



# 2013 NEM SUMMIT

*Implementing Future Media Internet towards New Horizons  
Maximizing the global value of Content, Media and Networks*

*Nantes - October 28-30, 2013*

## CONFERENCE PROCEEDINGS



Copyright © 2013 – Eurescom GmbH – On behalf of NEM Initiative – <http://www.nem-initiative.org>

All rights on Proceedings of 2013 NEM Summit (Nantes - October 28-30, 2013) reserved. All rights on individual papers, published in the proceedings, remain unaffected.

ISBN 978-3-00-043123-4

### **Publisher**

Eurescom – the European Institute for Research and Strategic Studies in Telecommunications – GmbH

Wieblinger Weg 19/4 - 69123 Heidelberg - Germany

Phone: +49 6221 989 0 - Fax: +49 6221 989 209 - <http://www.eurescom.eu>

For publisher: Halid Hrasnica

### **eBook and USB produced by Sigma Orionis**

1240, route des dolines - BP287 Valbonne - France

Phone: +33 (0) 493 001 550 - Fax: +33 (0) 493 001 560 - <http://www.sigma-orionis.com>

For producer: Roger Torrenti

### **Editors**

**Roger Torrenti**, Sigma Orionis

**Nga Tran**, Sigma Orionis

**Halid Hrasnica**, Eurescom



# Foreword

L'édition 2013 du NEM Summit se tiendra du 28 au 30 octobre prochain au Centre de congrès «La Cité» de Nantes, une des agglomérations françaises les plus créatives et les plus ouvertes sur le monde, élue «Capitale verte de l'Europe» pour 2013.

Dans la période difficile que traverse aujourd'hui l'économie européenne, la recherche et l'innovation apparaissent plus que jamais comme des priorités essentielles, afin de soutenir le développement économique et social, et de conduire à la création de richesse et d'emplois.

C'est d'autant plus vrai dans le domaine «Networked and Electronic Media» (NEM), ce domaine à la croisée du «multimédia» et des réseaux de communication (au premier plan desquels l'Internet), qui joue un rôle de plus en plus central dans nos activités professionnelles et de loisirs, et notre vie de tous les jours.

Le NEM Summit, organisé par la Plate-forme Technologique Européenne NEM sous l'égide de la Commission européenne, et aujourd'hui dans sa 6ème année, est devenu au fil des ans un événement-clé pour tous les acteurs européens et mondiaux de la filière NEM.

La conférence, l'exposition et l'ensemble des événements associés au NEM Summit 2013 offriront aux participants une occasion unique d'apprendre, de découvrir, de partager, de rencontrer. Il permettra aussi de construire ou renforcer des partenariats, et de préparer dans les meilleures conditions les réponses aux appels d'offres à venir, en particulier ceux lancés dans le cadre du programme européen Horizon 2020.

En vous remerciant très cordialement pour l'intérêt que vous portez à nos activités et à ce recueil des dernières évolutions technologiques du domaine du contenu.

Bonne lecture!

The 2013 edition of the NEM Summit will take place from October 28 to 30 in the «La Cité» Congress Center in Nantes, one of the most creative French urban areas, fully open to international exchanges, and selected to be «The European Green Capital» for 2013.

In the present difficult period that European economy has been facing, research and innovation are more than ever top priorities in any policies striving for economic and social development, and top drivers for the creation of wealth and jobs.

This is particularly the case in the «Networked and Electronic Media» (NEM) domain, located at the crossroads of «multimedia» and communication networks (first and foremost the Internet) and playing an increasingly central role in our everyday life, at work, on the move, and at home. The NEM Summit, organized by the NEM European Technology Platform and today in its 6th year, has become over the years the not-to-be-missed event for all NEM stakeholders in Europe and beyond.

The 2013 NEM Summit conference and exhibition, and all co-located events, will together offer event delegates with a key opportunity to learn, discover, share, and network.

It will also give them the right occasion to build or reinforce partnerships, and get ideally prepared to answer the upcoming calls for projects and proposals, namely those to be launched in the framework of the EU-funded Horizon 2020 programme.

Thank you for the interest you are showing in our initiative and in this proceedings of the latest technological evolutions in the content field.

Enjoy!



**Jean-Dominique Meunier**  
**NEM Chairman**

*NEM Summit 2013 General co-Chair*



**Gérard le Bihan**  
**Images & Réseaux CEO**

*NEM Summit 2013 General co-Chair*

# Table of content

|   |    |
|---|----|
| <b>Programme Committee</b> .....  | 6  |
| <b>Keynote addresses</b> .....  | 7  |
| <b>Enhanced Media Content Generation, Transmission and Consumption I</b> .....  | 8  |
| <i>"High Efficiency Video Coding for Ultra High Definition Television"</i> .....  | 9  |
| <b>Rajitha Weerakkody, Marta Mrak</b> (BBC)   |    |
| <i>"Long-distance Human-Robot Interaction with 3D UHDTV 60p video supported by VISIONAIR"</i> .....   | 15 |
| <b>Artur Binczewski, Maciej Glowiak, Bartlomiej Idzikowski, Maciej Strozyk, Eryk Skotarczak</b> (Poznan Supercomputing and Networking Center)                                       |    |
| <i>"Content-Adaptive Color Transform For HEVC"</i> .....  | 19 |
| <b>Philippe Bordes, Pierre Addrivon</b> (Technicolor)   |    |
| <i>"Enhancing MPEG for Model Based Coding"</i> .....  | 23 |
| <b>Christopher Haccius, Sukhpreet Khangura, Thorsten Herfet</b> (Universität des Saarlandes)  |    |
| <i>"Entropy Constrained Scalar Quantization for Laplacian Distribution"</i> .....   | 28 |
| <b>Michael Ropert</b> (Envivio), <b>François Ropert</b> (Inouco)  |    |
| <b>Enhanced Media Content Generation, Transmission and Consumption II</b> .....   | 34 |
| <i>"From Raw Data to Semantically Enriched Hyperlinking: Recent Advances in the LinkedTV Analysis Workflow"</i> .....   | 35 |
| <b>Stein Daniel, Öktem Alp</b> (Fraunhofer IAIS), <b>Apostolidis Evlampios, Mezaris Vasileios</b> (CERTH-ITI), <b>Redondo Garcia Jose Luis</b> (EURECOM)                            |    |
| <i>"Think Before You Link – Meeting Content Constraints when Linking Television to the Web"</i> .....   | 41 |
| <b>Stein Daniel, Eickeler Stefan, Bardeli Rolf</b> (Fraunhofer IAIS), <b>Apostolidis Evlampios, Mezaris Vasileios</b> (CERTH-ITI)   |    |
| <i>"Social Backup and Sharing of Video using HTTP Adaptive Streaming"</i> .....   | 47 |
| <b>Hans Stokking, Victor Klos</b> (TNO), <b>Jin Jiang, Claudio Casetti</b> (Politecnico di Torino)  |    |
| <i>"A Novel Scene Representation For Digital Media"</i> .....   | 53 |
| <b>Christopher Haccius, Thorsten Herfet, Victor Matvienko</b> (Intel Visual Computing Institute), <b>Adrian Hilton</b> (University of Surrey), <b>Peter Eisert</b> (Fraunhofer HHI) |    |
| <i>"HbbTV a powerful asset to alert the population during crisis"</i> .....   | 58 |
| <b>Ralf Pfeffer, Sebastian Siepe, Benedikt Vogel</b> (IRT), <b>Roberta Campo</b> (Eutelsat), <b>Cristina Párraga Niebla</b> (DLR)   |    |

|  |     |
|--|-----|
| <b>Experience, Inclusion and Environmental Responsibility and Networked Media Analytics</b> .....  | 62  |
| <i>"Remote Presence: Communicating deictic gestures through handheld multi-touch devices"</i> .....  | 63  |
| <b>Clinton Jorge</b> (University of Madeira), <b>Jos van Leeuwen</b> (The Hague University of Applied Sciences),<br><b>Dennis Dams, Jan Bouwen</b> (Bell-Labs, Alcatel-Lucent) |     |
| <i>"Automatic 3DTV Quality Assessment Based On Depth Perception Analysis"</i> .....  | 69  |
| <b>Juan Antonio Rodrigo, Juan Pedro López, David Jiménez, José Manuel Menéndez</b> (UPM)   |     |
| <i>"Hybrid TV services for work integration of people with disabilities"</i> .....   | 75  |
| <b>Carlos Alberto Martín, José Manuel Menéndez, Guillermo Cisneros</b> (UPM)   |     |
| <i>"A Self-organising Isolated Anomaly Detection Architecture for Large Scale Systems"</i> .....   | 80  |
| <b>Emmanuelle Anceaume</b> (RISA / CNRS), <b>Erwan Le Merrer</b> (Technicolor), <b>Romarc Ludinard, Bruno Sericola</b> (INRIA), <b>Gilles Straub</b> (Technicolor)             |     |
| <i>"Technology Enablers for a Future Media Internet Testing Facility"</i> .....  | 86  |
| <b>Michael Boniface, Stephen Phillips</b> (IT Innovation), <b>Athanasios Voulodimos</b> (NTUA), <b>David Salama</b> (ATOS)   |     |
| <b>Application, Experimentation, and Market</b> .....  | 92  |
| <i>"OnEye – Producing and broadcasting generalised interactive videos"</i> .....   | 93  |
| <b>Alain Pagani, Christian Bailer, Didier Stricker</b> (German Research Center for Artificial Intelligence)  |     |
| <i>"KoKoo (Kontent Kooration) – Evolving a Content Curation System to a comprehensive editorial backend platform"</i> .....  | 98  |
| <b>Fabio Luciano Mondin</b> (Telecom Italia), <b>Daniele Merola, Lucia Longo</b> (Politecnico di Torino), <b>Maurizio Belluati</b> (Telecom Italia)                            |     |
| <i>"Networked Visualisation Systems in Professional markets: Prospects &amp; Challenges"</i> .....   | 102 |
| <b>Augustin Grillet</b> (Barco)  |     |
| <i>"Composite Media; A new paradigm for online media"</i> .....  | 105 |
| <b>Ingar Arntzen, Njål Borch</b> (Northern Research Institute, Motion Corporation)   |     |
| <i>"Impact of new technologies and social networks on a secondary education theatre project"</i> .....   | 111 |
| <b>Juan Pedro López</b> (UPM), <b>Pablo Ballesteros</b> (I.E.S. Al-Satt), <b>David Jiménez, José Manuel Menéndez</b> (UPM)   |     |
| <b>Supporting organisations</b> .....  | 117 |

# 2013 NEM Summit Programme Committee

## 2013 NEM Summit General co-Chairs

---

Jean-Dominique Meunier (Technicolor)  
Gérard le Bihan (Images & Réseaux cluster)

## Programme Committee Chair

---

Thorsten Herfet (Intel)

## Programme Committee Board members

---

Jovanka Adzic (Telecom Italia)  
Jose Manuel Menendez (Universidad Politecnica de Madrid)  
Rowena Goldman (BBC)

## Programme Committee Coordinator

---

Halid Hrasnica (Eurescom)

## Organisation Committee co-Chairs

---

Pierre-Yves Danet (France Telecom)  
Roger Torrenti (Sigma Orionis)

## Further Programme Committee members

---

Jon Arambarri (Virtualware)  
Andy Bower (BBC)  
Christoph Dosch (Institut für Rundfunktechnik GmbH)  
Manuel Gorius (Saarland University)  
Hadmut Holken (Holken Consultants & Partners)  
Richard Jacobs (BT)  
David Jiménez (Universidad Politécnica de Madrid)  
Amela Karahasanovic (SINTEF)  
Damla Kilicarslan (Turk Telekom)  
Jochen Miroll (Saarland University)  
Marta Mrak (BBC)  
Katy Noland (BBC)  
Francesco Saverio Nucci (Engineering SpA)  
Goran Petrovic (Saarland University)  
Jukka Salo (Nokia Siemens Networks)  
Miguel Angel Santiago (Telefonica)  
Gwendal Simon (Telecom Bretagne)  
Alexandru Stan (IN2 Search Interfaces Development Ltd.)

