTML tags and performs top-down segmentation applying the proposed content distance function. However, their content distance does not satisfy our needs by two reasons: (i) the distance measure actually does not reflect the distance values that correspond to the HTML structure; (ii) the triangle inequality is not met, which is very important for our clustering-based approach. There is a variety of other approaches to page segmentation, i.e. [13], [14], but since they were developed for other applications, their adaptation to image context extraction is of high complexity.

III. Structural Information in Web pages

In a DOM tree of a Web document, we distinguish two kinds of elements: inner nodes and leaf nodes. The leaf nodes represent basic content units of a Web page, thus image as well as text nodes are elements of this kind. They are arranged from left to right in the order as they appear in the document source. On the other hand, the inner nodes correspond to tags that define the structural as well as functional properties of the contents in their subtree. They further group the underlying contents to DOM blocks. All these hints can be utilized to estimate a structural distance of the content units, which will be motivated by an example.

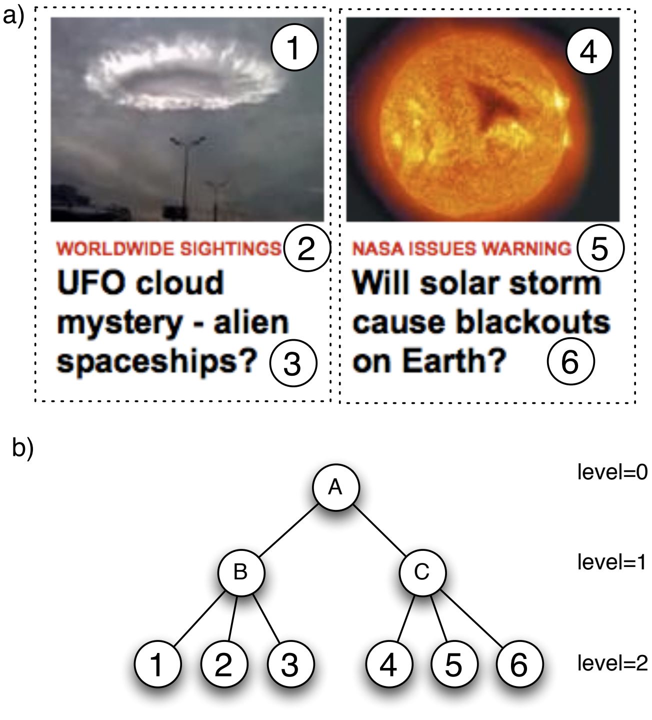

a)
1
4
WORLDWIDE SIGHTINGS
2
UFO cloud mystery - alien spaceships?
3
NASA ISSUES WARNING
5
Will solar storm cause blackouts on Earth?
6
b)
A
level=0
B
C
level=1
1
2
3
4
5
6
level=2

Figure 1. Example snipped of a Web page: a) shows the visual representation and b) the simplified DOM tree.

Figure 1 contains a small excerpt of a Web page and a simplified version of its corresponding DOM subtree. From both representations, we can simply infer that contents 1 - 3 form a structural block, and respectively contents 4 - 6 do the same.

This example can now be used to introduce the basic ideas for a DOM-based distance measure. Starting with the contents 1 and 2, we may set their distance to

$$d(\textcircled{1},\textcircled{2}) = c,$$

where $c > 0$ is an arbitrary constant. Because contents 2 and 3 are on the same level and both under the same parent, we set the distance

$$d(\textcircled{2},\textcircled{3}) = c.$$

However, between content 1 and content 3 there is content 2 and thus the distance between content 1 and 3 can be computed transitively

$$d(\textcircled{1},\textcircled{3}) = d(\textcircled{1},\textcircled{2}) + d(\textcircled{2},\textcircled{3}) = 2c.$$

The same rules can be applied in the right subtree resulting in following three equations:

$$d(\textcircled{4},\textcircled{5}) = c; d(\textcircled{5},\textcircled{6}) = c; d(\textcircled{4},\textcircled{6}) = 2c.$$

The only missing distance is that one between contents 3 and 4. As these elements belong to different blocks, we must ensure that their distance is greater than the maximum distance of contents in the left or right block. In this example, this maximum distance of siblings on level 2 is $2c$ and therefore the distance between content 3 and 4 might be set to

$$d(\textcircled{3},\textcircled{4}) = 4c,$$

which is a high distance, that separates the blocks at this position. The described distance measure can be further used to map content units of a Web page in a 1D space, simply by setting the point of origin at content 1 as depicted in Figure 2.

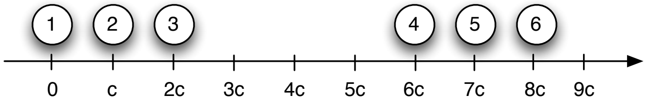


Figure 2. Six content elements spread over 1D space by proposed distance metric.

Following the ideas in this example, we formulate a general DOM distance metric, which is applicable to any DOM tree.

Definition: DOM Distance. Let d be a Web document and $D_d(N, E)$ the corresponding simplified DOM tree, with N as a set of nodes and E as a set of edges. Further let $P = (p_1, ..., p_n)$ be the sequence of all nodes in N ordered by traversing D_d in preorder traversal. The index $i \in \{1, ..., n = |N|\}$ gives the order of $p_i \in N$ in the sequence P. Based on this formulation, we define the *DOM Distance Metric* $d : N \times N \rightarrow \mathbb{R}$ for two nodes $p_a, p_b \in N, 1 < a < b < n$ (if $b < a$, wlog. switch p_a, p_b) as follows:

$$d(p_a, p_b) = \sum_{i=a+1}^{b} w_{p_i}, \tag{1}$$

where w_{p_i} is the *block weight* of element p_i, which has to be further specified with the knowledge from Section III. In general, the metric d is a sum of block weights.

The block weights w_{p_i} can be explained as the cost needed to reach the content block of p_i from its predecessor p_{i-1}. Therefore w_{p_i} corresponds to the distance $d(p_{i-1}, p_i)$ of two neighbored contents p_{i-1}, p_i. For nodes on the same tree level l, the block weights are all equal. On a lower level $l - 1$ the block weight has to be at least greater than the maximal distance of two sibling nodes at level l. The maximal distance of two sibling nodes on level l corresponds to the degree of the nodes on level $l-1$, which is the maximal count of children of the nodes at level l.

With these considerations, the block weight function $w : \mathbb{N} \rightarrow \mathbb{R}$ can be formulated by following recurrence:

$$w(l) = \begin{cases} c & : \quad d_l = 0 \\ d_l \cdot w(l+1) & : \quad d_l > 0, \end{cases} \tag{2}$$

where d_l refers to the maximal degree of nodes at level l and c is an arbitrary constant value. To apply w in Equation 1, we define function $l : N \rightarrow \mathbb{N}$ that delivers the tree level of a node. Thus the block weight w_p for a node $p \in N$ is $w_p = w(l(p))$.

To demonstrate the applicability of the defined distance, we consider again the example from Figure 1. First, we determine the preorder sequence of the nodes:

$$\textcircled{A} - \textcircled{B} - \textcircled{1} - \textcircled{2} - \textcircled{3} - \textcircled{C} - \textcircled{4} - \textcircled{5} - \textcircled{6}$$

Secondly, we compute the level weights w_l according to Equation 2. For level 2 we get $w_2 = c$ since the level degree d_l at level 2 equates to 0. At level 1 we have $d_1 = 3$ and consequently $w_1 = w_2 \cdot d_1 = 3c$. For level 0 the level weight equates $w_0 = w_1 \cdot d_0 = 3c \cdot 2 = 6c$. After this initialization steps, the distance between contents can be computed by summing up appropriate weights as defined in Equation 1. We can verify, that the distances correspond to that assumed in our preliminary thoughts.

Time complexity. The proposed distance measure consists of two steps, an initialization step, which has to be executed only once for a complete Web page, and the distance computation step. The initialization step includes a preorder traversing of the DOM tree consuming linear time depending on the *total number of nodes* n, and further the weights computation for l levels while l equates in average $log(n)$. Thus the time complexity for initialization is in $O(n + log(n)) = O(n)$. The distance computation includes a sum over n node weights at maximum and thus can also be computed in $O(n)$. By an additional step in initialization, where we map all nodes in the metrical space derived by the proposed metric, we can minimize the effort for distance computation to one subtraction computable in O(1). The additional effort costs $O(n)$ and thus does not affect the time complexity of initialization.

IV. Image Context Extraction

In this section, we will present our proposed method to Web context extraction by clustering contents in 1D space supplied by the proposed distance measure.

Problem Formulation. Given an image I in a Web document d, Web Image Context Extraction (WICE) denotes the process of determining the textual contents t_i of document d that are associated with the image I. The proposed DOM distance maps the basic content units of a Web page to a 1D space. By setting cuts at appropriate positions in this space, contents are partitioned (or clustered) to content blocks. The image I can now be associated with the textual contents t_i of the block, I belongs to. Thus the WICE problem is reduced to clustering in 1D space, or in other words, to estimating suitable positions in this space to separate contents.

1D-Clustering based on Distance Threshold. The idea to a clustering method for 1D data points will be motivated by the example page excerpt from Section III. Consider the simple one-dimensional optimization problem in Figure 2: we want to find a good clustering for the six data points so that the variance of the distances between each pair of adjacent points in same cluster is minimized. It is not hard to understand that by cutting at the largest distance between a pair of adjacent points, we will find a good solution to the clustering problem. Actually, if we are able to define threshold that the distance of adjacent contents should not exceed, the clustering can be done as described in Algorithm 1.

```
Input: Sequence of Web contents $S = (s_1, ..., s_n)$,
       threshold $t$
Result: Set of computed clusters $C$
$c$ = newCluster($s_1$);
for $i = 2$ to $n$ do
    if $d(s_{i-1}, s_i) > t$ then
        $C$.add($c$);
        $c$ = newCluster($s_i$);
    else
        $c$.add($s_i$); ;
    end
end
```

Algorithm 1: 1D-clustering in by thresholding

The algorithm starts by initializing a new cluster c with the first element s_1. Then a loop iterates over the sequence S, in which the elements are ordered by their appearing in the document, and computes the distance between every pair of adjacent contents. If the computed distance is greater than a predefined threshold t, the actual cluster c is put in the set of clusters C and c is initialized again with content s_i. Otherwise, content s_i expands the actual cluster c.

Threshold estimation. A static threshold value t could be computed by averaging over all distances of adjacent content pairs:

$$t = \frac{1}{n-1} \sum_{k=2}^{n} d(s_{k-1}, s_k).$$

This baseline threshold might work well for Web documents that consist of content clusters with similar density. However, since Web contents are usually distributed over different levels, the cluster density can significantly differ among different clusters. Thus the threshold should be more adaptive to distances in the environment.

To meet these requirements, we propose to use a gaussian weighted threshold function $t : \mathbb{N} \rightarrow \mathbb{R}$:

$$t(k) = \frac{1}{\sum_{i=2}^{n} G(i,k,\sigma)} \sum_{i=2}^{n} G(i,k,\sigma) d(s_{k-1}, s_k),$$

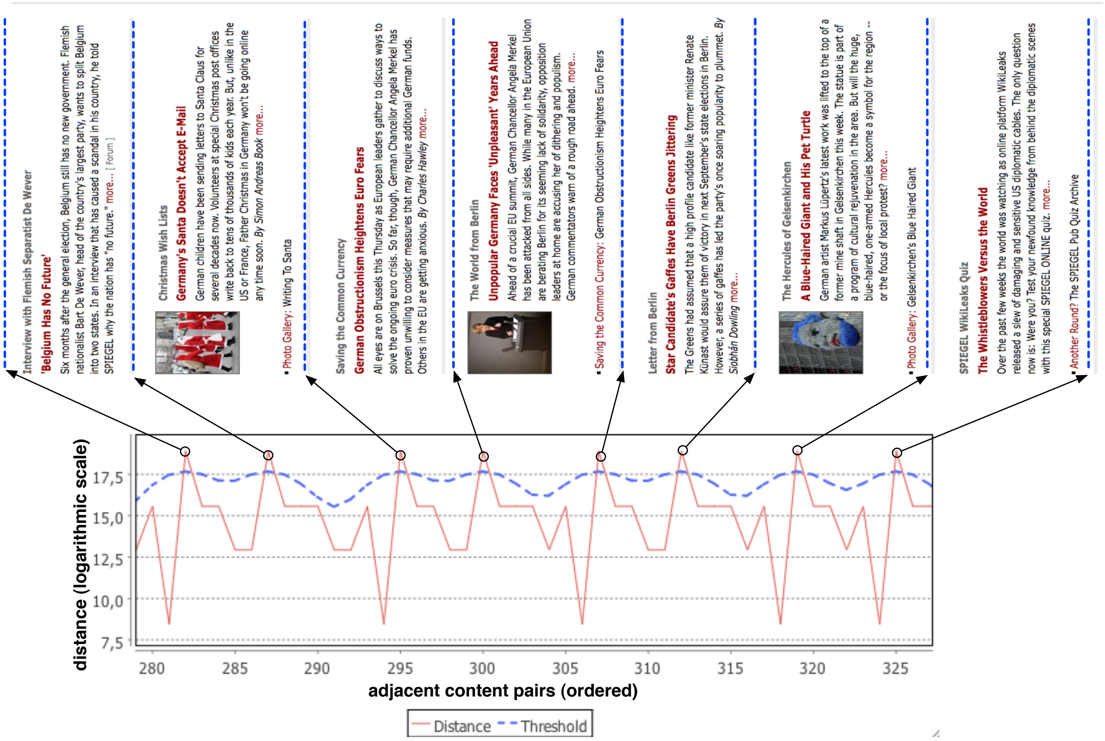
with the gaussian function $G(x, \mu, \sigma) = \frac{1}{2\pi\sigma^2} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$. The remaining parameter σ^2 is the variance (the measure of the width of the gaussian peak) and has to be considered empirically.

Visual Representation. To give a more intuitive description to the proposed algorithm, we visualize its main components. The solid-line curve in Figure 3 refers to the distances of adjacent Web contents, while the values on the x-axis correspond to the index of the contents in the contents sequence (e.g., x-value 280 means the distance of adjacent contents 280 and 281). The dashed-line curve in the same plot is the gaussian smoothed version of the red function, and corresponds to the threshold t. For each peak of red (distances) function exceeding the blue (threshold) function, we have drawn a circle at this function value and pointed with an arrow to the corresponding position in the browser representation of the document. This example shows empirically the quality of our method, since all blocks were properly recognized.

V. Evaluation

The accuracy of the proposed method was evaluated using the evaluation framework proposed in [10]. The ground truth data consists of different test collection gathered from real Web servers. Table I comprises the main information about the test collections. The *diverse* collection consists of 79 documents for which the context extraction was performed manually. The other collections were created by recalling the main page of a Web site and storing the gathered document whenever a significant change of the content to the previously stored was detected. In this way, we collected a large amount of documents based on the same template. A rule based extractor was then implemented for each collection, that extracted the image context. This resulted in a large amount of real testdata, consisting of 12,907 documents is and 155,565 context to image pairs.

Quality measure. In order to compare the extracted image context with the ground truth data we understand both texts as sequences of words and measure their overlap by computing the longest common subsequence. By treating



Interview with Flemish Separatist De Wever
'Belgium Has No Future'
Six months after the general election, Belgium still has no new government. Flemish nationalist Bart De Wever, head of the country's largest party, wants to split Belgium into two states. In an interview that has caused a scandal in his country, he told SPIEGEL why the nation has "no future." more... [Forum]

Christmas Wish Lists
Germany's Santa Doesn't Accept E-Mail
German children have been sending letters to Santa Claus for several decades now. Volunteers at special Christmas post offices write back to tens of thousands of kids each year. But, unlike in the US or France, Father Christmas in Germany won't be going online any time soon. By Simon Andreas Book more...
▪ Photo Gallery: Writing To Santa

Saving the Common Currency
German Obstructionism Heightens Euro Fears
All eyes are on Brussels this Thursday as European leaders gather to discuss ways to solve the ongoing euro crisis. So far, though, German Chancellor Angela Merkel has proven unwilling to consider measures that may require additional German funds. Others in the EU are getting anxious. By Charles Hawley more...

The World from Berlin
Unpopular Germany Faces 'Unpleasant' Years Ahead
Ahead of a crucial EU summit, German Chancellor Angela Merkel has been attacked from all sides. While many in the European Union are berating Berlin for its seeming lack of solidarity, opposition leaders at home are accusing her of dithering and populism. German commentators warn of a rough road ahead. more...
▪ Saving the Common Currency: German Obstructionism Heightens Euro Fears

Letter from Berlin
Star Candidate's Gaffes Have Berlin Greens Jittering
The Greens had assumed that a high profile candidate like former minister Renate Künast would assure them of victory in next September's state elections in Berlin. However, a series of gaffes has led the party's once soaring popularity to plummet. By Siobhán Dowling more...

The Hercules of Gelsenkirchen
A Blue-Haired Giant and His Pet Turtle
German artist Markus Lüpertz's latest work was lifted to the top of a former mine shaft in Gelsenkirchen this week. The statue is part of a program of cultural rejuvenation in the area. But will the huge, blue-haired, one-armed Hercules become a symbol for the region -- or the focus of local protest? more...
▪ Photo Gallery: Gelsenkirchen's Blue Haired Giant

SPIEGEL WikiLeaks Quiz
The Whistleblowers Versus the World
Over the past few weeks the world was watching as online platform WikiLeaks released a slew of damaging and sensitive US diplomatic cables. The only question now is: Were you? Test your newfound knowledge from behind the diplomatic scenes with this special SPIEGEL ONLINE quiz. more...
▪ Another Round? The SPIEGEL Pub Quiz Archive

Figure 3. Distances of adjacent contents and the threshold values computed by weighted averaging with gaussian. The peaks at which the distance exceeds the threshold point to the corresponding separators in the browser output of the document.

Table I
TEST COLLECTIONS WITH TOTAL NUMBER OF DOCS AND IMAGES.

Collection	#Documents	#Images
BBC	1077	7878
CNN	874	11612
Golem	789	3061
Heise	79	1403
MSN	375	9264
New-York Times	556	10927
Spiegel	1076	36310
Telegraph	530	10503
The Globe and Mail	735	15808
Wikipedia	3000	6728
Yahoo! (english)	3737	41170
diverse (manual)	79	901
total	12907	155565

Table II
EVALUATION RESULTS SHOWING THE AVERAGE F-SCORES OF CONTEXT EXTRACTION METHODS ON DIFFERENT COLLECTIONS.

	10 terms	20 terms	monash	siblings	full text	vips 5	vips 6	vips7	DOM-dist
cnn	0,40	0,27	0,75	0,91	0,02	0,16	0,28	0,29	**0,92**
golem	0,51	0,62	**0,96**	0,95	0,10	0,15	0,47	0,47	0,95
heise	0,43	0,52	0,93	0,95	0,03	0,31	0,77	0,77	**0,96**
msn	0,40	0,47	**0,95**	0,93	0,04	0,16	0,23	0,23	0,89
nytimes	0,41	0,45	0,86	0,76	0,03	0,16	0,46	0,63	**0,89**
spiegel	0,45	0,40	0,90	0,84	0,03	0,10	0,25	0,19	**0,98**
telegr.	0,63	0,61	0,92	0,23	0,03	0,12	0,62	0,80	**0,93**
gl.&m.	0,55	0,50	0,94	0,97	0,03	0,22	0,35	0,49	**0,99**
wiki	0,42	0,33	0,92	**0,94**	0,02	0,07	0,66	0,32	0,89
yahoo	0,54	0,59	0,91	0,49	0,04	0,05	0,14	0,26	**0,89**
diverse	0,41	0,41	0,81	0,80	0,05	0,21	0,36	0,36	**0,85**
overall	0,47	0,48	0,90	0,80	0,04	0,15	0,42	0,45	**0,93**

the extracted context as retrieved data and the ground truth as relevant data we can apply standard information retrieval concepts of *precision* P, *recall* R and *F-score* 1 as performance measures.

Parameter estimation. Our proposed method to context extraction has one open parameter σ^2 that has to be specified. σ^2 is part of the gaussian smoothing kernel and considers its variance. In order to avoid overfitting, the parameter was trained iteratively on a smaller subset of our testdata consisting of five documents of each collection. The maximal average F-score was reached when σ^2 was set to 2.25.

Results. Table II contains the average F-scores computed on different test collections. Besides the proposed algorithm, we have extracted image context with two DOM-based methods – Monash [7] and [6] Siblings; a heuristics based method –

N-terms [5], [4]; and a vision-based method – VIPS [11]. To show the benefit of extraction methods, we have further included a full text extractor in the evaluation as baseline.

As N in the N-terms extractor we chose 10 and 20 since these are the frequently used parameters in the literature. The PDoC value of the VIPS algorithm was set to 5, 6, and 7 during the evaluation. As observable, there are some results missing for VIPS on the bbc collection. The reason for this lies in the implementation of the VIPS library that is based on the Internet Explorer 6 (IE6). However, IE6 was not able to properly display the crawled documents from the BBC Web page due to javascript errors.

As a first results, we find out that the baseline extractor has a significantly low performance compared to other extraction methods. This is because in most Web documents the length of the image context is significantly smaller than the length of the full text. Therefore it is worth investigating image context extraction algorithms.

The heuristics-based methods extracting the text within a frame of N terms surrounding the image achieve both F-scores around $0,5$. A possible reason is the fact that images are often placed next to the borders of articles. As a consequence, half of the associated text does not belong to the image. Both parameters deliver similar results and thus no one can be preferred.

VIPS is traditionally a page segmentation algorithm that was frequently used for context extraction in the past. However, the performance of VIPS ranges over the lower third. While the segments that VIPS extracts with a PDoC value of 5-7 are too broad, higher PDoc values yield to segments that contain only the image and no text.

The DOM-based methods – monash and siblings, as well as our proposed method – perform best on all collections. While the F-score of the siblings method varies for different test collections, the other two reach constantly high values with a small advantage for our method compared to the monash extractor.

VI. Conclusion and Future Work

This work presents a new method for image context extraction based on the distances of Web contents. Distance is computed using structural clues from the document. Using this distance function, the complex problem of image content extraction is reduced to the familiar 1D clustering. The results of the evaluation task show that the proposed method delivers highest accuracy on almost all test collections. As future work, other traditional clustering approaches could be applied to the 1D clustering problem. Further, we plan to estimate the impact of our method to applications, like image ranking or image classification.

References

[1] J. R. Smith and S.-F. Chang, "Image and Video Search Engine for the World Wide Web," in *Storage and Retrieval for Image and Video Databases (SPIE)*, 1997, pp. 84–95.

[2] H. T. Shen, B. C. Ooi, and K.-L. Tan, "Giving meanings to www images," in *Proceedings of the eighth ACM international conference on Multimedia*, ser. MULTIMEDIA '00. New York, NY, USA: ACM, 2000, pp. 39–47.

[3] H. Feng, R. Shi, and T.-S. Chua, "A bootstrapping framework for annotating and retrieving www images," in *Proceedings of the 12th annual ACM international conference on Multimedia*, ser. MULTIMEDIA '04. New York, NY, USA: ACM, 2004, pp. 960–967.

[4] T. A. S. Coelho, P. P. Calado, L. V. Souza, B. Ribeiro-Neto, and R. Muntz, "Image Retrieval Using Multiple Evidence Ranking," *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 4, pp. 408–417, 2004.

[5] S. Sclaroff, M. L. Cascia, and S. Sethi, "Unifying textual and visual cues for content-based image retrieval on the World Wide Web," *Computer Vision and Image Understanding*, vol. 75, no. 1-2, pp. 86–98, 1999.

[6] T. Yong-hong, H. Tie-jun, and G. Wen, "Exploiting multi-context analysis in semantic image classification," *J. Zhejiang Univ. SCI. 6A(11)*, pp. 1268–1283, 2005.

[7] F. Fauzi, J.-L. Hong, and M. Belkhatir, "Webpage segmentation for extracting images and their surrounding contextual information," in *ACM Multimedia*, 2009, pp. 649–652.

[8] X. He, D. Cai, J.-R. Wen, W.-Y. Ma, and H.-J. Zhang, "Clustering and searching www images using link and page layout analysis," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 3, no. 2, p. 10, 2007.

[9] D. Cai, X. He, Z. Li, W.-Y. Ma, and J.-R. Wen, "Hierarchical clustering of WWW image search results using visual, textual and link information," in *MULTIMEDIA '04: Proceedings of the 12th annual ACM international conference on Multimedia*, New York,USA, 2004, pp. 952–959.

[10] S. Alcic and S. Conrad, "Measuring performance of web image context extraction," in *Proceedings of the Tenth International Workshop on Multimedia Data Mining*, ser. MDMKDD '10. New York, NY, USA: ACM, 2010, pp. 8:1–8:8.

[11] D. Cai, S. Yu, J.-R. Wen, and W.-Y. Ma, "VIPS: a Vision-based Page Segmentation Algorithm," Microsoft Research (MSR-TR-2003-79), Tech. Rep., 2003.

[12] G. Hattori, K. Hoashi, K. Matsumoto, and F. Sugaya, "Robust web page segmentation for mobile terminal using content-distances and page layout information," in *WWW '07: Proceedings of the 16th international conference on World Wide Web*. New York, NY, USA: ACM, 2007, pp. 361–370.

[13] C. Kohlschütter and W. Nejdl, "A densitometric approach to web page segmentation," in *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*. New York, NY, USA: ACM, 2008, pp. 1173–1182.

[14] D. Chakrabarti, R. Kumar, and K. Punera, "A graph-theoretic approach to webpage segmentation," in *WWW '08: Proceeding of the 17th international conference on World Wide Web*. New York, NY, USA: ACM, 2008, pp. 377–386.

Virtual Reality Technology used to Support the Buildings Inspection Activity

Alcínia Z. Sampaio, Augusto M. Gomes
Dep. Civil Engineering and Architecture
Technical University of Lisbon
Lisbon, Portugal
e-mail: zita@civil.ist.utl.pt, augusto@civil.ist.utl.pt

Ana Rita Gomes, Daniel P. Rosário
Dep. Civil Engineering and Architecture
Technical University of Lisbon
Lisbon, Portugal
e-mail:ritagomes05@hotmail.com,
derosario@gmail.com

***Abstract*— A Virtual Reality model was created in order to help in the maintenance of exterior closures and interior finishes of walls in a building. It allows the visual and interactive transmission of information related to the physical behavior of the elements, defined as a function of the time variable. To this end, the basic knowledge of material most often used in walls, anomaly surveillance, techniques of rehabilitation, and inspection planning were studied. This information was included in a database that supports the periodic inspection needed in a program of preventive maintenance. The results are obtained interactively and visualized in the virtual environment itself. This work brings an innovative contribution to the field of maintenance supported by emergent technology.**

Keywords- Construction, Maintenance, Inspection Virtual reality, Human-Interaction

I. Introduction

The main aim of a research project, now in progress at the Department of Civil Engineering of the Technical University of Lisbon, is to develop virtual models as tools to support decision-making in the planning of construction management and maintenance. A first prototype concerning the lighting system has already been completed [1]. A second prototype concerning the maintenance of the closure of walls, both interior walls and façades, is now being developed. This paper describes this part of the project.

The interactive model integrates Virtual Reality (VR) technology, the EON system [2], and an application implemented in Visual Basic (VB) language. The model allows interaction with the 3D geometric model of a building, visualizing components for each construction. It is linked to a database of the corresponding technical information concerned with the maintenance of the materials used as exterior closures and interior finishes. The principal objective of the interactive VR prototype is to support decision-making in the maintenance domain.

The present project aims to bring important contributions to this domain, through the implementation of virtual models able to relating the behavior of materials, their characteristics, anomalies and repair work to each other. During this work the basic knowledge of the topics involved, such as aspects related to the materials, the techniques of rehabilitation and conservation and the planning of maintenance is outlined and discussed in addition, methods of interconnecting this knowledge with the virtual model are explored.

The prototype for walls was trialed in a concrete project. This kind of building element has a continuous lifestyle, so requires the study of preventive maintenance (the planning of periodical local inspections) and of corrective maintenance (with repair activity analysis). The model facilitates the visual and interactive access to results, supporting the definition of inspection reports.

Actually, the VR technology is presented in works concerning construction, but there is a lack in the application of this technique in maintenance. Two VR models were developed supporting the maintenance activity. They are innovative applications in this subject.

II. VR in Construction and Maintenance

The performance of the maintenance of a building has been increased through the application of new modeling concepts, particularly the incorporation of VR techniques and the addition of time as a factor to be considered in the strategy of building conception. In the same way, 3D models have been developed, related to the time parameter, designated 4D models [3], focused in the beginning, basically, on planning the construction process. The geometric model of construction is presented as a progression of steps in its physical evolution following planning. The University of Stanford [4], and the Finnish Centre of Investigation VTT [5], have presented concrete applications in the design phase with considerable befits relevant to communication between specialists, constructors and promoters. In the construction domain, the VR models are used to show the physical evolution of the building, through 4D models, in different phases of its construction following specific planning [6] and the simulation of the operational evolution of the associated construction processes [7]. In the area of architecture, VR models are generally applied to the visualization of static physical models in the definition of itineraries of walk-through, as a means of transmitting the functional and geometric aspect of the building.

In addition, VR technology has also been applied as a complement to 3D modeling, leading to better communication between the various stakeholders in the process, whether in training or in professional practice. This task is particularly relevant to the presentation of processes which are defined through sequential stages as generally is the case in the learning of new curricular subjects. In professional contexts, note the contribution in Architecture/Engineering, to support for conception, presenting the plan or following the progress of construction [8].

In the maintenance domain some researches have been including visual interaction: Anna-Liisa Linholm describes the creation process of an interactive model for identifying the added value of corporate real estate management and implement it in a case organization, testing whether it works in practice [9]; Visualization of building maintenance through time is the topic of the researching activity of Rad [10]; Khosrowshahi focus the research VR application on lighting and paintings of interior wall maintenance [11].

One of the more recent targets of investigation is in fact, research into the sharing of data between applications, which can be manipulated by means of a common interface, as a way of rendering 4D tools efficacious and of wide use. Virtual reality is seen today as an integrating technology, with great potential for communication between project participants, and most recently, as a tool for the support of decision-making, made possible by the integration of distinct computer applications in the virtual model. In this context, the present work presents the development of a system concerning maintenance based on VR technology, involving knowledge of the physical aspects of materials, in particular, those which refer to wear and tear (a function of time, use and environmental factors), integrating them in digital spatial representations. In this way, the indisputable advantage of the ease interpretation and perception of space provided by the visualization of 3D models, and the technical content underlying the real characteristics of the observed elements are brought together.

III. Interactive And Collaborative Models

Virtual Reality technology can support the management of data that is normally generated and transformed or replaced throughout the lifecycle of a building. This technology becomes an important support in the management of buildings allowing interaction and data visualization. At present, the management of building planning can be presented in a 3D form and various materials can be assigned to the fixtures and furnishing enabling the user to be placed in the virtual building and view it from inside as well as outside. This study contemplates the incorporation of the 4th dimension, that is, time, into the concept of visualization. The focus of the work is on traveling through time: the ability to view a product or its components at different points in time throughout their life. It is envisaged that the incorporation of the time dimension into 3D visualization will enable the designer/user to make more objective decisions about the choice of the constituent components of the building. In maintenance the time variable is related to the progressive deterioration of the materials throughout the building's lifecycle.

The present prototype incorporates interactive techniques and input devices to perform visual exploration tasks [12]. To support this system a data base was created which included a bibliographic research support made in regard to the closure materials used in the interior and exterior walls of a building, anomalies concerning different kinds of covering material, and corrective maintenance. Repair activities were also studied. The programming skills of those involved in the project had to be enhanced so that they could achieve the integration of the different kinds of databases needed in the creation of the interactive model.

The 3D model linked to a database concerning maintenance produces a collaborative virtual environment, that is, one that can be manipulated by partners interested in creating, transforming and analyzing data in order to obtain results and to make decisions. For example, inspection reports can be defined and consulted by different collaborators. The process of developing the prototype interface considers these purposes [13]. The developed prototype associates the characteristics of the coating component of the exterior and interior walls to activities concerning the maintenance of buildings (Fig. 1).

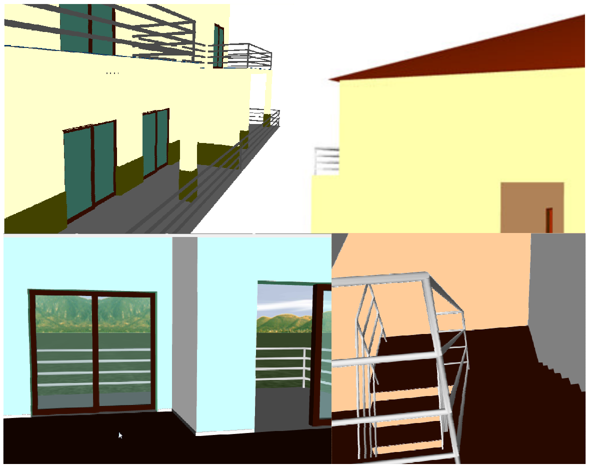


Figure 1. 3D model of the house: coating elements of exterior and interior walls.

IV. Wall Maintenance Prototype

Façade coatings play an important role in the durability of buildings, since they constitute the exterior layer that ensures the protection of the wall against the aggressive actions of physical, chemical or biological nature. Naturally, they should also give the façade the required decorative effect. Since this building component is exposed to bad atmospheric conditions it frequently shows an evident degree of deterioration, requiring maintenance interventions. To perform maintenance activities a survey of failures in the building must be conducted in order to arrive at the best solution for repair and maintenance.

In order to better understand the operation of façade coating, bibliographic research of materials usually applied to façade coatings was carried out and a table of characteristics of these was drawn up. Subsequently, a

Copyright (c) IARIA, 2011. ISBN: 978-1-61208-129-8

survey was made of anomalies, probable causes, solutions and methods of repair for each of the coatings studied.

The visualization of the maintenance data of a building and the impact of time on the performance of these exterior closure materials require an understanding of their characteristics [14] (Fig. 2):

- Types of material: painted surfaces, natural stone panels and ceramic wall tiles;
- Application processes: stones (panel, support devices, adherent products, …); ceramic tiles (fixing mechanism, procedures, …); painted surfaces (types of paint products, prime and paint scheme surface, exterior emulsion paints, application processes);
- Anomalies: dust and dirt, lasting lotus leaf effect, covering power, insufficient resistance to air permeability or weatherproof isolation, damaged stones or ceramic tiles, alkali and smear effect, efflorescence, fractures and fissures ….;
- Repair works: surface cleaning, wire truss reinforcing, cleaning and pointing of stonework joints, removing and replacement of ceramic wall tiles, removing damaged paint and paint surface, preparing and refinishing stone panels, ...

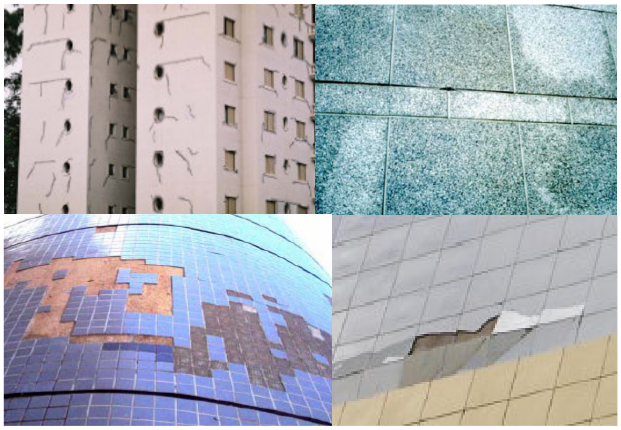


Figure 2. Different type of materials applied as façade coatings.

A. *Characteristics of the Materials used in Façades*

Depending on the role that the façade coatings play on the wall as a whole they can be classified as finishing, sealing or thermal insulation. The most frequent materials used as coating finishes are painting, tiling and, as sealing coating of the natural stone:

- Paint coating contributes to the aesthetic quality of the building and its environment and also protects the surface of the exterior wall against corrosion, deterioration and penetration of aggressive agents [15]. In order to obtain a good performance as an exterior coating, several aspects must be considered, such as covering power and resistance to water, sunlight, chemical products and to the development of micro organisms;
- The ceramic coating consists essentially of tiling panels, cement and adhesive and the joints between the slabs. The application of ceramic tiling to building façades has considerable advantages particularly as some degree of waterproofing is afforded by the glazed surface along with a great resistance to acids, alkalis and vapor. Other advantages are that it does not need painting and that it can be easily applied or substituted during repair work [16];
- The use of natural stone in the coating of façade surfaces is a good solution both technically and aesthetically. The stone coating is composed of slabs of stone attached to the wall by a support system. The principal characteristics of the stones are: reduced water absorption, sufficient mechanical resistance to bending and impact, abrasion and shearing parallel to the face of the slabs [17].

B. *The Database*

The most frequent anomalies that occur in the coated façades were analyzed in order to create a database linked to the virtual model that could support planning of inspections and maintenance strategies in buildings. The database contains the identification of anomalies that can be found in each type of material used in façades and the corresponding probable cause. For each kind of anomaly the most adequate repair solutions are also selected and included in the data base. The following example concerning deficiencies in tiles presents the methodology implemented in this virtual application (Tab. 1).

The characteristics related to anomalies, causes, repair solutions and rehabilitation tasks were included in a database of each type of material and linked to the 3D model of the building. Thus, the virtual model gives users the ability to transmit, visually and interactively, information related to the closure properties of exterior walls, allows them to analyze the anomalies observed in an inspection of the real building and to predict the corresponding repair work. The 3D virtual model can be seen, therefore, as an important tool for anomaly surveillance in structures and for supporting decision-making based on the visual analysis of alternative repair solutions.

C. *The Interface*

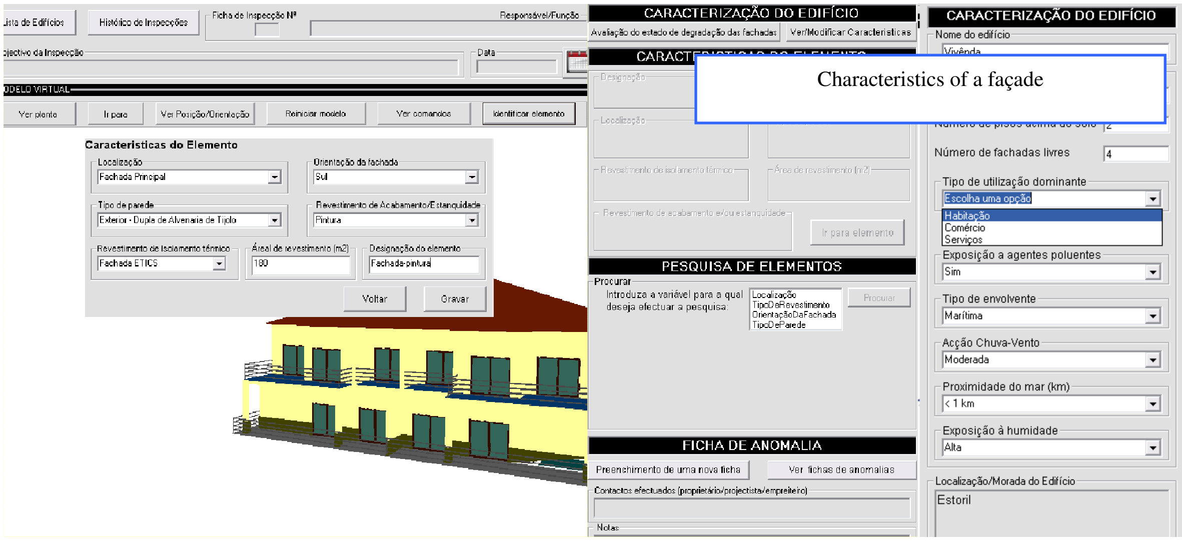
The implementation of the prototype system makes use of graphical software programming, *Visual Basic 6.0 Microsoft*, software to establish an adequate database, *Microsoft office access,* graphical drawing system*, AutoCAD Autodesk* and VR technology based software, *EON Studio*.

Human perceptual and cognitive capabilities were taken into account when designing this visualization tool so the model is easy to use and does not require sophisticated computer skills, as many potential users are not computer experts. It uses an interactive 3D visualization system based on the selection of elements directly within the virtual 3D world. Furthermore, associated with each component, there are integrated databases, allowing the consultation of the required data at any point in time.

Copyright (c) IARIA, 2011. ISBN: 978-1-61208-129-8

Table 1. Example of anomalies and the associated repair solution.

Anomaly	Specification of the anomaly	Repair solution	Repair methodology
Detachment 	Fall in areas with deterioration of support	Replacement of the coat (with use of a repair stand as necessary)	1º Removal of the tiles by cutting grinder with the aid of a hammer and chisel; 2º Timely repair of the support in areas where the detachment includes material constituent with it; 3º Digitizing layer of settlement; 4º Re-settlement layer and the tiles.
Cracking / Fracturing 	Failure of the support (wide cracks with well-defined orientation)	Replacement of the coat (with repair of cracks in the support)	1º Removal of the tiles by cutting grinder; 2º Removal of material adjustment in the environment and along the joint; 3º Repair of cracks, clogging with adhesive material (mastic); 4º Settlement layer made with cement in two layers interspersed with glass fiber; 5º Re-settlement layer and the tiles.



Caracteristicas do Elemento

Localização: Fachada Principal

Orientação da fachada: Sul

Tipo de parede: Exterior - Dupla de Alvenaria de Tijolo

Revestimento de Acabamento/Estanquidade: Pintura

Revestimento de Isolamento térmico: Fachada ETICS

Áreal de revestimento (m2): 180

Designação do elemento: Fachada-pintura

CARACTERIZAÇÃO DO EDIFÍCIO

Nome do edifício: Vivênda

Número de fachadas livres: 4

Tipo de utilização dominante: Escolha uma opção

Exposição a agentes poluentes: Sim

Tipo de envolvente: Marítima

Acção Chuva-Vento: Moderada

Proximidade do mar (km): < 1 km

Exposição à humidade: Alta

Localização/Morada do Edifício: Estoril

Characteristics of a façade

Figure 3. The main interface of the interactive application.

The interface is composed of a display window allowing users to interact with the virtual model, and a set of buttons for inputting data and displaying results (Fig. 3).

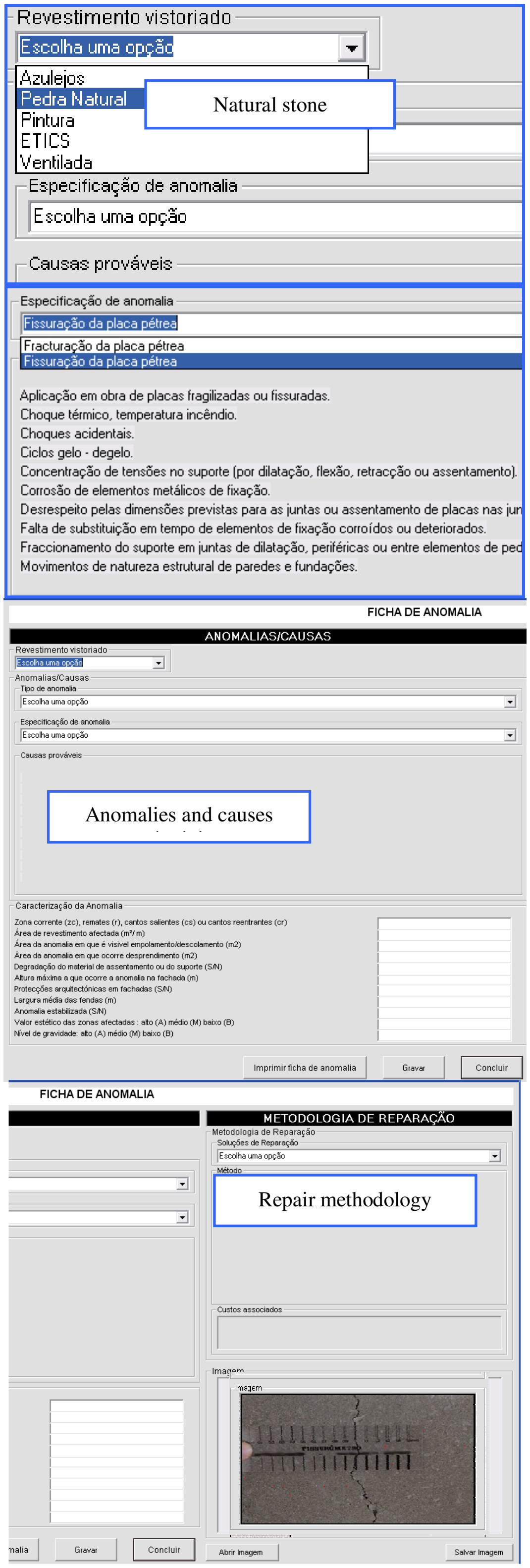
For each new building to be monitored the characteristics of the environment (exposure to rain and sea) and the identification of each element of the façades must be defined. The data associated to each element are the building orientation, the type of exterior wall (double or single), and the area and type of coating.

Once each monitored element has been characterized, several inspection reports can be defined and recorded and thereafter consulted when needed. An inspection sheet is accessed by the main interface (Fig. 4).

Using the drop-down menus allowed by the interface, the user can associate the characteristics of the observed anomaly to: a façade element; the type of anomaly, the specification, details and the probable cause of the anomaly, an adequate repair solution and pictures taken in the building. After completing all fields relating to an anomaly, the user can present the report as a pdf file.

D. *The Case Study*

First, the 3D geometric model of a building case was created. The building consists of a ground-floor, a 1st floor and an attic with dwelling space shown. The coating elements of the walls were then modeled as independent geometric objects (Fig. 5). In this way, each element can then support characterization data of the applied material and different kinds of information related to maintenance.



Aplicação em obra de placas fragilizadas ou fissuradas.
Choque térmico, temperatura incêndio.
Choques acidentais.
Ciclos gelo - degelo.
Concentração de tensões no suporte (por dilatação, flexão, retracção ou assentamento).
Corrosão de elementos metálicos de fixação.
Desrespeito pelas dimensões previstas para as juntas ou assentamento de placas nas jun
Falta de substituição em tempo de elementos de fixação corroídos ou deteriorados.
Fraccionamento do suporte em juntas de dilatação, periféricas ou entre elementos de ped
Movimentos de natureza estrutural de paredes e fundações.

Zona corrente (zc), remates (r), cantos salientes (cs) ou cantos reentrantes (cr)
Área de revestimento afectada (m²/ m)
Área da anomalia em que é visivel empolamento/descolamento (m2)
Área da anomalia em que ocorre desprendimento (m2)
Degradação do material de assentamento ou do suporte (S/N)
Altura máxima a que ocorre a anomalia na fachada (m)
Protecções arquitectónicas em fachadas (S/N)
Largura média das fendas (m)
Anomalia estabilizada (S/N)
Valor estético das zonas afectadas : alto (A) médio (M) baixo (B)
Nível de gravidade: alto (A) médio (M) baixo (B)

Figure 4. Inspection sheet interface.

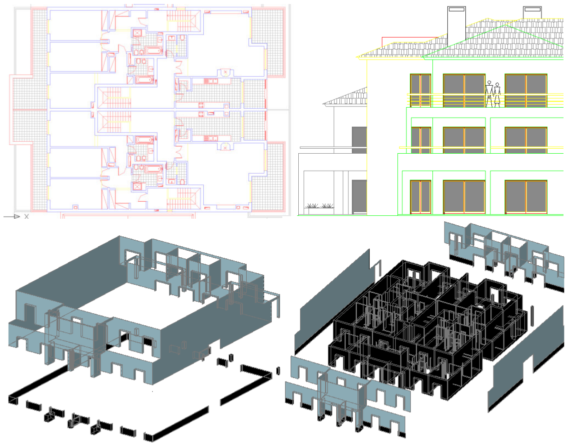
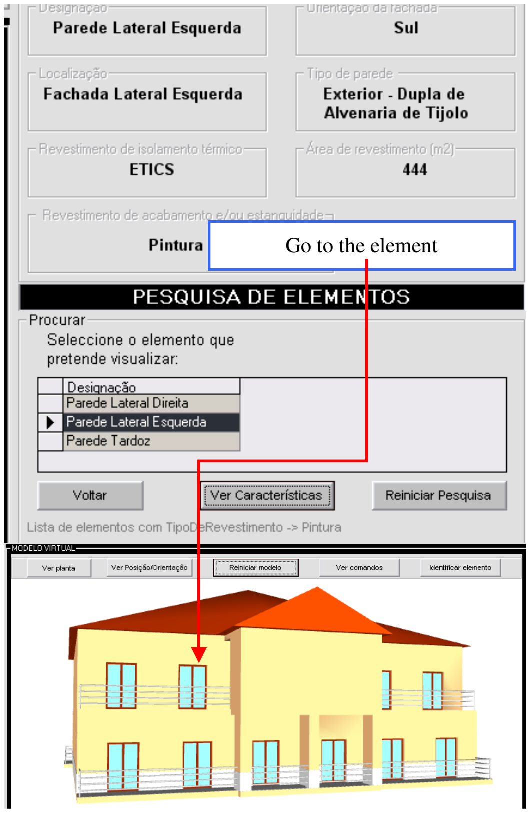


Figure 5. Steps of the geometric modelling process.

The 3D model is previously created and then inserted into the EON studio software in order to prepare each building component (3D object) that is going to be monitored for the analysis survey propose.

All coatings studied were considered in this case-study. Thus it was assumed that the main façade is covered with tile and the remaining façades are painted while hall façades are of natural stone. Fig. 6 shows how to identify a façade in the virtual model of the building. The Fig. 7 includes the inspection report of the anomaly considered in Table 1.

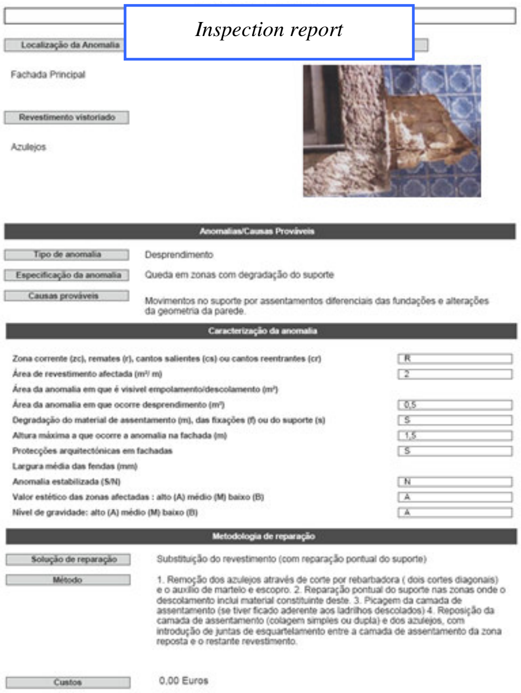


Designação: Parede Lateral Esquerda
Orientação da fachada: Sul
Localização: Fachada Lateral Esquerda
Tipo de parede: Exterior - Dupla de Alvenaria de Tijolo
Revestimento de isolamento térmico: ETICS
Área de revestimento (m2): 444
Revestimento de acabamento e/ou estanquidade: Pintura

Lista de elementos com TipoDeRevestimento -> Pintura

Figure 6. Identification of a façade element.

Copyright (c) IARIA, 2011. ISBN: 978-1-61208-129-8



Inspection report

Localização da Anomalia

Fachada Principal

Revestimento vistoriado

Azulejos

Anomalias/Causas Prováveis

Tipo de anomalia	Desprendimento
Especificação da anomalia	Queda em zonas com degradação do suporte
Causas prováveis	Movimentos no suporte por assentamentos diferenciais das fundações e alterações da geometria da parede.

Caracterização da anomalia

Zona corrente (zc), remates (r), cantos salientes (cs) ou cantos reentrantes (cr)	R
Área de revestimento afectada (m²/ m)	2
Área da anomalia em que é visível empolamento/descolamento (m²)	
Área da anomalia em que ocorre desprendimento (m²)	0,5
Degradação do material de assentamento (m), das fixações (f) ou do suporte (s)	S
Altura máxima a que ocorre a anomalia na fachada (m)	1,5
Protecções arquitectónicas em fachadas	S
Largura média das fendas (mm)	
Anomalia estabilizada (S/N)	N
Valor estético das zonas afectadas : alto (A) médio (M) baixo (B)	A
Nível de gravidade: alto (A) médio (M) baixo (B)	A

Metodologia de reparação

Solução de reparação	Substituição do revestimento (com reparação pontual do suporte)
Método	1. Remoção dos azulejos através de corte por rebarbadora (dois cortes diagonais) e o auxílio de martelo e escopro. 2. Reparação pontual do suporte nas zonas onde o descolamento inclui material constituinte deste. 3. Picagem da camada de assentamento (se tiver ficado aderente aos ladrilhos descolados) 4. Reposição da camada de assentamento (colagem simples ou dupla) e dos azulejos, com introdução de juntas de esquartelamento entre a camada de assentamento da zona reposta e o restante revestimento.
Custos	0,00 Euros

Figure 7. Inspection sheet report.

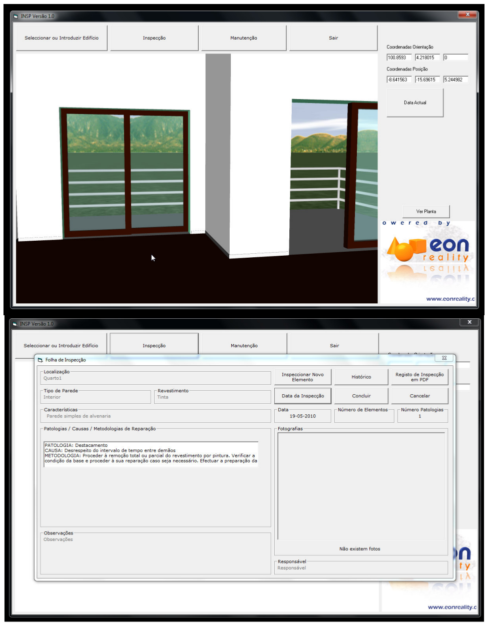
An inspection to the building local is usually report only as an anomaly survey and based only in paper sheets. The degree of deterioration is reported and some photos are usually obtained to support the diagnostic. After that a management plane is established. It includes the type and quantity of rehabilitation work and an estimated cost. Frequently it is realized at office by other collaborators. The presented virtual model support:

- An interactive identification of the element in analyses over the 3D model (the type of material applied in the element, the geographic orientation and the characteristics of the environment are automatically associated to the selected element);
- Consequently the anomalies that, in the database, were linked to that closure material type are listed in the PC of the collaborator, in the building place. In addition a predict cause is also listed. It supports the collaborator to choose the most adequate anomaly;
- In addition, an adequate solution and repair methodology are associated to the selected anomaly. An estimated cost is also provided and digital photos can be taken are inserted in the VR inspection sheet.
- A final report of the inspection visit to the local can be then obtained included the anomalies observed in all the analyzed building façades.

So, the benefits of using this VR model are evident. The database includes an oriented, vast and accurate knowledge (support in specialized bibliography) concerning the identification of anomalies and the most adequate rehabilitation work. The VR help the specialist to link the survey anomaly to a specific part of the building and allows a better communication (supported in visualization and interaction) with the building owner and other partners.

V. THE INTERIOR WALLS

An identical analysis of the characteristics, anomalies and repair works concerning the interior finish materials was carried out. With this information a database was created (Fig. 8).



Localização: Quarto1

Tipo de Parede: Interior

Revestimento: Tinta

Características: Parede simples de alvenaria

Data: 19-05-2010

Número de Elementos:

Número Patologias: 1

Patologias / Causas / Metodologias de Reparação

PATOLOGIA: Destacamento
CAUSA: Desrespeito do intervalo de tempo entre demãos
METODOLOGIA: Proceder à remoção total ou parcial do revestimento por pintura. Verificar a condição da base e proceder à sua reparação caso seja necessário. Efectuar a preparação da

Fotografias: Não existem fotos

Observações: Observações

Responsável: Responsável

Figure 8. Painted coating of interior walls of the building and the inspection interface.

Another inspection interface was defined for the painted surfaces of interior walls, also associated to repair solutions and corresponding methodologies of rehabilitation.

In addition the model identifies the period of time between the application of new paint and the predicted time when the next paint will be needed. The color changes between white (new) and red (when an area needs to be painted again, Fig. 9.

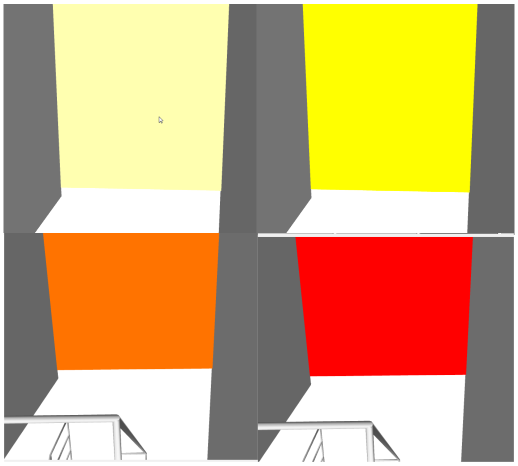


Figure 9. Changing colours related to the maintenance of the painted interior wall.

ISBN: 978-1-61208-129-8

The data of periodic on-site inspections included in preventive maintenance is taken into account for each monitored element of the interior wall (Fig. 10). Thus, when the date of interaction with the prototype is compared with the date predicted for the new paint application, the correspondent RGB (red, green and blue) values are calculated.

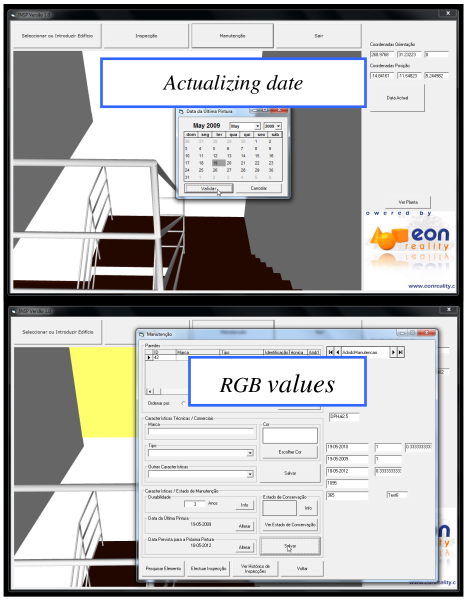


Figure 10. Changing date of interaction and the correspondent RGB values.

Different RGB values generate different color, using this virtual model. Therefore, by using data visualization supported by this VR technology, it is possible to estimate if the wall needs painting immediately or not

VI. CONCLUSIONS

A VR model to support the maintenance of walls in a building was developed within a research project. It enables the visual and interactive transmission of information related to the physical behavior of the elements, defined as a function of the times variable. The model shows the characteristics of each element of the building in the model and the information related to inspection, anomalies and repair works. As the 3D model is linked to a database in an interactive environment and has a friendly interface to deal with this knowledge, it allows a collaborative system. The work is still in progress. With this application the user may fully interact with the program referring to the virtual model at any stage of the maintenance process and analyze the best solution for repair work. It can also support the planning of maintenance strategies. The developed software is easy to handle and transport for on- site inspections and comprises information of the causes, solutions and methods for repairing.

ACKNOWLEDGMENT

The authors gratefully acknowledge the financial support of the Foundation for Science and Technology, a Governmental Organization for the research project PTDC/ ECM/67748/ 2006, now in progress.

REFERENCES

[1] A.Z. Sampaio, M.M. Ferreira and D.P. Rosário, “Interactive virtual application on building maintenance: The lighting component”, Proc. IRF2009, 3rd International Conference on Integrity, Reliability and Failure: Challenges and Opportunities, Symposium Visualization and human-Computer Interaction, Porto, Portugal, July 20-24, 2009, abstract pp. 221-222, paper 11 pgs.

[2] Introduction to working in EON Studio. EON Reality, Inc. http://www.eonreality.com/ [Accessed May 03, 2010].

[3] M. Fisher, J,. Haymaker and K. Liston, “Benefits of 3D and 4D models for facility managers and AEC service providers", in book: 4D CAD and visualization in construction: developments and applications, A.A. Balkema Publishers. Lisse, pp. 01 – 32, 2003.

[4] Stanford University, Dept. of Civil and Environmental Engineering, Group 4D CAD Research, accessed on March 2011.. http://www.stanford.edu/group/4D/projects/projects.shtml,

[5] VTT Technical Research Centre of Finland, Building & Built Environment, http://www.vtt.fi/services/cluster6/index.jsp, accessed on March 2011.

[6] J Leinonen and K.A.A. Kähkönen, “New construction management practice based on the Virtual Reality technology”, in book: 4D CAD and visualization in construction: developments and applications, A.A. Balkema Publishers. Lisse, pp. 75 – 100, 2003.

[7] F. Petzold , O. Bimber and O. Tonn, “CAVE without CAVE: on-site visualization and Design Support in and within existing building”, Proc. eCAADe 07, 25th Conf. of Education and Research in Computer Aided Architectural Design in Europe, Frankfurt, Germany, pp. 161-168. 2007.

[8] A. Khanzode, M. Fisher and D. Reed, “Challenges and benefits of implementing virtual design and construction technologies for coordination of mechanical, electrical, and plumbing systems on large healthcare project”, Proc. CIB 24th W78 Conference, Maribor, Slovenia, pp. 205-212, 2007.

[9] A.L. Lindholm, ” A constructive study on creating core business relevant CREM strategy and performance measures”, in Facilitie journal, 26(7/8), pp. 343-358, 2008.

[10] H. N. Rad, “Visualisation of building maintenance through time”, Proc. of the IV'97, IEEE 1st International Conference on Information Visualisation, 308-314, 1997.

[11] F. Khosrowshahi and E. Banissi, “Visualisation of the degradation of building flooring systems”. IEEE, Information Visualisation & Computer Society, pp. 507-513, 2001.

[12] A.Z. Sampaio and P.G. Henriques, “Building activities visualized in virtual environments”, Proc.eCAADe 07, 25th Conf. of Education and Research in Computer Aided Architectural Design in Europe, Frankfurt, Germany, pp. 85-89, 2007.

[13] A.R. Gomes, “Virtual Reality technology applied to the maintenance of façades”, Integrated Master Degree Thesis in Construction, TU Lisbon, Portugal, 2010.

[14] A.M. Gomes and A.P. Pinto, “Didactic text of construction materials”, Technical University of Lisbon, IST, Lisbon, Portugal, 2009.

[15] C. Lopes, ” Colour anomalies in painted closure of exterior walls”. Integrated Master Degree Thesis in Construction, TU Lisbon, Portugal, 2008.

[16] L. Ferreira, J. Coroado, V. Freitas and I. Maguregui, “Causes of the fall aplied in exterior walls (1850-1920)”, Conf. Patorreb, 3º Meeting of Pathologies and Rehabilitation in Buildings, Porto, Portugal, 2009.

[17] M. Veiga and S. Malanho, ”Closure in natural stone: Methdology of diagnostic and rehabilitation”, Conf. Patorreb, 3º Meeting of Pathologies and Rehabilitation in Buildings, Porto, Portugal, 2009.

Cyber Physical Multimedia Systems: A Pervasive Virtual Audio Community

Markus Duchon, Corina Schindhelm, Christoph Niedermeier
Siemens AG, Corporate Technology CT T DE IT 1, Otto-Hahn-Ring 6, 80200 Munich
{*markus.duchon.ext, corina.schindhelm, christoph.niedermeier*}@*siemens.com*

***Abstract*—In recent years, pervasive systems have gained importance in the context of home automation. Social online communities are becoming more and more popular, and indoor positioning techniques have made considerable progress. Thus, creation of virtual environments integrated with the physical world that enhance the user's perception and cognition, becomes feasible. In this paper, we propose a system combining these concepts into something we call a cyber physical multimedia system. This technology leverages pervasive audio communities that facilitate social activities for people with limited mobility, such as the elderly or handicapped.**

***Keywords*-Social Networks, Multimedia, Pervasive Computing, Indoor Positioning, Cyber Physical Systems;**

I. Introduction

Nowadays being virtually connected with friends, business partners, or items of interests has already become an inherent part of our lives. Via Twitter, Facebook, Ebay and Amazon, we follow news feeds of our favorite musicians or political happenings, we keep in contact and chat with friends and we go shopping while sitting on the couch, respectively. The web has opened a new world of interconnection and communication; however, only the young and computer savvy generation is able to grasp its benefits. People with limited mobility, such as the elderly or handicapped, would benefit greatly from social communities due to their naturally tends to suffer from isolation, yet they have only few possibilities because of their lack of computer knowledge and/or lack of special in-/output assistance for those features. For Germany, the demographic forecast shows that by the year 2035 more than two-thirds will be over 60 years old. Many projects (e.g. Smart Senior) are based on this fact and its resultant problems, such as rising health care costs. The main goal is to create comprehensive cyber physical systems that allow elderly people to stay in their home environment as long as possible, for example, by enabling remote healthcare or community platforms for friends and family to prevent isolation. Cyber physical systems (CPS) [1] are physical and engineered systems whose operations are monitored, coordinated, controlled and integrated by a computing and communication core. We will extend this by the term multimedia reinforcing the utilization of audio communication to a Cyber Physical Multimedia System (CPMS).

In this paper, we will focus on the design of a CPMS, which enables handicapped and elderly people to communicate in an easy way. The remainder is organized as follows. In Section II, we will introduce a scenario and detail basic requirements we consider indispensable for a suitable CPMS enabling a pervasive virtual audio community. Our system architecture is presented in Section III. Section IV discusses related work before Section V concludes the paper.

II. Requirements

Before discussing requirements of a CPMS, we will outline an application scenario. Imagine three different apartments like illustrated in Figure 1 of Peter, Mary and Paul. They agreed on a coffee party within their own kitchens. Peter and Mary are already in their kitchens and have a conversation about the latest news, whereas Paul resides in another room at that time (A). As soon as he enters his own kitchen, it is recognized by the system and an audio channel is opened to a virtual kitchen where he participates in the conversation of Peter and Mary (B). Seeking for a CPMS to enable a distributed, virtual audio community, we must consider some basic requirements a proper solution has to fulfill.

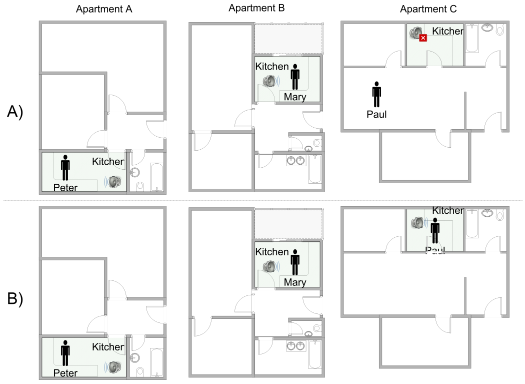


Figure 1. Three different apartments forming a virtual audio community

A. System Configuration

The system should provide different levels of configuration mechanisms. This is necessary because such a complex system as the proposed virtual audio community, in some cases, requires fine-tuning involving a considerable amount of user interaction whereas during normal operation it should be possible to configure basic behavior in a simple and intuitive way. The system must provide a detailed configuration

ISBN: 978-1-61208-129-8

interface that could be accessed via a Web browser. This kind of interface may be used for fine-tuning, in particular during initial configuration of the system for a user that has just joined the community or wants to perform major changes in his configuration. For instance, the web-based configuration interface may be used for specification of a mapping function for a physical actor to be integrated into the virtual environment. The system must provide a basic configuration interface that can be controlled via voice commands. This kind of interface may be used to change settings during normal operation of the system. For instance, a user could temporarily map a corner of his living room to the virtual kitchen by issuing a respective voice command.

B. Physical Actors

The system may integrate with the physical environment of the real rooms which are mapped to virtual rooms of the pervasive audio community wherefore some requirements are essential. Integration of physical actors into the virtual environment must be restricted to certain contexts. The virtual actor should only have an effect on real actors belonging to a community user if the respective user has authorized and activated a corresponding context. For instance, lighting control should only occur in a particular room if the user has actually entered the corresponding virtual room. Integration of physical actors requires a mapping between states of the physical actors and the virtual actor associated with them. For a simple actor that just has an ON and an OFF state, this is trivial. But, if a whole spectrum of states (e.g. light intensity or volume of background music) is to be addressed, a mapping function is required. When a user leaves a virtual room, the actors in the real room must be restored to their original states (i.e. the states that they had before the user entered the virtual room) unless the user explicitly decides otherwise.

C. Audio System

Since we want to convey the impression that physically separated people are together in an audio community the audio system should support surround sound capabilities for each room. Since the system should not be restricted to a single room but rather provide coverage for several rooms or even a complete apartment, several input and output devices are necessary. These resources need to be connected and individually responsive to an audio management component. A dynamic change of audio components in terms of volume adjustment and device handover is required to provide an accurate impression when moving around. Also, the exchange of audio communication data among remote participants has to be achieved especially for conference-like community communication.

D. Maps and Indoor Positioning

The system must be able to map physical rooms onto virtual rooms and also be able to find the current position of the user. Therefore, two aspects have to be considered:

Maps of the home environment must be available in digital form and context must be added to these maps. The maps serve as a basis for the positioning system, and the additional context is necessary to identify a room e.g., as living room etc. making a mapping to virtual rooms possible.

A positioning system must be installed. The higher the precision of this system, the better the features developed for it will be. Room level accuracy is the minimum accuracy the system requires. WLAN positioning systems offer a precision around 1-2 meters, which would be sufficient to determine with high probability in which room the user resides. If high precision positioning is available, for example with ultra-wideband (UWB), special effects are possible. As an example, we would like to describe sound fading. Imagine a user group in the virtual room ”kitchen” standing around the stove. One user physically moves away from this stove but doesnt leave the room. The voices of the remaining users would become softer out of the speakers, which creates a more realistic experience.

III. Architecture

Driven by today’s ambient assisted living approaches our system utilizes several mature technologies and combines them to a CPMS enabling a `Pervasive Virtual Audio Community` for elderly, handicapped or disabled people to enhance their participation in social life including a realistic experience. An overview of our system architecture is illustrated in Figure 2. Because our goal is to facilitate a community among physically separated users, a platform represented by a `Community Server` (CS) that provides virtual rooms and offers community capabilities is used. Connected through broadband internet, the platform can be accessed by several `Home Gateway Servers` (HGS), which represent the basis on the user side and are able to control and sense activities within flats. To ensure the user’s privacy the proposed system has to be activated and deactivated explicitly by the user. Several physical actors are connected to the gateway and can be controlled regarding the community context - imagine light, window shutters or coffee makers. The required hardware, speakers and

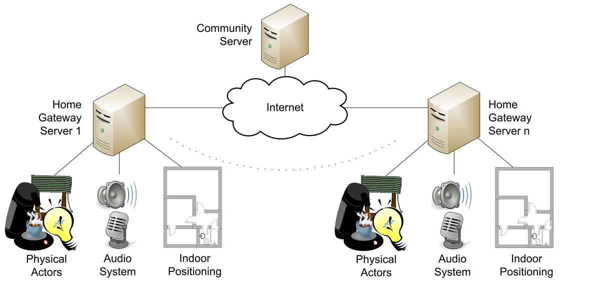


Figure 2. System architecture for a pervasive virtual audio community

microphones, for audio system is deployed and also linked to the control entity. Since the system heavily depends on the current location of the users moving around in their homes, a fine-grained indoor positioning system is utilized.

A. Community Server

The CS provides virtual rooms where participants can exchange audio signals in the first step. The main idea is sketched in Figure 3, whereby the kitchens of three different flats are unified in a virtual kitchen. These virtual rooms represent certain areas of a habitation where multiple users can participate similar to a telephone conference. Thereby the CS needs to provide certain group functionalities in order to provide multiple virtual kitchens for several independent circles of friends as well as user registration capabilities.

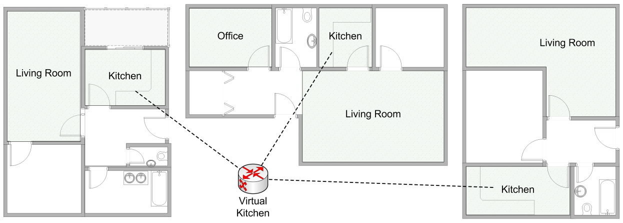


Figure 3. The CS provides a virtual room (kitchen) to which the real kitchens in several flats are interconnected through their HGSs.

B. Home Gateway Server

Within our system the HGS has to fulfill several tasks, which will be explained in turn.

Interaction and Communication: Considering our target group the handling during operation should be kept as simple as possible wherefore our system utilizes voice commands. These commands can be easily defined and extended following a simple syntax: `System: <command>`. For example, the user can use the command `System: start/stop` to start or stop the community system. This command structure can also be used for simple system configuration tasks. For example, the command `System: map to living room` can be used to set the current room the user is in, to the living room. After the recognition of a command the gateway maps the corresponding orders into machine readable commands by applying a ruled based approach. Audio communication within the community is crucial, since we want to achieve interaction with other users. We decided to utilize audio communication in the first step wherefore a VoIP system is sufficient for the desired functionality. The user's command `System: call kitchen` initiates a connection to the CS to join the virtual kitchen of that group the user is associated with. Once the registration process at the CS is completed an audio channel is set up to the specified virtual room. In addition to the transmission of audio signals, notification messages among remote participants need to be exchanged which include information that will be utilized by the connected HGS to adapt the local environment. For these messages we utilize the SUBSCRIBE request dialog [2], [3], which can be used for status notifications to remote users and therefore can be interpreted by the remote HGSs.

Physical Actors, Audio- and Positioning System: The HGS is also responsible for managing the user's environment by the interpretation of notification messages. According to the content of these messages a command is generated and passed to the corresponding home appliance. These messages are created according to the status of the real home appliances which are connected to their virtual counterpart. This will enhance the user's perception and cognition when participating in a pervasive virtual audio community. For instance, community users have attached their room lighting to a virtual lighting control system which than can be used to dim the lighting in the virtual room. As a consequence, the lighting of the real rooms mapped to the virtual room is dimmed accordingly. Another example is a coffee party that has agreed to meet in a virtual kitchen at a certain time. A community user starts all coffee machines mapped to a virtual coffee machine by starting his own machine so that coffee is ready when the other users show up in the virtual kitchen. Besides the adaptation of the physical environment we want our users to be informed about the communication status. Therefore, controllable status lights (i.e. small LEDs) are deployed in appropriate rooms. These lights indicate the remote activity in a specific virtual room by changing color and intensity.

The audio system mainly consists of several microphones and speakers deployed in the household, which represent different sources and sinks in terms of acoustic signals and can be accessed and controlled individually by the HGS. It should be emphasized that an actual voice transmission only takes place when the user is nearby. A user who leaves the kitchen physically, also leaves the virtual kitchen and therefore does not participate in a conversation anymore which takes place there. However, when a user enters the (virtual) living room, where another group resides, the HGS needs to switch to the corresponding room at the CS as well as change the microphone source and the speaker sinks to the user's new location. Furthermore, by using a more precise positioning system it is possible to adjust the surround volume at the remote sides considering the user's position in the real room in terms of a cardinal direction. As highlighted, the audio system needs to be highly coupled to a positioning system. Whenever the user changes his location, the HGS takes action: a) to inform remote gateways about the activity change (status lights, position update) and b) switching the communication in-/output in accordance with the user's physical location. The coupling is performed by the HGS, which holds and manages an indoor map, an overlay map of the defined virtual rooms, and an overlay map for the audio components and their corresponding volumes. The better the precision of the positioning system, the more fine-grained the audio control performance is, the more realistic a virtual conversation will be.

IV. Related Work

In the following, we will outline the state of the art of technologies and subcomponents that are part of virtual audio communities.