## Keynote addresses



*"Vision and User experience"*

**Philip Corriveau** (Intel Corporation, USA)

---



*"Revisiting QoE for Future media"*

**Patrick Le Callet** (Université de Nantes)

---



*"Personalising Socially-Aware Multimedia: Is Bigger Better?"*

**Dick Bultermna**, FXPAL (Palo Alto, California, USA) /CWI (Amsterdam, The Netherlands)

---



*"Enabling Ultra-HbbTV - A vision on future converged media"*

**Oskar van Deventer** (TNO, The Netherlands)

---



*"The challenges of large scale collaborative social sensing",*

**Paul Lukowicz** (DFKI and University of Kaiserslautern Germany)

---



*"Access for all – the convergence of TV and Internet as a new means"*

**Christoph Dosch**, (IRT, Germany)



# Enhanced Media Content Generation, Transmission and Consumption I

# High Efficiency Video Coding for Ultra High Definition Television

Rajitha Weerakkody and Marta Mrak

British Broadcasting Corporation, Research and Development Department,  
London, W12 7SB, United Kingdom.

E-mail: Rajitha.Weerakkody@bbc.co.uk, Marta.Mrak@bbc.co.uk

**Abstract:** The emergence of higher resolution video content beyond the now widespread HDTV format demands more efficient video compression technologies. This paper presents an analysis of the new High Efficiency Video Coding (HEVC) standard, in the context of High Definition and Ultra High Definition content coding. The performance is compared against its predecessor, H.264/AVC video coding standard. Further, this paper examines the individual coding tools that contribute to the compression gain in HEVC, with particular focus on the larger coding block sizes introduced in the new standard. Experimental results are presented to demonstrate the benefit of using large coding units. A number of further observations are made in the paper on correlating the new coding tools to the characteristics of different classes of video sequences.

**Keywords:** H.265/HEVC, H.264/AVC, UHD TV, Coding unit

## 1 INTRODUCTION

The High Efficiency Video Coding (HEVC) standard is developed by the Joint Collaborative Team on Video Coding (JCT-VC), which is a collaboration between ITU-T Video Coding Experts Group (VCEG) and ISO/IEC Moving Picture Experts Group (MPEG) [1]. HEVC, also known as H.265, has been designed to be a more efficient video coding technology compared to its predecessor H.264/AVC [2, 3], with a target to halve the bit rate of H.264/AVC, while retaining the same objective video quality.

Since its publication in 2003, the AVC standard has seen widespread usage in the distribution of video content including internet and terrestrial content delivery. In the past few years, there has been an increasing interest in video resolutions beyond HD, generally termed Ultra High Definition (UHD). While AVC can be used to compress these new video formats, which have up to 16 times more pixels per frame than HD, the significantly higher amounts of data to be carried stresses the distribution infrastructure. Therefore, HEVC with higher compression efficiency is expected to play a major role in future UHD deployments. However, so far little is known about the performance of HEVC for UHD content. During the standard development the coding tools were optimised for a broad range of content and the reported evaluations [4, 5, 6] used a very limited number of

sequences that were above HD resolution. These studies indicate that the performance of HEVC for above HD resolutions is higher than for HD and lower resolutions. The aim of this paper is to provide a brief overview of the new video compression standard and a performance evaluation with particular focus on investigating the benefits of HEVC in coding UHD content. The key new coding tools in HEVC are introduced, with special focus on the use of larger coding block sizes. An evaluation of the effects of varying the maximum coding unit size is also presented.

The remainder of this paper is organised as follows: Section 2 presents a brief overview of the video signal formats larger than HD and Section 3 provides an introduction to some of the coding tools in HEVC. A performance evaluation of HEVC is presented in Section 4 and finally Section 5 concludes the paper.

## 2 BEYOND HD VIDEO FORMATS

HDTV incorporates two spatial resolutions: 1280×720 and 1920×1080 pixels/frame, both of which are used in different contexts. It supports both interlaced and progressive scanning and a number of frame rates. In television broadcasting the most common formats are 1080i50 (1920×1080; interlaced; 50 fields/sec) and 720p25 (1280×720; progressive; 25 frames/sec). Research has been undertaken into higher picture resolutions and frame rates for several years, with particular interest in two specific levels of spatial resolution: 3840×2160 pixels/frame and 7680×4320 pixels/frame, both of which are integer multiples of the 1920×1080 picture size. UHD TV is specified in the ITU-R Recommendation BT.2020 (Rec. 2020) [7] covering both UHD picture sizes noted above. Table 1 shows a brief comparison of the UHD TV and HDTV (ITU-R BT.709) specifications.

Super Hi-Vision (SHV) [8] is one of the first systems to publicly demonstrate UHD TV functionalities. SHV has been developed by Japan Broadcasting Corporation (NHK, Japan) as a future broadcast system that will give viewers a much greater sensation of reality increasing the quality of experience. It also incorporates a 22.2 multi-channel three-dimensional audio system which also adds to an immersive experience. A demonstration of the SHV system took place during the London 2012 Olympic Games [9]. It used AVC as the video compression technology, as the HEVC specification was still under development at the time.

**Corresponding author:** Rajitha Weerakkody, BBC R&D, UK, Rajitha.Weerakkody@bbc.co.uk

**Table 1: Characteristics of UHD TV and HDTV.**

| Parameter   |             | UHD TV<br>(ITU-R Rec. 2020)                                 |       | HDTV<br>(ITU-R Rec. 709)                               |       |
|---|-------------|---|-------|--|-------|
| Pixel count<br>(horizontal×vertical)                              |             | 7680 × 4320<br>3840 × 2160                                  |       | 1920 × 1080  |       |
| Picture aspect ratio  |             | 16: 9   |       | 16: 9  |       |
| Standard viewing<br>angle (horizontal)                            |             | 100°  |       | 30°  |       |
| Frame frequency<br>(Hz)   |             | 120, 60, 60/1.001,<br>50, 30, 30/1.001,<br>25, 24, 24/1.001 |       | 60, 60/1.001, 50,<br>30, 30/1.001, 25,<br>24, 24/1.001 |       |
| Scan mode   |             | Progressive   |       | Interlaced,<br>Progressive                             |       |
| Primary<br>colours:<br>chromaticity<br>coordinates<br>(CIE, 1931) | R<br>G<br>B | x   | y     | x  | y     |
|   |             | 0.708   | 0.292 | 0.640  | 0.330 |
|   |             | 0.170   | 0.797 | 0.300  | 0.600 |
|   |             | 0.131   | 0.046 | 0.150  | 0.060 |
| Pixel aspect ratio  |             | 1:1<br>(square pixels)                                      |       | 1:1<br>(square pixels)                                 |       |

### 3 HEVC OVERVIEW

Similar to previous generations of video coding standards, HEVC uses a block-based hybrid coding scheme while introducing a number of new coding tools to improve the compression efficiency. Some of the key innovations include: more flexible coding tree structures and sub-partitioning mechanisms, extended directional intra prediction and adaptive prediction of motion parameters. It also employs an improved deblocking filter and an enhanced version of CABAC [10].

In HEVC, the frames are split into a collection of Coding Tree Units (CTU) and their sub partitions called coding units (CU). Each CU is inter or intra coded using a Prediction Unit (PU) partition scheme. The PU partitioning in HEVC allows asymmetric rectangular sub division, in addition to the square and rectangular symmetric partitions which are used in previous standards. HEVC uses enhanced algorithms in both intra and inter prediction, compared to AVC. There are 35 intra prediction modes (33 angular modes, Planar and DC) used in HEVC, compared to only 9 in AVC. In inter

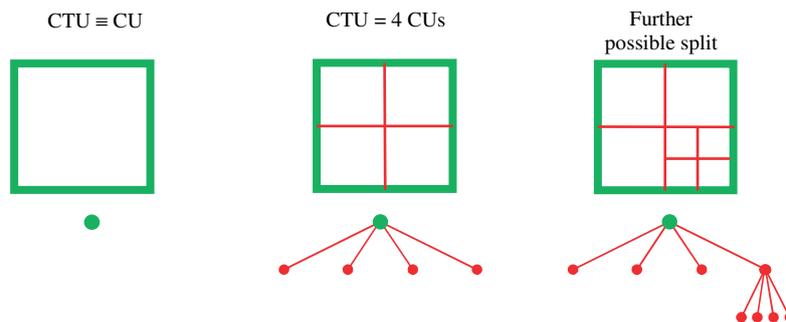
prediction, HEVC introduces Adaptive Motion Vector Prediction (AMVP) in which probable candidates are derived from adjacent prediction blocks. In a merge mode, motion vectors can also be inherited from neighbouring PBs. Quarter pixel precision is used in motion estimation, where HEVC incorporates improved interpolation filters compared to its predecessor.

CUs are also split into Transform units (TU) using a residual quad-tree (RQT) partitioning scheme. The residual signal contained in TUs is the difference between original picture and the prediction signal. The residual signal is spatially transformed and quantised. HEVC utilises discrete cosine transform (DCT) based integer transforms over transform block (TB) sizes of 32×32 and 16×16, in addition to 8×8 and 4×4 used in AVC. Further, HEVC uses a new integer 4×4 transform based on discrete sine transform (DST), only on intra predicted luma residuals. There are two in-loop filtering algorithm in the new standard, in the forms of a de-blocking filter improved from AVC, and a new Sample Adaptive Offset (SAO) technique. SAO is aimed at improving the reconstruction of signal amplitudes using a histogram analysis and a lookup table. The reconstructed CTUs are assembled to produce the pictures, which are stored in a decoded picture buffer for use in inter prediction of neighbouring frames. The entropy coding in HEVC is improved from the CABAC algorithm in AVC, particularly in terms of throughput speed for parallel processing architectures.

One of the key new features of HEVC is the use of larger blocks (coding unit) compared to AVC. In addition, it has a flexible block partitioning scheme, as described in the next subsection.

#### 3.1 Block structure in HEVC

A CTU contains one Coding Tree Block (CTB) for the luma component, and two CTBs for the two chroma components. These are further split into a number of Coding Units (CUs). A CU consists of square blocks of luma and chroma pixels that can be of the size of the associated CTU blocks, or recursively split into four smaller equally sized CUs (down to 8×8 luma samples). CTU splitting into a number of CUs is defined by a quad-tree, forming a recursive structure. The CU is used as the basic unit of region splitting, which is then used for



**Figure 1: Examples of CTU split into CU and related quadtrees describing splitting.**

prediction and spatial transformation. Intra or inter prediction may be used for each CU. Examples of CTU splitting into CUs are shown in Figure 1. The maximum CU size in HEVC is 64×64 pixels, compared to the 16×16 macroblock size in AVC. Larger coding block sizes has the potential to improve compression efficiency, at the possible expense of higher computational complexity. In this regard, the coding gain achieved by the use of higher block sizes is evaluated in this paper by constraining the maximum CU size.

## 4 PERFORMANCE EVALUATION

A performance evaluation of HEVC in comparison to AVC is presented below in Subsection 4.2, followed by an evaluation of the effects of the maximum CU size in Subsection 4.3. The experimental results are presented for test sequences of different frame sizes in order to analyse the incremental performance gain of HEVC at higher spatial resolutions.

### 4.1 Coding conditions and test material

The common test conditions and six classes of sequences defined in JCT-VC [11] are used for experiments reported in Subsections 4.2.1 and 4.3.1. UHD experiments are performed on seven sequences provided by the European Broadcasting Union (EBU) [12]. The details of all sequences are provided in Table 2. All sequences are in YCbCr 4:2:0 format. Four quantization parameters (QP), 22, 27, 32 and 37, were used in all experiments. The compression efficiency comparisons for the proposed methods are presented as Bjøntegaard Difference in bit rate (BD-rate) [13].

Sequences from classes A to E as well as UHD are camera-captured, while Class F sequences include computer-generated content. Sequences in Classes A, B, E and UHD are in native resolution although some Class A sequences are cropped from higher resolution originals. Sequences in Classes C and D include down-sampled content. All video sequences, including those for UHD experiments are in Rec. 709 colour space, as sequences in Rec. 2020 colour space are not widely available.