A. Physical interaction

Regarding physical control, so called home automation servers are used for processes like adjusting the heating according to the current temperature or adjusting the lighting system as well as the background music either by voice commands or proactively by context recognition. Lots of attention has been put on voice controlled systems [4]–[6] to facilitate their utilization. Also several communication standards and network protocols are widely used. An overview of relevant standards is given in [7].

B. Audio Communities

Prominent representatives like Skype or, GoogleTalk are also capable of conference calls, but are inconvenient to operate, require direct computer interaction and lack of spatial perception of human interaction. Healy et al. [8] present a prototype of audio spatial augmentation headphones in order to take advantage of the innate psycho acoustical perception of sound source locations. Hyder et al. [9] outline the difficulties to identify the speaker of a teleconference and present a solution which adds a virtual acoustic room simulation. Kim et a. [10] propose a 3-dimensional VoIP system for two user groups, whereby the participants can hear the voice of remote users as if each remote user speaks at his or her corresponding position.

C. Positioning Systems

One key information is the position of the user within his home environment. When there is only one person living in a household, the identification of the person is not actually necessary, and as a result, positioning could be accomplished via smart home equipment, such as motion sensors in each room or contact sensors on doors. However, when more than one person lives within a household, the identification of the person using the system is essential. A vast number of positioning technologies and systems exist which support different levels of accuracy. An overview is provided in [11]. If radio-based technologies are used, the user needs to wear or carry a tag. For room level accuracy, WLAN positioning is sufficient and can be enhanced by adding motion sensors. If a higher level of accuracy is needed, UWB systems enable positioning in the range of centimeters.

V. Conclusion

As for now, most of the systems dealing with pervasive computing and home automation put the focus on an easy interaction with the user, by utilizing gesture control, voice control, or on fully automated approaches which analyze the user's context in order to support him in his daily life. As a result, only one facility is addressed which is inhabited by one or more users. Our approach enhances pervasive systems by adding the community aspect. This is achieved by adopting a mixed-reality approach that associates several real and a single virtual environment. At the current stage of implementation we identified future research directions and possible enhancements. The system could be extended by integrating augmented reality aspects both acoustically and optically. This could further enhance the user experience and intensify the illusion of actually spending time together rather than communicating via the internet. This could mean a significant improvement to the lives of people not able to leave their homes and thus being subject to social isolation.

References

[1] E. A. Lee, "Cyber Physical Systems: Design Challenges," in *ISORC*, 2008.

[2] G. Camarillo, A. Roach, and O. Levin, "Subscriptions to Request-Contained Resource Lists in the Session Initiation Protocol (SIP)," RFC 5367 (Proposed Standard), IETF, October 2008.

[3] A. B. Roach, "Session Initiation Protocol (SIP)-Specific Event Notification," RFC 3265, IETF, June 2002.

[4] I. Mporas, T. Ganchev, T. Kostoulas, K. Kermanidis, and N. Fakotakis, "Automatic Speech Recognition System for Home Appliances Control," in *PCI*, 2009, pp. 114–117.

[5] J. Zhu, X. Gao, Y. Yang, H. Li, Z. Ai, and X. Cui, "Developing a voice control system for ZigBee-based home automation networks," in *IC-NIDC*, 2010, pp. 737–741.

[6] A. Gárate, N. Herrasti, and A. López, "GENIO: an ambient intelligence application in home automation and entertainment environment," ser. EUSAI, 2005, pp. 241–245.

[7] W. Kastner, G. Neugschwandtner, S. Soucek, and H. M. Newmann, "Communication systems for building automation and control," vol. 93, no. 6, pp. 1178–1203, 2005.

[8] G. Healy and A. F. Smeaton, "Spatially augmented audio delivery: Applications of spatial sound awareness in sensor-equipped indoor environments," in *MDM*, 2009, pp. 704–708.

[9] M. Hyder, M. Haun, and C. Hoene, "Placing the participants of a spatial audio conference call," in *CCNC*, 2010, pp. 1–7.

[10] C. Kim, S. C. Ahn, I.-J. Kim, and H.-G. Kim, "3d voice communication system for two user groups," in *ICACT*, 2005, pp. 100–105.

[11] Y. Gu, A. Lo, and I. Niemegeers, "A survey of indoor positioning systems for wireless personal networks," *Communications Surveys & Tutorials, IEEE*, pp. 13–32, 2009.

Queue-based scheduling for soft real time applications

Fabrizio Mulas
University of Cagliari
Cagliari, Italy
Email: fabrizio.mulas@unica.it

Salvatore Carta
University of Cagliari
Cagliari, Italy
Email: salvatore@unica.it

Andrea Acquaviva
Politecnico di Torino
Torino, Italy
Email: andrea.acquaviva@polito.it

***Abstract*—Modern multitasking multimedia streaming applications impose tight timing requirements that demand specific scheduling policies. General purpose operating systems such as Linux (widely diffused even in embedded systems) are not specifically designed for such applications as they must ensure an overall performance level for a wide range of user processes. Realtime versions of general purpose kernels can be used, however since they are designed for hard real-time applications, they impose explicit knowledge of deadlines for all tasks composing the application to set their priorities.**

In this work a novel streaming-oriented scheduling algorithm is proposed, that relies on a standard interprocess communication support for applications composed by multiple pipelined stages communicating by means of message queues. It determines the scheduling order depending on the queue occupancy, for this reason does not require explicit deadline information. It has been developed in Linux OS as a new real time policy, showing that it is relatively easy to integrate in it and, worthily, it does not require modifications of existing applications.

***Keywords*-scheduling; Linux; soft realtime; multimedia streaming.**

I. Introduction

Multimedia applications are increasingly complex and demanding in terms of both computational power and time constraints. A significative example is given by the increasing resolution and frame rate requirements of video streaming applications. When these applications run on top of a general purpose operating system their requirements become very challenging. Indeed, these OSes are currently used in system with demanding networking capabilities, where multiple network flows must be managed. This is true not only for desktop PCs, but also in embedded networking systems such as media gateways, where general purpose OSes are widely used for cost and flexibility reasons. Besides typical network processing, these systems must perform various general purpose processing at line rate such as video decoding, video transcoding, image processing and encryption. In general purpose OSes, the scheduler is not specifically designed for handling real-time requirements even if a standard real-time support does exist in well known general purpose OSes such as Linux or Windows. However, this support is not enough to fulfill the application requirements, basically consisting on giving, to a process defined as “real-time”, a static priority higher than any other “conventional process”.

Current multimedia applications are composed by a cascade of multiple dependent tasks communicating by means of message queues. For instance, a H.264 decoder is composed by several steps including motion compensation, entropy decoding, dequantization, inverse Discrete Cosine Transform (DCT). Furthermore, multimedia frameworks such as GStreamer create complex multimedia applications by chaining several stages [1]. In both cases, the frame rate (i.e., QoS) requirements are backward propagated from the last stage to the previous ones. A general purpose scheduler, such as the Linux one, is not aware of task dependencies and timing constraints, but only looks at how much a task is demanding in terms of CPU utilization.

The “conventional process” scheduler is designed to promote the so called I/O bounded applications, by giving them a high dynamic priority. These are characterized by small (compared to the timeslice) CPU bursts interleaved to large I/O access periods. CPU bounded ones, instead, are characterized by much larger CPU bursts, and thus are given a smaller dynamic priority. This is because I/O applications are supposed to interact with the user and hence the OS attempts to reduce their latency. On the other side, the real-time process scheduler in Linux implements either a FIFO or a Round-Robin policy. Both of them, as it is going to be shown in this paper, do not take into account actual requirements of tasks, leading to QoS degradation especially in high CPU utilization conditions.

An additional limitation of general purpose OSes arises in presence of multiple real-time applications running simultaneously, as in the context of media gateways, where several streams need to be decoded at the same time to feed multiple network connections. Here the computational power must be allocated to multiple decoding applications having heterogeneous QoS requirements, such that all they perceive a degradation proportional to their QoS requirements. This can be hardly achieved using general purpose OSes that lack the concept of fairness related to the QoS.

Putting it all together, general purpose schedulers are not longer suitable to modern multimedia applications ([2], [3]). Nevertheless, they are still common in Windows family, Linux, and all other variants of Unix such as Solaris, AIX and BSD (see [4] for further details).

An alternative solution would be to adopt hard real-time schedulers, that are specifically developed for scenarios where deadlines must be strictly respected (e.g., life-safety critical applications). The counterpart is that they are hard to manage and require to explicitly provide the scheduler with timing constraints of applications (i.e., deadlines) that must be hence modified accordingly. There are situations where the requirements about the deadlines are not strict, that is, a certain amount of them can be tolerated (for example, in multimedia streaming): it is the case of soft-realtime applications.

In this paper a variant of the Linux scheduler is proposed, called *queue-based scheduler* (QBS) that deals with soft real-time streaming applications composed by multiple pipelined stages. QBS is inherently aware of QoS requirements of multitask applications similarly to real-time schedulers, but does not require application modifications, as general purpose ones. In order to achieve this goal, it monitors the intertask communication, thus requiring the instrumentation of the communication and synchronization library.

QBS implicitly assumes that applications are composed by multiple pipelined stages that communicate by means of queues of messages. Such applications follow a data-flow paradigm, where tasks continuously process frames arriving in their input queue and produce frames on their output queue for the next processing stage. Figure 1 shows an example of such paradigm (H.263 decoder). Most modern multimedia applications are realized in such a manner (e.g., audio/video decoders). The application output queue is read at fixed time intervals (by a *consumer*) and if it is found empty a deadline miss occurs.

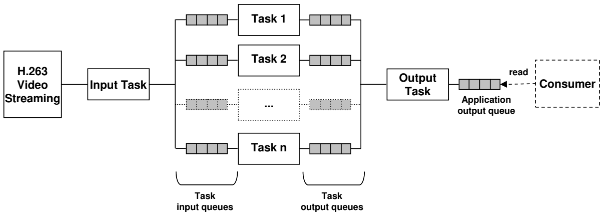


Fig. 1. Pipelined multi-stage application scheme (H.263 Decoder)

The main idea behind QBS is to monitor the queue occupancy level of all queues in the system and to take scheduling decisions based on this information. Basically, QBS seeks the emptiest queue in the system and schedules the process or task writing into it (given that it is in *running* state). Thus QBS can quickly react to situations that may lead to deadline misses, exploiting the feedback from the queues.

In the considered application model, QoS is preserved as long as there are data items available in the application output queue (that is, the last queue of the application) when they are needed by the final consumer stage. This leads to two very important considerations. First, the application output queue can be even empty in some periods of time without necessarily having misses (that is, there is not a miss if the output queue is empty when the consumer does not read data from it). Second, in general intermediate stages have less stringent timing requirements (because they do not generate misses directly) .

The queue feedback approach ensures a more effective CPU time allocation to each task, based on its real and actual QoS requirements. From a practical point of view, the occupancy level of the output queue of a task is used as a measure of its CPU utilization needs. A deep explanation of that, together with a detailed description of the proposed algorithm, is provided in Section III. To test its effectiveness, the scheduler has been implemented inside the Linux OS and the standard System V message queue library has been instrumented to support monitoring features. Thanks to this implementation, various sets of experiments have been carried out, using multiple video decoding applications. Experiments compare the deadline miss rate of QBS w.r.t. both default real-time and conventional process scheduler in case of single and multiple decoding applications having heterogeneous QoS requirements. Results demonstrate that QBS improves the deadline miss rate in high CPU utilization conditions and provides better CPU resource allocation, that is, proportional to frame rate requirements.

The rest of the paper is organized as follows: Section II describes related work in the area of scheduling for real-time and multimedia applications. Section III full details the QBS algorithm, Section IV explains why Linux has been chosen as testbed platform, Section V describes the implementation while Section VI shows the experimental results. Section VII concludes the paper.

II. Related Work

In literature many approaches have been proposed to manage soft real-time applications in commodity OSes. [5] performs a deep evaluation of how clock interrupt frequency influences the response time of multimedia applications. Their study aims at helping tuning existing schedulers. Similarly, other techniques as soft timers [6], firm timers [7] and one-shot timers have been proposed to significantly enhance response time. However, none of them proposes a new scheduler algorithm but rather latency reduction techniques.

On the other side, many real time schedulers have been proposed. SMART [8] is a scheduler for multimedia and real time applications implemented in UNIX-like OSes. It schedules real time tasks even trying to avoid the starvation of conventional processes, but it requires deep modifications of existing applications. In fact, applications have to communicate their deadlines to the scheduler, which can also return feedbacks to enable some proactive countermeasure (e.g., remodulate their workload in order to meet the deadline). On Linux, some examples are Linux/RK [9], RTE-Linux [10], Linux-SRT [11] and RTLinux [12]. These all have the same general drawbacks of real-time schedulers (i.e., programmers must use a dedicated interface to exploit these services). Other approaches explicitly require user intervention to specify the needs (in terms of priority) of the processes or of a class of processes (e.g., multimedia applications) [11] [13].

The algorithm proposed in this paper (QBS) provides QoS sensitive scheduling without requiring explicit user awareness and modification of existing applications, given that they follow the message queue paradigm. As mentioned in the introduction, this model adheres with the one of modern multimedia applications and frameworks.

III. Queue-based Scheduling Algorithm

The idea behind the proposed algorithm is to exploit the level of the interprocess communication queues as indication of task requirements and consequently to grant CPU time proportionally to that. To better explain that, let us consider a simple example of two applications, A and B, with a CPU need of 65% and 55% respectively (that is, the system is overloaded). Running them in a standard operating system, without any knowledge of application requirements, A and B will receive more or less the same treatment (i.e., about 50% of CPU each), thus A will experience a worse QoS with respect to B. From the point of view of the queues, those of A will be more empty, in average, than those of B. Instead QBS monitors all queues in the system and tries to level them. As a consequence, comparing to the previous case, A will receive more CPU time than B, thus reducing the QoS gap between the two applications (i.e., A will have less deadline misses than before and B a little more than before) and assuring a CPU time sharing proportional to their needs (i.e., both applications will be penalized in a proportional manner rather than in the same way). Furthermore, it is worth noting that QBS, exploiting the feedback from the queues, is able to quick react to situations that potentially lead to deadline misses. For example, if a queue suddenly becomes empty, QBS notices that and properly reacts to fill it.

Algorithm 1 describes how QBS functions. Let Q_n be the nth queue, Q_{Ln} be its level (by definition, Q_L is an integer non-negative number) and let N be the total number of queues in the system, at any moment. Let T_n be the last scheduled time of Q_n's producer. QBS basically finds the most empty queue in the system and schedules the task that writes in it (the producer). Note that in the paradigm used each queue has only one producer and one consumer. If as a result of the search two or more queues are found at the same minimum level, QBS chooses the oldest scheduled producer, that means the process that less recently has been executed in CPU. The $scheduleProducerOf()$ function schedules the producer of the queue passed to it as argument.

Algorithm 1 Queue-based scheduler algorithm

```
Every decision instant do:
$Q_{min} = Q_1$
$T_{min} = T_1$
**for** $n = 1$ to N **do**
    **if** $(Q_{Ln} < Q_{Lmin})\ OR\ (Q_{Ln} = Q_{Lmin}\ AND\ T_n < T_{min})$ **then**
        $Q_{min} = Q_n$
        $T_{min} = T_n$
    **end if**
**end for**
$scheduleProducerOf(Q_{min})$
```

The last point to analyse is how frequently QBS should be executed. There is clearly a trade-off here, indeed: choosing a high frequency achieves a better leveling of the queues, but, on the other hand, it increases the number of context switches, thus causing a higher overhead. Thus, it has been chosen to maintain the concept of Linux *timeslice*: every process can consecutively use the CPU till a maximum amount of time (i.e., the timeslice), at the end of which the scheduler is called and the current process (most of the times) is preempted and another one is scheduled.

A. QBS Complexity

The algorithm's complexity is related to the need of scanning all queues in the system to find the most empty one. Thus QBS would have a linear complexity, that is $O(n)$ (where n is the total number of active queues in the system). Given that the scheduler is called very frequently, it is mandatory to reduce its complexity as much as possible. Then it has been reduced to $O(1)$, that means it no longer depends on the number of the queues. This result has been achieved adopting a special data structure to keep trace of all queues and considering that, at any moment in time, the only ones that could change are those read and written by the task currently in execution. So, when the scheduler is invoked, it quickly updates in the structure the information about the only queues that could have been changed. Hence, the time taken for this operation is constant ($O(1)$). The details of how it is implemented are described in Section V

IV. Testbed System Description

QBS has been implemented in Linux 2.6, thanks to its open source nature and widespread diffusion. Indeed it is used in desktop PCs, many server systems (e.g., web, mail, dns, routers, etc.) and, recently, in mobile platforms too. One of the most notable examples of that is probably Android [14], the Google OS for smartphones, based on Linux and widely thought to reach a leading position in the market very soon. QBS aims at be adopted in above systems and even in small/medium multimedia servers (e.g., audio/video on demand, voip, etc.), where expensive high specific solutions (e.g., real time OSes) are not affordable and commodity operating systems are the usual choice. Thus, in all these systems the standard Linux scheduler is adopted. For all these reasons it has been decided to compare QBS versus Linux standard policies. The following section (IV) details these policies.

Linux standard distributions come with three policies (some slight variations are possible depending on kernel versions, but they are basically the same): SCHED_NORMAL, SCHED_RR and SCHED_FIFO. The first one is the default policy for all tasks. It is a relatively complex algorithm that deals with conventional processes (i.e., not real time processes). It continuously attempts to identify interactive applications from CPU intensive ones, using the common mechanism (common to many OSes) described early (in the Introduction): processes that spend most of their time waiting for I/O operations are supposed to be interactive, while those that heavily exploit the CPU fall in the second category.

Then the scheduler grants more priority to the interactive ones, in order to reduce their latency. Unfortunately nowadays interactive multimedia applications are CPU greedy too, thus they are penalized by this mechanism ([2], [3]). For this reason this policy is not adequate for managing modern CPU-demanding interactive applications (this is demonstrated in Section VI-C).

SCHED_RR and SCHED_FIFO are both real time algorithms: basically the former (round robin policy) equally shares the CPU time among tasks, while the latter (fifo policy) grants all CPU time to the first arrived process as far as it uses it, after that it schedules the next task in the FIFO queue. Thus the last one, given its fifo behaviour, is not adequate for multimedia applications (it does not treat all processes fairly). Instead round robin (RR) performs quite well and consequently has been chosen as the main algorithm to confront against (Section VI-C). It must be noted that Linux real time policies are intended to manage soft real time processes. To specify a task as a real time one, the programmer needs only to state that using a system call. No any other modification is needed. Alternatively, the user can set it using the *chrt* linux command, without any modification to the application code.

V. QBS Implementation Details

This section describes how QBS has been implemented in a standard Linux kernel. In particular, all details are referred to kernel 2.6.20.16.

The Linux scheduler picks up the next task to be executed from the top of a specialized task queue. Thus, the main routine of QBS (i.e., the code that implements the algorithm and chooses the next task to be scheduled) is called just right before this choice, in such a manner to put the process selected by QBS on top of that queue. In this way, the standard Linux scheduler will find in it the task chosen by QBS.

In Section III the core algorithm has been described and in Section III-A it has been stated that its complexity is $O(1)$. All above has been accomplished using the structure showed in Figure 2. It is an array of simply linked lists, where MAX represents the maximum possible number of items in a queue (i.e., a System V message queue). Each element of the lists is a *queue identifier*, a special structure that points to an allocated queue. The key point here is that, at any moment in time, each element in the nth list (i.e., that at position n in the array) points to a queue that has n items in it (a that time). Thus the algorithm described in Section III is implemented in this way: it scans the array starting from 0 and selects the first element found. Hence, it points to the most empty queue in the system, as requested by the algorithm. Using this structure, the algorithm needs to scan at maximum MAX array items, resulting in a constant seek time (i.e., $O(1)$ complexity).

The queue identifier is composed by three fields: (i) lid is a pointer to a queue; (ii) $timestamp$ represents the last scheduled time of the producer of that queue; (iii) $next$ is a pointer to the next element in the linked list. It must be noted that in each list, all elements are ordered in a temporal way using the timestamp, from left to right, where on the left there is the oldest one. Hence this assures that the first element found during the scan of the array represents both the producer of the most empty queue in the system and, among all queues at the same level, the oldest scheduled one. This structure assures that the time spent for selecting a task is constant ($O(1)$), because it depends on neither the number of the tasks, nor the number of the queues.

This structure is updated every time the scheduler is called: as a further optimization, it checks only the queues modified by the last executed task and, if needed, moves the corresponding identifiers in the correct array position.

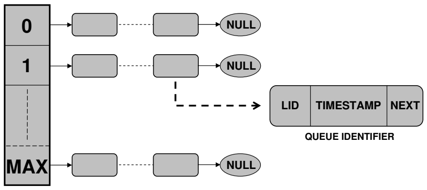


Fig. 2. Array of simply linked lists of queue identifiers

VI. Experiments

This section describes the experimental setup (Section VI-A), the objectives of experiments (Section VI-B) and finally the tests that have been performed (Section VI-C).

A. Experimental Setup

A dedicated machine has been set up for all experiments, equipped with a CPU Athlon XP 1100 GHz and with 512 MB of RAM. For the reasons explained in Section IV, the Linux standard round robin policy (SCHED_RR) is the primary algorithm QBS is compared against. Nevertheless, some comparisons versus the SCHED_NORMAL (conventional) algorithm have been performed too. Several experiments have been set up using many instances of two different applications, both following the message queue paradigm described early in Section I.

The first one, depicted in Figure 3, is composed by synthetic tasks (i.e., they perform some useless work). Each of the first three tasks puts data in its output queue, while Task 4 reads data from all its input queues, performs some elaboration, and puts the result in its output queue.

Instead the second application is a real H.263 decoder, already showed in Figure 1. The movie to be decoded is fully loaded in RAM before the start of experiments, in order to avoid possible bottlenecks reading it from the hard disk. Then the memory is locked to prevent swapping (that could alter the results). All these operations are done by the *Input Task* (see Figure 1), that then decomposes each frame of the video in *n* parts and puts them in the next proper queue. Each following task (*Task 1 to n*) elaborates the nth part of the frame. In the end, the *Output Task* reassembles the decoded frame, performs some elaboration and puts it in the application output queue.

It must be noted that both applications use the System V message queue library. All operations on queues (read and write) are blocking, that means if a process attempts to read in a empty queue or to write in a full one, it is suspended and automatically woken up as soon as this situation changes.

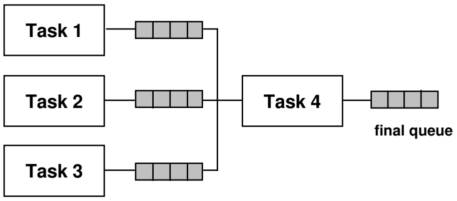


Fig. 3. Synthetic task application

B. Objectives

In the kind of applications that are being considered, an important metric to be taken into account is the quality of service (QoS). Indeed the output queue of each application instance is read at a fixed frequency, depending on the wanted frame rate, and if it is found empty a deadline miss occurs. The fewer the deadline misses, the greater the QoS, thus they should be as few as possible.

As above noted, the total QoS is a significant metric, nevertheless is not the most important one. Indeed, a more valued characteristic is its uniformity, both per and among applications. In order to explain that, consider the case where all instances are perfectly identical: it is not desirable to have a decoder that performs very well while another is working very bad, but rather to have all them with the same QoS level (uniformity among applications), at any time (uniformity per application, i.e., the performance of each application is constant in time). Generally speaking, considering application instances with different requirements, ideally each one should get a QoS proportional to its needs.

The experiments aim at demonstrating that QBS, with respect to standard Linux policies, is able to achieve a better total QoS, a better QoS performance uniformity (both per and among applications) and to provide a QoS proportional to application requirements.

C. Tests and Results

In order to compare the algorithms in real-world situations, a media server has been set up using many instances of the two decoders described before. Thus, many experiments with several instances of such applications running in parallel have been carried out, varying their parameters, as task workload, frame rate, and so on.

1) Synthetic Decoder: Some experiments with the synthetic application have been executed, comparing against both SCHED_RR and SCHED_NORMAL. Figure 4 shows the deadline misses (in percentage with respect to the total number of reads at the application output queue) versus the frame rate, running two application instances in parallel. The miss rate plotted is the average between the two values (note that each application has its own number of deadline misses). In these experiments QBS performs better than the others, having always less misses. Furthermore it sustains a higher frame rate without having QoS worsening (namely, 26.4 fps versus 25.8 fps for SCHED_RR and 22.9 fps for SCHED_NORMAL).

This experiments revealed that SCHED_NORMAL is not adequate for comparing versus QBS: indeed numerical results (not reported in the paper) show that there is a great gap of performance between the two application instances. For example, it can happen that one application has zero misses for a very long time while the other has 15% of it. This is because SCHED_NORMAL is not thought to deal with soft real time processes and furthermore it continuously tries to prioritize interactive tasks (this mechanism is described in Section IV). This is the reason (for fairness) why it has been chosen to not compare against it anymore .

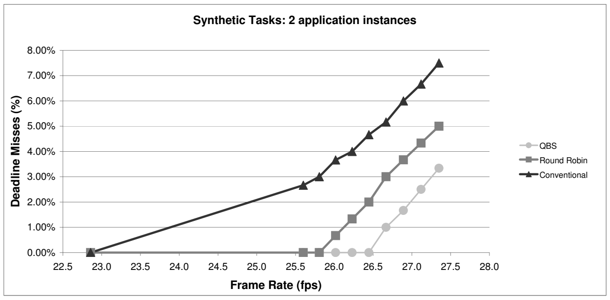


Fig. 4. Synthetic tasks: deadline misses versus frame rate

Using a debug monitoring infrastructure, the behaviour over the time of all queues have been carefully analyzed, observing that QBS is able to level them (in average) while RR shows great differences. It is possible to observe this behaviour in Figure 5. For example (RR case), some queues are totally full while others are completely empty. This suggests the idea that if a queue is always almost empty (in average) and another is in the opposite condition, probably the CPU time could be more fairly distributed (i.e., more CPU time than strictly needed is granted to the task which output queue is fuller). Instead QBS shows the capacity to better level all queues in the system, in average, suggesting a smarter CPU repartition among tasks.

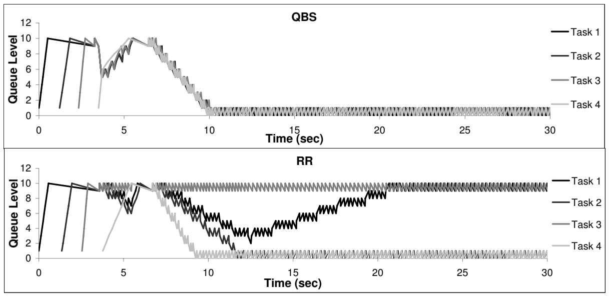


Fig. 5. Queue levels over time

2) H.263 Decoder: In the following examples both algorithms have been much more put under stress, using many instances of the H.263 decoder. Experiments have been carried

ISBN: 978-1-61208-129-8

out ranging from six parallel instances up to eighteen (note: in this set of experiments all instances are perfectly identical and the input file is the same). Figure 6 shows the deadline misses (in percentage) versus the frame rate for six applications (note: the value is the average among all applications). It is possible to see that QBS performs slightly better (similar results apply for the other above mentioned cases, that is with more than six decoders).

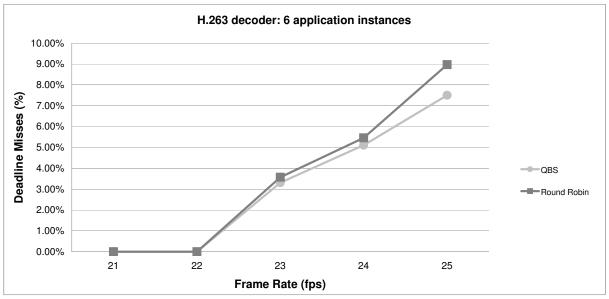


Fig. 6. H.263 decoder: deadline misses versus frame rate

However, the differences are really tiny. But, as it has been explained in Section VI-B, it is more important to assess the QoS uniformity. The two plots in Figure 7 show the miss percentage over the time for each application (eighteen decoders at the same fixed frame rate). Even if it is not possible to distinguish every single application, its aim is to display how QBS is able to much better level the QoS among decoders. Indeed the lines in the QBS plot appear closer each others. To numerically quantify this behaviour, the standard deviation of deadline misses among decoder instances has been calculated, at fixed interval times. The results reveal that standard deviation values in the RR case are roughly three times higher (the average values are 2.0 and 6.1 for QBS and RR, respectively). Thus, RR at any moment in time causes quite big differences among decoders, meaning that some applications are performing much better that others. Another important aspect, not clearly distinguishable from the plots, is that this not uniformity changes also in the time (for RR). That is, given a certain decoder, its QoS oscillates a lot over the time (this is not a desirable behaviour). This happens much less in QBS. Table I numerically points out that, showing the standard deviation of deadline misses (in percentage) of each decoder instance. It is worth noting that has been plotted the case with eighteen decoders, the most stressing one for the algorithm: with less instances QBS performs even better. Carefully observing the Figure 7 in the QBS case, it is possible to see a sort of periodic trend. This is due to three main reasons: (i) the workload varies from frame to frame, depending on their complexity; (ii) all decoders read from the same source file and their application output queue is read at the same instant, hence all tasks have a similar workload at any moment in time (with a certain flexibility due to the queues that intrinsically function as a buffer); (iii) it has been previously stated that for avoiding bottlenecks the video is full loaded in RAM before the starting of experiments, but due to memory space restrictions, a longer duration is simulated re-reading the same movie several times. The first two points explain why all decoders have always similar workloads and their variations over the time, while the last one justifies the periodic trend. To prove that an experiment similar to the above one has been executed, loading only one frame in RAM: RR continues to behave as before (as in Figure 7) while QBS, plotted in Figure 8, now shows a flat trend, without peeks and periodic shapes.

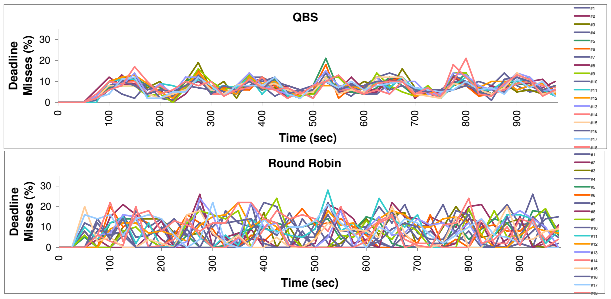


Fig. 7. Eighteen H.263 decoders: deadline misses over time

TABLE I
STANDARD DEVIATION OF EACH H.263 DECODER INSTANCE

	# Decoder Instance																		
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	Average
QBS	3.7	4.0	4.0	3.5	4.1	4.0	3.4	3.2	4.3	3.4	3.7	3.5	4.1	3.7	3.6	3.2	3.4	4.4	3.7
RR	7.1	7.7	6.3	6.2	5.8	6.0	6.5	7.0	6.9	6.0	6.9	6.0	6.9	6.5	6.2	5.9	6.2	7.4	6.5

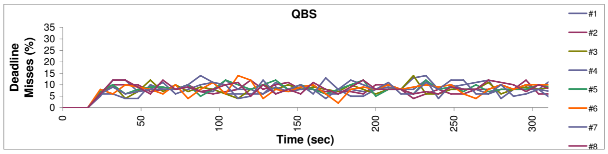


Fig. 8. Eight H.263 decoders: deadline misses over time

Previous experiments have been performed using several instances of the same decoder (either synthetic or real), with the same workload of internal tasks and the same frame rate. They aimed at more easily pointing out some characteristics of both algorithms. In order to assess their behaviour in real scenarios, where applications can have any possible combination of workload and frame rate, other experiments have been carried out, varying these parameters too. Plots in Figure 9 sketch the deadline misses over the time for a case in which there are twelve H.263 decoders at 10 fps and one at 20 fps, for each algorithm. RR causes a higher number of deadline misses in the faster instance (32.8% in total) while none of them in the slower ones (0.0% in total). This is because RR equally shares the CPU time among tasks, without knowledge of their requirements. That means that each decoder, being composed by the same number of tasks, receives the same slice of CPU time. Instead QBS is fairer, indeed observing the queues it recognizes that the faster decoder has a higher CPU need and grants it more CPU time. Hence QBS reduces the gap in QoS between the two application categories (with respect to the previous case), causing less QoS worsening in one case (10.9% in total) and more in the other one (2.95% in average among decoders).

In order to confirm this positive behaviour of QBS, other experiments have been realized, using eleven identical decoders all at the same frame rate, but with one of them with a much higher workload of its internal tasks (i.e., its tasks perform heavier elaboration). The results (not reported here) are very similar to the previous case, confirming such behavior.

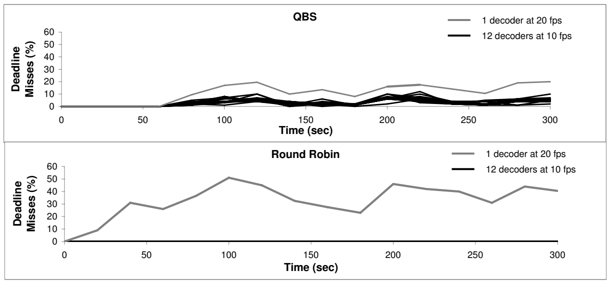


Fig. 9. Thirteen H.263 decoders with different frame rate: deadline misses over time

Finally, one last experiment has been set up, using twelve decoders with incremental workload: the second decoder has a higher workload than the first one, the third one a higher workload than the second one, and so on. Both algorithms show a step results among QoS of applications, as expected, but QBS distributes the performances in a more uniform manner (with respect to RR). Figure 10 plots the results whilst the numerical values are in Table II.

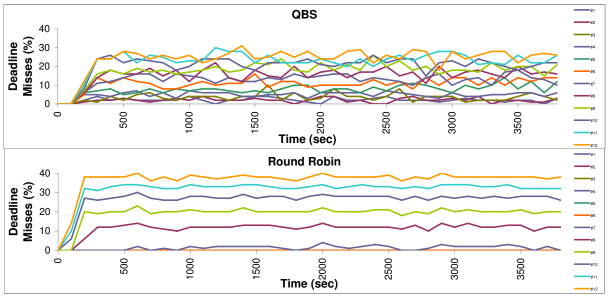


Fig. 10. Twelve H.263 decoders with incremental workload: deadline misses over time

TABLE II
TOTAL DEADLINE MISSES (%) OF EACH H.263 DECODER INSTANCE

	# Decoder Instance											
	1	2	3	4	5	6	7	8	9	10	11	12
QBS	2.1	2.0	3.5	4.9	7.8	10.7	12.8	15.3	17.7	20.5	22.6	24.0
RR	0.0	0.0	0.0	0.0	0.0	0.0	1.4	11.4	19.3	26.3	31.6	36.4

VII. CONCLUSIONS AND FUTURE WORK

Nowadays multimedia applications are widespread in several fields and there are many situations where they are executed in commodity operating systems, such as devices for playing audio/video or small/medium voip servers. General purpose OSes do not provide adequate support to them. The proposed scheduling algorithm (QBS) outperformed standard Linux policies, both in QoS and uniformity performance among application instances. QBS has been validated against various utilization scenarios, using both real and synthetic multimedia applications. Finally, it is relatively easy to integrate in a standard distribution and does not require any modification of existing applications.

We are working to further improve it in several ways, for instance experimenting priorities among queues. We also plan to extend it for multiprocessor systems.

REFERENCES

[1] GStreamer, "Gstreamer multimedia framework." [Online]. Available: http://www.gstreamer.net/

[2] J. Nieh, J. G. Hanko, J. D. Northcutt, and G. A. Wall, "Svr4unix scheduler unacceptable for multimedia applications," 1993.

[3] Y. Etsion, D. Tsafrir, and D. Feitelson, "Desktop scheduling: how can we know what the user wants?" in *NOSSDAV '04: Proceedings of the 14th international workshop on Network and operating systems support for digital audio and video*. New York, NY, USA: ACM, 2004, pp. 110–115.

[4] Y. Etsion, D. Tsafrir, and D. G. Feitelson, "Human-centered scheduling of interactive and multimedia applications on a loaded desktop," ,, Tech. Rep., 2003.

[5] Y. Etsion, D. Tsafrir, and D. Feitelson, "Effects of clock resolution on the scheduling of interactive and soft real-time processes," *SIGMETRICS Perform. Eval. Rev.*, vol. 31, no. 1, pp. 172–183, 2003.

[6] M. Aron and P. Druschel, "Soft timers: efficient microsecond software timer support for network processing," *ACM Trans. Comput. Syst.*, vol. 18, no. 3, pp. 197–228, 2000.

[7] A. Goel, L. Abeni, C. Krasic, J. Snow, and J. Walpole, "Supporting time-sensitive applications on a commodity os," in *OSDI '02: Proceedings of the 5th symposium on Operating systems design and implementation*. New York, NY, USA: ACM, 2002, pp. 165–180.

[8] J. Nieh and M. S. Lam, "The design, implementation and evaluation of smart: a scheduler for multimedia applications," in *SOSP '97: Proceedings of the sixteenth ACM symposium on Operating systems principles*. New York, NY, USA: ACM, 1997, pp. 184–197.

[9] S. Oikawa and R. Rajkumar, "Linux/rk: A portable resource kernel in linux," in *In 19th IEEE Real-Time Systems Sumposium*, 1998.

[10] Y.-C. Wang and K.-J. Lin, "Enhancing the real-time capability of the linux kernel," in *Real-Time Computing Systems and Applications, 1998. Proceedings. Fifth International Conference on*, Oct 1998, pp. 11–20.

[11] S. Childs and D. Ingram, "The linux-srt integrated multimedia operating system: Bringing qos to the desktop," in *RTAS '01: Proceedings of the Seventh Real-Time Technology and Applications Symposium (RTAS '01)*. Washington, DC, USA: IEEE Computer Society, 2001, p. 135.

[12] M. Barabanov and V. Yodaiken, "Real-time linux," *Linux Journal*, 1996.

[13] M. A. Rau and E. Smirni, "Adaptive cpu scheduling policies for mixed multimedia and best-effort workloads," in *MASCOTS '99: Proceedings of the 7th International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems*. Washington, DC, USA: IEEE Computer Society, 1999, p. 252.

[14] Google, "Android operating system." [Online]. Available: http://www.android.com/

Email as Electronic Memory: A Spatial Exploration Interface

Florian Müller, Martin Guggisberg, Helmar Burkhart
Computer Science Department
University of Basel
Basel, Switzerland
{*florian.mueller, martin.guggisberg, helmar.burkhart*}*@unibas.ch*

***Abstract*—Recently, electronic memory (e-memory) applications have come into research focus. Using personal data ranging from email to actual life logs, they are to provide us with an interface that facilitates functions such as retrieving, reminiscing, and reflecting information from our past – functions that we know well from our biological memory. We present an exemplary e-memory application based on personal email archives that supports reflection and reminiscence by providing a spatial layout of email communication data. The spatial layout is derived using a physical force relaxation simulation. In order to emphasize various properties of the communication network, the communication is represented as a weighted, directed graph. This allows analysis in terms of various metrics.**

***Keywords*-Multimedia, Information retrieval, Data visualization, Electronic memory, Lifelogging**

I. Introduction

In his Memex vision, Vannevar Bush imagined an information system that would allow the effective storage, editing and retrieval of all information encountered throughout a lifetime (see [1]). Today, his vision is often cited by proponents of lifelogging. In lifelogging, data arising from everyday activities both in virtual and real spaces is persisted and used for further processing. The recording of visual information has been a dominant component of lifelogging since its early application by Steve Mann, who described his first capture activities as *personal imaging*. However, he also noted that the result of having a lifelog is the ubiquitous availability of a *personal information domain* (see [2]). In this perspective, lifelogging provides the basis for an *electronic memory* (e-memory). Specific applications of e-memory cover memory deficit compensation (recall names and faces, retrieve lost objects), memory-related medical conditions (amnesia, dementia) and applications for reminiscence and self-reflection, which could be called explorative e-memory applications.

While lifelogging is still often an explicit activity carried out only by a few enthusiasts such as Gordon Bell of Microsoft (see [3], [4]), every-day use of computer technology without special focus on lifelogging already provides users with a vast corpus of data that could serve as the basis of a considerable e-memory. Note that a mere collection of data cannot be considered an e-memory. It is only when these data are made accessible and comprehensible (similar to our biological memories) that we can speak of e-memory. In this work, we show how the use of visualization techniques can aid users in exploring their digital corpora with the example of email communications. The visualization of the activities in a user's communication network is based on a spatial layout derived from a physical force relaxation simulation. Contacts are represented as particles, and interactions (emails) are represented as springs between particles. In order to obtain interesting views and insights of the user's communication, these communications are represented as a directed, weighted graph on which various computations can take effect.

This paper is organized as follows. In the remainder of this section, we introduce the notion of electronic memory and describe the status quo of electronic mail interfaces. In the second section, we detail the architecture of our electronic memory system and the data used. Sections three and four provide a detailed account of the mechanisms on which our system is based. In the final section, we conclude our results.

A. Electronic Memory

Specific applications of e-memory cover memory deficit compensation (recall names and faces, retrieve lost objects), memory-related medical conditions (amnesia, dementia) and applications for reminiscence and self-reflection, which could be called explorative e-memory applications. A classification of electronic memory applications was recently proposed by Sellen (see [5]). She defines five classes of electronic memory applications, which she labels as the five "R"-requirements. These are:

- Recollecting (re-experiencing past memories for the purpose of locating specific information items)
- Reminiscing (re-experiencing past memories for emotional reasons)
- Retrieving (retrieving some specific information, without re-experiencing)
- Reflecting (analyzing behavior over time and deducing conclusions from it)
- Remembering intentions (prospective memory, i.e. remembering to execute a decision taken in the past)

Our work focuses on the two aspects of retrieving and especially reflecting. One of the major challenges for reflecting is the reduction of complexity in the available data.

ISBN: 978-1-61208-129-8

It will be shown that the aggregation of spatial and grouping information, combined with several specific communication metrics, are valuable tools in achieving such a reduction and can help in making sense of large amounts of data through the use of a visual interface.

B. Interfaces for Email

Since we have chosen to use personal email communication as the basis for our e-memory application, a short review of state-of-the-art email interfaces is in order. Most current email interfaces are relatively unsuitable for providing e-memory interfaces to personal communication. They are mostly list-based, i.e. the messages are displayed in a chronological (or otherwise sorted) list, and details about the currently selected message are displayed in a separate view component. While search functionalities enable users to retrieve messages that match specific criteria, the list-based view is not suitable for reminiscing or recollecting: going through the list item by item is time-consuming, and the view does not aggregate the information contained in multiple messages – only one at a time is displayed in detail.

Several interfaces have been proposed to support higher-level views of email communication. Frau and others (see [6]) have proposed a dynamic email interface ('Mailview') which displays plots of email communication over time. The interface focuses on visually aggregating existing message attributes such as its size or the folder it is stored in. Viégas has developed several visualizations for email, including the PostHistory interface (see [7], [8]). It presents users with an overview of their email communications through the use of a timeline overview in the form of a calendar and a visualization of the contact network of the user. Depending on time windows and contact selection, the communication with one or several persons is displayed over time in the calendar view, and relevant social network context is displayed in the contact view.

Such proposed visualizations are more suitable for e-memory applications such as reminiscing and recollecting than list-based interfaces. They aggregate the information contained in multiple messages and present them in a single, at-a-glance view. The fundamental advantage of such visualizations is that they abstract from individual messages and display information on the time-aggregated structure and context of email. In this contribution, we would like to provide a new basis for email visualization for e-memory applications, namely a spatial layout for email which, intrinsically, does not have a representation in this domain. As we will show, deriving a spatial layout makes large email corpora comprehensible by providing an overall-view of email communication. Examples of such approaches can be found in social network visualization (see [9], [10]).

II. Architecture and Dataset

The overall architecture of the visualization system is depicted in Figure 1. In a first step, the email communication from several mailboxes is extracted via the IMAP protocol and preprocessed. In the preprocessing, various irregularities resulting from non-standard-conform email clients are eliminated, such as incorrect representation of dates and different character encodings. Once the email data has been regularized, it is stored in a relational database, using one table for the sent messages, and another table for a per-message and per-destination listing of the recipients.

In the next step, which is depicted in the top part of Figure 1, a spatial layout for all communication participants, including grouping information, is derived (*grouped contact map*). This is based on a physical force relaxation simulation, for which details are described in section III. The grouped contact map is the first part of the input for the actual visualization. The second part is generated through a graph processor (depicted in the bottom part of the figure). The graph processor uses the open source JUNG Graph API (see [11]) to represent the entire email communication as a directed, weighted graph. Based on this graph representation, several metrics can be applied and later used in the visualization. Details are described in section IV. The output of the graph processor is a *graph-metric map*, which contains the weights of the nodes and edges of the communication graph according to the applied metric.

The grouped contact map as well as the graph-metric map are used by the visualization unit to generate a graphical representation of the email communication. The visualization unit is written in Processing (see [12]), a Java-based language specifically engineered to support visualizations. It obtains the spatial layout from the contact map, and draws nodes and edges according to the graph-metric map. While the grouped contact map is loaded at start-up, the graph-metric map can be generated on demand in order to switch between various metrics.

Explored Dataset Statistics	
Unique email addresses	3.999
Total recipients	16.692
Total direct recipients	14.480
Total copy recipients	2.212
Avg. recipients/message	1,67
Messages with in-reply-to	5.016
Total messages	**10.218**

Table I
STATISTICAL OVERVIEW OF THE EMAIL COMMUNICATION DATA USED IN THIS WORK. THE DATA IS FROM ONE SINGLE USER, SPANNING OVER SEVERAL EMAIL ACCOUNTS.

As stated, the data used in our work is extracted from several email accounts of a user. It spans over a period of 4 years, from 2006 to 2010. It contains a total of over 10.000 messages, and of almost 4.000 different contacts. In

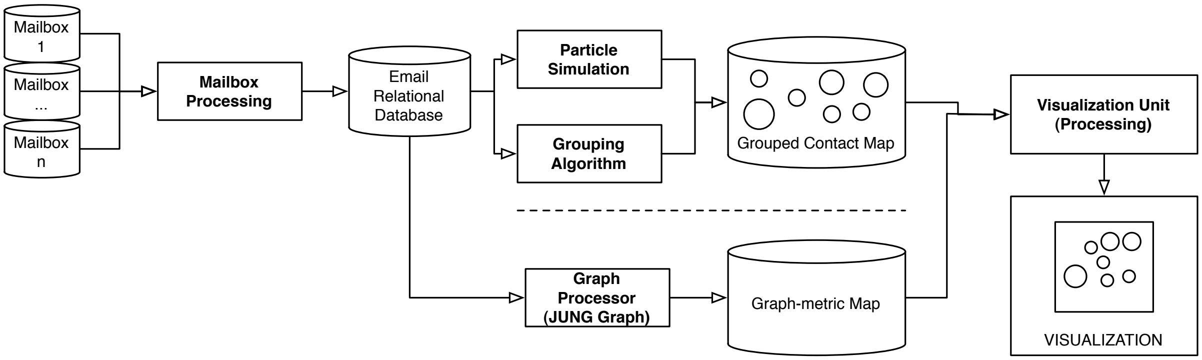


Figure 1. Overall system architecture. In a first step, the spatial layout and the grouping information is derived using the physical force relaxation simulation and a grouping algorithm. In a second step, this spatial layout is used in combination with a graph-based analysis of email communication to obtain a view focusing on a specific aspect of the communication.

order to efficiently use the email data, all the email header information (without the message bodies) was retrieved once using the IMAP4 protocol and then stored in a database. While the total size of all email messages including message bodies is approximately 6 gigabytes, the header information itself contains around 50 megabytes and can be processed quickly. Table I provides a statistical overview of the email communications used for the visualizations.

III. Spatial Email Layout

All communications extracted from the email accounts are processed and brought into a unified form, the resulting set of all messages is called $\mathbb{M}$. The spatial layout is generated by running a physical force relaxation simulation. Every contact (i.e. email address, with the special case that several email addresses belonging to the same person are collapsed into one single contact) in $\mathbb{M}$ is represented by a particle in a particle system. In analogy to the physical world, each particle is assigned a *mass*. The mass of the particle depends on the number of times the contact occurs in a communication (either as a sender or as a recipient). All particles in the simulation are repulsive towards one another (i.e. they have negative attraction). Messages define relations between contacts, and these relations are represented as elastic springs in the simulation. For every direct relation between two contacts (i.e. for every message where contact A is the sender, and B is the or a recipient, and v.v.), a weight update is performed for the involved particles: a spring is created between the two particles representing the contacts, and the weight of the particles is adapted. Both the particle mass and the strength of the springs between them depend on the frequency and nature of the communication they originate from.

$$\Delta m = \frac{1}{n_R} \cdot \frac{\gamma_i}{m} \qquad (1)$$

The mass update of the particles, Δm, is calculated according to the following formula, where n_R is the number of recipients of a message, and γ_i is a communication specific growth constant (e.g. direct messages are rated higher than messages received as a carbon copy), and m is the current mass (see Equation 1). The strength update of the spring is calculated in an analogous manner, but with a different growth constant. The weight and spring update is illustrated in Figure 2. Note that for obtaining better clarity, the weight and strength updates are calculated using only the term $\frac{1}{n_R}$.

Once the particles with their respective weights and the springs between them have been created, the particles are initially placed at a random location in the space of the simulation. Then, the simulation is started, and we wait until the effect of the default mutual repulsion and the attraction through the springs have lead to a stable state after a certain relaxation time. In this state, the particles have a stable position, and the spatial map can be generated accordingly.

Through the simulation procedure, we gain a spatial model of the email communication network. As will be shown in the results section, the spatial layout already provides the user with an overview of her email communication that shows important clusters of contacts and their interrelation. Note that the absolute position of a node is not relevant (since it is arbitrary, for in every simulation run, it may be different). However, the relative position reveals important information, such as distance and closeness to neighbors and other clusters of nodes.

In addition to the clustering implicit in the relaxation simulation, an algorithm for grouping the contacts is employed during the simulation. Nodes that communicate with each other are assumed to influence one another, and for every communication between a sender and one or several recipients, the sender exerts a certain amount of influence on the recipients. After all messages have been processed, every

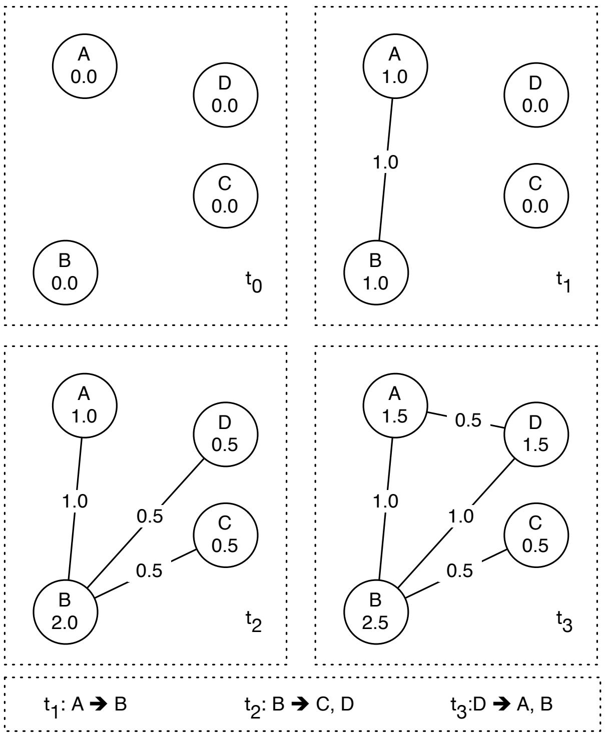


Figure 2. Illustration of the particle simulation using unit weights. The three communications ’A $\rightarrow$ B’, ’B $\rightarrow$ C, D’, and ’D $\rightarrow$ A, B’ occur one after the other at times t_1, t_2, t_3. As they occur, the particle weights and the springs between them are updated accordingly.

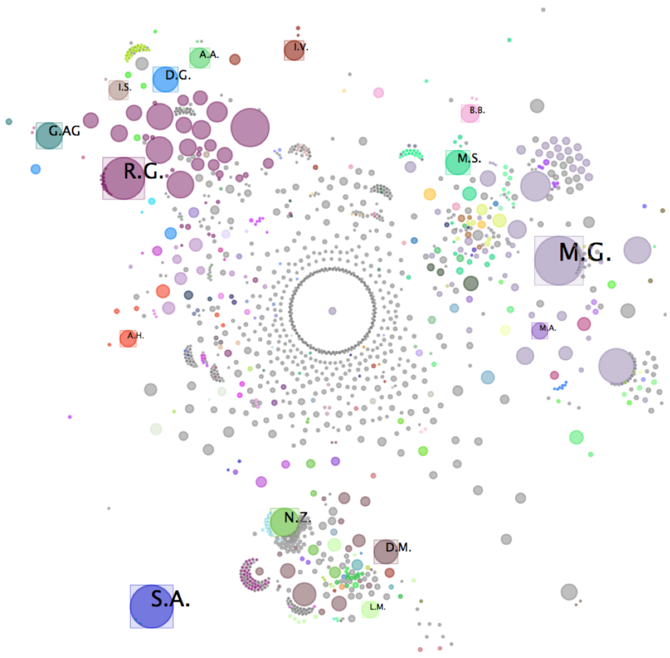


Figure 3. Spatial layout and grouping as resulting from relaxation simulation and grouping algorithm. The weights of the nodes reflect the number of overall occurrences in communications.

contact node is influenced by one or several other nodes. If for any node, the relation between its own mass and the largest amount of influence exerted upon is below a certain threshold, the node is said to belong to the node with the largest influence on it. This relation is applied recursively, and results in the creation of groups within the contact network. The grouping is used to apply different colors to different groups, which results in an additional simplification of the interface. The spatially clustered and colored email contact network layout is the basis for the next step, the graph-based analysis of the communication.

IV. Graph-based Analysis

Similar to the case of the particle system, the entire email communication can be represented as a directed, weighted graph. It is constructed in a similar manner as the particle system used to derive the spatial layout: every contact is represented as a vertex, and every communication relation (sender to recipient) is represented as a directed edge. In a default case, the weight of the vertices and the weight of the edges can be derived as in the case of the particle simulation. A more interesting approach is to find various metrics according to which the vertices and edges are weighted. Depending on the metric chosen, different aspects of the email communication can be shown. While typical social network metrics such as *betweenness centrality* are more interesting for social networks that comprise several ego-networks, for our case, where we look at one single ego-network, other metrics have proven to be of more relevance. We have used the information contained in the email header information, especially: (a) whether a message is a direct reply to another message, and to which, (b) how deep threads run (a back and forth of replies and replies to replies), and (c) who forwards information. The results of these metrics will be shown in the results section. Note that since we can look at email communication as a graph, we can apply any metric we wish.

V. Results

Figure 3 shows the base layout derived from the relaxation simulation. The owner of the mailboxes is located in the center. As can be seen, three major clusters have been formed: one in the south, one in the northwest, and one to the east. Spatial proximity suggests knowledge of one another and communication with one another. The colors show additional structural information for clusters. The ring of small nodes around the owner of the mailbox are contacts that have only occurred few times in communication and that are not networked (i.e. they have never occurred in messages that were destined to multiple recipients). The size of each node is directly dependent on the number of times a contact has occurred in any communication.

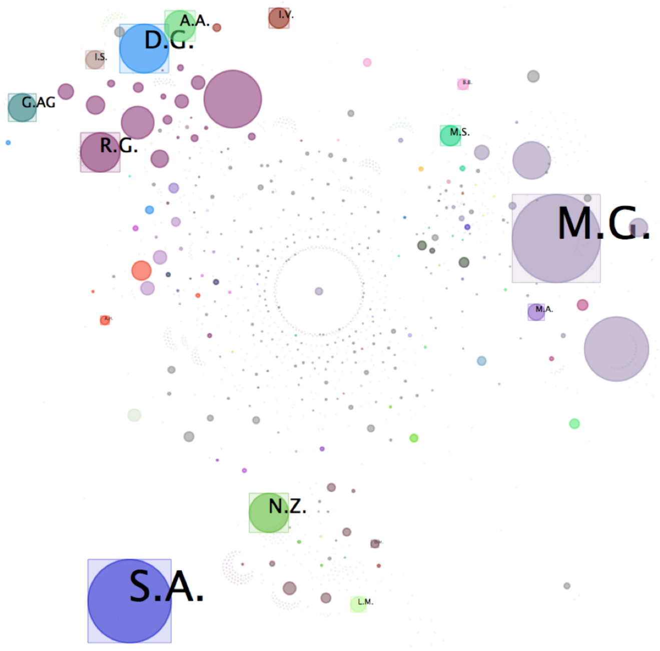


Figure 4. Weighting based on counting frequency and depth of replies between contacts. Three contact clusters can be clearly identified. The internal structure of each cluster can also be assumed: in two cases, one main communicator is identifiable, while in the case of the cluster to the northwest, there are several equal contributors.

Figure 4 shows a layout derived from the base layout, where the weighting of the graph depends on the reciprocity of communication. The header fields ’Message-in-reply-to’ and ’References’ are used to determine if a message was sent in reply to another message. A high number of replies (of various depths) in communication between two contacts suggests that they cooperate more closely than others and that the flow of information between them is two-sided. Within the three clusters and compared to the base layout, we can see how contacts with whom the owner of the mailbox communicates reciprocally are prioritized. Finally, in Figure 5, the layout is derived from information forwarding activity. The subject as well as forwarded headers are analyzed to determine whether a message is a forwarded message, and the nodes are weighted according to the number of messages they have forwarded. In addition, the edges between the nodes are weighted according to the number of messages forwarded between the two nodes. As can be seen, other nodes are prioritized than in the reciprocity example. The forwarding of information suggests organizational hierarchy, which may be formal or informal. Nodes that are weighted high are likely to be important organizers and distribute information in the network.

VI. Conclusion

We have shown how a spatial layout for email communication data can be derived based on a physical force relaxation simulation. The spatial layout is the basis for a two-dimensional interface for email that is targeted at e-memory applications. The base layout functions as a summarization of information in two manners: first, it generates clusters of related nodes and shows these clusters in relation to other clusters. Second, through the use of an grouping algorithm that calculates mutual influence of nodes, these can be further grouped by applying different colors.

Since the email communication data is represented as a graph, several metrics can be applied to it. This allows the analysis and visualization of various aspects of the communication, such as cooperation, information flow, and hierarchy. It was initially stated that our aim is to develop an e-memory application focusing on retrieving information and, especially, reflecting on it. If the application of various metrics on the graph-represented communication is seen as reflecting on that communication, our system provides not only reflecting capabilities, but also an interface to visualize them. Using a graphical representation of the results obtained allows users to gain insights into their communication patterns and networks. Apart from this user experience-centric view, further possibilities arise.

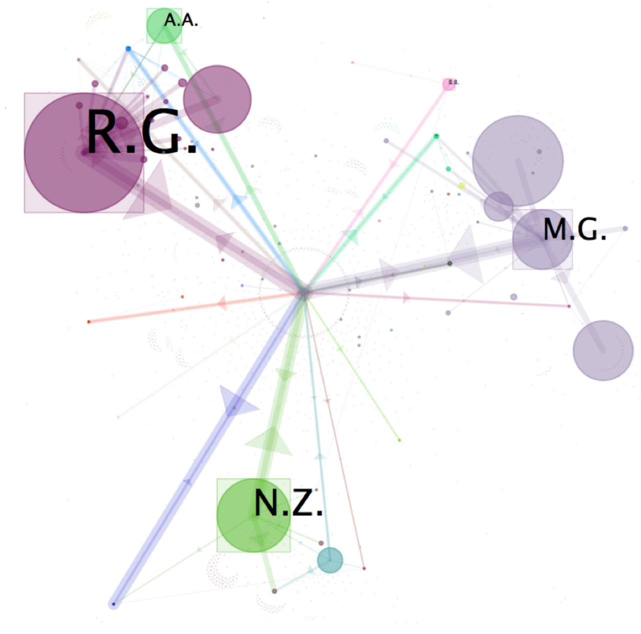


Figure 5. Weighting based on number of forwarded emails, with the direction of the forwarding indicated by arrows. This visualization shows who forwards information, which allows the assumption that this person has an important organizational role within the network.

The structural information derived from the communication network can be used to communicate more efficiently.

For example, features such as the recently introduced *priority inbox* from Google (see [13]) can be based on our system. It allows the classification of message priority relative to the importance or function of the sender in an email communication network. We are currently also evaluating our system using multiple ego-networks of email communication. In such applications, the inferred knowledge is not limited to a single user, but is situated at an organizational level. In such a context, the scope of possible applications is even wider.

References

[1] Bush. V.: As We May Think. In: Atlantic Monthly, vol. 176, 1, pp. 101–108 (1945)

[2] Mann, S.: Wearable computing: A first step foward personal imaging. In: ACM Computer, vol. 30, 2, pp. 25–32 (1997)

[3] Gemmell, J., Bell, G., and Lueder, R.:MyLifeBits: a personal database for everything. In: Communications of the ACM, vol. 49, 1, pp. 88–95 (2006)

[4] Gemmell, J., Bell, G., Drucker, S., and Wong, C.: MyLifeBits: fullfilling the Memex vision. In: Proc. ACM International Conference on Multimedia, pp. 235–238 (2002)

[5] Sellen, A.J. and Whittaker, S.: Beyond total capture: a constructive critique of lifelogging. In: Communications of the ACM, vol. 53, 5, pp. 70–77 (2010)

[6] Frau, S., Roberts, J.C., and Boukhelifa, N.:Dynamic Coordinated Email Visualization. In: Proc. WSCG, pp. 187–193 (2005)

[7] Viegas, F.B., Golder, S., and Donath, J.: Visualizing Email Content: Portraying Relationships from Conversational Histories. Proc. CHI 2006, pp. 979–988 (2006)

[8] Viegas, F. B., Boyd, D., Nguyen, D.H., Potter, J., and Donath, J.: Digital Artifacts for Remembering and Storytelling: PostHistory and Social Network Fragments. In: Proc. System Sciences, pp.10–19 (2004)

[9] Hansen, D., Shneiderman, B., and Smith, M.: Visualizing Threaded Conversation Networks: Mining Message Boards and Email Lists for Actionable Insights. Active Media Technology, pp. 47–62 (2010)

[10] Freire, M., Plaisant, C., Shneiderman, B., and Golbeck, J.: ManyNets: An Interface for Multiple Network Analysis and Visualization. In: Proc. CHI, pp. 213–222 (2010)

[11] JUNG Graph API. http://jung.sourceforge.net, last retrieved 16 December 2010.

[12] Processing. http://processing.org, last retrieved 16 December 2010.

[13] Gmail Priority Inbox. http://mail.google.com/mail/help/priority-inbox.html, last retrieved 16 December 2010.

ISBN: 978-1-61208-129-8

Changing the Middleware System for IPTV Services Telecom Operators Based on the Methodology of the Change Management Process

Anel Tanovic
Department for IT development of multimedia services
BH Telecom d.o.o. Sarajevo
Sarajevo, Bosnia and Herzegovina
anel.tanovic@bhtelecom.ba

Fahrudin Orucevic
Department of Computer Science and Informatics
University of Sarajevo, Faculty of Electrical Engineering
Sarajevo, Bosnia and Herzegovina
forucevic@etf.unsa.ba

***Abstract* – Managing the change plays the most important role in any IT business organization. Managing changes is important for strategic components of the existing IT service. One of such strategic components is the alteration of the Middleware system in the IPTV service of the Telecom Operator. The Middleware system is a central system which controls all other IPTV systems of the Telecom Operator: Video on Demand system, Encrypting system, Headend system, Monitoring system and Database system. That is why, before we uninstall the old and install the new Middleware system we require a set of activities which have to be clearly stated. The reason for swapping the Middleware system can be the common problems and also the inefficiency in the function of the old system. This document describes the steps which need to be implemented in the process of changing the Middleware system in the IPTV system of BH Telecom, the leading Telecom Operator in Bosnia and Herzegovina. The result of the work should be a reflection upon the improvements of the quality parameters of the IPTV system of BH Telecom which are a contribution of the newly implemented Middleware system, which is as a final result going to lead to a increase of users and profit income. This work represents the continuation of the investigation in BH telecom which are conducted with a goal to improve the management over the IT servers.**

Keywords-Service Management; ITIL V3; ISO 20000; Change Management; IPTV

I. INTRODUCTION

Today’s management of alterations is one of the biggest challenges in the IT industry [1]. Replacement in the IT industry occurs as a result of external needs and problems which have been created in the common operation of the existing IT service [3], [4]. Every successful IT organization has to define how to manage all changes regardless of being operative or strategic. That is why it is needed to define the process of managing all changes (Change Management process) [2].

Various standards of Service Management praxis differently define the Change Management process. ITIL V3 defines Change Management as „ a process which as a goal has to assure that standardized methods and procedures use for an efficient and quick management of all changes which as a goal has to decrease the impact of incidents linked to the change of service quality“ [2], [3], [4]. ISO 20000 defines Change Management as „ a process which has to assure that all made changes have to be evaluated, confirmed, implemented and reviewed in a controlled way“ [5], [6]. A important thing to emphasize is that the Change Management isn't a process of identifying strategies or IT service designs, but rather the process of implementation of the IT service. Regardless of the standards the key activities of the Change Management process are [1], [2], [3], [4], [5], [6], [9], [11]:

- Controlling and managing the process of changing.
- Recording, evaluating, confirming and rejecting admitted requests.
- Assembling team meetings for the implementation of the change.
- Coordination of the development and implementation of change.
- Evaluation of results made by change and approach to the finalization of change if it has been successfully completed.

Earlier published contributions from this research area are: [7] and [8]. These contributions were written as a part of the implementation of ITIL V3 processes in BH Telecom, the leading Telecom Operator in Bosnia and Herzegovina. The result of the first contribution is a description of the implementation of ITIL V3 Service Design processes in an information system of Telecom Operator, and the result of the second contribution is a percentage of the implementation of ITIL V3 Supplier Management process in IPTV/VoIP system of BH Telecom.

This contribution is connected to Multimedia conference because the paper was submitted under topic: “Multimedia applications” because Middleware system has all IPTV applications for: Live TV, Radio, Internet, Video on

ISBN: 978-1-61208-129-8

Demand, EPG, TV Mosaic, Interactive chat and Messaging and this document describes the changing of Middleware system (or in other words: Changing Middleware applications). Middleware is the heart of IPTV service, and IPTV service is the main Multimedia system in every Telecom Operator – Today Telecom Operators offer their x-play services (Dual-play, Tripple-play etc.) in the way that IPTV is the central module of some x-play service.

The second chapter explains nine steps which BH Telecom should implement towards the change of the old with a new Middleware system. In the conclusion a list of improvements is presented which should guides us into a new Middleware system of the IPTV service of BH Telecom and also a time table for completing all nine phases [7]. The conclusion also contains a description of the testing process of the change of Middleware system in BH Telecom's IPTV service which has been done by using test Middleware software versions of some Middleware vendors.

II. Alteration of the old Middleware system with a new Middleware system

Because of the replacement with the old Middleware system with a new Middleware system it is needed to implement the following 10 steps, where an external company is responsible for the realization of the step 4:

1. The specification of the new database has as a goal to define their compatibility with older data bases
2. Define user and administrator lists of specifications which the new Middleware system has to have
3. Criteria for the choice of a external company that is going to design and implement the new Middleware system
4. Defining specifications of the network plan and installation of the new Middleware system by the hand of the external company
5. Implementation of the new Provisioning system towards connecting the new Middleware system with the centre informational system of the Telecom Operator
6. Implementation of the new Billing system
7. Testing the new Middleware system with other IPTV systems
8. Data migration from the old to the new Middleware system
9. Releasing into production the new Middleware system

A. Specification of the new database towards the identification of its compatibility with older database

Figure 1 shows database scheme of the new Middleware system [15]. The number of tables in this database scheme is 20 and main tables are: subscriber, channel, device, vod_contents, program and billingiptv.

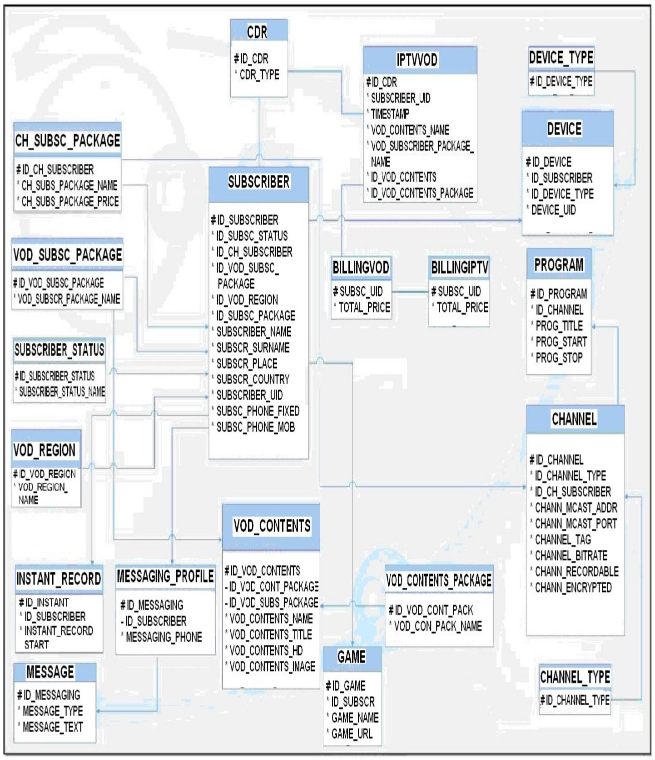


Figure 1. Database scheme of the new Middleware system

The new Middleware system should have an identical database as the old system. The assignment of the team is to specify the design of the new database which has to be compatible with the old database as show on Figure 1. Next to this specification, the teams task is to expand this database with new functions which are defined through consumer and administrative listings of specifications which need to have the new Middleware system [12], [13], [14], [15].

B. Defining the user and administrator listing specification that the new Middleware system has to have

In this phase two teams should be working: one for defining user specifications of the new Middleware system and a team for defining administrative specifications of the new system [7].

The team for defining user specifications of the new Middleware system should be defining the following listed specifications [8], [10], [15], [16]:

- Enable emitting live TV channels as well as SD channels. The new platform should have unlimited number of channels which can be configured. The platform should be secured by arranging the TV channels into certain categories like: all channels, favourite channels, domestic channels, regional channels, informative channels, sports and music TV channels.

Copyright (c) IARIA, 2011. ISBN: 978-1-61208-129-8

- Enable browsing shows which are defined in Electronic Program Guides (EPG). The module for browsing has to be realised so that on demand with a minimum of 2 characters with which the search is going to take place in the EPG. The search results should give information about: starting time of the show, exact name of the show and the name of the TV channel to which the searched show belongs.
- Realize Electronic Program Guide (EPG). EPG should be support for all TV channels. The time period of the EPG should be a minimum of 10 days prior and 10 days in advance. EPG should enable recording all shows from all TV channels. The maximum playing time of one recording should be 6 hours. All shows which have passed should be automatically taped. For all shows in the EPG certain features should be defined like: the name of the show, starting time, ending time, show description and total playing time.
- Support the existence of TV Mosaic. TV Mosaic should give the possibility to watch many different TV channels on the same screen at the same time. The TV Mosaic needs to enable defining an unlimited number of even TV channels on one screen.
- Realize the module Video library in which movies need to be categorized in certain categories like: action, adventure, animated, children, documentary, domestic, drama, history, horror, crime, science fiction, comedy and thriller movies. Every category has to have a unlimited number of movies. Every movie has to have their price, playing time, producers name, main actors name and a short description listed. The purchase of a movie should be made based on the identification PIN which is given to every user.
- Enabling the option of direct recording of TV shows using remote control. Direct recordings will be stored in a file named recordings. Maximum playing time of one recording should be 6 hours. For all recorded shows an option of moving and stopping shows should also be enabled. For all taped shows an option should be enabled so the user can start and stop the recording. All taped shows stay permanently encrypted for one user on his Set Top Box.
- Enable additional option on a user menu like: Radio channels, Internet which enables access to contents from the menu and games where the user can save unlimited number of interactive 3D games.
- Realising the option Settings where the user will have a opportunity to change his PIN, set parental control on a chosen TV channel, set the language on which the TV channel will be emitting, restart their Set Top Box and also have basic settings about their account as well as what an STB IP address is, STB MAC address is and also the name of the software which can be found on his STB.
- Enable the option Help where the management of the IPTV service will be explained. The option help should be full of unlimited help content.
- Enable the option Telephone where all the calls of a user will be inputted which he received on his VoIP (If he is next to being a IPTV user also a VoIP user). Option that he should have are: all calls, outgoing calls, missed calls and received calls.
- Realization of the option Address book where inputting, editing and erasing telephone contacts which number should be infinite is possible.
- Implement the option Messaging where every user that has a VoIP telephone will be able to exchange messages with other users that also have engaged a VoIP telephone service option. The Messaging option should include options of creating new messages, viewing existing once and viewing received messages. The number of existing and sent messages should be unlimited.
- Implementation of the option Chat that should enable the users to between themselves exchange messages. Identification parameter by which the users are going to differ from each other is subscriber_uid that every user has to get when they activate the service. The option Chat should have a realized option of authentication respectively the possibility of accepting and rejecting users for chat. In this option a listing of all chats which one user had till a certain point should be found.
- Support the option of exhibiting the VoIP number on the TV for all users which have the additional option of VoIP. The VoIP number should be shown on the TV when the user is receiving a call from another user.
- Support the option of Message Waiting Indicator (MWI) that should enable displaying textual messages on the TV when a user leaves it on the VoIP telephone which supports the option of leaving voice mails.

The team for assigning administrative specification to the new Middleware system should define the following list of specifications [16]:

- Enable all defined provisioning functions.
- Enable adding, editing and erasing TV and Radio channels.
- Enable adding, editing and erasing TV and Radio channel packages.
- Enable the option of marking TV channels which can be recorded.

- Enable adding, editing and erasing movie contents
- Enable adding, editing and erasing movie categories.
- Enable adding, editing and erasing priced movie categories.
- Enable adding, editing and erasing the movie distributors.
- Enable the option of adding, editing and erasing contents linked to the Electronic Program Guide (EPG).
- Enable the option of adding, editing and deleting interactive games.
- Enable the option of adding, editing and erasing contents linked to the Help option attended for the users.
- Enable the option of adding, editing and deleting contents linked to the Internet option.

C. Criteria for the choice of an external company that needs to design and implement the new Middleware system

In the purpose of choosing a external company which is responsible for the implementation of the new Middleware system it is needed to release a public announcement that consists from two phases: phase of prequalification and the phase of the final partner company choice. In the phase of prequalification all candidates have to satisfy all legal and economic conditions [8]. In regards to the technical and professional requirements, the partner companies have to satisfy the following requirements:

- A minimum of 20 employed IPTV consultants.
- A minimum of one reference of implementation of the IPTV system in the telecom industry.
- A minimum of one reference regarding the exchange of the Middleware system in some IPTV system in the telecom industry.
- A statement that the new Middleware system basis is going to be compatible with the existing basis which is defined in Figure 1.
- A statement that the new Middleware system is going to have a specification listing that is going to be defined in chapter III.
- A statement that the partner company has its own hardware and software infrastructure which is enough for the realization of this project.
- A statement that the partner companies new Middleware system is compatible with the Set Top Box of the following 5 manufacturers: Motorola, Amino, Milinet, Albis Technologies and Technotrend which are already certified for the old system.