The configurations used in reported experiments are All Intra (AI), Random Access (RA) and Low Delay (LD) with B slices configurations defined in [11]. In AI all frames are coded as I slices. The RA configuration is typically applied in a broadcasting environment, which uses pyramidal picture reordering with a random-access picture approximately every one second. The LD configuration is typical for video conferencing where picture reordering is not applied and only the first frame is encoded as an I slice. HEVC defines Main and Main 10 profiles, where Main supports 8 bits per sample and Main 10 supports 10 bits per sample. Both profiles are used in the test.

Since JCT-VC test conditions do not include tests of typical video conferencing content (Class E) for broadcast-like configuration (RA) and high resolution content (Class A) for video-conferencing configuration (LD), these points are not evaluated and are marked as "N/A" in reported results.

Table 2: Details of test sequences

| Class | Sequence name         | Resolution | Bit depth | Frame rate (fps) |
|-------|-----------------------|------------|-----------|------------------|
| A     | Traffic               | 2560×1600  | 8         | 30               |
| A     | People On Street      | 2560×1600  | 8         | 30               |
| A     | Nebuta                | 2560×1600  | 10        | 60               |
| A     | Steam Locomotive      | 2560×1600  | 10        | 60               |
| B     | Kimono                | 1920×1080  | 8         | 24               |
| B     | Park Scene            | 1920×1080  | 8         | 24               |
| B     | Cactus                | 1920×1080  | 8         | 50               |
| B     | BQ Terrace            | 1920×1080  | 8         | 60               |
| B     | Basketball Drive      | 1920×1080  | 8         | 50               |
| C     | Race Horses           | 832×480    | 8         | 30               |
| C     | BQ Mall               | 832×480    | 8         | 60               |
| C     | Party Scene           | 832×480    | 8         | 50               |
| C     | Basketball Drill      | 832×480    | 8         | 50               |
| D     | Race Horses           | 416×240    | 8         | 30               |
| D     | BQ Square             | 416×240    | 8         | 60               |
| D     | Blowing Bubbles       | 416×240    | 8         | 50               |
| D     | Basketball Pass       | 416×240    | 8         | 50               |
| E     | Four People           | 1280×720   | 8         | 60               |
| E     | Johnny                | 1280×720   | 8         | 60               |
| E     | Kristen And Sara      | 1280×720   | 8         | 60               |
| F     | Basketball Drill Text | 832×480    | 8         | 50               |
| F     | China Speed           | 1024×768   | 8         | 30               |
| F     | Slide Editing         | 1280×720   | 8         | 30               |
| F     | Slide Show            | 1280×720   | 8         | 20               |
| UHD   | Candle Smoke          | 3840×2160  | 8         | 50               |
| UHD   | Lupo Boa              | 3840×2160  | 8         | 50               |
| UHD   | Park Dancers          | 3840×2160  | 8         | 50               |
| UHD   | Pendulus Wide         | 3840×2160  | 8         | 50               |
| UHD   | Studio Dancer         | 3840×2160  | 8         | 50               |
| UHD   | Veggie Fruits         | 3840×2160  | 8         | 50               |
| UHD   | Wind Wool             | 3840×2160  | 8         | 50               |

### 4.2 Comparison of HEVC and AVC

#### 4.2.1 Results for JCT-VC common test conditions

The BD-rate results for sequences from common test conditions in JCT-VC are shown in Table 3. Note that the negative BD-rate values mean a gain in compression efficiency as the application of HEVC reduces the bit rate, compared to AVC. It is noted that the compression gain in HEVC, compared to AVC, varies significantly across the different classes and also across the coding configurations (AI, RA, and LD). Further, some notable variations were observed for each sequence within each class as well. These differences may be attributed to a number of potential properties, including (but not limited to): size of the frame, frame rate, spatial and temporal complexity and any pre-processing (scaling, cropping, artistic effects etc.).

Studying the results for the classes B, C, and D, a trend of increased gains could be observed with the increasing picture size for all coding configurations. This is because of the fact that lower resolution sequences have more details that are less correlated, which is also the property of sequences that were used in AVC development. It can also be observed that the bit rate reductions introduced by HEVC are generally larger for configurations which involve motion compensation (RA and LD).

**Table 3: Average BD-rate gains of HEVC compared to AVC (Luma only) for Classes A to F**

| Class | AI-Main | RA-Main | LD-Main | AI- Main10 | RA- Main10 | LD- Main10 |
|-------|---------|---------|---------|------------|------------|------------|
| A     | -23.7%  | -36.7%  | N/A     | -25.2%     | -38.1%     | N/A        |
| B     | -22.7%  | -39.8%  | -42.2%  | -23.8%     | -41.0%     | -43.5%     |
| C     | -19.7%  | -30.3%  | -32.6%  | -20.2%     | -31.0%     | -33.3%     |
| D     | -16.4%  | -27.8%  | -29.7%  | -16.8%     | -28.3%     | -30.2%     |
| E     | -28.9%  | N/A     | -44.0%  | -30.2%     | N/A        | -46.4%     |
| F     | -28.4%  | -30.8%  | -33.7%  | -28.3%     | -30.9%     | -32.9%     |

**Table 4: BD-rate gains of HEVC compared to AVC (Luma only) for UHD sequences**

| Sequence       | AI-Main       | RA-Main       | LD-Main       | AI-Main10     | RA-Main10     | LD-Main10     |
|----------------|---------------|---------------|---------------|---------------|---------------|---------------|
| Candle Smoke   | -31.6%        | -62.7%        | -67.0%        | -35.8%        | -66.4%        | -70.2%        |
| Lupo Boa       | -31.1%        | -40.2%        | -40.9%        | -33.0%        | -41.9%        | -42.6%        |
| Park Dancers   | -20.1%        | -36.9%        | -46.0%        | -20.8%        | -38.3%        | -47.7%        |
| Pendulus Wide  | -18.2%        | -32.9%        | -29.8%        | -19.2%        | -34.0%        | -31.1%        |
| Studio Dancer  | -35.4%        | -57.5%        | -59.9%        | -39.4%        | -60.2%        | -62.3%        |
| Veggie Fruits  | -27.1%        | -57.9%        | -58.8%        | -29.6%        | -60.2%        | -61.3%        |
| Wind Wool      | -32.0%        | -56.3%        | -57.6%        | -35.0%        | -58.3%        | -59.6%        |
| <b>Average</b> | <b>-27.9%</b> | <b>-49.2%</b> | <b>-51.4%</b> | <b>-30.4%</b> | <b>-51.3%</b> | <b>-53.6%</b> |

In the light of the above, a similar performance comparison is performed on UHD content which has a significantly larger frame size (4 times more pixels per frame) compared to HD content.

#### 4.2.2 Results for UHD sequences

The BD-rate results for each UHD sequence compared to AVC are shown in Table 4 along with the average for the selected sequences. The results show higher average gain for UHD sequences than for other classes shown in Table 3. This observation suggests that HEVC yields larger bit rate savings compared to its predecessor AVC for content with larger frame sizes.

Having analysed a number of new coding tools deployed in HEVC, the larger maximum Coding Unit (CU) size is seen to be one of main contributors to the above observation. The next subsection supports this observation and presents results for experiments that limit the maximum CU size.

### 4.3 Evaluation of the Effects of the Coding Unit Size

The HEVC standard allows a maximum CU size of 64x64 pixels, compared to the maximum macro block size of 16x16 pixels in AVC. In this regard, the effect of the maximum CU size on the compression efficiency is evaluated in this subsection for each class of test video sequences. In the experiments reported below, first the maximum CU size is constrained to 16x16, and then a further step to 8x8. It should be noted that in HEVC this can be enabled in two ways: by constraining the luma CTBs to a smaller size (e.g. 16x16) or by keeping the maximum CTB size (64x64) and restricting the maximum CU size. This choice will influence rate-distortion performance. In the experiment reported in this paper the maximum CTB size is used (64x64) and maximum CU size is reduced since this choice introduces smaller losses compared to reduction of CTB size.

#### 4.3.1 Results for JCT-VC common test conditions

Table 5 presents the results for limiting the maximum CU size in HEVC to 16x16 pixels, compared with the same codec using the default setting of 64x64 pixels. Note here that the positive values of BD-rate mean a performance loss.

Table 6 presents the results for a further strict restriction of the maximum CU size in HEVC to 8x8 pixels, again compared with the same codec using the setting that leads to best performance (maximum CU size of 64x64 pixels). This step is used to demonstrate that the smaller block sizes lead to an inferior performance.

**Table 5: Average BD-rate (luma) for HEVC with maximum CU size of 16x16, compared to maximum CU size of 64x64**

| Class | AI-Main | RA-Main | LD-Main |
|-------|---------|---------|---------|
| A     | 3.5%    | 21.1%   | N/A     |
| B     | 2.4%    | 15.4%   | 14.0%   |
| C     | 0.6%    | 6.8%    | 6.5%    |
| D     | 0.5%    | 3.7%    | 3.5%    |
| E     | 3.8%    | N/A     | 32.9%   |
| F     | 1.2%    | 6.7%    | 9.5%    |

**Table 6 Average BD-rate (luma) for HEVC with maximum CU size of 8x8, compared to maximum CU size of 64x64**

| Class | AI-Main | RA-Main | LD-Main |
|-------|---------|---------|---------|
| A     | 14.4%   | 63.1%   | N/A     |
| B     | 11.1%   | 47.8%   | 46.9%   |
| C     | 3.6%    | 24.3%   | 24.6%   |
| D     | 3.0%    | 15.9%   | 17.2%   |
| E     | 16.4%   | N/A     | 87.3%   |
| F     | 6.8%    | 22.3%   | 29.9%   |

From the above results it is evident that there is a significant performance loss (as indicated by large positive values) incurred in restricting the maximum CU size to 16×16, which is the same value as in AVC. Further reducing to 8×8 worsens the performance, thus confirming the trend. The impact is particularly large in RA and LD configurations, compared to AI configuration. This is caused by the reduction of the temporal prediction efficiency at the smaller block sizes. Also, considering Classes A to D, it is noted that the efficiency loss is increasing with the increase of the picture size.

The same experiment is repeated in the next subsection for a number of UHD sequences

### 4.3.2 Results for UHD sequences

Table 7 and Table 8 present the results for limiting the maximum CU size in HEVC to 16×16 pixels and 8×8 respectively, compared with the same codec using a maximum CU size of 64×64 pixels, for UHD sequences. Based on the tabulated results, a number of key observations could be made, as discussed below.

First, there is a significant performance loss in constraining the maximum CU size, and the average figures for this class of sequences are notably higher than that for other classes. This extends on the earlier observation in Subsection 4.3.1 that the larger block sizes are more important for larger picture sizes.

Moreover, significant variations of BD-rate results from Tables 7 and 8 can be observed for different sequences. Sequences with large smooth areas (such as Candle Smoke) benefit more from the application of larger CUs. On the other hand, sequences that have more spatial details (Park Dancers and Pendulus Wide) rely on compression using smaller block sizes, yet also benefit to some extent from the usage of larger blocks.

Figure 2 illustrates the comparison of HEVC at the maximum CU sizes considered, against AVC, for a number of video sequences representing each class. Here, the curves for constrained maximum CU size, show varying amounts of deviation from the default HEVC curve for different classes of sequences. The deviation is shown to increase with the increasing picture size. The UHD graph shows the maximum deviation from default HEVC, where the curve for the constrained maximum CU size is approaching the AVC curve.

It is observed here that the HEVC performance with maximum CU size limited to 16×16 luma pixels is still significantly better than AVC with same maximum block size. That is identified as the contribution from a number of other new coding tools in HEVC.

## 5 CONCLUSIONS

Following the reasonably wide adoption of HDTV services, even higher resolution video formats are emerging, in the form of UHD TV. With some content already available on the internet, and the increasing likelihood of the broadcasters providing UHD TV services within a few years, the new video compression standard, HEVC is expected to see wide adoption in near future.

This paper presented a performance evaluation of HEVC with particular focus on its impact on coding UHD content. It was experimentally shown that the new codec approached its intended target of doubling the compression efficiency compared to its predecessor AVC and the comparative performance is best for larger picture sizes such as UHD. Further, this paper identified some of the new coding tools used in HEVC, with particular attention to those that help coding larger picture sizes. The results presented suggested that the use of larger maximum coding block size plays a significant role in this regard. Further research is on-going in order to identify the correlations of the effects of other coding tools to the various characteristics of video sequences.