The final choice of partner companies that pass the prequalification stage should be conducted on the basis of the following three parameters: the lowest price (with a percentage share of 50%), the total number of changes of IPTV Middleware systems in the telecom industry systems (with percentage share of 30%) and the total number of realization IPTV systems in the telecom industry (with a percentage share of 20%). The company, which is first in terms of these parameters should get a contract to change the Middleware system. In the event that two or more firms were to have the same number of points, the contract is awarded to the firm which has a large number of references to change the Middleware System.

D. Defining the specification of the network plan and the installation of a new Middleware system by foreign firms

Figure 2 shows the relationship between the Middleware system and other systems of the IPTV service: Video on Demand system, Real Time Encryption system, Verimatrix system, RAC system, Load Balancing system and Monitoring system [8]. The scope of the IP address is 10.120.0.x where x is the number indicated for each server in Figure 2. The team for the installation of a new Middleware system must do so by the same scheme indicated on Figure 2.

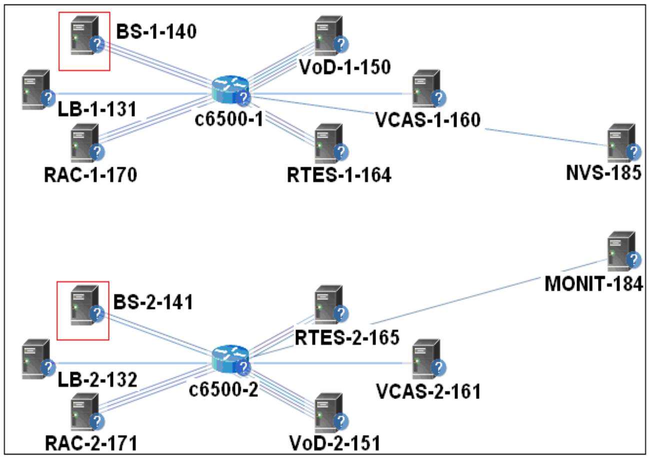


Figure 2. Relationship between a new Middleware system and other systems of the IPTV service

BH Telecom should form a special team to monitor the installation of the system which is being worked on by a foreign firm.

E. Implementation of the new Provisioning system

The Provisioning system represents a link between the central information system of the Telecom Operator and the new Middleware System. The new Provisioning system must support the following operations:

1. Adding new Set Top Box's
2. Changing existing Set Top Box's
3. Deleting existing Set Top Box's
4. Creating a new user

ISBN: 978-1-61208-129-8

5. Assigning a basic package of channels to the end user
6. Assigning an additional package of channels to the end user
7. Assigning a package of HD channels to the end user
8. Suspension of users (temporarily turned off)
9. Reconnection of users
10. Permanently delete users
11. Deleting an additional package of channels for the end user
12. Deleting an end users HD channel packages
13. Changing the name of the user
14. Changing the surname of the user
15. Changing the address of the user
16. Changing the location of the user
17. Changing the postal code of the user
18. Changing subscriber_uid of the user
19. Adding a VoIP number for the user
20. Changing the VoIP number of a user
21. Deleting the VoIP number of a user
22. Adding a fax for the user
23. Changing the fax of a user
24. Deleting the fax of a user
25. Adding a phone number for the user
26. Changing the phone number of a user
27. Deleting the phone number of a user
28. Adding a mobile phone number for the user
29. Changing the mobile phone number of a user
30. Deleting the mobile phone number of a user
31. Turing on the option for leaving VoIP voice messages on a TV
32. Turing off the option for leaving VoIP voice messages on a TV

F. The implementation of a new Billing system

The basic value changes depending on whether you choose to go with IPTV along with the VoIP service, ADSL service, VoIP and ADSL service, or VoIP, ADSL and mobile services together, which is already defined in the existing Billing system and is not necessary to change when you change the Middleware (Table I).

On the final value of consumption Video on Demand (VOD) service for one of the users indicated in the table billingvod impact parameters from the tables: subscriber, vod_subscriber_package, vod_subscriber_contents and cdr. In the cdr table are written records of each individual purchase of a movie. A software script goes through the database tables and based on the prices defined in Table I form a final price for VOD consumption in one month for one user (Table II). The program script needs to take into consideration during calculations, that every fifth purchase of a movie in one month by one user is free.

The final consumption value of the IPTV services indicated in the table billingiptv are affected by the data in tables: subscriber, billingvod, ch_subscriber_package and whose values are indicated in Table II. Adding values from the tables billingvod and ch_subscriber_package for each user from the table subscriber you get the total value of consumption of IPTV services, which are kept in the table billingiptv.

TABLE I. PACKAGES FOR VIDEO ON DEMAND SERVICES

vod_subscriber_ package_name	**vod_subscriber_ package_uid**	**vod_subscriber_ package_price**
Category A	VOD_package_a	0.25 E
Category B	VOD_package_b	0.50 E
Category C	VOD_package_c	1.00 E
Category D	VOD_package_d	1.50 E

TABLE II. PACKAGES FOR IPTV SERVICES

ch_subscriber_ package_name	**ch_subscriber_ package_uid**	**ch_subscriber_ package_price**
Basic	basic	12.50 E | 27.50 E | 32.50 E | 47.50 E
Plus	plus	4.00 E
HD	hd	3.00 E

G. Testing the new Middleware system with different IPTV systems

The new Middleware system must be fully compatible with all IPTV systems: Video on Demand system, Real Time Encryption system, Verimatrix system, Monitoring system, RAC system and Headend system [8].

From the central information system all the provisioning functions and all billing scenarios must be thoroughly tested depending on the combination of all the VOD packages (Table I) and IPTV packages (Table II).

H. The migration of information from the old to the new Middleware system

The overall data migration from the old to the new database needs to be done in order for the following tables: subscriber, device, device_type, program, channel, channel_type, ch_subscriber_package, vod_contents, vod_contents_package, iptvvod, vod_subscriber_package, subscriber_status, game, cdr, instant_record, message, messaging_profile, billingvod, billingiptv, vod_region. What needs to be manually configured during this migration is that the value of the id_ch_subscriber_package field needs to be marked CH_SCP_1, the value of field id_vod_subscriber_package is marked as VOD_SCP_1, and the value of field id_subscriber_package is marked as SCP_1. These three values are required, together with data for ch_subscriber_package_name and for Video on Demand field vod_subscriber_package_name, which are defined in

Tables I and II. All other data can be automatically switched from one system to another system.

I. Releasing into production the new Middleware system

If all test results from phase G were positive, a new Middleware system can be put into production. Coordination of the release of the new system into production needs to be done by a special team composed of 3-5 IT professionals.

III. CONCLUSION

The newly implemented Middleware system should solve some existing problems that exist in BH Telecom's IPTV service, and implement some additional functionality [15]:

- The Middleware system's resistance to failure due to the changes that have occurred in other systems of the IPTV service.
- Display a callers VoIP number on TV.
- Display the option Messaging Waiting Indicator on TV.
- Continuous operation of all provisioning functions.
- Interactive chat between users.
- Interactive message exchanging between users.
- Unlimited number of users and TV channels with the option of recording.
- Increasing the bitrate level output signal from the headend of the system from 2.5 Mbps to at least 4.0 Mbps.

Each of the nine-defined phases that need to be changed from the old to the new Middleware system requires formation of an independent team [7]. Each team starts with their activities once the previous team has finished their activities. Previous Middleware systems realizations and other supporting systems such as Provisioning and Billing systems showed that a longer period is needed for the implementation of the phases. Activities of the first team should last one month, and the other team a maximum of 15 days. Activities of the third team should last for 3 months in which time the whole period until the final partner company is chosen and the signing of a contract with them is included. The time frame for the fourth phase is 1 month and 15 days in which time 15 days are included to define the specifications of the network plan and one month for the installation of the new Middleware system. Phases 4 and 5 should take 2 months each because of the complexity of connecting a new system with a central information system. The testing phase of the new system and data migration from the old to the new system should each last 15 days. The last phase, the phase of commissioning the new system into production, is purely informative and should last 1 day.

The total time for the realization of the nine phases is 11 months. However, this temporal analysis does not include non-working days, which means that the total time for the realization of these phases should be extended 1 month, so the total time for the realization of the project is 12 months.

The testing process of the change of Middleware system in BH Telecom’s IPTV service has been done by using test Middleware software versions of some Middleware vendors and these test software versions have included only 10 IPTV users. The test results fully agree with the expected results of a new Middleware system.

Further research in this area is related to the Problem Management process, where any incidents or problems that may arise with the IPTV service after the implementation of a new Middleware system should be investigated [8]. Research should be done at least 6 months after installing the new system and should show the level of efficiency of the new system compared to the old system.

ACKNOWLEDGMENT

The authors wish to thank experts from some of the leading manufacturers of Middleware systems such as Nokia Siemens Networks, Ericsson, BeeSmart, Dreampark and Zapware for advice during this project.

REFERENCES

[1] J. van Bon, A. de Jong, A. Kolthof, M.Pieper, R. Tjassing, A. van der Veen and T. Verheijen, ”Foundations of IT Service Management Based on ITIL V3”, Third edition, September 2007.

[2] S.Taylor, S. Lacy and I. Macfarlane, “ITIL Version 3 Service Transition”, First edition, The Office of Government Commerce, May 2007.

[3] I. Menken and G. Blokdijk, “The Change Management Guide: The Missing it Change Management Planning, Process, Theory and Tools Guide – Itil Compliant”, Second Edition, December 2009.

[4] I. Menken and G. Blokdijk, “Software Testing and Quality Assurance with It Change Management Transition”, First Edition, December 2008.

[5] J. Dugmore and S. Taylor, “ITIL V3 and ISO/IEC 20000”, First Edition, March 2008.

[6] J. v. Bon, M. Nugteren and S. Polter, “ISO/IEC 20000”, First Edition, May 2006.

[7] A. Tanovic and F. Orucevic, “Implementation of the Information System of the Telecom Operator Using the ITIL V3 Methodology for the Service Design Phase”, 2nd International Conferences on Advanced Service Computing, pp. 82 – 91, November 2010.

[8] A. Tanovic and F. Orucevic, “Comparative analysis of the practice of Telecom operators in the realization of IPTV systems based on ITIL V3 reccomendations for the Supplier Management Process”, 6th International Conference on Systems ICONS 2011, pp. 115 – 121, January 2011.

[9] U. K. Tripatji, K. Hinkelmann and D. Feldkamp, „Life Cycle for Change Management in Business Processes using Semantic Technologies“, Journal of Computers, vol. 3, pp. 24-31, January 2008.

[10] J. Kramer and J. Magee, “The evolving Philosofers Problem: Dynamic Change Management”, IEEE transactions on software engineering, vol. 16, pp. 1293 - 1306, November 1990.

[11] H. Cheng, Y. Xia and X. Hu, “Requirements Change Management of Information System Based on the Keyword Mapping”, 6th Wuhan

International Conference on E-Business, vol. 3, pp. 135 – 140, May 2007.

[12] S. Ahn and K Chong, “Requiremets Change Management on Feature-Oriented Requirements Tracing”, Computational Science and Its Applications – ICCSA, vol.10, pp. 296-307, May 2007.

[13] R. Weber, T. Helfenberger and R. K. Keller, „Fit for Change: Steps towards Effective Software Maintenance“, 21st IEEE International Conference on Software Maintenance, vol. 10, pp. 26-33, September 2005.

[14] S. Adesola and T. Baines, “Developing and evaluating a methodology for business process improvement”, School of Industrial and Manufacturing Science, Cranfield University, vol.11, pp. 37-46, May 2005.

[15] K. Oh, Y. Park and S. Park, “Middleware Architecture of Hybrid Digital Cable Receiver for Cable Broadcast and IPTV Service based on OCAP”, International Conference on Consumer Electronics 2008, ICCE 2008, pp. 1-2, January 2008.

[16] G. M. Lee, C. S. Lee, W. S. Rhee and J. K. Choi, “Functional Architecture for NGN-Based Personalized IPTV Services”, IEEE Transactions on Broadcasting, vol. 55, pp. 329 – 342, June 2009.

3D Objects Watermarking and Tracking of Their Visual Representations

Mireia Montañola Sales*, Patrice Rondão Alface† and Benoît Macq*
*ICTEAM, Université Catholique de Louvain, Louvain-la-Neuve B-1348 Belgium
mireia.montanola@uclouvain.be, benoit.macq@uclouvain.be
†Alcatel-Lucent Bell Labs, Copernicuslaan 50, B-2018, Antwerp, Belgium
patrice.rondao_alface@alcatel-lucent.com

***Abstract*—In the context of 3D watermarking, most of the state-of-the-art techniques analyze a 3D/3D approach where the insertion and the extraction of the mark take place over the object itself, whereas the most common use of 3D objects is done through its 2D projections or eventually stereovision. In this paper we present a work in progress in the context of 3D watermarking that introduces an asymmetrical approach 3D/2D which will allow the extraction of the mark without access to the 3D object. The extraction will be carried out from one or several 2D views with the aim of protecting the Intellectual Property Rights associated to the object in its projections.**

***Index Terms*—3D mesh; watermarking; visualization.**

I. Introduction

Over the last years the “digital rights management” (DRM) issue has been addressed for different types of information in order to protect the data from piracy [11]. 3D watermarking is a well known technique that is being introduced and combined with cryptography in current DRM systems. It consists in protecting the Intellectual Property Rights (IPR) related to a content by means of insertion of a secret mark in an imperceptible and also in a robust way.

Watermarking techniques provide rights protection for multiple contents: audiovisual documents, images, videos and 3D objects. 3D models are distinguished firstly by their usually associated irregular sampling and secondly by the fact that they deal with arbitrary geometry varieties (manifold) immersed in a threedimmensional Euclidian space. As a consequence of this particular nature the usual tools for signal processing commonly used by regular signals such as audio, image or video, cannot be directly applied to 3D objects. Accordingly, 3D watermarking techniques do not present the maturity of their equivalents for other types of contents.

In addition, the current watermarking algorithms generally insert a secret mark into the 3D model, and need this watermarked object to be able to detect or read the mark [5]. However, the most common use of 3D models is done through their visualization (2D or eventually stereo). For instance, 3D models are commonly used in image or video content, for example in 3D cinema via stereo or home entertainment applications using 3D meshes. While it is possible to reconstruct a 3D object from several of its 2D views, to decode the mark in this context is presently difficult or even impossible (see Fig. 1).

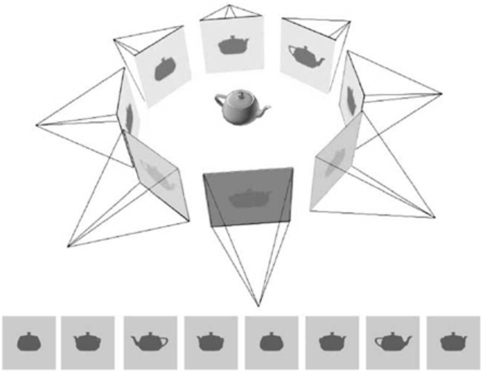


Fig. 1. The Teapot model and a set of projections depending on the angle of visualization. The model can be reconstructed from a certain number of its 2D representations.

This is the problem that we propose to address in this work. Contrary to the ordinary 3D watermarking techniques, for which the insertion and the extraction of the mark take place over the 3D model itself (3D/3D approach), we plan to develop techniques following an asymmetric approach 3D/2D in which the extraction of the mark could be done over one or several 2D projections of the 3D object without having access to the latter (see Fig. 2).

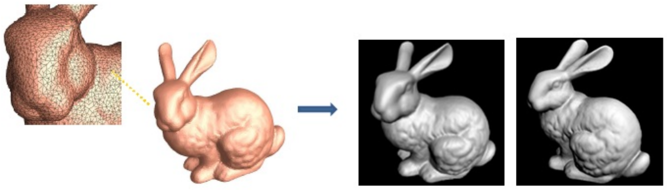


Fig. 2. The original Bunny 3D mesh and two of its 2D representations. The scheme we want to develop should provide the capability of reading the watermark from one or several representations of the object.

In this particular setting, we can name the pionneer work of J. Bennour et al. [1], who propose an insertion in the silhouette of a 3D object in several of its views belonging to image or video contents using this 3D object. This technique is able to detect the watermark only for certain views. However, given the fact that a minimal change in the visualization angle leads to a different silhouette along with the lack of knowledge of

ISBN: 978-1-61208-129-8

the orientation angles of the object in its visualizations, the object may not be protected. On the other hand, in the work presented by E. Garcia et al. [2], the 3D model is protected via its texture. The texture is defined by a set of images projected onto the object surface. However, this technique requires a texture rich enough to provide a robust watermark. Hence this results to a compromise between the texture richness and the robustness of the 2D hidden mark. These original methods still present limitations in both theoretical and application levels which this work intends to study.

This paper is organized as follows. The second section introduces the industrial applications of the techniques we will develop. The third section is devoted to present the proposed work and the way we plan to address it. The last section presents the conclusions.

II. Applications

Digital watermarking finds already applications related to image and video content. The most suitable application is copyright protection, where the watermark identifies the buyer of the digital content. This mechanism allows to prevent piracy by discouraging people to make illegal copies of a protected content.

3D multimedia content has also to be protected. 3D watermarking finds industrial applications such as 3D cinema, 3D videoconference, video games, home entertainment or augmented reality applications in smartphones. The idea is to protect the intellectual property rights of 3D meshes as well as their visualization.

In this particular work, the proposed watermarking schemes of 3D content respond to several cases of different applications. Among them, 3D watermarking based on Quantization Index Modulation (QIM) is suitable in a blind and semi-fragile context, and it is useful in data authentication applications.

The 3D watermarking technique making possible a detection from a certain number of 2D representations of the object is useful for copyright protection tools in the case of a 3D object immersed in 2D video or 3D stereo content. This need suits several domains: digital cinema, video games or home entertainment, audiovisual content players (for example Blue-Ray) or screens and stereovision glasses used to visualize 3D content.

On the other hand, this work will study the projections more suitable to watermark in order to protect the whole 3D model. In particular, the analysis of feature points or lines, roughness properties and local texture on the surface will be performed. This offers new perspectives in the watermarking of 3D physical objects either immersed or not in an augmented reality environment, like 3D videoconference or mobile applications in smartphones related to the 3D navigation.

III. Proposed Work

Presently there exist several schemes which provide a robust protection of the intellectual rights of a 3D object by embedding a secret and imperceptible mark [4]. But there is a lack of protection of the digital rights of visual representations of 3D models.

This work will focus on the extraction of the information (originally hidden into a 3D model) from one or several views of the object, with the interest of protecting the intellectual rights associated to the model in its 2D visual representations. Figure 3 summarizes the scheme of our work.

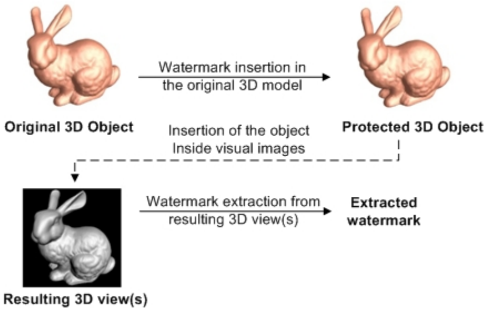


Fig. 3. A secret mark is first inserted in the 3D original model, protecting the content. The watermarking scheme has to be able to resist the visualization transformation and recover the mark from one or several 2D representations of the object.

Given the uncertainty related to the angles (2D views) in which the 3D object will be visualized, the use of silhouettes as in the work presented by J. Bennour et al. [1] does not seem as an optimal choice, since a slight variation in the direction of view would completely modify the silhouette.

We plan to perform an automatic detection of robust feature points in the 3D object around which robust neighbourhoods will be defined, performing local watermarks along the surface. Feature points detection together with neighbourhood definition could be based on the results presented in the work by P. Rondao Alface et al. [4]. To insert the mark we plan first to make use of the Spread Transform Dither Modulation (STDM) technique [3]. STDM is a variant of QIM technique and present good properties in the frame of 3D blind watermarking. Other variants of QIM will be tested for performing the insertion, as well as other state-of-the-art techniques, either in the spatial or the spectral domain, evaluating their pertinence in the frame we are dealing with.

Since there are many parameters for visualizing a 3D mesh (rendering, angle, texture, resolution, etc), we will address the problem step by step, defining a scenario with the simplest conditions in the first instance, and building increasing complexity models afterwards. The impact of the rendering process can be seen in Figure 4, which shows the rendering when using different shading techniques. As a matter of fact, different parameters in the rendering (shading, lighting models, etc) will have different impacts over the watermark which will be evaluated. We plan to develop a model of the visualization attack that the 3D object undergoes in the 2D to 3D transformation. We will study the distortion the model undergoes and the impact on the watermark. This model will extend the works discussed in [1, 2] and will be based on the analysis of the whole transformation chain that a 3D object undergoes from the watermarking process until the complete or partial reconstruction from its visualization in 2D.

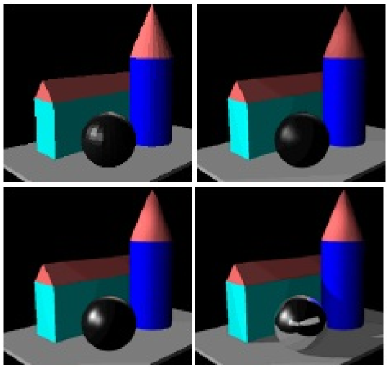


Fig. 4. From left to right: the same scene rendered with flat shading, Gouraud shading, Phong shading and Ray Tracing respectively (data courtesy of [9]).

One of the important steps is based on the use of depth maps, which can be for example obtained from stereovision. Depth maps give information of the depth of every pixel of the 2D view. Hence these depth maps provide part of the 3D mesh structure and should help in recovering the mark. The issue can then be brought to the case of ”cropping” of the 3D object as well as a regular remeshing (resampling of positions of 3D points), depending on the resolution and accuracy of the depth estimation (see Figure 5). However, whereas the state-of-the-art blind and robust 3D watermarking schemes already withstand combinations of a wide variety of attacks (noise addition, simplification, smoothing, etc), there is a lack of blind schemes which can withstand the cropping attack and the subsequent de-synchronization. We will study in particular the state-of-the-art methods for depth map reconstruction from several projections as well as the impact of the resolution (image) of the captured views, and possibly the impact of their compression [6, 7, 8]. As a matter of fact, topological errors may appear due to obstructions when there is not enough number of used views, such as the presence of holes or disconnected parts of the mesh. Few watermarking schemes provide resistance to this kind of errors in the mesh reconstruction [5].

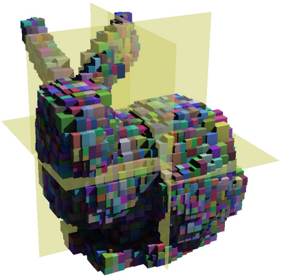


Fig. 5. A simulation of the Bunny model after the visualization process. The depth component for every pixel of the image resulting from the projection transformation is represented by blocks. The problem can be brought to the case of ”cropping” of the 3D Bunny model together with a regular re-sampling attack (data courtesy of [10]).

Other factors likely to influence the reconstruction accuracy, the selection of the views or the watermarking robustness, such as the presence of textures, the surface curvature and roughness and the presence of edges, feature points and feature lines, will be as well examined in detail subsequently in order to refine the watermarking algorithms.

The goal is thus to determine if the mark can be detected from a number of views lower than what is required to reconstruct the 3D object (without watermark) with good quality.

Subsequently this technique could be extended to animated 3D objects as well as the study of the impact of the 2D views resolution and their compression (for example via JPEG) over the robustness of the watermark. Another possible extension consists in the evaluation of the watermarking of solid objects which are later numerized (from one or several 2D representations) with the intention of their use into augmented reality applications.

IV. Conclusion

In this paper, we presented a work in progress based on the protection of the intellectual rights of 3D objects through their 2D projections. Since the most common use of 3D models is done through their visual projections in 2D or stereovision, we plan to develop 3D watermarking schemes able to resist the visualization process undergone by the object during the 3D to 2D transformation. This will help to protect the 3D models in its 2D representations.

References

[1] J. Bennour and J.-L. Dugelay. *Protection of 3D object visual representations*. ICME 2006, IEEE International Conference on Multimedia & Expo, July 9-12, 2006, Toronto, Canada, pp. 1113-1116

[2] E. Garcia and J.-L. Dugelay. *Texture-based watermarking of 3D video objects*. IEEE Transactions on Circuits and Systems for Video Technology, vol. 13, no. 8, pp. 853-866, Aug. 2003.

[3] R. Darazi, R. Hu, and B. Macq. *Applying spread transform dither modulation for 3D mesh watermarking by using perceptual models*. In International Conference on Accoustic Speech and Signal Processing 2010. ICASSP 2010. IEEE International Conference on, pp. 1-5 Mar. 2010.

[4] P. Rondao Alface, B. Macq, and F. Cayre. *Blind and robust watermarking of 3D models: How to withstand the cropping attack?*. In Image Processing, 2007. ICIP 2007. IEEE International Conference on, 5, pp. V-465 - V-468, Sept. 2007.

[5] P. Rondao Alface and B. Macq. *From 3D Mesh Data Hiding to 3D Shape Blind and Robust Watermarking: A Survey*. T. LNCS Transactions on Data Hiding and Multimedia Security 2:91-115, 2007.

[6] C.H. Esteban and F. Schmitt. 2004. *Silhouette and stereo fusion for 3D object modeling*. Computer Vision Image Understanding, vol. 96, no. 3 (Dec. 2004), pp. 367-392, 2004.

[7] N.D.F. Campbell, G. Vogiatzis, C. Hernandez, and R. Cipolla. *Automatic 3D object segmentation in multiple views using volumetric graph-cuts*. In 18th British Machine Vision Conference, vol. 1, 2007.

[8] Y. Yemez and C.J. Wetherilt. *A volumetric fusion technique for surface reconstruction from silhouettes and range data*. Computer Vision and Image Understanding, vol. 105, pp. 30-41, 2007.

[9] A. Oebbeke. http://www.glossar.de.

[10] V. Lu. https://netfiles.uiuc.edu/victorlu/www.

[11] S. Haber, B. Horne, J. Pato, T. Sander, and R. E. Tarjan. *If Piracy is the Problem, Is DRM the Answer?*. in E. Becker, W. Buhse, D. Gnnewig and N. Rump (Ed.): Digital Rights Management - Technological, Economic, Legal and Political Aspects. Heidelberg: Springer-Verlag, pp. 224-233, 2003.

Representative Picture Selection from Albums

Gábor Szűcs, Tamás Leposa, Sándor Turbucz
Dept. of Telecommunications and Media Informatics
Budapest University of Technology and Economics
Budapest, Hungary, 2nd Magyar Tudósok Krt., H-1117
e-mail: szucs@tmit.bme.hu, lepi_t@sch.bme.hu, sandor.turbucz.work@gmail.com

Abstract—**The paper is concerned with managing image album, where the picture set (containing very similar or different images) in each album is given. The goal has been to select the most representative pictures from the album. Our solution is based on the clustering of the images. The developed clustering procedure takes the large variety of the pictures and different type of image features into account. We have solved the incomplete feature value problem as well. The central pictures of the largest clusters are selected for representing the album.**

Keywords - representative picture; k-means++ clustering; qualitative and quantitative features; content features; EXIF data;

I. INTRODUCTION

There are lot of solutions and systems (e.g., FotoFile [3]) at the multimedia organization and retrieval; and demands are growing with new functionalities in the future too. A large part of these deals with personal photograph retrieval [2].

For a good organization the human persons usually put the pictures into albums as they wish (may be based on users' feeling). But there is a problem at the large set of pictures and albums: it is not easy to give representing issues (e.g., titles) for these albums. A representative image and its thumbnail is an ideal solution for this problem. The goal of our work has been to select the most representative picture from each album automatically without human interaction.

We consider very realistic situations, where the images come from different sources (camera, edited or created by software), the resolutions are various, the qualities are also different, so the variety of them is large.

For the above mentioned problem with realistic situations we present a solution in this paper. The structure of the paper is the following: Section II describes the background, our solution with clustering is detailed in Section III, brief conclusion and future works can be found in Section IV.

II. BACKGROUND

Finding the interesting photos from collections is a similar task to our goal, but the selection of them is always based on user feedbacks. (i) Commercial systems such as Flickr use an interaction mechanism for sampling the collection, it relies on social activity analysis for determining the notion of interestingness. Photo album creation can benefit from leveraging information learned from many users in regard of the album's content, structure, and semantics [4]. (ii) An alternative technique [1] is based on content analysis, the solution uses the combination of visual attention models and an interactive feedback mechanism to compute interestingness.

For representative photo selection and smart thumbnailing an other solution [5] uses the results of near-duplicate detection. Near-duplicate photo pairs are first determined, and the relationships between them are modeled by a graph. The most typical one is then automatically selected by examining the mutual relation between them. For smart thumbnailing, the region-of-interest of the selected representative photo is determined based on locally matched feature points, which is a view different from conventional saliency-based approaches [6][7].

The related works in the topic of representative images have solved the problem in three different ways: textually interesting [4], mutually distinct [5] and presence of faces in the image [12] or combination of them [1]. These works have used content features of the images. Only one or two papers have mentioned a few EXIF data, but these have been the *time* and the *camera name* [12] only. The works have not dealt with all EXIF data from camera; these metadata could be equal to content features.

III. SOLUTION WITH CLUSTERING

In picture selection procedure different types of features can be considered. If we consider only content features of the images, then the search space for the most representative picture is narrow. If we take both the content and the metadata features (from the camera) into account, then the search space will be wider. In this wide space the search procedure may find easier the most representative picture. The consequence of the narrow space is the possibility to take a bad decision in selection of the most representative picture. E.g., if all photos – except one or few – are taken by flash, then users probably will not consider a *photo without flash* as a representative picture. Another example is about the focal length: if all photos – except one or few – are taken with ordinary focal length (tableau), then a picture with small focal length (portrait) will not representative. Thus our solution considers both the content and the metadata features (EXIF data from the camera).

A. *Overview of the picture selection procedure*

The first idea for selecting the most representative picture in an album is choosing the central picture in the place of the pictures, where the place of the pictures is a vector space, and each picture is transformed into a point in this space. In this space the central picture can represent the whole set in an album.

But it occurs many times, that the album consists of different larger groups of images, where the pictures are similar in a group and far away between groups. In this case a strange situation can occur, where the distance between the central picture and the others is large (this image is alone), and the central picture will not be representative.

In order to avoid this often occurring situation we suggest a solution using clustering. After the clustering the procedure suggests the nearest picture of central point of the largest cluster. The solution contains some phases:

- Content feature values are extracted from the pixel data of the picture.
- Metadata features are the EXIF (Exchangeable Image File Format [9]) data.
- Clustering algorithm calculates the clusters of the pictures.
- Central point is determined of each cluster.
- The closest picture to the central point – namely the central picture – is marked as candidate for selection.
- The central picture of the largest cluster is selected for representative picture.
- If it is necessary more than 1 picture for representing the whole album, then central pictures of the second largest, third largest, etc. cluster are selected.

B. *Features for clustering*

In our picture selection solution the challenge has been taking both the content and the metadata features into account (correlation may occur between features). The content features are based on the statistics of RGB values of the picture points: mean, variance, mode, range, quartiles. There are 3 features for each statistical type: a feature related to red (R), one related to green (G) and one related to blue (B). These content features characterize the image and the values of them are indifferent from the orientation and size of the pictures.

The metadata features are the EXIF data of the pictures made by cameras. These data are not always available because an album can contain not only photos, but drawn, animated, edited pictures as well. (The absent metadata features naturally may influence the goodness of the result.) The used metadata have been the all accessible EXIF data: exposure program, contrast, flash, light source, metering mode, saturation, scene capture type, sharpness, white balance, image orientation, exposure time, F number, focal length in 35 mm film, ISO speed ratings. Some of these features are qualitative, others are quantitative. E.g., exposure program may be portrait, landscape, sea, mountain, etc. (so this is a qualitative feature), flash is a binary feature with two values: yes or no.

C. *Clustering algorithm with content and metadata features*

The content features are always available, nevertheless metadata features may be partly or totally deficient, which leads to problem in the comparison of pictures.

A distance (similarity) value needs for every picture pair for the clustering. The difficulties come from the different feature types (content and the metadata features), the different scales (qualitative and quantitative), and the deficient metadata.

The quantitative values of the pictures can be presented in a vector space, where each coordinate axis is a quantitative feature; and each picture is a point in this vector space. The distance between two pictures is calculated by the Euclidean distance, the number of all features gives the dimension.

If one of the values (of two pictures) at some features is missing, then the Euclidean distance formula can not be used. Let us omit the squared differences in the sum, where one of the two values is missing. The number of the rest of features is *k*. Instead of the Euclidean distance formula we have used normalized version of the distance (related to 1 dimension) for the missing value problem.

We have used normalized values (between 0 and 1) in x_i and y_i for the quantitative features, but we should have solved the other problem – distance calculation for qualitative features – as well. In the clustering the most frequent work is calculation the distance between a point and a cluster. For this we have used an idea about the histogram of the given cluster. Let us denote the mode of the histogram by *mode*, the frequency of the given x_i feature by $f(x_i)$. The distance only in the examined coordinate axis is defined in (1), where C is the examined cluster.

$$d_i(x_i, C) = \frac{f(mode) - f(x_i)}{f(mode)} \qquad (1)$$

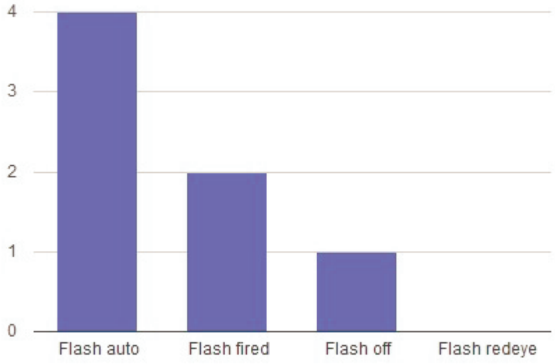


Figure 1. Example histogram for a qualitative feature

If the cluster contains only one element, then this will be zero or one. Fig. 1. shows an example histogram for a flash qualitative feature, where the values can be “flash auto”, “flash fired”, “flash off” or “flash redeye”. The mode is the “flash auto” and the frequency of this is 4. In the comparison of this cluster and an image four different results can be: if at the examined picture the feature is “flash auto” (“flash fired”, “flash off”, “flash redeye”), then the distance is 0 (1/2, 3/4, 1 respectively).

In order to tune the relative importance of the features, particularly the balance between the content and the metadata features, we have introduced weights for each feature, and the distance between two pictures is modified as can be seen in (2). In the current phase of our implemented system the w_i weights have been determined manually (balanced between content and metadata features) based on the results of thousand pictures (there was a fine tuning), but we have intend to estimate the weights by automatically using supervised learning, where albums and their most representative pictures are given as training set.

$$d(x, y) = \sqrt{\frac{1}{k}\sum_{i=1}^{k} w_i (x_i - y_i)^2} \qquad (2)$$

D. The k-means++ clustering algorithm

The k-means method is a widely used clustering technique that seeks to minimize the average squared distance between points in the same cluster. Although it offers no accuracy guarantees, its simplicity and speed are very appealing in practice (it is standard practice to choose the initial centers uniformly at random from *X* space). By augmenting k-means with a simple, randomized seeding technique, a new algorithm, so called k-means++ [10] has been outlined with the optimal clustering. Preliminary experiments show that the augmentation improves both the speed and the accuracy of k-means.

The k-means algorithm begins with an arbitrary set of cluster centers, but k-means++ algorithm uses a specific way of choosing these centers. At any given time, let D(x) denote the shortest distance from a data point x to the closest center we have already chosen; so k-means++ algorithm is the following:

- 1a. Choose an initial center c_1 uniformly at random from *X*.
- 1b. Choose the next center c_i, selecting $c_i = x' \in X$ with probability p, where p can be calculated by (3).

$$p = \frac{D(x')^2}{\sum_{x \in X} D(x)^2} \qquad (3)$$

- 1c. Repeat Step 1b until we have chosen a total of *k* centers.
- 2. For each $i \in \{1, . . . , k\}$, set the cluster C_i to be the set of points in *X* that are closer to c_i than they are to c_j for all $j \neq i$.
- 3. For each $i \in \{1, . . . , k\}$, set c_i to be the center of mass of all points in $C_{i,}$, as can be seen in (4), where ${}_m c_i$ and ${}_m x$ is the m^th^ coordinate of the c_i point and x point respectively.

$${}_m c_i = \frac{\sum_{x \in C_i} {}_m x}{|C_i|} \qquad (4)$$

- 4. Repeat Steps 2 and 3 until C no longer changes [10].

Choosing the number of the clusters in k-means++ algorithm is a sensitive parameter for the goodness of the results. We have used the rule of thumb formulated in (5) for the determination of the clusters.

$$k = \sqrt{n/2} \qquad (5)$$

After the clustering the closest picture to the cluster central point of the largest cluster is selected for the most representative image. Our solution is able to select more than 1 picture for representing the whole album with choosing central pictures of the second largest, third largest, etc. clusters. This will be very useful at characterization of large image sets, where not only one picture characterize the all images. At this case similar pictures at the selection would be a wrong result, which is avoided in our solution because of very different pictures.

E. Results

We have implemented our ideas and solution described above in Python programming language. The Python Imaging Library (PIL) [11] has been used for the image handling. This library contains useful functions for basic content features. The extraction of EXIF features has been solved also in Python.

The method has been just now implemented, the evaluation can be subjective (users’ decisions may be based on emotion anyway). There is subjective evaluation of images in other works (e.g. consumer photography [8]) as well.

We have used the implemented program for personal images. In Fig. 2. a little part of the album can be seen: two rows are the results of the clustering and the pictures with different border are the central images. The cluster of the pictures in the bottom row is largest cluster, so the 6th image is the most representative picture in the album.

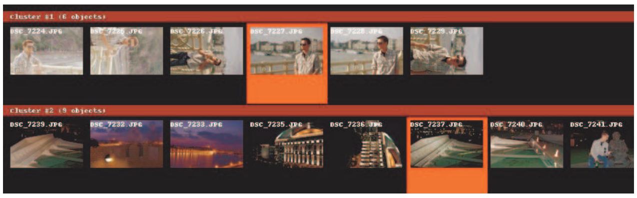


Figure 2. Examples for clustering and picture selection

F. *Experimental evaluation*

We have collected 600 pictures from different sources (camera with EXIF data, other sources without EXIF data), and we have organized them in 20 albums with different topics (party, holiday, town, or unified mood, etc.). Three human evaluators have selected the most representative pictures (as first in the order), then second ones, etc., so they have sorted the pictures in each album. Our implemented solution has also selected a picture (as most representative one) in each album. These machine results have been compared with the aggregated order of three human decisions (the aggregation is based on Borda method). The machine results are not always the first in the humans' order, but at 25% of albums they are in the best 5 representative pictures (denoting by p_5=25%). Furthermore we have summarized how many cases, where the machine results are in the best 10 representative pictures, we have counted 14 cases in 20 albums, so this is 70% (denoting by p_{10}=70%). These figures are not excellent, but good enough. We have investigated the humans' order, and we have concluded that humans' decisions are dispersing. With cross-validation only two humans' order were considered and aggregated, then were compared with the most representative pictures of third person. The comparison results of cross-validation for the first person: p_5=30%, p_{10}=85%, for the second person: p_5=40%, p_{10}=75%, for the third person: p_5=45%, p_{10}=70%. These figures present that our automatic solution is almost good as humans' decisions.

Conclusion and Future Work

This paper presents a description of a work in progress with new idea. The aim is to find the best way to automatically choose a picture from an album in order to be the best representation of it. The new idea is the consideration (in clustering) of different type of image features (content and EXIF data) with incomplete feature value possibilities. After clustering the central pictures of the largest clusters will be selected for representing the album. We have implemented this idea in Python and

In calculation of distances for clustering many features are considered, but the set of features can be expanded. We are at the beginning of this research, we intent to take more features – like texture, local features, time-based features – into account. Further development will be the automatic calculation of weights in distance formula.

Acknowledgement

The research was supported by ETOCOM project (TÁMOP-4.2.2-08/1/KMR-2008-0007) through the Hungarian National Development Agency in the framework of Social Renewal Operative Programme supported by EU and co-financed by the European Social Fund.

References

[1] K. Vaiapury and M. S. Kankanhalli, "Finding interesting images in albums using attention", Journal of Multimedia, Vol. 3., Num. 4., October 2008, pp. 2-13.

[2] P. D. B. Bujac and J. Kerins, "Developing and implementing a sparse ontology with a visual index for personal photograph retrieval" AI & Society, 2009, Vol. 24, Num. 4, pp. 383–392, DOI: 10.1007/s00146-009-0221-6.

[3] A. Kudhinsky, C. Pering, M. L. Creech, D. Freeze, B. Serra, and J. Gvvizdka, "FotoFile: a consumer multimedia organization and retrieval system", CHI '99 Proceedings of the ACM SIGCHI conference on Human factors in computing systems: the CHI is the limit, May 15-20, 1999, Pittsburgh, Pennsylvania, USA , pp. 496-503, DOI: 10.1145/302979.303143.

[4] S. Boll, P. Sandhaus, and U. Westermann, "Semantics, content, and structure of many for the creation of personal photo albums", ACM Multimedia '07, Proceedings of the 15th international conference on Multimedia, Augsburg, Germany, September 24-29, 2007, pp. 641-650, DOI: 10.1145/1291233.1291385

[5] W.-T. Chu and C.-H. Lin, "Automatic selection of representative photo and smart thumbnailing using near-duplicate detection" MM '08, Proceeding of the 16th ACM international conference on Multimedia, Vancouver, Canada, October 26-31, 2008, pp. 829-832, DOI: 10.1145/1459359.1459498.

[6] L. Itti and C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention" Vision Research 40, 2000, pp. 1489–1506

[7] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for Rapid Scene Analysis", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 20, Issue 11, 1998, pp. 1254-1259, DOI: 10.1109/34.730558.

[8] A. E. Savakis, S. P. Etz, and A. C. Loui, "Evaluation of image appeal in consumer photography", In Proceedings SPIE Human Vision and Electronic Imaging San Jose, CA, 2000, pp. 111–120.

[9] Japan Electronic Industry Development Association, Digital Still Camera Image File Format Standard, (Exchangeable image file format for Digital Still Camera:Exif) Version 2.1, Dec. 1998, http://www.exif.org/dcf-exif.PDF (last access date: 2011-01-24).

[10] D. Arthur and S. Vassilvitskii, "k-means++: the advantages of careful seeding", SODA '07 Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, 2007, pp. 1027–1035.

[11] http://www.pythonware.com/library/pil/handbook/ (last access date: 2011-01-24).

[12] E. Potapova, M. Egorova, and I. Safonov, "Automatic Photo Selection for Media and Entertainment Applications", GraphiCon'2009, Proceedings of The 19th International Conference on Computer Graphics and Vision, October 5-9, 2009, Moscow, Russia, pp. 117-124.

Accelerating Image Processing in Flash using SIMD Standard Operations

Chamira Perera[†], Daniel Shapiro[‡], Jonathan Parri[‡], Miodrag Bolic[‡], Voicu Groza[‡]
[†]*Systems and Computer Engineering, Carleton University*
[‡]*Computer Architecture Research Group, University of Ottawa*
[†]*cperera@sce.carleton.ca,* [‡]*{dshap092, jparr090, mbolic, groza}@site.uottawa.ca*

Abstract— **Flash applications have played an integral role in shaping the interactivity of the Internet. Desktop Flash applications feature vector-based processing such as image and video processing to enhance the user experience. In response to these needs, Adobe has added graphics card based acceleration for vector processing in Flash applications starting with Flash Player 10. This solution is limited to computer systems that have the proper graphics card. In this paper, we investigate the possibility of making explicit use of Single Instruction Multiple Data instructions, specifically SSE in the Intel x86-64 platforms, to accelerate vector operations in a Flash application. We also discuss certain limitations of the Flash virtual machine. The data reveals that a 90-92% speedup can be achieved by using SSE instructions to accelerate the alpha blending image processing algorithm in a Flash application. The SSE instructions are accessed by providing a standardized limited native interface to the Flash application.**

Keywords-SIMD; image processing; native code interface; image processing acceleration; virtual machine; Flash

I. INTRODUCTION

Since the advent of Web 2.0, Adobe Flash has played an important role in making websites interactive and fun to use. An example of this is the well-known YouTube service, which allows users across the world to share and stream videos. ActionScript, currently in version 3, is the programming language used to create Flash programs. Flash's ActionScript, just like Java, is an interpreted language. The ActionScript Virtual Machine (AVM) performs this interpretation. The advantage here is that the AVM allows Flash applications to run in a platform independent manner.

The inclusion of various image, video, and audio type processing makes Flash applications feature rich. The data input to the application must be processed in a timely manner. Otherwise, the application will not be considered to be enjoyable. Flash applications are interpreted and so they are not expected to perform as quickly as applications that are compiled and linked for a particular hardware platform, but are often expected to meet a user's perceived real-time perspective.

The Single Instruction Multiple Data (SIMD) instructions allow programs that feature vector-based processing such as image, video, and audio to be accelerated. The Intel x86 and x64 platforms support SIMD instructions inside the Central Processing Unit (CPU), and these instructions are known as Streaming SIMD Extensions (SSE). SSE instructions have been added to the instruction sets of modern CPUs to offer fast vector processing possibilities.

Currently, image and other types of vector processing can be accelerated in Flash programs using hardware acceleration from the Graphics Processing Unit (GPU). This acceleration is only available when used with certain graphics cards from Nvidia and ATI [1]. In addition to the GPU, SIMD instructions available on the CPU can also be used to accelerate vector processing for Flash applications. Unlike the expensive GPU, which is a computer system add-on, CPUs with SSE are more widely available and have no extra price tag.

In this paper, we show that SSE instructions available on an Intel x86 CPU can be used to accelerate the processing of images in a Flash application. This work follows closely in the footsteps of [2], which used SSE to accelerate Java applications using available SIMD instructions. The acceleration for Flash applications is performed by providing a limited native interface to a standalone Flash application so that it can access SSE instructions directly. This study focuses on accelerating the alpha blending image processing algorithm for color images in the RGB color space using SSE instructions.

Figure 1 shows the high-level view of the components involved in accelerating image processing in Flash applications using SSE. The native interface opened by the Flash application allows it to invoke functions in a Dynamic Linked Library (DLL), which in turn uses SSE instructions to accelerate the processing of images. We found no explicit support in the AVM to make use of SSE instructions from within Flash applications and it is not discussed in the available AVM literature.

The remaining sections of the paper are organized as follows: Section 2 provides background information on Flash applications, SSE, and alpha blending, Section 3 discusses related work, Section 4 discusses the implementation work of this study, Section 5 discusses the experimentation that was performed to evaluate the effectiveness of the proposed method and results, and we conclude by summarizing our results in Section 6.

ISBN: 978-1-61208-129-8

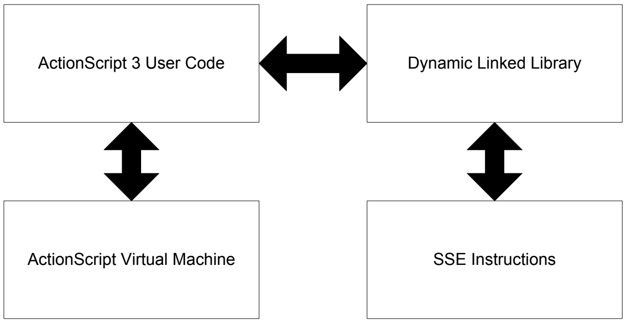


Figure 1. High-level organization of the method to accelerate image processing in Flash applications.

II. Background

This section provides background information on the technologies used in the study. Section 2.1 briefly outlines Flash, Section 2.2 discusses Intel's SIMD version: SSE, and finally Section 2.3 illustrates the alpha blending image processing algorithm.

2.1 Flash

An Adobe Flash program compiled into a .SWF file (pronounced as "swiff") runs on any web browser that has a Flash Player plug-in installed. Flash media can also run as a standalone application outside a web browser when packaged into an executable file on various operating systems (OS). These are known as Flash projectors. Having Flash Player allows Flash programs to run on any processor architecture and OS, which is similar to how Java programs run on top of a Java Virtual Machine.

The low level processing in a Flash program can be implemented using ActionScript 3. The ActionScript 3 code is compiled into ActionScript Byte Code (ABC) and packaged into a .SWF file, which is interpreted by the AVM when the Flash application is executed [3].

The Flash platform does not provide a native interface for Flash applications to invoke native code. The lack of native access can be attributed to the perceived security risks that it imposes, and the cross-compatibility that may result from calls to native code. For example, a rogue Flash application running in a web browser could hack the target machine, and the Flash VM of a cell phone is likely not equipped to execute native code for a desktop PC. We leave the problem of security and cross-compatibility for future work. The Zinc 3.0 Flash builder tool by MDM provides the ability to create a standalone Flash projector from a .SWF file and includes functionality in the projector to load a DLL [4]. This is the basis for our native interface.

2.2 SSE

SSE is Intel's version of SIMD instructions for their current and recent x86-64 CPUs and is the evolution of MMX. AMD processors also feature support for SSE. Support for SIMD initially started with the introduction of MMX and until today SSE has had many revisions and currently it is at revision 4.2 (SSE4.2) [12]. The SSE architecture features eight 128-bit wide vector registers. In 2010, Intel showcased its Advanced Vector Extensions (AVX), which feature 256-bit wide vector registers [6]. The first version of SSE provided more support for floating point operations compared to integer operations. In addition, it does not allow an instruction to pack smaller units of data (8-bit values) to vector registers.

SSE2 enhanced SSE by providing more support for integer arithmetic including operations to handle 8 and 16-bit data in SSE registers. This is ideal for image processing as the smallest size of data (e.g., a color channel of a pixel) in an image is usually an 8-bit value [5]. The net result of using SSE2 for image processing rather than SSE is that more pixels can be loaded into a single vector register and more parallelism can be achieved and exploited. For this study, SSE2 instructions were used to accelerate image processing.

2.3 Alpha-Blending

Simply put, alpha blending is an image processing algorithm that allows two images to be blended together [7]. An example is blue screen matting, where a newscaster can be superimposed in front of a particular background to give the illusion that he/she is actually in front of the background. To perform alpha blending, Eq. (1) can be used. FG and BG correspond to foreground and background images respectively. F is the resulting image from the blending, and α ranges from 0 to 1 inclusively. By changing the value of α, the contributions of the pixels from the foreground and background to the blended image can be controlled. To apply this equation to color images, all three color channels, i.e., red, green, and blue values have to be updated in the same manner.

$$F\,pixel = \alpha \times FG\,pixel + (1.0 - \alpha) \times BG\,pixel \qquad (1)$$

An optimized version of the alpha blending equation is shown in Eq. (2). This equation was derived from an equation shown in [10]. Unlike in (1), the values of α in this equation range from 0-255. Therefore, this equation assumes that the α values are read from an image. This image is known as a matte.

$$F\,pixel = \alpha \times (FG\,pixel - BG\,pixel) >> 8 + BG\,pixel \qquad (2)$$

Note that the right shift by 8 (divide by 256) introduces a certain amount of error in the blending as the value of α should be divided by 255 to ensure that the values are within the proper range (0 to 1). This error cannot easily be detected by the human eye. An example of a matte image is shown in Figure 4. Figure 5 shows an example of how alpha blending was applied to blend the background shown in Figure 2 and foreground shown in Figure 3 using the matte available in Figure 4. SIMD can easily be applied to accelerate alpha blending because the same operation is applied to all pixels and each color channel in the pixel.

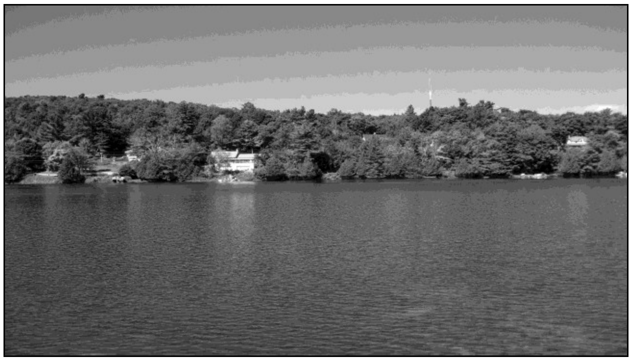


Figure 2. Example background image.

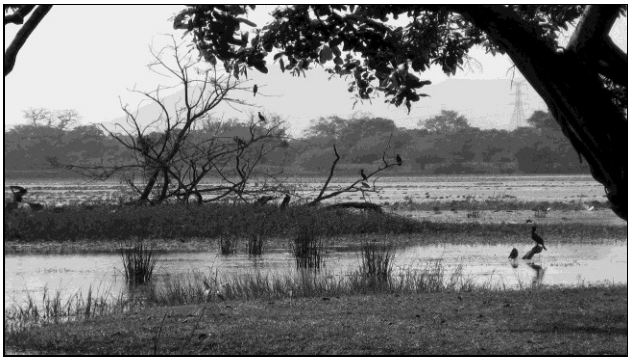


Figure 3. Example foreground image.

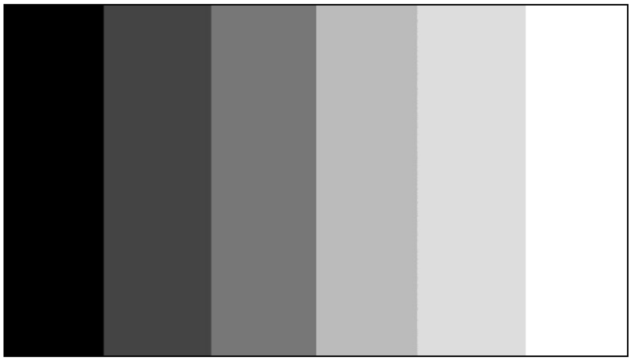


Figure 4. Example custom matte.

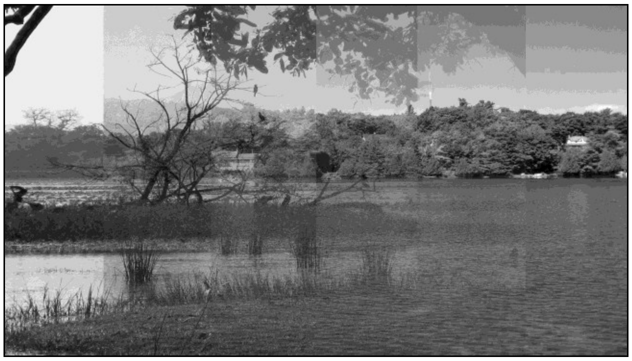


Figure 5. The alpha blended image.

III. Related Work

The work done in this paper is in line with the work of [2] to accelerate Java applications that perform vector operations using SSE instructions. Their work used the Java Native Interface (JNI) to access SSE instructions and accelerate vector operations in Java. Their vector operation was completely implemented using SSE instructions and wrapped in a function, which is part of a DLL and presented as a standardized API. Their results showed a 1.0x-4.0x speed up of the Java application with SSE acceleration over an application that implements the operation only in Java.

Usage of SSE instructions to accelerate image processing is not a new concept. This has been demonstrated in the Intel whitepaper [5], which makes use of SSE, SSE2 and AVX instructions to accelerate cross-fade and sepia filters in image processing. Since processors that support AVX instructions are still not mass produced, Intel used a processor simulator to simulate the AVX instructions. In [9], SSE2 instructions were used to speed up a Harris corner detector used in image processing.

Adobe provides the ability to accelerate image and other vector processing algorithms using the GPU in their Pixel Bender Toolkit [1, 11]. Using this toolkit a developer can implement a vector processing algorithm and compile it into byte code. In the Adobe literature, the vector processing algorithm produced by the toolkit is known as a shader [11]. This byte code is imported into the Flash application by the ActionScript code and is used to perform the vector processing. This capability was added starting with Flash Player 10. The usage of the GPU is opaque to the developer and there is no way to say for certain that specific operations are performed in the GPU or not. Another disadvantage of this approach is that the acceleration can only be achieved with certain GPUs: only a subset of those available from Nvidia and ATI. Overcoming this weakness, the work in this paper shows that the vector unit is a viable alternative to the GPU. Hence, a user should not have to invest in an additional

Copyright (c) IARIA, 2011. ISBN: 978-1-61208-129-8

graphics card but instead experience optimal acceleration from the already available vector unit in the processor.

IV. Implementation

To invoke SSE instructions in a Flash program, a limited native interface was added to Flash applications (see Figure 1). This approach is similar to the JNI, which allows Java applications to call native functions in a DLL written in other languages such as C++. The Flash platform does not provide a native interface that Flash programs can use. To overcome this issue the MDM Zinc 3.0 Flash Builder tool was used.

Prior to implementing an image processing algorithm, a generic vector processing algorithm was implemented in a manner similar to [2]. In their work, the data that is processed by the Java application is generated by the Java application and is processed by the native function. Due to severe limitations in the native interface provided via the Zinc 3.0 tool, no speedup of the algorithm was achieved when the Flash application relied on SSE instructions to accelerate the vector operation. This is because the native interface only allows the programmer to pass either primitive types (i.e., different size integers) or a string of characters, therefore, if a vector of integers have to be passed to the DLL to process and another vector has to be returned then each element in the vector has to be converted to a string of characters. Furthermore, this string of characters must be decoded and the original vector has to be reconstructed so that the data can be used. This operation of encoding and decoding of the elements of the vector consumes a large amount of time and this time alone was enough to offset the execution time improvement, such that the implementation of the vector operations in Flash was faster. We look to greatly improve on this in future work.

Instead of operating on vectors using data that are generated in the Flash application, a different approach was applied where the Flash application performs vector processing on file based data such as image files or audio files. The native code in the DLL performs the vector processing as before. The Flash application does not load the image file into its local memory space but instead it delegates that task to the DLL by passing it a path to the location of the file. After the images are loaded, the Flash application invokes a function that performs the image processing, i.e., the alpha blending. The native function saves the resulting blended image to a location specified by the Flash application. Once the control returns back to the Flash application, it loads the blended image to its own local memory space and displays the image to the user.

The native code to perform alpha blending was implemented using SSE2 instructions. The pseudo implementation of the algorithm using SSE2 instructions is shown in Figure 6. The underlying assumption in the implementation is that the total number of the pixels in each image (i.e., the foreground, background, and matte images) should be a multiple of 16. This is a reasonable assumption as almost all of the mainstream image resolutions (e.g., 720p, 1080p) have this property. The matte image is usually gray-scale but it was converted to an RGB image so that the algorithm could be implemented with less complications. Note that the RGB channel values for a gray-scale image are the same.

Initially, 16 bytes worth of pixels of each image is loaded into three SSE registers. The Portable Image Library (PIL) was used as an image handling library [8]. The PIL stores a pixel as a 3-byte value (each color channel occupies one byte), as it loads an RGB image and returns a handle to the location that it is stored in memory [8]. Because 3 bytes per pixel are needed, there can be at most five full pixels and one partial pixel in an SSE register.

All three images have to be interleaved with zeros because the SSE2 multiply operation only multiplies 16-bit data. Note that, because of the interleaving, 8 bytes of pixel data will be in one SSE register and the other 8 bytes will be in another register. After the operations required for the algorithm are performed, the bytes that correspond to the pixel data have to be extracted one by one and written to the buffer that holds the resulting image.

At first glance, this operation may seem very inefficient due to 16 individual memory writes by software. However, the performance of this approach was compared against another implementation that makes use of an SSE3 instruction (`PSHUFB`). This instruction can be used to shuffle the bytes in an SSE register around. With this instruction the two sets of interleaved data can be placed into one single SSE register and the contents of it can be written to memory using a single SSE2 instruction. When the performance was evaluated, the execution time was exactly the same as the original solution, and since SSE3 is supported only on newer Intel CPUs (Core 2 Duo, Core i7), the SSE3 approach was not pursued any further.

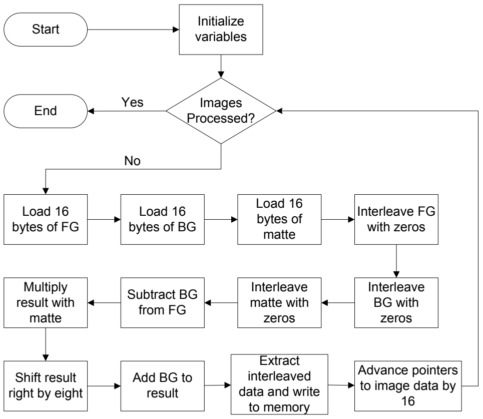


Figure 6. Pseudo SSE2 implementation of the alpha blend algorithm.

ISBN: 978-1-61208-129-8

V. Experimental Evaluation and Results

To evaluate the effectiveness of accelerating alpha blending in a Flash application using SSE instructions, the performance was measured for four different implementation possibilities, and the results compared. The approaches evaluated were: implementation of the algorithm with our accelerated Flash approach, only in Flash, only in C, and in C using SSE instructions. In the latter implementation, the DLL that was compiled for the Flash and SSE implementation was reused. The latter two implementations represent the best possible implementation in terms of performance as the program is compiled to run directly on the hardware without any interpretation. Alpha blending was performed on a pair of eight images with different image resolutions. In addition, eight matte images were also used for the blending. Only mainstream image resolutions were used for the evaluation as the total number of pixels is a multiple of 16.

The evaluation of all four implementations of alpha blending was done on an Intel Centrino Duo machine with 2.5GB of RAM. The OS used on the machine is Windows Vista. The C implementation and the DLL were compiled with optimization for speed enabled. In the Flash-only implementation, the entire set of pixels were extracted and stored in an instance of a vector class in ActionScript 3. This method provides the quickest possible implementation.

Fifty runs were performed on each resolution for each implementation and the execution time for each run was recorded. Note that only the image resolution and not individual pixel values affect the performance of the given operation. The average and the standard deviation were computed at the end of each run. Figure 7 shows pseudo code for the method that was used to measure the execution time of each implementation. The average execution times (in milliseconds) for each resolution used in the performance evaluation are shown in Table 1, while Table 2 shows the standard deviations of the execution times. Figure 8 depicts the average execution times shown in Table 1 in a graphical format.

```
start = get_sys_time_stamp()
alpha_blend()
end = get_sys_time_stamp()
execution_time = end - start
```

Figure 7. Method of measuring execution time.

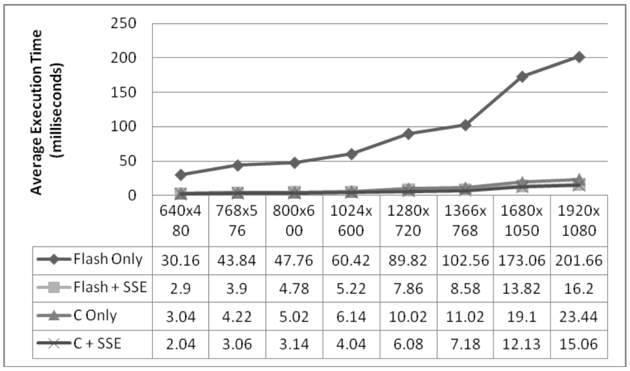
According to the results shown in Table 1 and Figure 8 it is evident that there is a speedup when a Flash program makes use of SSE instructions to perform vector operations over an implementation that solely relies on Flash to perform the computations. This speedup is summarized in Table 3. In addition, it shows the speedup of the implementation of the algorithm in C that uses SSE over the implementation that uses both Flash and SSE. We found that the Flash with SSE implementation has a speedup of 90-92% over a Flash-only implementation and a speedup of negative 7%-34% over a C and SSE implementation. The results were in line with expectations, as the native code was faster than the code running through the AVM. Since the standard deviations of the execution times were low it is evident that external factors such as garbage collection in the AVM and OS context switching have not influenced the variability of the runtimes in a drastic manner.

Table 1. Average execution times in ms for different image resolutions.

Image Resolution	***T_{Flash} (ms)***	***$T_{Flash+SSE}$ (ms)***	***T_C (ms)***	***T_{C+SSE} (ms)***
640x480	30.16	2.9	3.04	2.04
768x576	43.84	3.9	4.22	3.06
800x600	47.76	4.78	5.02	3.14
1024x600	60.42	5.22	6.14	4.04
1280x720	89.82	7.86	10.02	6.08
1366x768	102.56	8.58	11.02	7.18
1680x1050	173.06	13.82	19.1	12.13
1920x1080	201.66	16.2	23.44	15.06

Table 2. Standard deviations of the execution times for different image resolutions.

Image Resolution	***σ_{Flash}***	***$\sigma_{Flash+SSE}$***	***σ_C***	***σ_{C+SSE}***
640x480	0.65	0.416	0.282	0.282
768x576	1.076	0.364	0.418	0.24
800x600	1.733	0.932	0.141	0.405
1024x600	1.071	0.545	0.405	0.34
1280x720	1.119	0.808	0.141	0.274
1366x768	1.981	1.247	0.141	0.482
1680x1050	1.931	0.748	0.364	0.598
1920x1080	1.303	1.125	2.011	0.314



	640x480	768x576	800x600	1024x600	1280x720	1366x768	1680x1050	1920x1080
Flash Only	30.16	43.84	47.76	60.42	89.82	102.56	173.06	201.66
Flash + SSE	2.9	3.9	4.78	5.22	7.86	8.58	13.82	16.2
C Only	3.04	4.22	5.02	6.14	10.02	11.02	19.1	23.44
C + SSE	2.04	3.06	3.14	4.04	6.08	7.18	12.13	15.06

Figure 8. Graph of image resolution vs. average execution time.

TABLE 3. SPEED OF FLASH VS. FLASH+SSE AND FLASH+SSE VS. C+SSE.

Image Resolution	*Speedup Flash+SSE vs. Flash (%)*	*Speedup Flash+SSE vs. C+SSE (%)*
640x480	90.38	29.66
768x576	91.1	21.54
800x600	90	34.3
1024x600	91.36	22.61
1280x720	91.25	22.65
1366x768	91.63	16.32
1680x1050	92.01	12.22
1920x1080	91.97	7.03

VI. CONCLUSION

The purpose of this work was to prove that a Flash application that performs vector processing can be easily accelerated using SIMD instructions such as the SSE instruction sets on common x86-64 CPUs. Adobe already has a solution to accelerate vector-based processing using GPUs, however, this solution is only available to a limited number of GPUs, and moreover only to consumers who pay extra to have a GPU in their system. If the acceleration is provided via the vector unit in the CPU then it is available without having to invest money in a special GPU. The minimum system requirements to run the alpha blending algorithm outlined in the paper is a Intel Pentium 4 machine since it requires support for SSE2 and 128MB of RAM to run Flash Player 10.

This acceleration was proven only in a limited case where the data to be processed can be loaded from a file and the resulting data can be saved back into a file. In addition, the Flash application can only be executed as a standalone desktop application (Flash projector) and not in a web browser in its current state. The reason for this limitation is due to the fact that the current Flash platform does not have a native interface to call native functions in a DLL unlike Java, which has a fully featured JNI, and the fact that the Zinc 3.0 Flash builder only provides support to create standalone Flash desktop applications that feature a very limited native interface. The native interface is provided via loading a DLL and being able to invoke functions in it.

For this work, we implemented the alpha blending image processing algorithm using SSE2 instructions, compiled it into a DLL, and the Flash application invoked the algorithm using the native interface provided by the Zinc 3.0 tool. The experimental results show that there is a speedup of 90-92% when alpha blending is performed by a Flash and SSE implementation over a Flash-only implementation. The size of the image was orthogonal to the speedup. Even though only one image processing algorithm was considered in this study, due to the nature of most image processing and vector processing algorithms it can be concluded that by using Flash applications that have vector operations can be accelerated using SSE instructions. Such inherent parallel support should be included in the ActionScript specification and made available in the AVM.

The solution outlined in this paper neglects that a Flash based web application calling native code via a native interface can introduce a security risk and compatibility issues. To alleviate these security concerns the available options are to:

- Make Pixel Bender generate SSE instructions when it cannot detect the proper GPU on the computer system.
- Incorporate SSE directly into the AVM and augment ActionScript specification for better vector operation support.

In our upcoming work, we will address the security and compatibility concerns posed by exposing the vector instructions directly through generic calls to the VM. Mapped and complete SSE integration into the ActionScript specification and AVM will move ahead with an SSE bypass in the open source ActionScript virtual machine, Tamarin [13], produced by Mozilla and Adobe.

REFERENCES

[1] Adobe Systems Inc., “Pixel Bender release notes.” [Online]. Available: http://www.adobe.com/devnet/pixelbender/articles/releasenotes.html. [Accessed: Feb 1, 2011].

[2] J. Parri, J.M. Desmarais, D. Shapiro, M. Bolic, and V. Groza, “Design of a Custom Vector Operation API Exploiting SIMD Intrinsics within Java,” 23rd Canadian Conference on Electrical and Computer Engineering, pp. 1-4, May 2010.

[3] *ActionScript Virtual Machine 2 (AVM2) Overview,* Adobe Systems Inc., San Jose, CA, USA, 2007.

[4] MDM, “Zinc™ 3.0 - Rapid Application Development for Adobe® Flash.” [Online]. Available: http://www.multidmedia.com/software/zinc/. [Accessed: Feb. 1, 2011].

[5] *Image Processing Acceleration Techniques using Intel® Streaming SIMD Extensions and Intel® Advanced Vector Extensions,* Intel, Santa Clara, CA, USA, 2009.

[6] Intel, “Picture the future now Intel® AVX.” [Online]. Available: http://software.intel.com/en-us/avx/. [Accessed: Feb. 1, 2011].

[7] J.F. Blinn, “Compositing. 1. Theory,” *Computer Graphics and Applications, IEEE*, vol. 14, no. 5, pp. 83-87, September 1994.

[8] A. Whitehead, “Portable Image Library (PIL).” [Online]. Available: http://iv.csit.carleton.ca/~awhitehe/PIL/. [Accessed: Feb 1, 2011].

[9] J. Skoglund and M. Felsberg, “Fast image processing using SSE2,” in *Proc. SSBA Symposium on Image Analysis,* Malmö, Sweden, 2005.

[10] W. Shao, “Tip: An Optimized Formula for Alpha Blending Pixels.” [Online]. Available: http://www.codeguru.com/cpp/cpp/algorithms/general/article.php/c15989/Tip-An-Optimized-Formula-for-Alpha-Blending-Pixels.htm. [Accessed: April Feb. 1, 2011].

[11] *Programming ActionScript 3.0 for Flash,* Adobe Systems Inc., San Jose, CA, USA, 2009.

[12] S. Siewert, “Using Intel® Streaming SIMD Extensions and Intel® Integrated Performance Primitives to Accelerate Algorithms.” [Online]. Available: http://software.intel.com/en-us/articles/using-intel-streaming-simd-extensions-and-intel-integrated-performance-primitives-to-accelerate-algorithms/. [Accessed: Feb. 1, 2011].

[13] Mozilla, “Tamarin Project.” [Online]. Available: http://www.mozilla.org/projects/tamarin/. [Accessed: Feb. 1, 2011].

Scalable Video Coding Transmission over Heterogeneous Networks

Reuben A. Farrugia, Lucianne Cutajar
Department of Communications and Computer Engineering
University of Malta
Msida, MSD 2080, Malta
reuben.farrugia@um.edu.mt, lucycut88@hotmail.com

***Abstract*—Video streaming is currently occupying a huge chunk of the Internet bandwidth. This is mainly attributed to the wide variety of applications that are being transmitted over current Internet infrastructure, such as videoconferencing, mobile television (TV), Internet video streaming, and Internet Protocol TV (IPTV). These applications are generally encoded using the H.264/AVC codec which encodes the video content into a single layer stream with a fixed spatio-temporal video resolution. This poses a limitation for such applications since the same video content must be encoded into different streams in order to cater for heterogeneous devices demanding different spatio-temporal resolutions. This paper presents the performance evaluation of the recent H.264/SVC standard. The H.264/SVC encodes the video into different layers and the receiving device can decide to drop some layers in order to meet the required spatio-temporal resolution. This work shows that transmission of H.264/SVC using multicasting provides a substantial reduction in bandwidth requirement over traditional H.264/AVC. Simulation results further demonstrate that the H.264/SVC provides less congestion and is thus provides better Quality of Experience (QoE).**

***Keywords*-Computer Networks; H.264/SVC; Quality of Service; Scalable Video Coding; Video Streaming**

I. Introduction

Internet video is expected to consume 91% of the global consumer Internet traffic by 2014 [1]. The increase in popularity of multimedia content is accredited to the wide range of devices which make multimedia content available on several devices. Typical video streaming applications adopt the H.264/AVC standard [2] to deliver video content over the Internet. It achieves high compression efficiency relative to other standards and encodes the video content into a unique bitstream. Therefore, the generated bitstream is only suitable for a particular spatio-temporal resolution.

However, as shown in Fig. 1, different devices provide different requirements in terms of frame rate and image resolution. Therefore, the traditional H.264/AVC must generate different streams for different devices, thus becoming inefficient in terms of bandwidth utilization. For example, consider that the Main Video Server in Fig. 1 needs to transmit the same video content to two different devices; a mobile device and a High Definition (HD) Client. The standard H.264/AVC must encode two different streams, a lower resolution stream to mobile devices and a higher resolution stream to HD Clients.

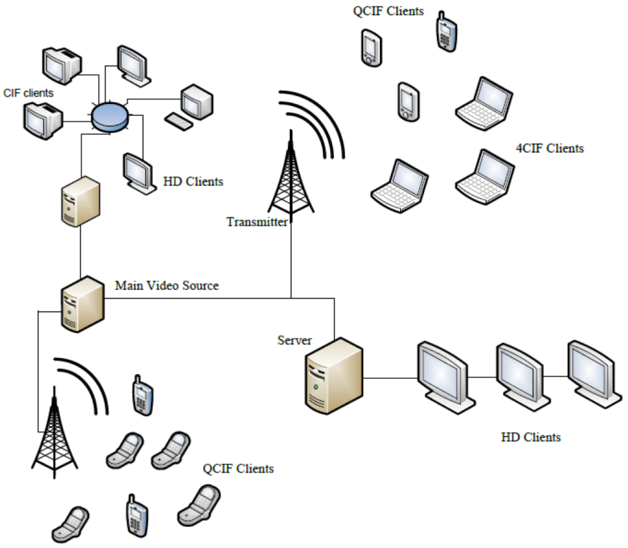


Figure 1. Typical Heterogeneous Network

Scalable Video Coding (SVC) [3] poses an attractive solution to the above mentioned problems encountered by the standard H.264/AVC codec and was recently introduced as an extension to the same standard. The H.264/SVC offers scalability by allowing the removal of parts of the video bitstream in order to comply with the various needs or preferences of the end user and to adhere to the network/device capabilities. Taking the above mentioned example, the H.264/SVC generates a unique bitstream that will be received by both devices. The HD client will decode the whole stream and thus recover the HD content, while the mobile device will drop part of the bitstream to recover a lower resolution version.