**Table 7: BD-rates for luma component for HEVC with max CU size of 16×16, compared to the maximum CU size of 64×64**

| Sequence       | AI-Main     | RA-Main      | LD-Main      |
|----------------|-------------|--------------|--------------|
| Candle Smoke   | 13.9%       | 81.6%        | 82.8%        |
| Lupo Boa       | 11.0%       | 25.2%        | 19.7%        |
| Park Dancers   | 2.9%        | 14.9%        | 21.7%        |
| Pendulus Wide  | 1.9%        | 10.7%        | 7.8%         |
| Studio Dancer  | 14.3%       | 41.2%        | 38.2%        |
| Veggie Fruits  | 5.8%        | 42.7%        | 46.9%        |
| Wind Wool      | 8.1%        | 40.5%        | 35.5%        |
| <b>Average</b> | <b>8.3%</b> | <b>36.7%</b> | <b>36.1%</b> |

**Table 8: BD-rates for luma component for HEVC with max CU size of 8×8, compared to the maximum CU size of 64×64**

| Sequence       | AI-Main      | RA-Main       | LD-Main       |
|----------------|--------------|---------------|---------------|
| Candle Smoke   | 60.7%        | 222.3%        | 216.4%        |
| Lupo Boa       | 38.8%        | 89.4%         | 72.9%         |
| Park Dancers   | 11.6%        | 47.2%         | 66.1%         |
| Pendulus Wide  | 10.3%        | 36.3%         | 32.2%         |
| Studio Dancer  | 67.2%        | 135.6%        | 119.6%        |
| Veggie Fruits  | 30.4%        | 102.0%        | 124.2%        |
| Wind Wool      | 45.3%        | 132.5%        | 122.0%        |
| <b>Average</b> | <b>37.8%</b> | <b>109.3%</b> | <b>107.6%</b> |

## REFERENCES

- [1] ITU-T VCEG and ISO/IEC MPEG, "Joint Call for Proposals on Video Compression Technology," document VCEG-AM91 and WG11 N11113, Kyoto, Japan, Jan. 2010
- [2] Wiegand, G. J. Sullivan, G. Bjøntegaard and A. Luthra, "Overview of the H.264/AVC video coding standard", IEEE Transactions on Circuits and Systems for Video Technology, vol. 13, no. 7, pp. 560-576, July 2003
- [3] ITU-T and ISO/IEC JTC 1, "Advanced Video Coding for Generic Audiovisual Services," ITU-T Recommendation H.264 and ISO/IEC. 14496-10 (MPEG-4 AVC).
- [4] P. Hanhart, M. Rerabek, F. De Simone, T. Ebrahimi, "Subjective quality evaluation of the upcoming HEVC video compression standard," in Proc SPIE Optics+Photonics 2012 Applications of Digital Image Processing XXXV, Aug. 2012.

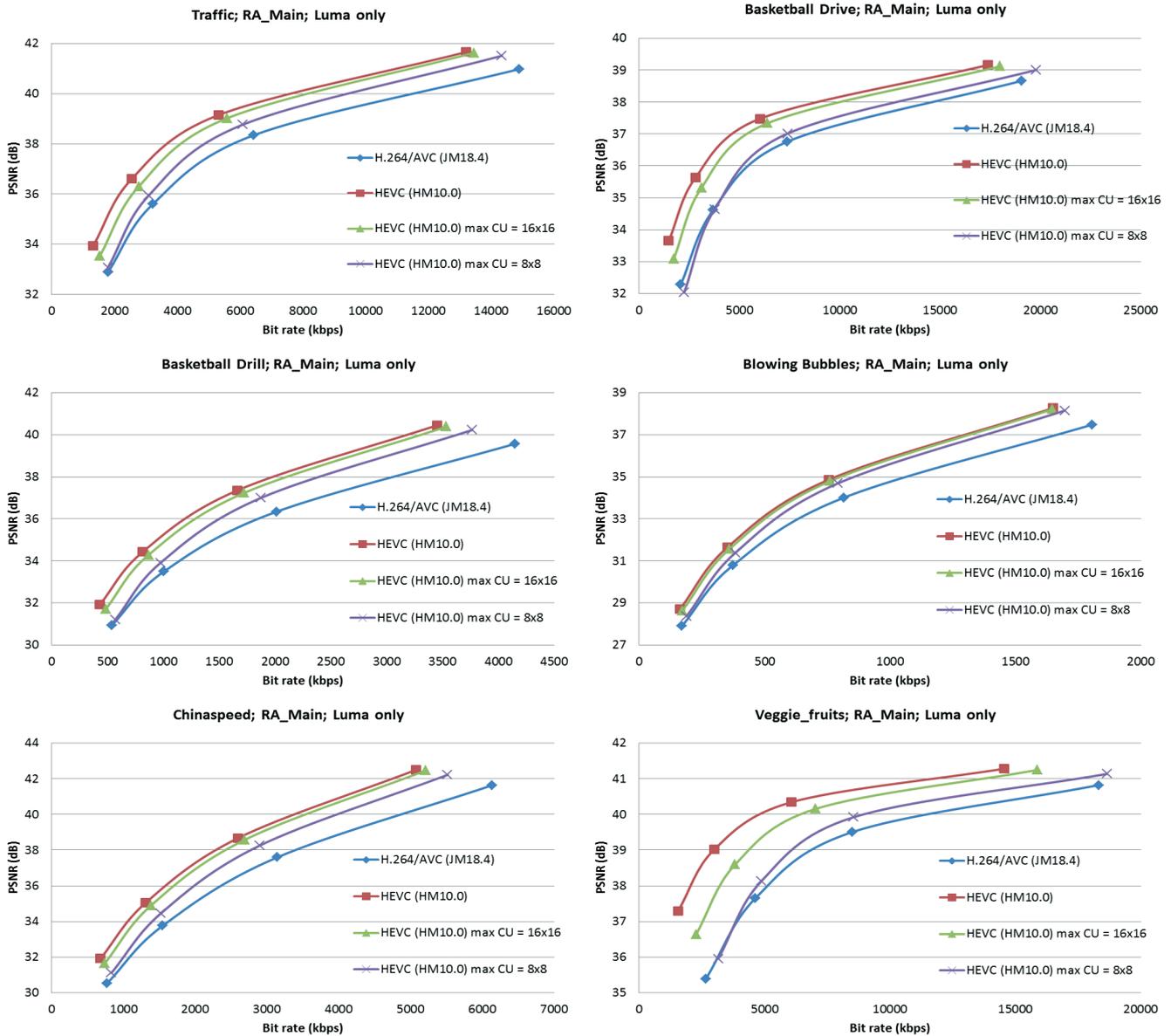


Figure 2: Comparison of HEVC and AVC coding performance (RA Main, Luma only). (1) Traffic (Class A), (2) Basketball Drive (Class E) (3) Basketball Drill (Class C), (4) Blowing Bubbles (Class D), (5) Chinaspeed (Class F), (6) Veggie Fruits (UHD)

[5] Y. Wang, C. Abhayaratne, M. Mrak, "High Efficiency Video Coding (HEVC) for next generation video applications", to appear in Elsevier E-Reference on Signal Processing 2013.

[6] J.-R. Ohm, G. J. Sullivan, H. Schwarz, T. K. Tan, T. Wiegand, "Comparison of the Coding Efficiency of Video Coding Standards – Including High Efficiency Video Coding (HEVC)", IEEE Transactions on Circuits and Systems for Video Technology, December 2012.

[7] ITU-R, "Parameter values for ultra-high definition television systems for production and international programme exchange" ITU-R Rec.BT.2020, August 2012.

[8] S. Sakaida, N. Nakajima, A. Ichigaya, M. Kurozumi, K. Iguchi, Y. Nishida, E. Nakasu, and S. Gohshi, B, "The Super Hi-Vision codec," in Proc. IEEE International Conference on Image Processing, 2007, pp. I.21–I.24.

[9] M. Sugawara, S. Sawada, H. Fujinuma, Y. Shishikui, J. Zubrzycki, R. Weerakkody, and A. Quedsted, "Super Hi-Vision at the Lond 2012 Olympics," SMPTE Mot. Imag. J; January-February 2012:(1) 29-39.

[10] F. Bossen, B. Bross, K. Suehring and D. Flynn, "HEVC complexity and implementation analysis," IEEE Transactions Circuits and Systems for Video Technology, December 2012.

[11] F. Bossen, "Common Test Conditions and Software Reference Configurations", JCTVC-L1100, 12th JCTVC Meeting, Geneva, January 2013.

[12] Test Sequences, <http://tech.ebu.ch/testsequences/uhd-1>.

[13] G. Bjøntegaard, "Calculation of Average PSNR Difference Between RD Curves", VCEG-M33, 33rd meeting, Austin, TX USA, April 2001.

# Long-distance Human-Robot Interaction with 3D UHDTV 60p video supported by VISIONAIR

Artur Binczewski<sup>1</sup>, Maciej Glowiak<sup>2</sup>, Bartłomiej Idzikowski<sup>3</sup>, Maciej Strozyk<sup>4</sup>, Eryk Skotarczak<sup>5</sup>, Migel de Vos<sup>6</sup>, Wladimir Mufty<sup>7</sup>, Florian Draisma<sup>8</sup>, Roy Damgrave<sup>9</sup>, Marc Lyonnais<sup>10</sup>

<sup>1,2,3,4,5</sup>PSNC, Poznan, PL; <sup>6,7,8</sup>SURFnet, Utrecht, NL; <sup>9</sup>University of Twente, Enschede, NL; Ciena, CA

E-mail: <sup>1</sup>artur@man.poznan.pl, <sup>2</sup>mac@man.poznan.pl, <sup>3</sup>idzik@man.poznan.pl, <sup>4</sup>mackostr@man.poznan.pl, <sup>5</sup>eryk@man.poznan.pl, <sup>6</sup>migiel.devos@surfnet.nl, <sup>7</sup>wladimir.mufty@surfnet.nl, <sup>8</sup>florian.draisma@surfnet.nl, <sup>9</sup>R.G.J.Damgrave@ctw.utwente.nl, <sup>10</sup>mlyonnai@ciena.com

**Abstract:** SURFnet, University of Twente, Poznan Supercomputing and Networking Center (PSNC) and Ciena performed a demonstration of human-robot interaction with stereoscopic UHDTV 60p video feedback over more than 2.000 km. The presented combination of a cutting edge technologies: all optical network, robotics, haptic force-feedback and ultra-high resolution compressed and uncompressed video, highlights the future potential of Human-Robot interaction in fields such as e-Health, e-Culture, e-Education and other real-time collaboration areas. UHDTV stereoscopic live video streams were transmitted through a 40G link established between Poznan, PL and Maastricht, NL with very low-latency. The demonstration took part during TERENA Networking Conference 2013 in Maastricht and was possible thanks to VISIONAIR project which calls for the creation of a European infrastructure for high level visualisation facilities that are open to research communities across Europe.

**Keywords:** UHDTV, Human-Robot interaction, 40G network, High Frame Rate, Holography, Low-latency, VISIONAIR

## 1 INTRODUCTION

Solving a specialized or riskful mechanical problem from the other side of the ocean? Performing a specialized surgery over a large distance? Or just have social interaction by shaking hands remotely? Possibilities and opportunities are getting bigger, but are we ready? Common collaboration of three European research institutions: SURFnet, PSNC and University of Twente with Ciena as a network technology vendor proved that integration of different technologies in order to perform such experiment is possible. Partners devised a demonstration to show what is required to control a remote environment. There were several objectives of such a demonstration. The main goal was to reach out to researchers, teachers and students and create an interactive show using the newest network and visualization technologies, that would be easily understandable for all participants, even those not related to ICT. On the other hand, the experiment should use cutting edge visualization, haptic and network technologies. The first step was to find partners with specific competence in

all required fields. This was possible through VISIONAIR project, which provides advanced visualization installations for researchers and users. VISIONAIR enabled the demonstration to be done during TERENA Networking Conference 2013.