The authors in [5] have proposed a rate adaptation mechanism for H.264/SVC. On the other hand, this paper is aimed to analyze the performance of the H.264/SVC standard relative to traditional H.264/AVC codec in both unicast and multicast scenarios. Simulation results show that the H.264/SVC multicast is the most promising solution since it poses a bitrate reduction of 72 % relative to the traditional H.264/AVC unicast. This work further demonstrates that the H.264/SVC multicast transmission generates less packet loss

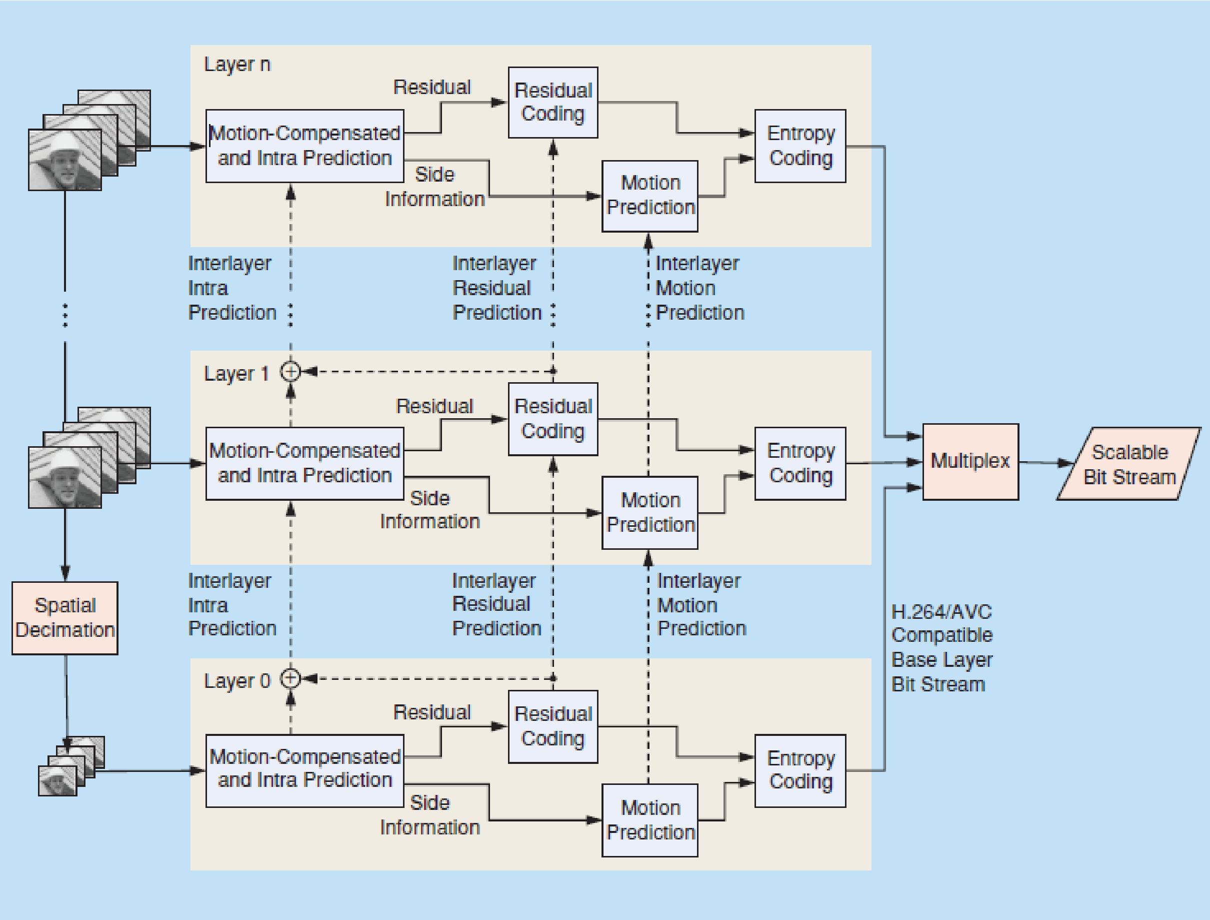

Layer n
Residual
Residual Coding
Motion-Compensated and Intra Prediction
Side Information
Motion Prediction
Entropy Coding
Interlayer Intra Prediction
Interlayer Residual Prediction
Interlayer Motion Prediction
Layer 1
Residual
Residual Coding
Motion-Compensated and Intra Prediction
Side Information
Motion Prediction
Entropy Coding
Multiplex
Scalable Bit Stream
Spatial Decimation
Interlayer Intra Prediction
Interlayer Residual Prediction
Interlayer Motion Prediction
H.264/AVC Compatible Base Layer Bit Stream
Layer 0
Residual
Residual Coding
Motion-Compensated and Intra Prediction
Side Information
Motion Prediction
Entropy Coding

Figure 2. Simplified H.264/SVC Encoder Structure (from [4])

due to congestion and thus provides a higher level of Quality of Experience (QoE).

This paper is organized as follows. The Scalable Video Coding paradigm is presented in Section II. Section III presents the methods and protocols adopted in order to transmit H.264/SVC content over the Internet. The simulation environment is described in some detail in Section IV followed by the simulation results in Section V. This paper is concluded with the comments and conclusion in Section VI.

II. Scalable Video Coding

The H.264/SVC is encoded using a layered structure that allows the user/device to derive the most appropriate spatio-temporal resolution. Therefore, scalable video coding enables the encoder to encode only once while decoding many times at different spatio-temporal resolutions. The same bitsteam is delivered to all devices (mobile devices, HDTV etc.). However, the required spatio-temporal resolution is achieved by dropping part of the bitstream. Fig. 2 outlines the basic encoding steps taken by the H.264/AVC encoder. Each representation of the same video content can be altered to various spatial and temporal resolutions. The number of layers utilized for decoder depends on the needs of the application i.e. higher spatio-temporal resolutions are achieved by increasing the number of enhancement layers. The following sub-sections introduce the theory after which the scalable video coding paradigm is based on. More information can be found in [3].

A. Temporal Scalability

Temporal scalability refers to the frame rate of the video representation. A higher temporal layer would imply a higher frame rate. The H.264/SVC is encoded to achieve the highest frame rate *T*. Applications which need a frame rate of *N*, where $N \leq T$, drop the temporal enhancement layers *M* within the range $N < M \leq T$.

The H.264/SVC employs the Hierarchical B-picture [6] for temporal scalability. Fig. 3 illustrates the three separate bitstreams which can be extracted and independently decoded to give three temporal layers: layer 0 $T0$ and enhancement layers $T1$ and $T2$. Devices which decode only the base layer achieves one ninth the full rate, while if layers $T0$ and $T1$ are decoded one third the full rate is achieved.

ISBN: 978-1-61208-129-8

On the other hand, in order to achieve full rate the base layer and all enhancement layers must be decoded.

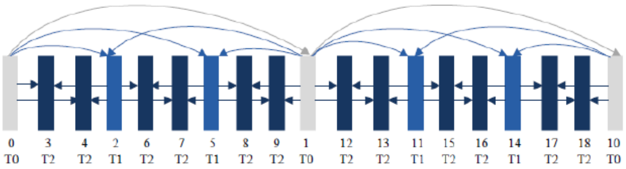


Figure 3. Hierarchical B-Picture structures for enabling temporal scalability (from [3])

The coding efficiency of H.264/SVC is dependent on the quantization parameters of each layer. This is because the motion-compensation prediction process of one layer is dependent on the other succeeding it. Therefore, the quantization parameters for the lower layers must not be very large in magnitude since this would result in reducing the image quality. Thus, the quantization parameters must be in increasing order of magnitude, with the uppermost layer having the largest quantization parameters.

B. Spatial Scalability

The Spatial Scalability process adopted by H.264/SVC employs the multilayer coding concept where each layer supports a particular spatial resolution. Similar to temporal scalability, the base layer provides enough information in order to reconstruct a low resolution video. Each enhancement layer enhances the video to a higher resolution. Therefore, the decoder can drop a number of enhancement layers in order to achieve the required spatial resolution. Fig. 4 illustrates typical spatial-scalability architecture. If only the base layer is decided the resulting frame will have an image resolution of 176×144. Increasing the number of layers will increase the spatial resolution to a maximum of 1704×576, which is the largest resolution supported by this system.

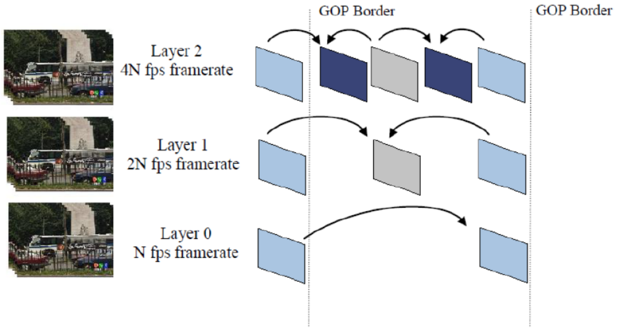


Figure 4. Inter-Layer Prediction for Spatial Scalability

In order to maximize coding efficiency, each spatial layer adopts both inter and intra predictions as for H.264/AVC. To improve coding efficiency, the inter-layer prediction is used to encode the enhancement layers. The lower resolution frames are upsampled and the similarities between the upsambled reference and the current frame are exploited using Inter-Layer prediction.

C. Spatio-Temporal Scalability

Both Spatial and Temporal scalable coding can be combined to form the spatio-temporal scalability. Fig. 5 shows a typical example of spatio-temporal scalability with a GOP of 8. The key pictures at the GOP borders are intra–coded. Higher data rates can be achieved by decoding more temporal enhancement layers. The frame structure illustrated in Fig. 5 adopts two spatial resolutions (spatial layer 0 at *QCIF* resolution and spatial layer 1 at *CIF* resolution). The upper stream adopts four temporal layers, with the top most enhancement layer being at a frame rate eight times the frame rate of the lower most base temporal layer. The lower stream employs three temporal layers.

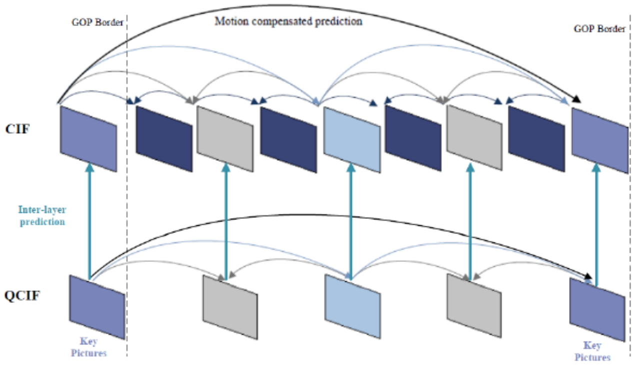
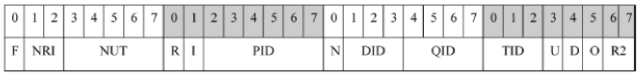


Figure 5. Spatio-Temporal Scalability (from [7])

III. Scalable Video Transmission

The H.264/SVC video related information is encapsulated within Network Abstraction Layer Units (NALUs). As illustrated in Fig. 6, the H.264/SVC standard employs a 4-byte header where the first byte is similar to the one adopted by the H.264/AVC, while the remaining bytes in the header indicate SVC related information.



0	1	2	3	4	5	6	7	0	1	2	3	4	5	6	7	0	1	2	3	4	5	6	7	0	1	2	3	4	5	6	7
F	NRI		NUT					R	I	PID						N	DID			QID				TID			U	D	O	R2	

Figure 6. 4-byte SVC NALU header structure (from [8])

The *forbidden_zero_bit* (*F*) is used to indicate an error in the particular NALU while the *nal_ref_idc* (*NRI*) is used as an indication of the visual importance of the particular NALU. The *nal_unit_type* (*NUT*) field indicates which type of payload is being used for the particular NALU i.e. whether the unit is a Video Coding Layer (VCL) or non-VCL.

Copyright (c) IARIA, 2011. ISBN: 978-1-61208-129-8

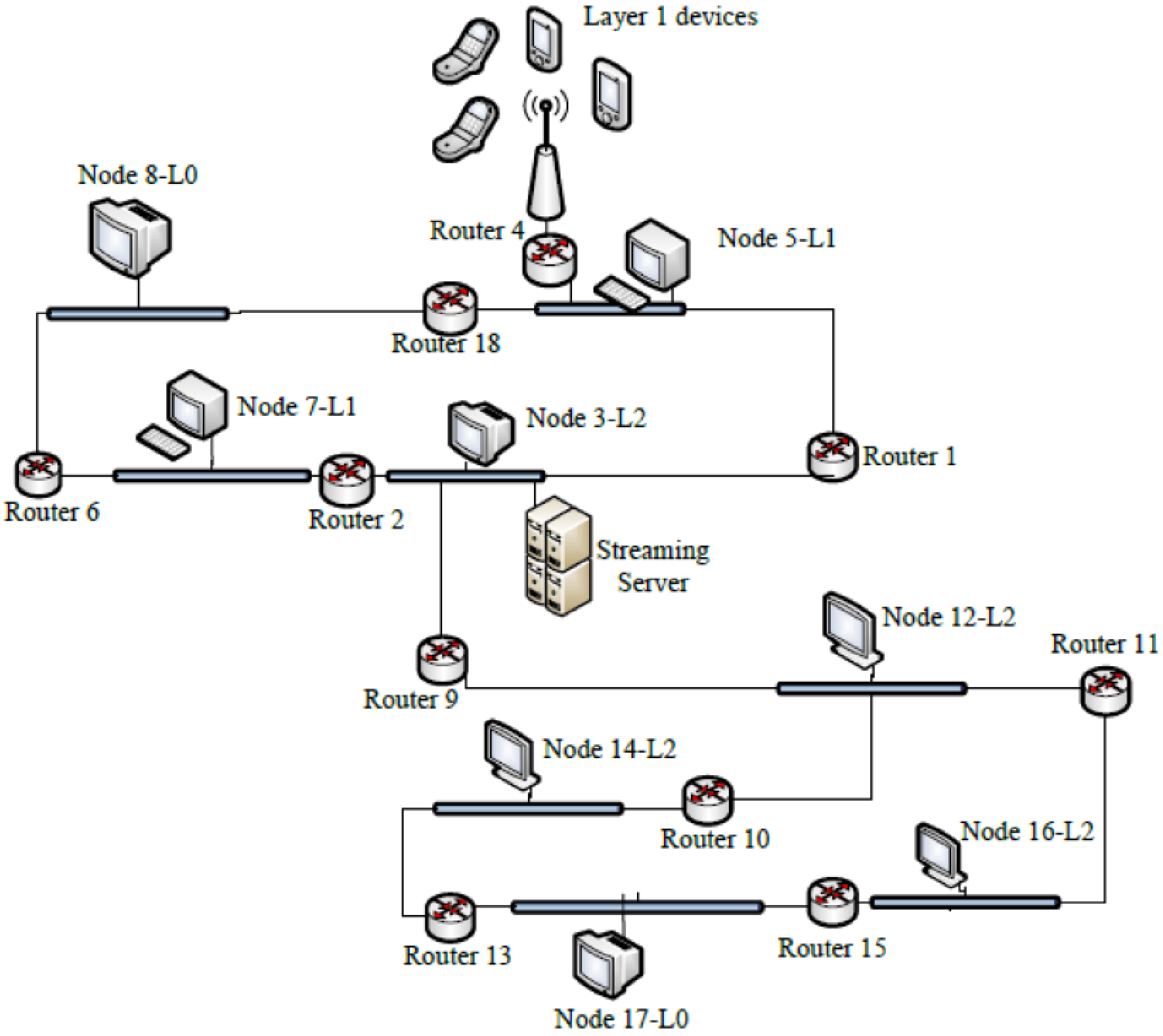


Figure 7. Network topology used in the following simulations

The remaining fields are used by the H.264/SVC codec. The *dependency_id* (*DID*) denotes the spatial scalability inter layer coding structure. The *temporal_id* (*TID*) indicates the temporal scalability hierarchically. The *quality_id* (*QID*) is used to define the quality scalability structure while the *priority_id* (*PID*) is used to assign priority to the stream. More information about each field is provided in [8].

Typical video streaming services are provided using the User Datagram Protocol (UDP) at the transport layer. UDP is a connectionless and unreliable protocol and therefore the sending node does not have a feedback channel. Therefore, the sender has no knowledge about the receiver nodes.

The UDP is a simple protocol which is convenient for real time applications, especially when using multicasting transmission. However, UDP does not ensure good Quality of Experience (QoE) and thus cannot be employed on its own. The Real–Time Transport Protocol (RTP) is a protocol that stands in between the transport and the application layers and provides additional functionalities such as timestamping and sequencing. These methods reduce the effect of jittering and enable the reordering of the received packets thus making transmission of real-time multimedia content feasible.

IV. Simulation Environment

The JSVM [7] software model was used as a reference for both the H.264/AVC and H.264/SVC. This software package contains libraries that can be used to generate both single and multiple layer video streams according to the respective standard. Every NALU is encapsulated within RTP/UDP/IP packets, where the single NALU packetization mode is adopted [9]. Unless otherwise specified, 100 frames at 60fps of the *City.YUV* sequence were encoded and transmitted using the JSVM reference model. The SVC stream was composed of three spatial layers and six temporal layers.

The packetized video content are simulated to be transmitted over the Internet using the Network Simulator 3 (NS-3) [10], which is a discrete-event network simulator. It allows the study of Internet-protocols and monitoring of data flow of large scale systems in a controlled environment. For this work, the network topology illustrated in Fig. 7 is used. The video stream was transmitted using both unicast and multicast transmissions.

V. Simulation Results

The ns-3 was combined with the JSVM codec to simulate the transmission of both H.264/AVC and H.264/SVC over heterogeneous networks. Fig. 8 shows the throughput at every router for both unicast and multicast transmission modes of H.264/SVC streams. The simulation results clearly demonstrate that under multicast transmission, a reduction of around 60% was achieved relative to unicast transmission. This can be easily explained since multicast transmission sends one stream to a group while unicast transmits a stream for each receiving node. Fig. 9 shows the bit rates at the video server when using different encoding and transmission modes. SVC multicast transmission results in a 92% decrease in bandwidth over the generated H.264/SVC unicast and a 72% decrease over .AVC unicast. This confirms that SVM multicast is the most appropriate mode for video streaming applications.

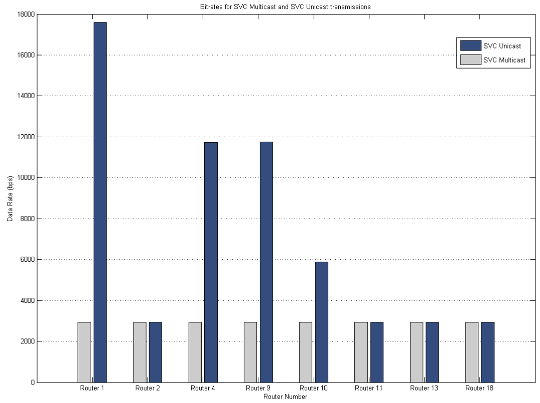


Figure 8. Throughput Analysis at routers for unicast and multicast transmissions

The H.264/SVC is inherently more robust to packet loss, especially when adopting multicasting. This is attributed to the fact that since the H.264/SVC multicast requires lower bitrates, thus reducing the probability of packet loss due to congestion. This can be easily observed from Fig. 10 where the packet loss for certain devices is much higher for H.264/AVC unicast than for H.264/SVC multicast. Furthermore, H.264/SVC is more robust to transmission errors relative to the H.264/AVC since the H.264/SVC sequence will either reduce the image resolution or else reduce the frame rate, thus providing minimal distortion. On the other hand, H.264/AVC has only one stream and therefore the only option is to conceal the damaged region of the frame which generally results in lower video quality.

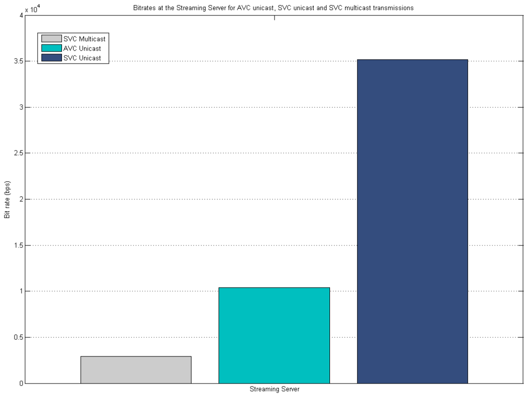


Figure 9. Bitrate of the Video Streaming Server at different modes of operation

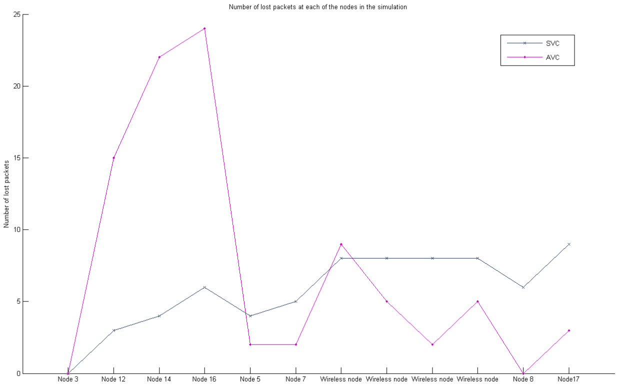


Figure 10. Number of lost packets by each receiver for AVC unicast and SVC multicast transmissions

VI. Conclusion and Future Works

This paper has presented a detailed analysis of the transmission of H.264/SVC over heterogeneous networks. It was shown that the best solution is to transmit the H.264/SVC stream using multicast transmission mode. This is mainly attributed to the fact that in multicast transmission the video stream is transmitted to a group of devices opposed to unicast transmission. Moreover, it was shown that H.264/SVC multicast encounters less congestion mainly due to the smaller data-rate required opposed to H.264/AVC unicast. Simulation results have shown that the congestion level using H.264/SVC is 75% smaller than when using

H.264/AVC. Furthermore, H.264/SVC is inherently more robust to transmission errors and thus making it ideal in packet loss scenarios such as IPTV. Future work involves the application of Overlay networks for H.264/AVC.

References

[1] Cisco. (2010, Jun) Cisco visual networking index: Forecast and methodology, 2009-2014 ONLINE. [Online]. Available: http://www.cisco.com/

[2] *Advamced Video Coding for Generic Applications*, ITU-T Std. H.264, 2005.

[3] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the scalable video coding extension of the h.264/avc standard," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 17, no. 9, pp. 1103 –1120, Sept. 2007.

[4] H. Schwarz and M. Wien, "The scalable video coding extension of the h.264/avc standard [standards in a nutshell]," *Signal Processing Magazine, IEEE*, vol. 25, no. 2, pp. 135 –141, March 2008.

[5] B. Zhang, X. Li, M. Wien, and J.-R. Ohm, "Optimized channel rate allocation for h.264/avc scalable video multicast streaming over heterogeneous networks," in *Image Processing (ICIP), 2010 17th IEEE International Conference on*, 2010, pp. 2917 –2920.

[6] H. Schwarz, D. Marpe, and T. Wiegand, "Analysis of hierarchical b pictures and mctf," in *Multimedia and Expo, 2006 IEEE International Conference on*, Dec. 2006, pp. 1929 –1932.

[7] *JSVM Software Manual*, JVT, 2009.

[8] W. Ye-Kui, M. Hannuksela, S. Pateux, A. Eleftheriadis, and S. Wenger, "System and transport interface of svc," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 17, no. 9, pp. 1149 –1163, Sept. 2007.

[9] D. Wenger, T. Stockhammer, M. Hannuksela, M. Westerlund, and D. Singer, "Rtp payload format for h.264 video," *IETF RFC3984*, Feb. 2005.

[10] *ns-3 Reference Manual*, ns-3 Project, 2009.

E-meeting Web-Interface Adaptive to Changing Context and Mobile Devices

Andrey Ronzhin, Viktor Budkov
Speech and Multimodal Interfaces Laboratory
SPIIRAS
St. Petersburg, Russia
e-mail: {ronzhin, budkov}@iias.spb.su

***Abstract*— Web-based collaboration using the wireless devices that have multimedia playback capabilities is a viable alternative to traditional face-to-face meetings. To provide quick and effective engagement to the meeting activity, the remote user should be able to perceive whole events in the meeting room and have the same possibilities like participants inside. The proposed logical-time model for compilation of multimedia content of web-based E-meeting system takes into account current situation inside the meeting room, states of audio-, video- and presentation equipment as well as constraints of user mobile devices and provides distant control by equipment and support of distributed meeting participants. The developed web-based application for remote user interaction with equipment of the intelligent meeting room and organization of E-meetings were tested with Nokia mobile phones.**

Keywords- E-meeting; smart space; remote collaboration; mobile communication; multimodal interfaces

I. INTRODUCTION

Distributed events organized via specialized web-applications are real alternative to traditional meetings and lectures, where participants interact "face to face." Development of multimedia technologies allows video conference systems not only recording and output of audio and video data, but they also use more sophisticated techniques, such as analysis, pattern recognition and structuring of multimedia data, that certainly enhances services for participants and leads to new ways of access to events in real-time and archive processing [1]. Internet applications for teleconference and distant learning (E-meeting and E-lecture) become more popular in business, research, education and other areas. Such systems allow us to save cost and provide self-paced education and convenient content access and retrieval.

However the main part of the secretary work on documentation and connection of remote participants is performed by a human-operator. Another constraint of E-meeting systems is capacities of communication network and performance of audio and video user devices, which influence on user interface features and sufficiency of remote participant possibilities.

One of the key issues of remote communication is high uncertainty, caused by the spatial and temporal distance between co-participants [2]. The physical and psychological barriers that exist in hybrid meetings make difficult for remote partners to attend a selected conversational flow, which the participants share the same room, and to initiate a new topic of discussion. Thus the main task of the research is to make remote meetings more engaging by giving remote participants more freedom of control in discussions and decision making processes.

Research projects AMI, CHIL, AMIGO, CALO [3,4] were targeted to study various aspects of arrangement of meetings or teleconferencing in smart environments and development of meeting support technology, multi-modal meeting browsers, as well as automatic audio-, and video-based summarization systems. Meeting support includes (semi-) automatic retrieval of information needed for arrangement of remote participation in hybrid meetings, in which one or more participants are remote and others stay in a shared room [5]. Development of technologies for automatic selection of the best camera view, switching on the projector or whiteboard output and selection of microphone of the current speaker is capable to improve audio-visual support for a remote mobile participant. The automatic analysis of audio-visual data of meetings or lectures is complicated by the fact that such events are usually held in large halls with lots of participants, who arbitrarily change positions of their body, head and gaze. Microphone arrays, panoramic cameras, intelligent cameras (PTZ - Pan / Tilt / Zoom) and distributed camera systems are used to improve the capturing and tracking of a group of participants.

In the system Cornell Lecture Browser [6], two video cameras with subsequent synchronization and integration of video streams are used. In the project eClass, videos of lectures were combined with the handwritten notes on a whiteboard. The system AutoAuditorium detects participants by the PTZ-camera, as well as carries out an automatic selection of video from one of three cameras, installed in the meeting room. The spatial sound source localization is used to control of the PTZ-cameras and to track speakers and listeners in the system [7]. Also a list of rules and recommendations for professional operators, assisting to select optimal positions of cameras in the hall, is defined in the paper. The system FlySpec implements the PTZ and omni-directional cameras, management of which is remotely carried out by several participants with the automatic control system adjusting the direction of PTZ-cameras. Application of the panoramic camera allows us to capture the images of all the events taking place in the room and to determinate locations of each participant.

Motion sensors and microphone arrays are used additionally to video monitoring, in order to detect participant positions and the active speaker. Sound source localization by microphone arrays is effective in small lecture or conference rooms only. To record audio signals in a large room, participants and speaker often use personal microphones or apply a system of microphone arrays, distributed in the room [8].

The various approaches have been proposed to record presentation slides projected on the screen during the event [9]. Some systems require loading presentation slides in advance or installation of special software on user's computer, which transmits the current slide to the server.

Parameters of the equipment located inside the intelligent meeting room, which are analyzed at changing the graphical content of the web-page, are considered in Section 2. A model of media stream fusion used for designing actual content of the web-based application is described in Section 3. Experimental results are presented in Section 4. The novelty of the proposed web-based application for E-meeting consists in automatic selection of web-camera of the current active speaker by multichannel speech activity algorithm [10] and designing a user interface adaptive to mobile device features of remote participant.

II. Analysis of the Current Situation Inside the Intelligent Meeting Room

The developed intelligent meeting room is equipped by the projector, the touchscreen TV for smart desk application, several cameras for video monitoring of audience and presentation area, and the personal web-cameras for analysis of behavior of participants sitting at the conference table [11]. Three T-shape microphone arrays mounted on the different walls serve for localization of sound sources, far-field recording and further speech processing. The personal web cameras mounted on the conference table have internal microphones and are used for record of video and speech of each meeting participant. A combination of the desktop web-cameras and microphone arrays provides both spatial localization of sound sources and record of participants' speech with high quality. The multimedia content compiled from audio and video signals captured by the referred devices is used in the web-based application for organization of hybrid E-meetings.

Table 2 contains parameters of objects (equipment, software, participants) located inside the intelligent meeting room, which is used for description of the current situation in the meeting room and taken into account during compilation of the graphical content of the web-page. Behavior of participants at the conference table, as well as the main speaker located in the presentation area is analyzed by developed technologies of sound source localization, video tracking of moving objects, detection and tracking of human face.

TABLE I. The Parameters Representing the Current Situation in the Meeting Room

Object in the meeting room	Parameters		
	Notation	***Values***	***Description***
Projector	p_1	0	Projector turned-off
		1	Projector turned-on
	p_2	0	Presentation is not started
		1	Presentation is started
	p_3	0	Current slide of a presentation is shown longer, than τ_{slide} , ($t_{cur} - t_{slide} > \tau_{slide}$, where t_{cur} is current time).
		1	Slide of a presentation was changed (time of the changing t_{slide} is saved)
Smart desk	d_1	0	Wide touchscreen TV is turned-off
		1	Wide touchscreen TV is turned-on
	d_2	0	Smart desk application is not loaded
		1	Smart desk application is loaded
	d_3	0	Touch input was not used longer than τ_{desk} , ($t_{cur} - t_{desk} > \tau_{desk}$).
		1	Touch input was used (beginning time of touchscreen using t_{desk} is saved)
Main speaker (presenter)	s_1	0	A speaker in the presentation area is not observed by the video monitoring system
		1	Speaker is found in the presentation area
	s_2	0	Speaker face is not found
		1	Speaker head is directed to (the face tracking system founded presenter face)
	s_3	0	Speech activity in the presentation area is not detected
		1	A presenter gives a speech (the sound source localization system detected an activity in the presentation area)
Personal web-cameras assigned with participants sitting at the conference table	c_1	0	Currently there are no speakers at the conference table
		1	Currently a participant at the conference table gives a speech comment
	c_{2i}	0	Personal web-camera i is turned-off
		1	Personal web-camera i is turned-on
	c_{3i}	0	No participant in front of the web-camera i
		1	There is a participant in front of the web-camera i (the video monitoring system estimates degree of changing image background recorded before the meeting)
	c_{4i}	0	No speech activity of the participant sitting in front of the web-camera i
		1	A participant sitting in front of the web-camera i gives a speech comment (the multichannel speech activity detection system determines useful signal in the audio channel of the web-camera i)
	c_{5i}	0	Face of the participant sitting in front of the web-camera i is not found
		1	Face position of the participant sitting in front of the web-camera i is detected (the face detection system found a participant face)

Values of the parameters of hardware and software are determined by a query of its states via TCP/IP protocol or by means of Object Linking and Embedding Automation.

III. E-MEETING WEB-INTERFACE DEVELOPMENT

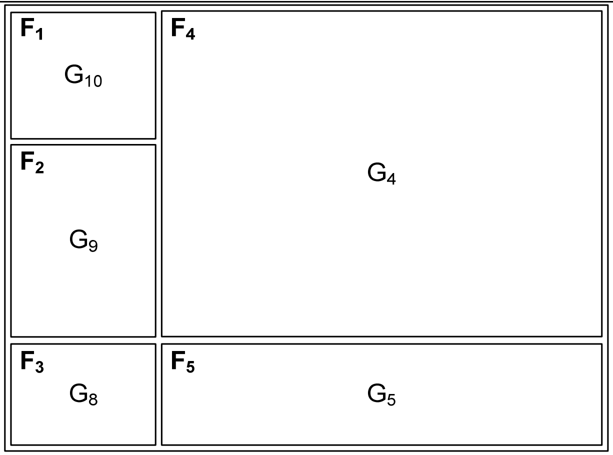
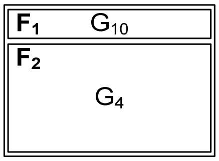
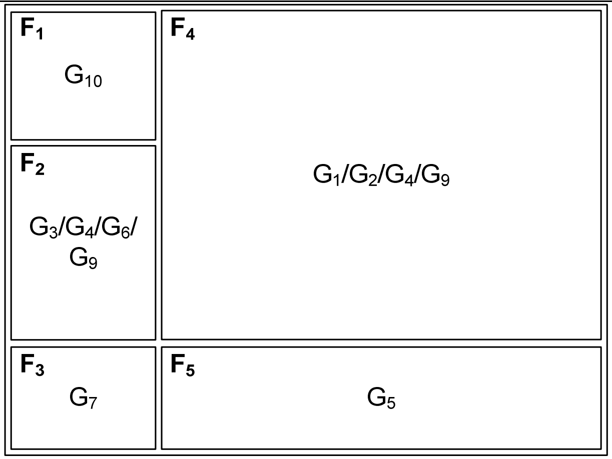
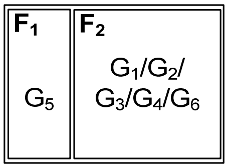
Graphical interface of the web-page, on which remote participants could observe a meeting organized inside the intelligent meeting room, contains several basic forms:

$$F = \{F_1, F_2, ...F_{N_F}\},$$

where N_F is a number of the forms depending on current meeting state and features of browser used in a client device. Two main states in meeting process were selected and corresponding notations of forms were used for the current version of web-page software: (1) registration (preparations before meeting), forms F^{reg}; (2) presentations (main part of meeting), forms $F^{meeting}$. Further the number of the meeting states will be increased taking into account peculiarities of participant behavior and necessity of use of specific technical equipment during the discussion, voting and other formal stages.

Another important factor effecting on the web-page content is a display resolution and correspondingly maximal size of browser window used for remote view of the meeting. So two classes of devices, which have especially different sizes of screen, and corresponding notations of the forms were selected: (1) personal computer, forms $F(PC)$, (2) mobile device, forms $F(MD)$. Table 2 shows the basic variants of the web-page layout depending on the current state of meeting and type of user device.

TABLE II. THE LAYOUT VARIANTS OF THE WEB-PAGE FOR E-MEETING

Meeting state	Screen of client device	
	Personal computer ($F(PC)$ ***)***	***Mobile device(*** $F(MD)$ ***)***
Registration (F^{reg})		
Presentations ($F^{meeting}$)	 F1 G10 F4 F2 G1/G2/G4/G9 G3/G4/G6/ G9 F3 G7 F5 G5	

Symbol "/" designates that several variants of graphical content are possible in the form. For instance, the current image on the projector or the current image on the smart desk will be represented in fourth form during presentations on the web-page browsed by a personal computer. Content of the forms could be changed during meeting, but it always includes a graphical component from a set:

$$G = \{G_1, G_2, ...G_{N_G}\},$$

where N_G is a number of used components (in the current version $N_G = 10$): G_1 is a component representing the current image on the projector; G_2 is a component representing the current image on the smart desk; G_3 is a component representing the current image captured by the video camera directed to main speaker; G_4 is a component representing the current image captured by the video camera directed to audience; G_5 is a component representing a assemble of the current images captured by personal web-cameras directed on participants sitting at the conference table in the meeting room; G_6 is a component representing the current image captured by a web-camera assigned with a participant, which currently gives a speech comment; G_7 is a component representing an indicator of speech duration; G_8 is a component representing a clock with time labels of the current meeting; G_9 is a component representing a logo of the current meeting; G_{10} is a component representing main data about the current meeting.

The enumerated components are connected with corresponding source, which transmits own graphical data (the projector – a presentation slide; the smart desk – window with handwriting sketches; the video and web-cameras – frames with an image; the software services – time indicators, logo and other data about meeting). Receiving new data on a source leads to updating content of corresponding form in the web-page.

Let us consider a process of graphical content compilation in the forms. Each graphical form F_j on the web-page is described by the following tuple:

$$F_j = \left\langle l_j, u_j, w_j, h_j, g_j \right\rangle,$$

where l_j is upper left corner position of the form at the abscissas axis, u_j is upper left corner position of the form at the ordinates axis, w_j is a form width, h_j is a form height, g_j is a graphical content of the form, which is actual and was chosen from the set G. Sizes of the forms could be changed depending on the current features of browser used in client device.

In the forms $F_2^{meeting}(PC)$, $F_4^{meeting}(PC)$, $F_2^{meeting}(MD)$ the number of the graphical components is changed depending on the parameter values mentioned in Table 1. In other forms the graphical component numbers are kept during the whole meeting. Selection of the current graphical component $g \in G$ for the referred forms is realized by a logical-temporal model of compiling the graphical interface web-page. The following set of logical rules is an essence of the model:

$$g_2^{meeting}(PC) = \begin{cases} G_3, s_1 \wedge s_2 \wedge s_3 \wedge \neg c_1, \\ G_4, \neg s_2 \wedge \neg c_1 \wedge ((p_1 \wedge p_2) \vee (d_1 \wedge d_2)), \\ G_6, \neg s_3 \wedge c_1, \\ G_9, \ \ overwise. \end{cases},$$

$$g_4^{meeting}(PC) = \begin{cases} G_1, (p_1 \wedge p_2 \wedge (\neg d_1 \vee \neg d_2)) \vee \\ \quad (p_1 \wedge p_2 \wedge p_3 \wedge d_1 \wedge d_2 \wedge \neg d_3) \vee \\ \quad (p_1 \wedge p_2 \wedge p_3 \wedge d_1 \wedge d_2 \wedge d_3 \wedge (t_{slide} > t_{desk})), \\ G_2, ((\neg p_1 \vee \neg p_2) \wedge d_1 \wedge d_2) \vee \\ \quad (p_1 \wedge p_2 \wedge \neg p_3 \wedge d_1 \wedge d_2 \wedge d_3) \vee \\ \quad (p_1 \wedge p_2 \wedge p_3 \wedge d_1 \wedge d_2 \wedge d_3 \wedge (t_{slide} < t_{desk})), \\ G_9, (\neg p_1 \vee \neg p_2) \wedge (\neg d_1 \vee \neg d_2), \\ G_4, \ \ overwise. \end{cases},$$

$$g_2^{meeting}(MD) = \begin{cases} G_1, ((p_1 \wedge p_2 \wedge (\neg d_1 \vee \neg d_2)) \vee \\ \quad (p_1 \wedge p_2 \wedge p_3 \wedge d_1 \wedge d_2 \wedge \neg d_3) \vee \\ \quad (p_1 \wedge p_2 \wedge p_3 \wedge d_1 \wedge d_2 \wedge d_3 \wedge \\ \quad (t_{slide} > t_{desk}))) \wedge \neg s_2 \wedge \neg c_1, \\ G_2, (((\neg p_1 \vee \neg p_2) \wedge d_1 \wedge d_2) \vee \\ \quad (p_1 \wedge p_2 \wedge \neg p_3 \wedge d_1 \wedge d_2 \wedge d_3) \vee \\ \quad (p_1 \wedge p_2 \wedge p_3 \wedge d_1 \wedge d_2 \wedge d_3 \wedge \\ \quad (t_{slide} < t_{desk}))) \wedge \neg s_2 \wedge \neg c_1, \\ G_3, \neg p_3 \vee \neg d_3 \wedge s_1 \wedge s_2 \wedge s_3 \wedge \neg c_1, \\ G_4, \neg p_3 \vee \neg d_3 \wedge \neg s_2 \wedge \neg c_1, \\ G_6, \neg p_3 \vee \neg d_3 \wedge \neg s_3 \wedge c_1, \\ current\ component\ is\ saved, overwise. \end{cases}.$$

Verification of the model was completed manually and during the experiments. The next version of the model will take into account behavior of participants sitting at the left side of the meeting room. Increase of participant number and zones of the meeting room, which should be analyzed, will required definition of some priorities in order to select image of the current speaker.

The component G_5, which consists of actual images of participants sitting at the conference table, is compiled by an analysis of states of personal web-cameras and presence of participants and faces on frame. Let us identify a set of images from the web-cameras as:

$$W = \{W_1, W_2, \ldots W_{N_W}\},$$

where N_W is a number of the web-cameras mounted on the conference table (in the developed meeting room $N_W = 10$). Then the component G_5 consists of the images captured by turned-on web-cameras, in which a participant is detected:

$$G_5 = \bigcup_{i=1}^{N_W} (W_i | c_{2i} \wedge c_{3i} = 1).$$

Taking into account the limited sizes of the forms used for representing the component G_5 , the number of displayed participants is reduced by an analysis of his speech activity c_{4i} and/or presence of his face in the frame c_{5i} . Particularly, the form $F_1^{meeting}(MD)$ for mobile device contains up to three participant images, so both parameters are used for selection of more active participants:

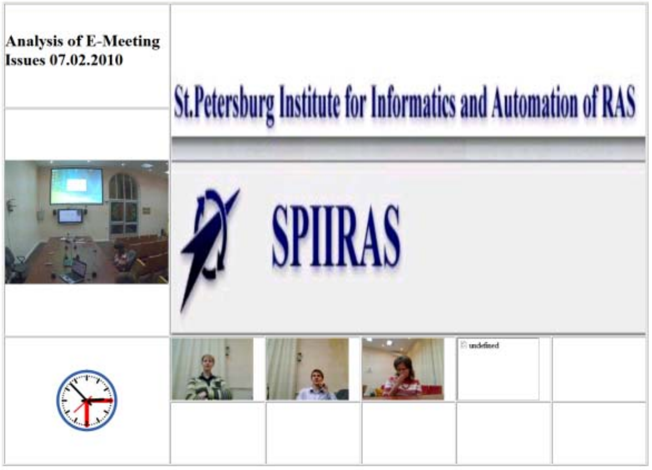
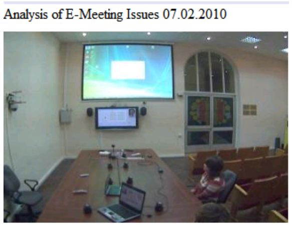
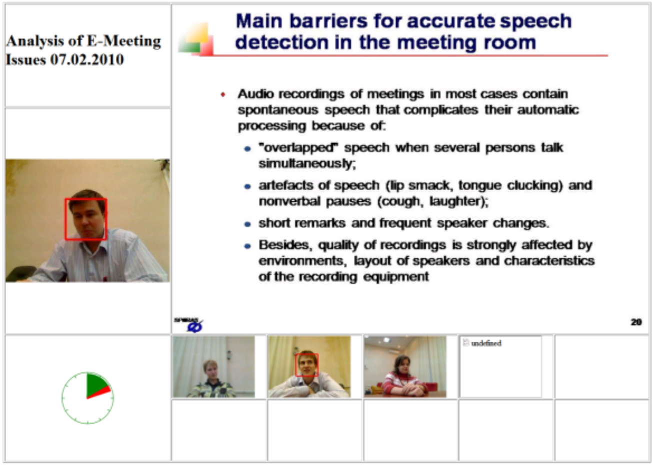
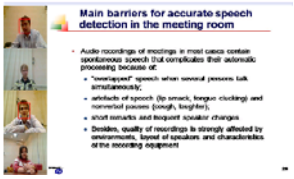
$$F_1^{meeting}(MD)=\bigcup_{i=1}^{N_W}(W_i|c_{2i}\wedge c_{3i}\wedge c_{4i}\wedge c_{5i}=1) \quad .$$

The proposed logical-temporal model of compilation of web-page graphical interface was tested on a personal computer and several models of Nokia smartphones. The fourteen different browser resolutions were applied for the tested devices, since the window size, which could be used for web page view, is significantly varied owing to different screen sizes and browser options [10].

IV. Experiments

Experimental results were received in a natural scenario where several people discuss a problem in the meeting room about forty minutes. One of the participants stayed in the presentation area and could use the smart desk and the multimedia projector. Other participants were located at the conference table. The main speaker started a talk, when all the participants bring together in the meeting room. Every participant could ask a question after finish of the presentation. Table 3 presents some examples of web-page content generated during the meeting for view on smartphone Nokia N95 with browser resolution 314x200 pixels and a monitor of personal computer with browser resolution 1280x768 pixels. Placement of elements in the web-page is specified using the CSS style tables. The resolution and orientation of screen are checked every 500 ms using Java Script. At the changing screen parameters the corresponding layout of the page is automatically selected and generated for a remote client.

TABLE III. Content examples of the web-based application for meeting

Meeting state	**Screen of client device**	
	Personal computer	***Mobile device***
Registration	 Analysis of E-Meeting Issues 07.02.2010 St.Petersburg Institute for Informatics and Automation of RAS SPIIRAS undefined	 Analysis of E-Meeting Issues 07.02.2010
Presentations	 Analysis of E-Meeting Issues 07.02.2010 Main barriers for accurate speech detection in the meeting room • Audio recordings of meetings in most cases contain spontaneous speech that complicates their automatic processing because of: • "overlapped" speech when several persons talk simultaneously; • artefacts of speech (lip smack, tongue clucking) and nonverbal pauses (cough, laughter); • short remarks and frequent speaker changes. • Besides, quality of recordings is strongly affected by environments, layout of speakers and characteristics of the recording equipment 20 undefined	 Main barriers for accurate speech detection in the meeting room • Audio recordings of meetings in most cases contain spontaneous speech that complicates their automatic processing because of: • "overlapped" speech when several persons talk simultaneously; • artefacts of speech (lip smack, tongue clucking) and nonverbal pauses (cough, laughter); • short remarks and frequent speaker changes • Besides, quality of recordings is strongly affected by environments, layout of speakers and characteristics of the recording equipment

The sound source localization system and the multichannel speech activity detection system were used for selection of source of audio stream transmitted to remote participant [10]. Speech of a presenter was recorded by the microphone, which is located over the presentation area. The built-in microphone of the web-camera assigned with the presently active participant sitting at the conference table was used for recording his/her speech.

The statistics of base events, which effect to changing situation in the meeting room and selection graphical components is presented in Table 4. During registration stage the all graphical components have own layout, so incorrect web-page content could be compiled during presentation only. Changing states of the smart desk and the projector was correctly detected. Most part of the errors arose at detection of speech activity of the participants sitting at the conference table. This type of errors leads to selection of the camera, which captures another participant, as result the image of the active participant is not displayed on the web-page. However, his speech captured by currently selected camera is transmitted with some attenuation of the sound signal.

TABLE IV. LIST OF EVENTS OCCURRED DURING THE MEETING

Event description	**Number of the event**	
	Determined manually	**Determined automatically**
Number of participants	**7**	**7**
Number of participants sitting at the conference table	6	6
Number of slide changes	14	14
Number of smart desk usage	5	5
Number of speech activity of main speaker	34	37
Number of changes of speakers sitting at the conference table	13	16
Number of temporal inactivity in audience	1	4

Also the miss of speech activity of main speaker led to selection of camera with view to the audience or other participant, whose speech was incorrectly detected. At whole about eighty percent of the graphical components were correctly selected during the analysis of the current situation in the meeting room. At this moment, the developed web-page layout model was tested at the task of support of passive remote participants. To enhance his potentials a toolbar allowing a participant located outside the meeting room to give a question to the presenter and to share the current discussion will be developed.

V. CONCLUSION

The developed intelligent meeting room is a distributed system with the network of software modules, actuator devices, multimedia equipment and audio-visual sensors. Awareness of the room about spatial position of the participants, their activities, role in the current event, their preferences helps to predict the intentions and needs of participants. Context modeling, context reasoning, knowledge sharing are stayed the most important challenges of the ambient intelligent design.

Assignment of easy-to-use and well-timed services, at that stay invisible for user, is one of another important feature of ambient intelligent. In the developed intelligent room all the computational resources are located in the adjacent premises, so the participants could observe only microphones, video cameras, as well as equipment for output of visual and audio information. Implementation of multimodal user interface capable to perceive speech, movements, poses and gestures of participants in order to determinate their needs provides the natural and intuitively understandable way of interaction with the intelligent room.

The proposed logical-temporal model of compilation of web-page graphical interface allows remote participants to perceive whole events in the meeting room via a personal computer or smartphones. The developed web-based application for remote user interaction with equipment of the intelligent meeting room and organization of E-meetings were successfully tested with Nokia mobile phones. Further efforts will be focused on enhancement of capabilities of remote participants during events conducted inside the intelligent meeting room.

ACKNOWLEDGMENT

This work is supported by The Russian Program "Research and Research-Human Resources for Innovating Russia in 2009-2013" (contract П2360), RFBR (projects 10-08-00199-a) and Nokia.

REFERENCES

[1] B. Erol and Y. Li. An overview of technologies for e-meeting and e-lecture. In: IEEE International Conference on Multimedia and Expo, 2005, pp. 6–12.

[2] N. Yankelovich, J. Kaplan, N. Simpson, and J. Provino. Porta-person: telepresence for the connected meeting room. In: Proceedings of CHI 2007, 2007, pp. 2789–2794.

[3] Computers in the human interaction loop. Ed. A. Waibel and R. Stiefelhagen, Springer, Berlin 2009. 374 p.

[4] R. Rienks, A. Nijholt, and P. Barthelmess. Pro-active meeting assistants: attention please! // AI & Society Vol. 23(2), Springer, 2009. pp. 213–231.

[5] R. Op den Akker, D. Hofs, H. Hondorp, H. Akker, J. Zwiers, and A. Nijholt. Supporting Engagement and Floor Control in Hybrid Meetings. Springer, LNAI 5641, 2009, pp. 276-290.

[6] S. Mukhopadhyay and B. Smith. Passive capture and structuring of lectures. ACM Multimedia, 1999, pp. 477–487.

[7] Y. Rui, A. Gupta, J. Grudin, and L. He. Automating lecture capture and broadcast: Technology and videography // Multimedia Systems, Vol. 10, 2004. pp. 3–15.

[8] M. Wienecke, G. Fink, and G. Sagerer. Towards automatic video-based whiteboard reading // Proc. ICDAR'2003, 2003. p. 87–91.

[9] A. Ronzhin and V. Budkov. Multimodal Interaction with Intelligent Meeting Room Facilities from Inside and Outside // Springer-Verlag Berlin Heidelberg, S. Balandin et al. (Eds.): NEW2AN/ruSMART 2009, LNCS 5764, 2009. pp. 77–88.

[10] A. Ronzhin, V. Budkov, and A. Karpov. Multichannel System of Audio-Visual Support of Remote Mobile Participant at E-Meeting // Springer-Verlag Berlin Heidelberg, S. Balandin et al. (Eds.): NEW2AN/ruSMART 2010, LNCS 6294, 2010, pp. 62–71.

[11] R. Yusupov and A. Ronzhin. From smart devices to smart space // Herald of the Russian Academy of Sciences, MAIK Nauka, Volume 80, Number 1, pp. 63-68.

Real-Time Multimedia Session Splitting and Seamless Mobility in Session Initiation Protocol Environments

Thomas Stähle, Thorsten Kaczmarz, Jürgen Müller, and Michael Massoth
Department of Computer Sciences
Hochschule Darmstadt University of Applied Sciences
Darmstadt, Germany
Email: {thomas.staehle, thorsten.kaczmarz}@stud.h-da.de, {juergen.mueller, michael.massoth}@h-da.de

***Abstract*—Multimedia applications have became more usable in recent years, since mobile broadband internet access became available. Additionally, the Session Initiation Protocol turned out to be the standard signaling protocol for Next Generation Mobile Networks. There is an increasing demand on solutions for seamless mobile communication. Two of the most important services are Voice and Video over Internet Protocol. This paper addresses issues for transferring an established session to another device without interruption as well as splitting multimedia streams over different devices. Finally, a description about the ongoing implementation of one transfer and one split mechanism for mobile phones is introduced.**

***Keywords*-Session Initiation Protocol; Voice over Internet Protocol; Seamless Mobility; Next Generation Mobile Networks**

I. Introduction

In recent years the global evolution of the internet and its bandwidth has enabled the development of many multimedia applications such as Voice over Internet Protocol (VoIP) or video streaming.

The daily used services are provided by separated networks until now. One of the main concepts in future networking is to merge all these networks into a single one, a Next Generation Mobile Network (NGMN). It will be based on the Internet Protocol (IP) and provide gateways to ensure compatibility to legacy systems [1].

The networks themselves have altered. For example, techniques and bandwidth for accessing the internet increased. The Global System for Mobile Communication (GSM) provided a much lesser bandwidth than the Universal Mobile Telephony System (UMTS) using High Speed Packet Access (HSPA). Long Term Evolution (LTE) will provide even more bandwidth in near future. This evolution enables almost all real-time multimedia services that are currently known [2].

Today, people are using social networks (e.g., Facebook) to exchange messages, videos, and photos in real-time with their friends all over the world. It is most likely that the users want to use their services wherever they are, even while traveling. Improvements in power consumption, processing power, and memory make mobile devices capable of using real-time multimedia services with high performance [3].

An increasing amount of services are transferred from local computers to the internet or to mobile phones. Therefore, users can transport their files and services and access them wherever they are. Seamless mobility is a precondition for mobile networks [4]. All these developments underline the need of users to use their services and devices everywhere.

This paper is structured as follows: First, an overview about different types of mobility is given in Section II. Some use cases that apply on our perspective are introduced in Section III. The following sections describe transfer and split mechanisms. In Section VI, the ongoing implementation is described. Finally, a conclusion is drawn and future work is suggested in the concluding Section VII.

II. Types of Mobility

An increasing number of real-time multimedia services are provided for mobile devices, since broadband internet became available for them [5]. Therefore, a concept of mobility is required. This can be divided into four categories as follows [6].

A. Personal Mobility

Personal mobility allows users to initiate and receive calls from any device and location. Therefore, they need a mechanism to be reachable on multiple devices, which is provided by a Session Initiation Protocol (SIP) registrar. A user is addressed by his SIP URI. There are two types of SIP URIs: The temporary SIP URI which is mapped to the permanent SIP URI. A permanent SIP URI is similar to an email address (e.g., sip:alice@home.de) and is used for a location independent addressing. Temporary SIP URIs contains a IPv4 or IPv6 address in the host part of the URI (e.g., sip:alice@192.168.1.100) and are used to address a specific user agent directly. A stateful SIP proxy could support call forking, which makes it possible to send an call attempt to multiple devices in parallel [7].

B. Service Mobility

Service mobility means that every service is usable with the same data across different devices. Such data include address book entries, call log, or speed dial settings. The idea is to maintain the data only once and changes will be synchronized with all participating devices. SIP does not provide this kind of mobility by default. Berger et al. proposed a seamless mobility architecture, based on multimedia, device integration, events, location-awareness, privacy, and invisible users [8]. Shacham et al. introduced another architecture with features such as support of heterogeneous devices, location-based configuration, and limited configurations [9].

C. Terminal Mobility

With terminal mobility a terminal can move between different networks without any interruption of the ongoing session. The best known solution is the concept of Mobile IPv4 [10] and Mobile IPv6 [11]. They enable mobile user equipment users to switch networks while maintaining a permanent IP address.

Another solution is mid-call mobility, where a moving device sends another INVITE request to the session partner after it obtains a new IP address [6]. This request is sent to inform the session partner about the new address.

D. Session Mobility

Session mobility makes it possible to transfer an ongoing session from one device to another without any interruption. The session can be transferred completely or only partial.

SIP has two ways for session mobility. The first approach is Third Party Call Control (3PCC) [12]. The second approach uses the REFER method to provide session mobility over multiple devices [13].

The restriction of these approaches is that an ongoing session is transferred completely. It is not supported to split a session into separate media streams such as audio and video [5].

The detailed possibilities for session mobility are described in more detail afterwards, since this paper focuses on it.

III. Use Case Description

Multiple scenarios are possible in the context of our work. First, let us assume that a representative user has a smart phone with camera in use. Additionally, the user has an SIP-enabled television and Hi-Fi components at home.

These are only the most obvious scenarios for the session transfer and session spit. In Figure 1 is a use case diagram that gives further illustration.

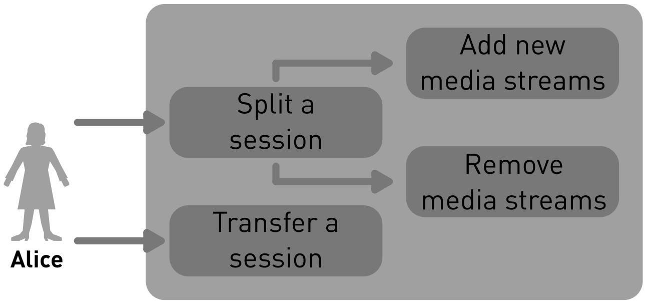


Figure 1. Use case diagram for session mobility.

A. Coming Home

Alice comes home while she has a phone call with Bob. She still wants to talk a bit with him but does not want to use her mobile phone any longer. She could have many reasons such as a low battery on her mobile phone or that she does not feel comfortable with the mobile any longer. However, she comes home and transfers the call to her devices at home. The voice transmission output is transferred to the Hi-Fi components while the voice input is done by a headset.

B. Coming to a Smart Home

This scenario is very similar to the one above. She transfers the complete phone call to a central device at her home, instead of transferring the parts of it one by one. This central device then splits the audio streams immediately to pre-configured devices (e.g., the Hi-Fi components and the headset).

Additionally, it is possible to configure the central device in order to act differently, depending on the time of day. This could be that it transfers the complete session to the handset of her cordless phone during the night. Therefore, she can continue to talk with Bob without the risk waking up her husband.

C. Session Extension

In addition to the scenarios above could it be desirable of to enrich the current session. Bob maybe want to explain Alice some facts, she cannot understand them just by hearing it. Therefore, it is possible to add a video stream to the phone call, even if this is not supported with her mobile phone. This video stream is transmitted to the televisions in her home. Therefore, she can see what he explains.

IV. Session Transfer

Schulzrinne and Wedlund discussed two possibilities of transferring a session from one device to another [6].

In 3PCC the transfer-initiating device invites the new device and changes the current media streams, so that they are redirected to another device. The second way is the REFER request, which is sent from the transfer-initiating device to the session partner. The session partner establishes a new session to the referenced device after he receives the REFER request. He transfer the current session to the new device and quits his old connection to the originating device after a successful session setup.

The following sections are illustrated by a example scenario, which is performed between the two users Alice and Bob. Alice uses her mobile phone and is on her way home. Bob uses a video phone at work. The third device is Alice's

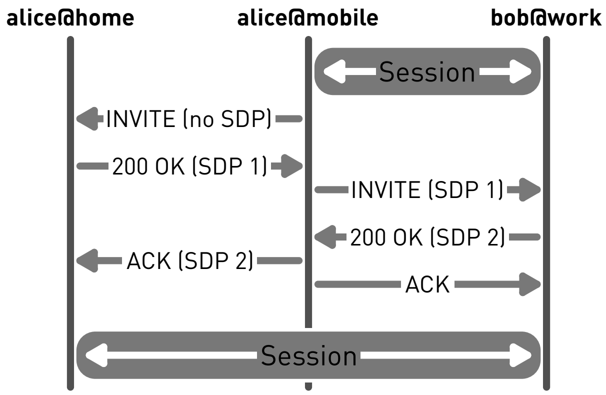


Figure 2. A session transfer using 3PCC.

video phone at home where the session should be transferred to.

A. Third Party Call Control

This method needs a third party that controls the call until the transferred session ends, as the name of this method indicates [12]. The session transfer with 3PCC is depicted in Figure 2.

Alice maintains a session on her mobile phone with Bob. Alice went home and wants to transfer the session to her home device. Therefore, Alice sends an INVITE request to it without Session Description Protocol (SDP) information. Following, Alice receives a 200 OK on her mobile phone from her home phone with an attached SDP offer (i.e., SDP 1). These SDP offer contain information about the supported media streams and sets of media codecs. Then, Alice updates the current session by forwarding this received SDP data to Bob in a reINVITE. Bob answers with a 200 OK and sends his SDP answer (i.e., SDP 2) including a selected media stream and codec. Finally, Alice sends an ACK request from her mobile phone with Bob's SDP offer to her home phone. An ACK request to Bob completes the transfer. Now, Bob sends his media streams to Alice's home phone and receives the media streams from there.

While the media streams are sent between Bob's and Alice's home phones, Alice's mobile phone still handles the signaling. Alice's mobile phone acts as the controller in this scenario. This is a disadvantage. Another variant is to let the controlling be done by a central controller that never keeps any media streams and only manages the session.

B. REFER Method

The REFER method [13] is the second possibility to transfer a SIP session to another device. The transfer-triggering device sends his session partner the SIP URI of the new device, where the session should be transferred, in the REFER request. The anchor device can retire from the session after a successful transfer. Figure 3 illustrates the transfer of a session via REFER method.

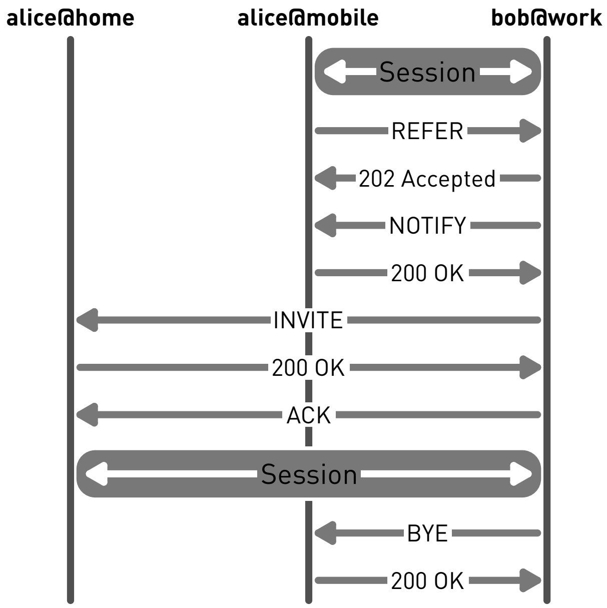


Figure 3. A session transfers using the REFER method.

Alice maintains a session on her mobile phone with Bob and wants to transfer the session to her home phone. Therefore, Alice sends a REFER request with the SIP URI of her home phone in the Refer-To header field to Bob. An exemplary REFER request with the Refer-To header field is presented in Listing 1. Bob establishes a new session with Alice's home phone via a new INVITE request if he accepts the REFER request. Bob quits the session with Alice's mobile phone by sending a BYE request, after the transfer is completed successfully

```
REFER sip:bob@work SIP/2.0
To: Bob <sip:bob@work>
From: Alice <sip:alice@mobile>;tag=193402342
Via: SIP/2.0/UDP proxy.mobile:5060
Call-ID: 0815@home
CSeq: 1 REFER
Refer-To: <sip:alice@home;method=INVITE>
Max-Forwards: 70
Content-Length: 0
```

Listing 1. An exemplary REFER request for a session transfer.

V. Session Split

Chen et al. proposed a mechanism to split a session over multiple devices, which uses the REFER request with the Mobility header field [14]. Another approach proposed from Shacham et al. groups all devices involved into a virtual device. Within this virtual device 3PCC is used to manage the session split [15].

In this section we assume that Alice uses an IP-based television with a camera instead of her internal phone.

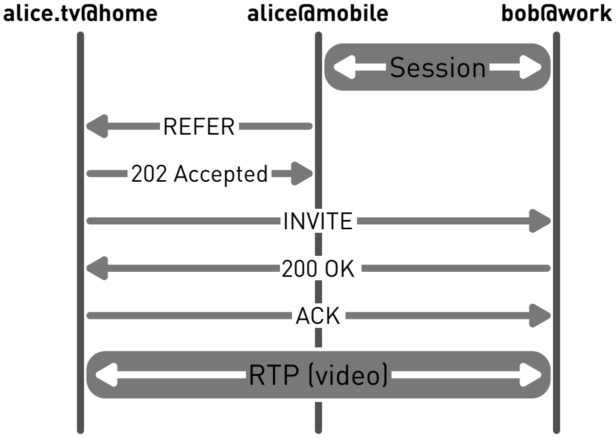


Figure 4. Split a session using SSIP.

A. Split Session over Multiple Devices

Chen et al. introduce an extension header field called Mobility [14]. The Mobility header field contains information about the current session and the media stream that should be split. Therefore, the header field contains the Call-ID of the session, an incoming split attempt belongs to. The header field is placed in REFER and INVITE requests. Figure 4 illustrates the session split with SSIP.

Alice maintains once more a multimedia session on her mobile phone with Bob. Alice now wants to split the video stream to her television. Therefore, she sends a REFER request to it. The Refer-To header field contains Bob's SIP URI and the Mobility header field contains the current Call-ID and the stream that should be transferred. In Listing 2 is an exemplary REFER request with Mobility header field. Then Alice's television sends an INVITE request to Bob in order to establish a video stream session. The Mobility header field in this INVITE request contains the Call-ID from the session between Alice's mobile phone and Bob. Therefore, Bob knows that this partial session belongs to his ongoing session with Alice's mobile phone [14].

```
To: Television <sip:alice.tv@home>
From: Alice <sip:alice@mobile>;tag=193402342
Via: SIP/2.0/UDP proxy.mobile:5060
Call-ID: 0816@home
CSeq: 1 REFER
Refer-To: <sip:bob@work;method=INVITE>
Mobility: 0815@home; media=video
Max-Forwards: 70
Content-Length: 0
```

Listing 2. An exemplary REFER request for a session split with SSIP.

B. Mobile Node Control

Shacham et al. combine several devices to a virtual device or Multi-Device System (MDS) [15]. One of these devices is the Multi-Device System Manager (MDSM). It is able to control the session and split certain media streams to other devices in the virtual device via 3PCC. This is called Mobile Node Control (MNC). Figure 5 shows the schematic design of a MDS with a MDSM, a television, and an audio device.

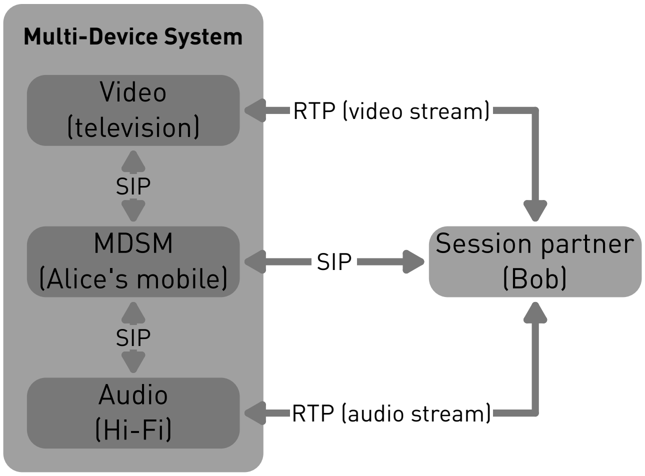


Figure 5. Schematic representation of a Multi-Device System.

If Alice comes home while she is in a video call with Bob on her mobile phone, she can use her mobile phone as MDSM and split the video stream (e.g., to her television). Figure 6 illustrates the split mechanism.

The mobile phone acts as MDSN and sends an INVITE request without SDP information to the television in order to initiate the session split. She receives a SDP offer from the television in the 200 OK response (i.e., SDP 1). The MDSM sends a reINVITE to Bob with this SDP offer, which invites Bob to send the video stream to the television. Then, Alice's mobile phone receives a 200 OK response from Bob, which contains Bob's SDP answer (i.e., SDP 2). The MDSM only has to forward this SDP answer to the television to conclude

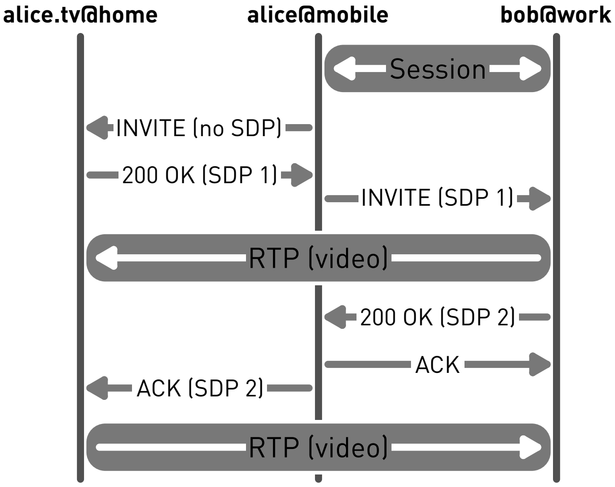


Figure 6. Split a session using a MDS in MNC mode [15].

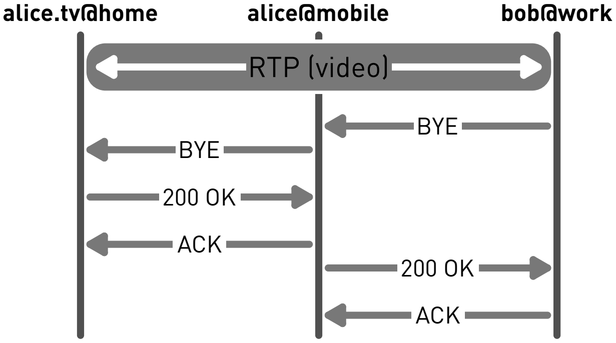


Figure 7. Tearing down a partial session using MDS in MNC mode.

the session split.

Alice only has to change the SDP information again if she wants to combine previously separated session again. Then, Bob sends the video stream to Alice's MDSM, which manages the session tear down with the television.

Alice has to forward an incoming BYE request from Bob to all devices in the MDS that are part of the current session, as depicted in Figure 7. Alternatively, she has to send a own BYE request if she decides to end the session.

VI. Implementation

This section introduces our ongoing implementation process for session mobility on mobile phones.

A. Method Decision

As described in Section II, session mobility include the session transfer and session split process. We chosed one of each aspects for the implementation.

Section IV explained two methods for transferring a SIP session (i.e., 3PCC and the REFER method). The disadvantage of 3PCC is that a central controller is always required. This is sometimes not desired (e.g., the battery of the initiating device runs out of energy) [14]. The REFER method lacks the ability to split up a session and transfer only certain media streams to other devices. The decision is made on the REFER method, because the session should be transferred completely from one device to another without any restrictions. Furthermore, the REFER method is already implemented in the most SIP stacks.

Section V explained two methods to split a SIP session (i.e., SSIP and MNC). The drawback of SSIP is that every user agent needs to understand the new Mobility header. According to this, every SIP user agent needs to be modified. The advantage of MNC is that only the user agent that initiates the split has to be modified. All other user agents involved can be regular SIP user agents. The MNC method has been chosen, because only the device that has to be capable of spliting and handle the partial sessions, have to be modified. In our opinion only few devices need such capabilities.

B. Testbed

To implement these methods the mobile phone operating system Android [16] was selected, because its application framework is well documented and easy to learn. The selected user agent is the open source application sipdroid [17]. It is released under the GNU General Public License (GPL) and uses the mjSip SIP stack [18]. mjSip is based on the SIP standard [7] and is released under the GNU GPL, as well. The REFER method is also still implemented in mjSip. Therefore, we can focus on the actual implementation of the selected methods.

sipdroid already has rudimentary preparations for the session transfer with the REFER method implemented, but there were bugs, which prevented the transfer from being successful. Furthermore, sipdroid has no proper multi session handling.

Unfortunately, the implementation process is not finished until this publication. Therefore, we are not able to present any evaluation results for now.

C. Security Considerations

Session mobility contains powerful procedures that are attracting abuse. An attacker could want to transfer a session to another destination of his or her use or interest. Another opportunity is that it could be used to hijack a session. The same for the session split functionality. An attacker could use this function to observe an ongoing session.

Therefore, it is important, if not necessary, to secure these procedures. This could be done by an authentication of the participating users (e.g., with Secure Multipurpose Internet Mail Extensions (S/MIME) [19]) [14]. Additionally, all security mechanisms that prevent an observation of the session, or even the session setup, could be applied to ensure higher security.

VII. Conclusion and Future Work

Mobility will be a very important application feature in the near future. This paper showed that SIP provides mechanisms to support significant session mobility by design. Nevertheless, SIP does not provide a way to split a session over multiple devices.

Some mechanisms are discussed to provide application layer mobility support for NGMNs in this paper. Therefore, different solutions were presented to show how a session can be transferred to other devices or split over multiple devices. Every solution has its benefits and drawbacks.

The most appropriate solutions were chosen for an implementation. The programming of the session transfer is already completed, while the implementation of the session split is still ongoing. We expect the completion by the end of February 2011.

Furthermore, security considerations will be taken into account after a successful implementations. A suitable solution to ensure the authenticity of the transfer- or split-initiating party should be specified. We will also start a first evaluation of the performance and reliability of the implemented sipdroid extension.

Acknowledgment

We would like to acknowledge Ronald Moore for his helpful comments.

References

[1] F. L. C. Ong, X. Liang, P. Pillai, P. M. L. Chan, G. Koltsidas, F. N. Pavlidou, E. Ferro, A. Gotta, H. S. Cruickshank, S. Iyengar, G. Fairhurst, and V. Mancuso, "Fusion of digital television, broadband internet and mobile communications – part i: Enabling technologies," *International Journal of Satellite Communications and Networking*, vol. 25, no. 4, pp. 363–407, Jul./Aug. 2007.

[2] "Next generation mobile life," BITKOM / KPMG, Berlin, Germany, 2008.

[3] "Gartner says worldwide mobile device sales grew 13.8 percent in second quarter of 2010, but competition drove prices down," Press Release, Gartner, Aug. 2010.

[4] N. Banerjee, A. Acharya, and S. K. Das, "Seamless sip-based mobility for multimedia applications," *IEEE Network*, vol. 20, no. 2, pp. 6–13, Mar.-Apr. 2006.

[5] M.-X. Chen and F.-J. Wang, "Session mobility of sip over multiple devices," in *Fourth International Conference on Testbeds and Research Infastructures for the Development of Netwoks & Communities and Workshops (TridentCom 2008), Innsbruck, Austria – March 17 - 20, 2008*. New York, NY, USA: ACM, 2008, pp. 23:1–23:9.

[6] H. Schulzrinne and E. Wedlund, "Application-layer mobility using sip," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 4, no. 3, pp. 47–57, Jul. 2004.

[7] J. Rosenberg, H. Schulzrinne, G. Camarillo, A. Johnston, J. Peterson, R. Sparks, M. Handley, and E. Schooler, "Sip: Session initiation protocol," RFC 3261, IETF, Jun. 2002.

[8] S. Berger, H. Schulzrinne, S. Sidiroglou, and X. Wu, "Ubiquitous computing using sip," in *Proceedings of the 13th international workshop on Network and operating systems support for digital audio and video (NOSSDAV 2003), Monterey, CA, USA, June 1-3, 2003*. New York, NY, USA: ACM, 2003, pp. 82–89.

[9] R. Shacham, H. Schulzrinne, W. Kellerer, and S. Thakolsri, "An architecture for location-based service mobility using the sip event model," in *The Second International Conference on Mobile Systems, Applications, and Services (Mobisys 2004), Boston, Massachusetts, USA, Hyatt Harborside, June 6-9, 2004*. New York, NY, USA: ACM, 2004.

[10] B. Patil, P. Roberts, and C. E. Perkins, "Ip mobility support for ipv4," RFC 3344, IETF, Aug. 2002.

[11] D. B. Johnson, C. E. Perkins, and J. Arkko, "Mobility support in ipv6," RFC 3775, IETF, Jun. 2004.

[12] J. Rosenberg, J. Peterson, H. Schulzrinne, and G. Camarillo, "Best current practices for third party call control (3pcc) in the session initiation protocol (sip)," RFC 3725, IETF, Apr. 2004.

[13] R. J. Sparks, "The session initiation protocol (sip) refer method," RFC 3515, IETF, Apr. 2003.

[14] M.-X. Chen, C.-J. Peng, and R.-H. Hwang, "Ssip: Split a sip session over multiple devices," *Computer Standards & Interfaces*, vol. 29, no. 5, pp. 531–545, Jul. 2007.

[15] R. Shacham, H. Schulzrinne, S. Thakolsri, and W. Kellerer, "Session initiation protocol (sip) session mobility," RFC 5631, IETF, Oct. 2009.

[16] "Android.com," http://www.android.com/ 05.02.2011.

[17] "sipdroid: Free sip/voip client for android," http://www.sipdroid.org/ 05.02.2011.

[18] "Mjsip," http://www.mjsip.org/ 05.02.2011.

[19] B. Ramsdell, "Secure/multipurpose internet mail extensions (s/mime) version 3.1 message specification," RFC 3851, IETF, Jul. 2004.