The collaboration between NRENs, a university, several hardware vendors and the VISIONAIR project created the rare opportunity to showcase various innovations in one holistic demonstration. During the TERENA Networking Conference demonstration a lot of participants were able to play a game: "the leaning tower of Pisa". The participant locally controlled a robot arm to place small objects of different size and weight on the tower construction. This tower construction was located remotely in Poznan, Poland. The participants had several visual feedbacks such as: UHDTV 60p, holographic illusion and 2D and 3D displays through various compressed and uncompressed video streams. They were able to feel the haptic response from the robot arm.



Figure 1: The leaning tower of Pisa game in Poznan studio driven by Haptic Master device

## 2 TECHNOLOGIES

To enable the demonstration, several innovative technologies were integrated and deployed together. The simplified demonstration scenario is depicted on Fig.3.

**Corresponding author:** Maciej Glowiak, PSNC, ul.Wichrowa 1a; 61-611 Poznan, Poland, +48 61 858 2024, mac@man.poznan.pl

## 2.1 Haptic Technology

To control a remote environment, a specific interface is required. A haptic robot arm gives humans the opportunity to make a virtual world tangible. Such an arm tries to mimic the real world. For this demo two Haptic Master devices from University of Twente were used. The haptic devices are able to perceive and transmit movements very precisely, and provide force-feedback. The person that controls one of the arms can feel weights, movements and pressure of the other arm. Both arms are network connected to each other and enable users to control precisely the position and movement of remote objects at a long distance with a very low delay.

## 2.2 Ultra High Definition video

Since the robot arms were not in the same room, even not in the same country, it was necessary to provide the users with visual feedback. By supplying different types of displays, resolutions and a low latency transmission of live video streams, users were able to watch the remote environment and interact with it. Participants could choose between large stereoscopic high frame rate UHDTV projection, 4K LCD displays as well as a holographic illusion display. The stereoscopic UHDTV projection at a rate of 60 frames per second was deployed using two professional 4K projectors (SONY SRX-T105) working together with polarization filters and a 5-meter-wide silver screen that gave an immersive experience to the audience. Another display was the holographic illusion booth. It was constructed using a 4K LCD display and a special holographic foil which was completely invisible to the naked eye and was placed in a special rig at 45° across the stage and then mirrored content off the LCD display. This view from the LCD was reflected upwards, reflected off the foil and gave the impression of a real 3D volumetric image on stage. As supportive devices three 4K LCD monitors from EyeVis and Astro Design were used as well as a 3D display from LG.

Video streams were transmitted from the PSNC studio in Poznan, where two 4K JVC (GY-HMQ10E) cameras on a special 3D rig were located and connected via special converters directly to Ciena's optical transmission system. Two separate live UHDTV video streams of 3840x2160 pixels enabled stereoscopic vision. As a support, several 2D and 3D camera streams were used in both directions to ease the experiment. All video signals from cameras were simultaneously streamed live without any video compression, however for quality comparison, one video signal was also encoded into JPEG2000. The total bit stream of transmitted video data reached about 30Gbit/s constantly over 5 days of the conference, what means that about 1,45 PB of data was transferred during the conference.

The latency added by the camera output-process, signal converting and displaying it through the projectors, was measured separately and took 3 frames (50ms). All devices in the chain: camera, video converters and projector were separated from the network and connected directly. Network transmission introduced additional 10 ms, so the total latency

of 60 ms meant that the delay was not noticeable to any of the participants. They could experience the direct visual and haptic feedback that was in line with the movements and expectations of the participant controlling the haptic robot arm.

## 2.3 Uncompressed and compressed video

The demonstration focused on low-latency video transmission. In order to guarantee the maximal video processing speed the uncompressed streams were sent from cameras directly to the optical network system, from which it was retrieved and connected to the projectors. It caused the smallest delay between acquiring video and projection on the other side, but required perfect network connectivity and a bandwidth of about 24 Gb/s for stereoscopic high frame rate UHDTV. It's impressive, however very difficult to achieve in real applications. One of the purposes of the demonstration was to see how the usage of video compression could result in a lower bandwidth usage without losing the real-time and high quality experience. It was step forward to make human-robot interaction more applicable for regular usage in the field of research and education. For comparison, the JPEG2000 codec based on hardware IntoPIX's Pristine4 solution was used. JPEG2000 has been a Digital Cinema Initiatives standard for video compression since July 2005 commonly used in digital cinema and the movie industry for both – 2K and 4K images. However, now HEVC becomes a new standard for Ultra-High Definition video, the presence of hardware and software tools capable for real-time encoding and decoding, especially for 4K is very poor or none. For a couple of years PSNC and SURFNet have been successfully using the intoPIX system for 4K 3D live encoding and network streaming with many own implementations and improvements. That was the main reason of using JPEG2000 for comparison uncompressed and compressed video.

During the conference, two 4K displays, one next to each other, showed uncompressed and compressed video. JPEG2000 compressed video of a single 4K stream required 20 times less bandwidth and reached 500Mbit/s. There was no visual degradation of the quality, as JPEG2000 at such bitrates is perceived as a visually lossless codec. The only visible side effect of introducing compression to the chain was additional latency of 4-6 additional frames, what was caused by internal input/output queues of the encoding and decoding devices. Using a high frame rate of 60fps it means less than 100 ms which makes the fast and good quality video compression system still possible in real-time appliances.

## 2.4 Network for Video Transport

The live high quality uncompressed video feeds, in a combination with the haptic robot arm data and other supportive data (audio, videoconference) could generate a dataflow of 30 Gbit/s, therefore it was decided to configure a dedicated 40 Gbit/s all-optical path between Maastricht, NL and Poznan, PL.

Although the road distance between two locations is a bit more than 900 km, the optical link established via Poland,

Germany (Frankfurt/Oder, Hamburg) and Netherlands reached about 2.000 km. The complete path was depicted on Fig.2.

An all optical dedicated path provided a low link latency connection (just 10 ms of one-way network latency!), required for almost real-time applications. Beforehand it was not sure whether a 40Gbit/s lightpath over a transmission distance of 2.000 kilometers was feasible. This was not only because of the huge distance, but also because SURFnet and PSNC are using different network vendor equipment. For experiment purposes Ciena provided an Ultra-Long Haul network system to establish the 40Gbit/s connection between SURFnet and PSNC. Error-free transmission of the 40G ULH alien wavelength using Ciena dual carrier BPSK coherent optical technology was deployed between Maastricht and Poznan. In total a distance of 2.000 km, mainly G.655 fiber, was covered without regeneration. Using SURFnet's existing Ciena uncompensated DWDM system between Maastricht and Hamburg and a compensated ADVA DWDM system between Hamburg and Poznan.

A major challenge with stereoscopic streams was that they consisted of several independent sub-signals (2 for HD 3D and 8 for UHDTV 3D). To keep these video signals in sync specific network hardware from Ciena was used. It was able to timestamp all Ethernet frames. This made it possible to guarantee perfect reconstructed, in sync, video streams. All the work with timestamping and synchronizing audio-video HD-SDI streams were done by Ciena network devices internally. The combination of the selected all optic network hardware and intensive communication between partner technicians gave this part of the project a unique and firm foundation that successfully enabled its extraordinary usage.

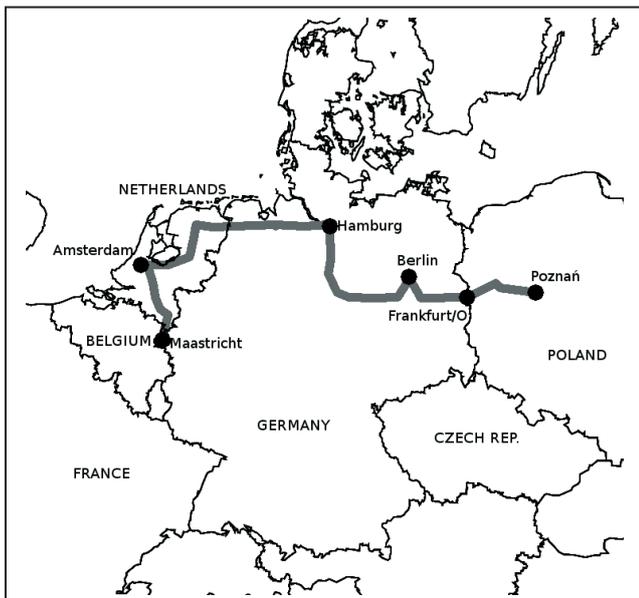


Figure 2: Optical path from Poznan to Maastricht

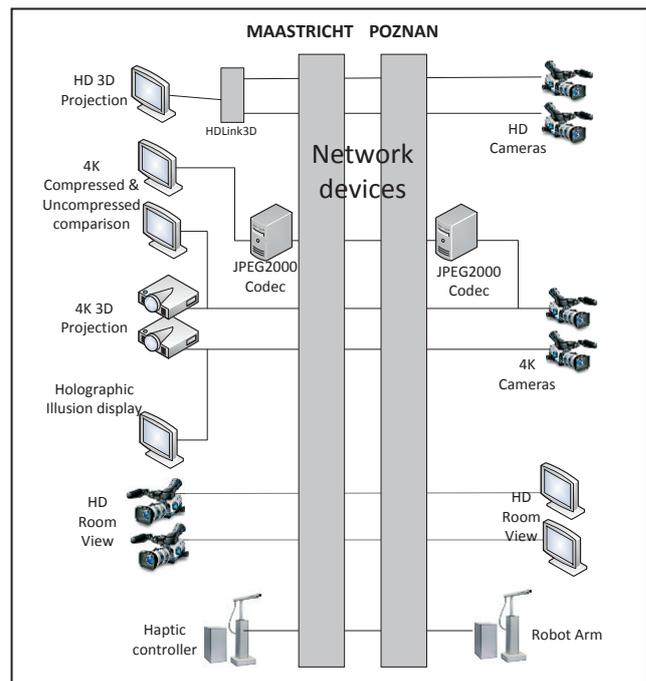


Figure 3: All technologies integrated in demonstration scenario

### 3 VISIONAIR SUPPORT

The preparation of the demonstration started several months before the conference. There was a lot of work to be done regarding the integration of video, network and haptic devices as well as establishing a dedicated optical path for the experiment. The physical access to PSNC and University of Twente visualization infrastructure was enabled by VISIONAIR project. VISIONAIR provides an open access to visualization infrastructure and services for the European research community. It comprises 4 main topics such as Scientific visualization, UHD networks, Virtual Reality and Collaborative Environments.

Scientists and researchers working on an interesting project related to visualization in one of mentioned areas are able to get access to one or several installations over Europe. Such access, even to the most advanced and innovative devices, is free of charge and a user may get back reimbursement for travelling and accommodation. It's also feasible for VISIONAIR partners to provide parts of their infrastructures at a distant location and then VISIONAIR project covers partial cost of transportation of scientific equipment. For human-robot interaction with UHDTV 60p video feedback the most relevant area was the UHD-NET activity, which focuses on tools and methods to visualize high quality images and allows sharing high definition images and movies over high bandwidth network. Visualization infrastructures from University of Twente and PSNC are part of UHD-NET activity, so SURFnet was able to apply for access and transportation of physical devices.

## 4 CONCLUSIONS

There were about 660 participants taking part in the TNC2013 conference. Most of them had an opportunity to play the game, successfully controlled the haptic robot arm. None of them experienced loss of video quality or video feedback delay as the total one way video latency took 60ms. Picture quality of all displays was excellent, however most of users particularly enjoyed an immersive stereoscopic large screen in UHDTV resolution. Users experienced the demonstration as if they were controlling a device in their local environment. Some of them even thought that the video screen was a high quality rendered video game, which shows the high quality and feasibility of the flawless and complete setup network, robotics, video and users that can actually use these together. The demonstration partners successfully combined multiple research and education aspects and will share this knowledge with others. Lessons learned from the experiment and the experience collected can be used in next collaboration scenarios for more complex use cases for e-Health, e-Culture or e-Education. The last but not least conclusion from the demonstration is that without VISIONAIR support such complex demonstration and collaboration of multiple partners would be much more complicated.

### References

- [1] UHDTV standard, ITU-R Recommendation BT.2020
- [2] VISIONAIR project, <http://www.infra-visionair.eu>
- [3] MUSION Eyeliner, <http://www.musion.co.uk/#!/products-services/eyeliner>
- [4] The HapticMaster, a new high-performance haptic interface, <http://www.eurohaptics.vision.ee.ethz.ch/2002/vanderlinde.pdf>
- [5] Binczewski, Głowiak, Idzikowski, Ostapowicz, Stroiński, Stróżyk, Research on stereoscopic 4K, TNC 2011, Prague
- [6] Binczewski, Głowiak, Idzikowski, Ostapowicz, Stróżyk, New generation high resolution media in PSNC (4K/3D), Cinegrid, 2011
- [7] Digital Cinema Initiatives specifications, <http://www.dcinemovies.com/>
- [8] JPEG2000 video codec, <http://www.jpeg.org/jpeg2000>



Figure 4: Maastricht demo room. From the left side: holographic-illusion display, haptic device stand, 4K 3D screen (rear), comparison of uncompressed and compressed video (right side).



Figure 5: Maastricht demo room. Haptic device in front, comparison of uncompressed and compressed video in the back

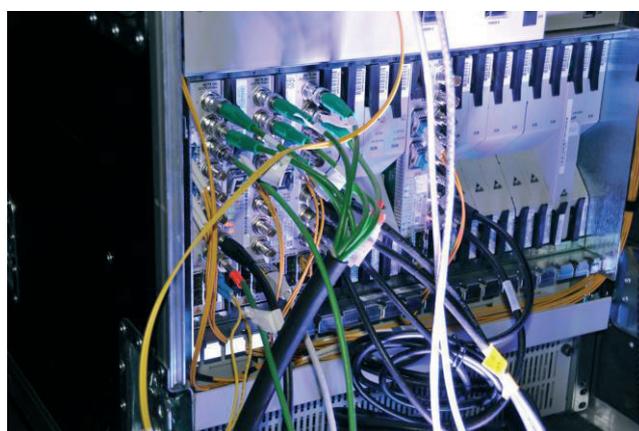


Figure 6: Ciena Ultra-Long-Haul for HD-SDI network transport



Figure 7: Poznan studio. Cameras and the Pisa game

# Content-Adaptive Color Transform For HEVC

Philippe Bordes<sup>1</sup>, Pierre Andrivon<sup>2</sup> and Patrick Lopez<sup>3</sup>

Technicolor, Cesson-Sévigné, France

E-mail: <sup>1</sup>philippe.bordes@technicolor.com, <sup>2</sup>pierre.andrivon@technicolor.com, <sup>3</sup>patrick.lopez@technicolor.com

**Abstract:** The adoption of the Main 10 Profile, a 10 bits consumer profile, in the new HEVC standard jointly developed by ISO/IEC MPEG and ITU-T VCEG opens the opportunity to provide to the consumer a new range of video content, both preserving the source characteristics and delivering content with potentially larger intrinsic quality like UHD or premium HD video. This paper presents a scheme based on a content-adaptive color space transform of input video sequences to improve the video coding efficiency. A simple and fast method to determine optimal transform parameters is proposed. Experimental results are provided on top of the HEVC reference software with Main 10 Profile demonstrating the efficiency of this approach.

**Keywords:** Video Coding, Color Transform, HEVC

## 1 INTRODUCTION

The new video compression standard called High Efficiency Video Coding (H.265/HEVC) [1] and developed by a Joint Collaborative Team of ISO/IEC MPEG and ITU-T VCEG (JCT-VC), has been finalized in January 2013. HEVC improves the coding efficiency by a factor of two compared to the former H.264/AVC compression standard, at least subjectively. To reach this level of performance, HEVC integrates a set of new tools, extending the existing hybrid coding concept rather than a real technology breakthrough. The luma and chroma components are still processed independently, despite the tentatively proposed Intra coding LM mode that predicts the chroma from the reconstructed luma, but was rejected because it introduced decoding delay and dependency.

Two profiles have been defined for the first version of the International Standard. The Main 10 Profile has the same constraints as the Main Profile except the input pictures can be provided with a bit-depth up to 10 bits rather than strictly 8 bits. A 10-bit consumer profile offers several advantages compared to traditional 8-bit coding such as less banding and contouring artifacts as well as an increase in coding accuracy in general.

This new 10 bits consumer-oriented profile aims mainly

This work is done as part of 4EVER, a French national project with support from Europe (FEDER), French Ministry of Industry, French Regions of Brittany, Ile-de-France and Provence-Alpes-Cote-d'Azur, Competitvity clusters "Images&Reseaux" (Brittany), "Cap Digital" (Ile-de-France) and "Solutions Communicantes Sécurisées" (Provence-Alpes-Cote-d'Azur).

at easing UHDTV (i.e. with Rec. 2020 [2] parameters) advent and deployment [3]. Support for 10-bit bit-depth is becoming available on consumer display systems today and the larger physical size of UHDTV displays, coupled with the wider color gamut and higher dynamic ranges supported by them, may more readily expose the visual artifacts of 8-bit video content.

In another hand, the potential usage of 10-bit precision in internal data paths (a.k.a Internal Bit-Depth Increase or IBDI) results in better prediction, smaller residuals and better overall visual experience. Experimental results conducted with HM software and coding 8 bits video contents using Internal 10 bits precision for reference frames showed significant BD-rate improvement [4].

## 2 CONTENT-ADAPTIVE COLOR SPACE TRANSFORM

The general principle of content-adaptive color transform for video compression is depicted in

Figure 1. Basically, the input video samples are transformed before coding and the inverse color transform is used by the decoder to output the reconstructed pictures in a standard display format. Then additional data allowing the decoder to build the inverse transform should be conveyed in the bit-stream. One advantage of this approach is that it does not need to modify the existing core codec except with pre/post processing steps. Other methods such as block adaptive transform may take advantage of the local color features of the images but are very much more invasive [5].

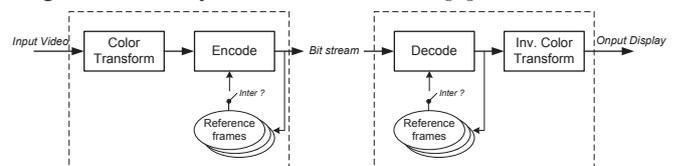
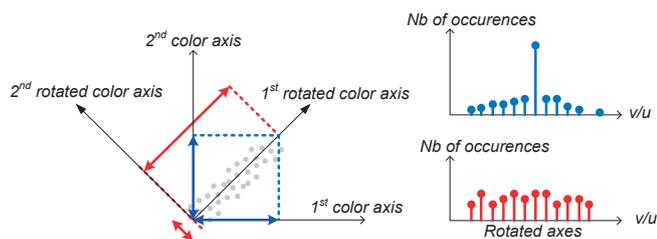


Figure 1: Color transform principle for video coding.

For the pictures with all the slices Intra encoded, one can set a color transform per picture. However, in case of Inter pictures coding, one has to define separate groups of pictures with their own transform. Because Inter prediction should use reconstructed reference pictures with same color space as the current picture to keep coding efficiency.

Several Color models for transforming RGB raw data into a more suited format for encoding exist. Their characteristics differ depending on the application requirements. The YCoCg transform and its derivations are good candidates when exact

reversibility is required [6]. However, the commonly used format in distribution video codecs originally designed for TV broadcasting is 4:2:0 Y'CbCr (a.k.a. YUV), with a chroma resolution twice less than luma one, both horizontally and vertically. This different sampling used for luma and chroma samples can be an issue when designing a color transform and the corresponding inverse transform. Hence, in this paper, we consider color transforms modifying only the chroma components. Additionally, if the luma component (Y') is built as a linear combination of R'G'B' primaries based on human psychovisual considerations, the chroma components Cr and Cb are simply derived as the weighted difference of R' and B' with Y', respectively [7].



**Figure 2: The rotated color axes allow to better de-correlate the (u,v) components.**

The basic idea behind using a color transform before coding is to re-align the color basis axes with the main color characteristics of the content, so that the video codec quantization artifacts are relatively lower on the reconstructed signal as depicted in a 2D example in Figure 2.

A second principle is to use a linear transform in order to not to distort the signal.

In our experiments we analyzed 2D chroma histograms  $H(u,v)$  of several video sequences sets, and for most of them we observed that they exhibit very particular characteristics. In Figure 3 we show 2D chroma histograms obtained with some of the HEVC regular sequences.

### 3 CHROMA SPACE ROTATION

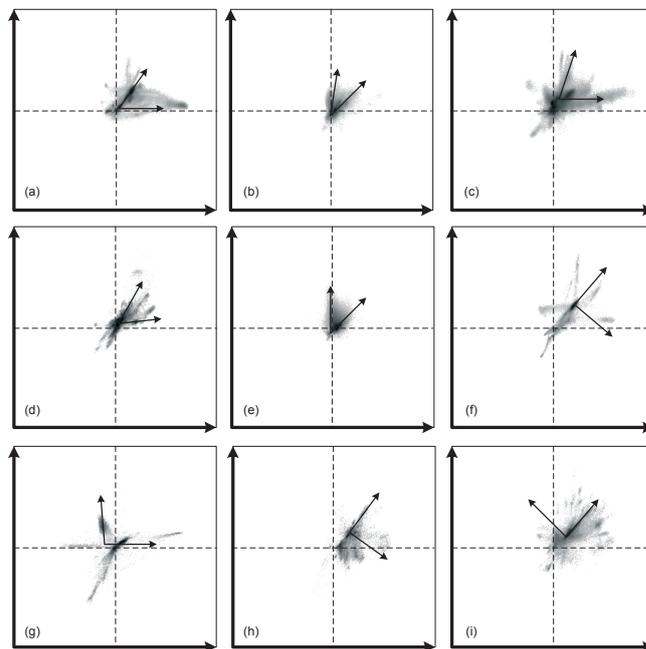
Through the observation of 2D chroma histograms (Figure 3), one can visually deduce two remarks. First they are not centered and second they may exhibit direction trends those are not the same as Cb and Cr axes. Then, the most straightforward transform to be considered in order to re-align the chroma axis and to re-center the samples is a translation plus a rotation as follows:

$$\begin{pmatrix} u' \\ v' \end{pmatrix} = \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix} \begin{pmatrix} u - C_u \\ v - C_v \end{pmatrix} \quad (1)$$

Where  $\theta$  is the rotation angle,  $(C_u, C_v)$  is the new basis center,  $(u,v)$  and  $(u',v')$  are the original and the transformed chroma coordinates respectively.

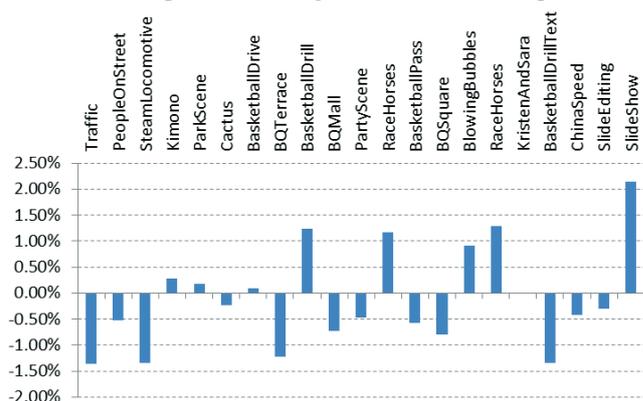
Principal Component Analysis (PCA) [8] allows to compute the orthogonal transform and the eigen-vectors corresponding to the rotation. This rotation is defined in such a way that the first principal component (major axis) has the largest possible variance. It is noticed that the second axis is orthogonal to the major axis by construction of the PCA process. The centering

by translation corresponds to the samples mean value subtraction.



**Figure 3: Chroma histograms of some regular HEVC sequences (see Table 1).**

We modified the HEVC reference encoder software (HM) [1] in order to include a pre-processing stage which computes the color transform as depicted in Figure 1. The rotation angle and the translation are encoded in the first slice of the CRA frames. Each picture is transformed before coding. In case of 8-bit content encoded in 10-bit, the transform is applied after the original samples have been 2 bits left shifted (two LSB set to 0). At the reconstruction stage (at the decoder side), the inverse transform (rotation and translation) allows to restore the chroma samples to the original Y'CbCr color space.



**Figure 4: BD-rate gains obtained using PCA with regular HEVC sequences (RA-MP10).**

We encoded various video sequences including the HEVC regular test set (21 sequences among 5 sequence classes are considered. A: cropped areas of size 2560x1600, B: 1080p, C: WVGA, D: WQVGA, F: Screen Content of misc. sizes), plus some from EBU and SVT. We used the Random Access Main

10 Profile (RA-MP10) test conditions specified in [9]. The PSNR are computed in the original Y'CbCr color space (reconstructed pictures with inverse color transform).

In Figure 4 we plotted the relative bit rate gain obtained (negative value is a gain) with our content adaptive color transform (PCA) using Bjontegaard BD-rate interpolation [10] and the combined (weighted) Y'CbCr PSNR as proposed in [11]:

$$psnr_{YUV} = \frac{w_Y \cdot psnr_Y + w_U \cdot psnr_U + w_V \cdot psnr_V}{w_Y + w_U + w_V} \quad (2)$$

with  $(w_Y, w_U, w_V)$  weights equal to  $(6, 1, 1)$  respectively.

The evaluation of the BD-rate gains may be difficult since the processing of the Y' component is unchanged. Indeed, by construction, an increase in bit rate will degrade BD-rate luma score even if chroma PSNR is improved significantly. In our case, combined Y'CbCr PSNR ( $psnr_{Yuv}$ ) and bit-rate allow to measure the overall (3 components) BD-rate gains.

We obtained a gain for half of the sequences only, corresponding to the cases when orthogonal transform is well fitted to the 2D chroma histogram shape, as shown in example of Fig.5. However, if the principal components are not orthogonal and/or the barycenter is far from their intersection, the PCA method is not well adapted. For these cases, an Independent Component Analysis (ICA) may be more appropriate [12].

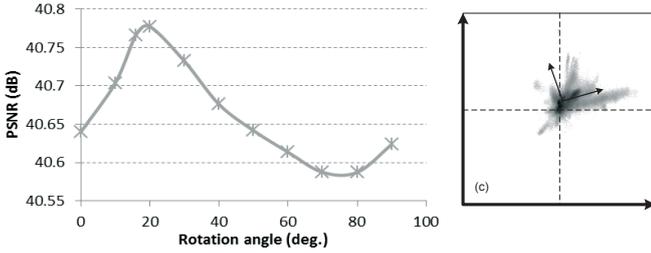


Figure 5: Evolution of YUV PSNR obtained for EBU CrowdRun sequence. The rotation angle given by PCA corresponds to the maximum PSNR.

## 4 NON-ORTHOGONAL COLOR TRANSFORM

To design a non-orthogonal transform, one has to express the new coordinates  $(u', v')$  corresponding to a non-orthogonal basis as a function of the original  $(u, v)$  chroma components (Figure 6).

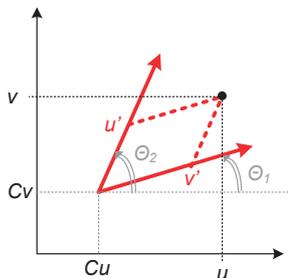


Figure 6: Non-orthogonal transform.

The corresponding chroma transform (encoding stage) and inverse chroma transform (decoding stage) are expressed by

(3) and (4) respectively, where  $\theta_1$  and  $\theta_2$  are the angles formed by the new basis vectors with the original chroma (u) basis vector, and  $(C_u, C_v)$  is the new basis center.

$$\begin{pmatrix} u' \\ v' \end{pmatrix} = \frac{1}{\sin(\theta_1 - \theta_2)} \begin{bmatrix} -\sin(\theta_2) & \cos(\theta_2) \\ \sin(\theta_1) & -\cos(\theta_1) \end{bmatrix} \begin{pmatrix} u - C_u \\ v - C_v \end{pmatrix} \quad (3)$$

$$\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} C_u \\ C_v \end{pmatrix} + \begin{bmatrix} \cos(\theta_1) & \cos(\theta_2) \\ \sin(\theta_1) & \sin(\theta_2) \end{bmatrix} \begin{pmatrix} u' \\ v' \end{pmatrix} \quad (4)$$

To determine the non-orthogonal transform, one can use an Independent Component Analysis (ICA) algorithm. ICA finds the independent components by maximizing the statistical independence of the estimated components. The way to define independence governs the form of the ICA algorithms [12]. In our case, one has to find the parameters  $(\theta_1, \theta_2, C_u, C_v)$  minimizing the distances of the chroma samples to one of the color axes ( $u'$  or  $v'$ ) using oblique projection: projection on one axis according to the other axis direction (Figure 6). For this purpose, we define an energy function (5) as the sum of the smaller distance of each chroma sample to one of the axis.

$$E_{\theta_1, \theta_2, C_u, C_v} = \sum_{u', v'} \min(|u'|, |v'|) \times H_{\theta_1, \theta_2, C_u, C_v}(u', v') \quad (5)$$

For a given set of parameters  $(\theta_1, \theta_2, C_u, C_v)$ , the corresponding transformed chroma histogram  $H_{\theta_1, \theta_2, C_u, C_v}(u', v')$  is computed by applying the color transform (3) on  $H(u, v)$ , without any need of re-scanning the images. One has to find the best parameter set that minimizes  $E_{\theta_1, \theta_2, C_u, C_v}$ . Even if the computation of  $E_{\theta_1, \theta_2, C_u, C_v}$  is very fast, the choice of initialization candidates for the values of  $(\theta_1, \theta_2, C_u, C_v)$  is of key importance to avoid local minima and to speed-up the refinement process. For  $(C_u, C_v)$ , we try the histogram barycenter and the default  $(0, 0)$  center values. For  $(\theta_1, \theta_2)$ , the angles corresponding to the rotation found with the PCA algorithm may be good initialization points.

## 5 EXPERIMENTAL RESULTS

The ICA algorithm has been implemented as a pre-processing stage in the HM7.0 [1]. The parameters are encoded with exponential Golomb entropy coding with a precision of 0.1 degree for the angles  $(\theta_1, \theta_2)$  and 1 pixel for the translation  $(C_u, C_v)$  in the first slice header of the CRA frames occurring every second (cf. RA-MP10 conditions [9]). Then only few additional bits are required.

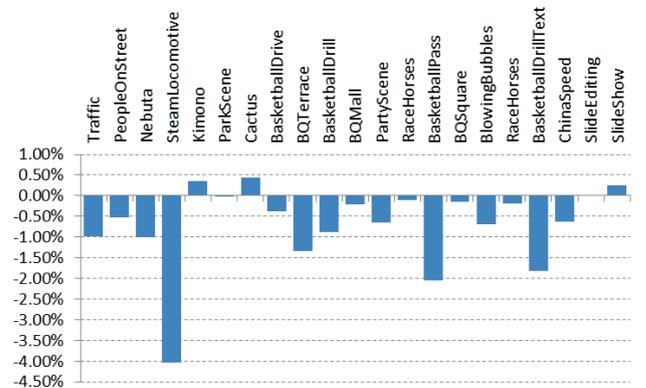


Figure 7: Gains obtained using ICA with regular HEVC sequences (RA-MP10).

We present in Figure 7 and Table 1 the results obtained with the ICA algorithm, using the same conditions as for PCA (Figure 4). The general trend shown in Figure 7 exhibits a clear improvement, with a gain up to 4%. We got similar results using Intra only configuration, but the gains are slightly smaller. The complexity increase (encoding/decoding time) is negligible both at encoder and decoder side (less than 0.5% enc./dec. time increase in average).

**Table 1: BD-RATE GAINS OBTAINED USING ICA WITH REGULAR HEVC SEQUENCES (RA-MP10).**

| Class A             | BD-rate (piecewise cubic) |        |        |       |
|---------------------|---------------------------|--------|--------|-------|
|                     | Y                         | U      | V      | YUV   |
| Traffic             | -0.23                     | -0.52  | -8.64  | -0.97 |
| PeopleOnStreet      | -0.09                     | -9.25  | 5.63   | -0.52 |
| Nebuta              | 1.14                      | -24.55 | -13.13 | -1.01 |
| SteamLocomotive     | -0.26                     | -24.85 | -32.17 | -4.03 |
| <b>Class B</b>      |                           |        |        |       |
| Kimono              | 0.18                      | -0.33  | 2.78   | 0.35  |
| ParkScene (e)       | 0.06                      | -0.37  | -0.49  | -0.02 |
| Cactus (c)          | 0.00                      | -8.55  | 9.51   | 0.44  |
| BasketballDrive (a) | 0.90                      | -29.62 | 17.07  | -0.38 |
| BQTerrace (b)       | -0.39                     | 16.06  | -24.51 | -1.35 |
| <b>Class C</b>      |                           |        |        |       |
| BasketballDrill (f) | -0.21                     | -17.96 | 11.12  | -0.89 |
| BQMall (d)          | -0.53                     | -10.87 | 12.67  | -0.22 |
| PartyScene (i)      | 0.36                      | -19.67 | 9.49   | -0.66 |
| RaceHorses          | -0.26                     | 1.40   | -0.08  | -0.11 |
| <b>Class D</b>      |                           |        |        |       |
| BasketballPass      | 1.39                      | -27.72 | -2.11  | -2.05 |
| BQSquare            | -0.10                     | -4.50  | 3.68   | -0.15 |
| BlowingBubbles (h)  | 0.00                      | -4.81  | -2.81  | -0.69 |
| RaceHorses (g)      | -0.16                     | -0.05  | -0.56  | -0.19 |
| <b>Class F</b>      |                           |        |        |       |
| BasketballDrillText | 0.02                      | -21.80 | 5.24   | -1.82 |
| ChinaSpeed          | 0.09                      | -1.84  | -6.47  | -0.63 |
| SlideEditing        | 0.49                      | -6.69  | 1.12   | -0.01 |
| SlideShow           | -0.74                     | 6.55   | 1.80   | 0.25  |
| Average Gain        |                           |        |        | -0.70 |

## 6 CONCLUSION

In this paper, a pre-processing technique of input video sequences to improve the video compression is studied. It is based on a content-adaptive color transform. Two methods are investigated to determine the best transform parameters. The results show one can increase the efficiency coding of HEVC Main 10 Profile using a simple algorithm based on ICA. Further improvements are still possible: if the de-blocking and SAO filters are applied after the inverse color transform (Figure 1) one can expect to correct the color transform rounding imprecision. In some cases, the limited chroma histogram envelop could be expanded using transform weighting to get similar benefits as chroma IBDI. At last, an

appropriate use of chroma QP-offsets may equilibrate the encoding gains in-between the 3 components.

## References

- [1] HM7.0, I.K.Kim, K.McCann, K.Sugimoto, B.Bross, WJ.Han, "HM67: High Efficiency Video Coding (HEVC) Test Model 67 Encoder Description," [http://phenix.int-evry.fr/jct/doc\\_end\\_user/documents/9\\_Geneva/wg11/JCTVC-I1002-v1.zip](http://phenix.int-evry.fr/jct/doc_end_user/documents/9_Geneva/wg11/JCTVC-I1002-v1.zip).
- [2] ITU-R BT2020, "Parameter values for ultra-high definition television systems for production and international program exchange", August 2012.
- [3] A.Dueñas et al., "On a 10-bit consumer-oriented profile in High Efficiency Video Coding (HEVC)," ITU-T/ISO/IEC Joint Collaborative Team on Video Coding (JCT-VC) document JCTVC-K0109, Oct. 2012.
- [4] M.Zhou, "Evaluation results on IBDI," ITU-T/ISO/IEC Joint Collaborative Team on Video Coding (JCT-VC) document JCTVC-D025, Jan. 2011.
- [5] A.Suhre, K.Kose, AE.Cetin, MN.Gurcan, "Content-Adaptive Color Transform for Image Compression," Optical Engineering, SPIE Digital Library, Jan. 14, 2012.
- [6] H.S.Malvar, G.J.Sullivan, "YCoCg-R: A Color Space with RGB Reversibility and Low Dynamic Range," ITU-T/ISO/IEC Joint Video Team (JVT) document JCTVC-1014r3, July 2003.
- [7] Recommendation ITU-R BT.709-5, (04/2002), Parameter values for the HDTV standards for production and international programme exchange, R-REC-BT.709-5-200204-I!!!PDF-E.pdf.
- [8] [http://en.wikipedia.org/wiki/Eigenvalue\\_algorithm](http://en.wikipedia.org/wiki/Eigenvalue_algorithm).
- [9] Frank Bossen, "Common test conditions and software reference configurations", Doc. JCTVC-I1100, JCT-VC of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, Geneva, CH, April 2012.
- [10] G. Bjontegaard, "Improvements of the BD-PSNR model," in ITU-T SG16 Q.6 Document, VCEG-A111, 2008.
- [11] B.Li, G.Sullivan, J.Xu, "Comparison of Compression Performance of HEVC Working Draft 5 with AVC High Profile," ITU-T/ISO/IEC Joint Collaborative Team on Video Coding (JCT-VC) document JCTVC-H0360, San Jose, CA, USA, Feb. 2012.
- [12] P. COMON, "Independent Component Analysis, a new concept?," Signal Processing, Elsevier, 36(3):287--314, April 1994, Special issue on Higher-Order Statistics.

# Enhancing MPEG for Model Based Coding

Christopher Haccius, Sukhpreet K. Khangura, Thorsten Herfet

Universität des Saarlandes, Saarbrücken, Germany

E-mail: {haccius, khangura, herfet}@nt.uni-saarland.de

**Abstract:** This paper describes an enhancement of an MPEG reference implementation to apply model based coding for more efficient video coding. While the maximum efficiency of standard coding schemes converges against the minimal entropy of a video signal, even lower data rates can be achieved by coding model based with the assumption of previous knowledge at the receiving end. With the help of computer graphical developments this approach promises a dramatic decrease in data rates for video content.

**Keywords:** Model Based Coding, Very high efficient video coding, MPEG enhancement

## 1 INTRODUCTION

Model Based Coding describes an idea to make use of rendered model information for video coding. As the compression capabilities of current video coding standards are converging against the minimal entropy of videos, model based coding offers the means to go even beyond that. The underlying idea is based on human communication and relies on a prior knowledge of scene content. If two persons communicate, the model identification of “a blue car” will already trigger a basic understanding of a certain object, and the description of special characteristics like “aluminium rims” or “a rear spoiler” requires little extra information. Model Based Coding in a similar fashion can exceed the coding capabilities of current video standards if we assume that some general knowledge is already available at the decoding side of the codec.

ITU-T H.264, also known as MPEG-4 part 10 has been an established video coding standard for the past decade. This standard was published with a reference implementation (part 5 of the MPEG-4 standard), which not only contains a rich set of features but is also very well documented. Very recently the next highly efficient video codec, H.265, contained in the MPEG-H standard, was released. This video codec is able to increase the coding efficiency of H.264 by another 50%. As this standard is just released the amount of features available in the reference implementation and its documentation are still not as complete as in the older reference implementation. In [1] Sullivan et al. have given an overview of this new standard. Part of the analysis in [1] is the structure of intra- and inter-picture coding, which is most relevant for our proposed implementation (see Section 3). According to Sullivan et al. “the video coding layer of HEVC employs the same hybrid approach (inter-

/intrapicture prediction and 2-D transform coding) used in all video compression standards since H.261” [1]. Therefore the H.264 reference implementation provides a valid test bed for Model Based Coding even with more efficient video coding standards already available.

In this paper we propose an enhancement of the MPEG reference implementation to enable model based prediction. The resulting enhanced codec can employ rendered model information as well as previously encoded frames for the prediction of coming video information. As the encoder decides between all options for the minimal data cost the implementation presents a hybrid model based video coder, which has, as the worst case scenario, the same data rate as current standard video codecs increased only by the rendering parameters necessary for model rendering. This increased efficiency comes at the cost of required storage capabilities for model data at both en- and decoding end as well as increased computational complexity for model rendering and additional prediction.

Using 3D models as additional information for video coding is not new. The idea of model aided coding has been exploited for coding of head-and-shoulder video sequences and significantly increased the coding efficiency [2]. Even for arbitrary models contained in a video sequence 3D models were employed for more efficient video coding. In [3] a 3D model is used to predict a video sequence with known motion from a starting frame. In [4] a coding scheme is proposed which allows prediction from frames that are synthesized from previous frames and 3D model information as an additional prediction source in a traditional hybrid waveform coder.

While current model based coders employ 3D model information for enhanced motion prediction of models in a video sequence, the focus of our work is usage of models for prediction of initial frames. As computer generated images become more and more photorealistic, such synthetic images can be used for I-Frame prediction. Further use of the 3D models for better motion compensation was already shown to increase the encoding performance in [2], [3] and [4].

In the next section we describe the architecture of the H.264 video codec that we enhanced to enable model based coding. Focus is on prediction structure and reference order, as these components of the implementation need to be modified to enable the proposed enhancement. In the consecutive section the necessary enhancements are described in detail. Experimental results in Section 4 underline the strength of

**Corresponding author:** Christopher Haccius, Telecommunications Lab, Saarland University, Campus C6.3, 9.06, 66123 Saarbrücken, +49 681 302 6544, haccius@nt.uni-saarland.de

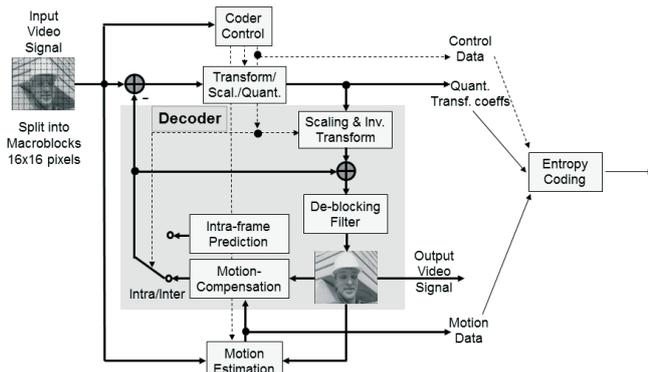


Figure 1: Basic coding structure for H.264 [5]

model based coding compared to standard encoding schemes. Finally, a conclusion is drawn and future work is outlined.

## 2 ARCHITECTURE OF THE ENCODER

The basic coding structure for H.264/AVC is shown in Figure 1. First the input video signal is split into macroblocks. The video coding layer employs temporal and spatial prediction, in conjunction with transform coding. The residual of the prediction (either intra- or inter-) which is the difference between the original and the predicted block is transformed into frequency domain. The resulting transform coefficients are scaled and quantized. In the next step these quantized transform coefficients are entropy coded and transmitted together with additional information for either intra-frame or inter-frame prediction. The encoder already contains a full decoder to enable prediction for the next blocks or the coming pictures. Therefore, in the decoder the quantized transform coefficients are inverse scaled and inverse transformed, parallel to the process at the decoder side, resulting in the decoded prediction residual. The decoded prediction residual is added to the prediction. The result of that addition is fed into a deblocking filter. The deblocking filter removes blocking artifacts resulting from the blockwise processing of the transform and prediction. It improves the video quality and provides the decoded video as its output [5]. Frames or fields are encoded to form coded pictures. A coded picture is composed of one or more slices. These slices are further decoded to produce decoded pictures which are stored in a Decoded Picture Buffer (DPB) from which these pictures may be used for inter prediction of further coded pictures and/or output for display. To enable model based input it is important to distinguish between the different picture orders.

**Decoding order:** The decoding order describes the order in which coded pictures are decoded from the bitstream by the video decoder. It is indicated by the parameter *frame\_num*.

**Display order:** The order in which pictures are output for display is given by the display order. This order of pictures is determined by the parameters *TopFieldOrderCount* and *BottomFieldOrderCount*, collectively described as Picture Order Count, POC.

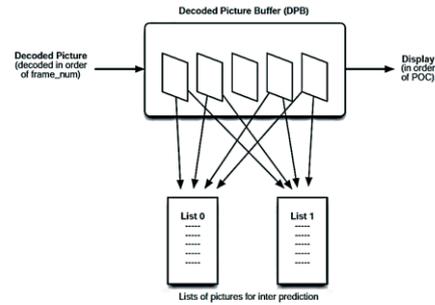


Figure 2: Decoded Picture Buffer and picture orders [6]

**Reference order:** The order in which pictures are arranged for inter prediction of other pictures. The reference order of pictures is determined by one or two lists, each of which is an ordered list of all the available decoded pictures. A P slice uses a single list, list0 (L0) and a B slice uses two lists, list0 (L0) and list1 (L1).

An Instantaneous Decoder Refresh (IDR) frame is a special kind of I frame used in H.264/AVC encoding. A coded video sequence begins with an IDR (made up of I- or SI-slices) and ends when a new IDR slice is received. On receiving an IDR coded picture, the decoder marks all pictures in the reference buffer as “unused for reference” i.e., it clears the contents of the reference picture buffer. All subsequently transmitted slices can be decoded without reference to any frame decoded prior to the IDR picture [6].

## 3 ARCHITECTURE ENHANCEMENTS

The main idea of the proposed enhancement is to modify the reference implementation in such a way, that rendered model information can be used in addition to already encoded video frames, as depicted in Figure 3.

The origin of the models for rendering is explicitly not part of this paper. In [7] we have shown that simplified model transmission is possible at negligible extra cost in the video stream. A vision is a centralized model store where numerous detailed models are made available for model based encoding. In [8] we present a novel scene representation for image and video data that allows object representations as part of video content. Additionally, novel capturing procedures and processing algorithms are presented, which results in segmentation and object detection that can be used for model based coding as presented here.

In order to inform the decoder about the required model renderings, parameters need to be transmitted. These parameters contain information on which models need to be rendered with corresponding transformations as well as extrinsic and intrinsic camera parameters. With the help of these parameters the decoder can render the required

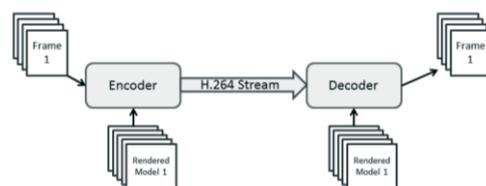


Figure 3: Simplified concept of Model Based Prediction



Figure 4: Frames of the encoded video sequence



Figure 5: Rendering of a Mazda 3 MPS model

model views. In order to use these rendered models for inter prediction of further coded pictures, the generated views need to be inserted in the Decoded Picture Buffer. We therefore allocate additional space in the decodable picture buffer according to the number of model renderings and the size of each model frame. The easiest way to make use of the models is to present the model frames to the encoder as already encoded video frames. This can be done by adjusting the picture order parameters introduced in Section 2 for the model frames. After successfully including the model frames in the decodable picture buffer and adjusting the picture order parameters the video can be encoded following the usual encoding order, with the single change of adjusting the frame order according to the before added number of model frames.

In the encoded H.264 stream model information shall not be included. This can be most easily done by explicitly excluding model frames from being added to the output stream. The resulting stream is of course no valid MPEG stream any more, as the MPEG decoder expects a fully self-contained video stream. Therefore, adjustments in the decoder need to be made as well.

On the decoding side the enhancements are parallel to the changed made on the encoding side. First of all, the model parameters are retrieved and the requested model views are rendered. Afterwards the decoded picture buffer is initialized according to the video parameters and extended by the space requirements of the model views. In the consecutive steps the model renderings are fed into the buffer, and the decoding process of the video information is then executed. Having the same model knowledge as the model based encoder the decoder can fully reconstruct the video stream.

## 4 EXPERIMENTS

Our current implementation includes the enhancement of the MPEG reference software without the renderer. For our experimental setup we have therefore created modified frames as pseudo-models, as they would be created by an embedded renderer. As the actual renderer is not part of the implementation, the rendering parameters are not exact.

The video to be encoded is given by the 6 frames shown in Figure 4; a Mazda 3 MPS sports car at a beach. The video resolution is adjusted to the sample video input of

the H.264 reference implementation of 176 x 144 pixels. We simulate a scene change by including a black frame before these six frames. For prediction we employ a 3DS Max model [9] of the same car type (see Figure 5). To employ this model for prediction of the video content the model was transformed to fit the content of the first video frame and a single light source was adjusted to match the direction of light in the video.

The car model has a size of roughly 2.12MB plus an additional 9.4MB for its textures. This amount of data already clearly exceeds the data requirements of the short video file. However, as we assume the model being available at the receiver transmission of transformation parameters for the model is comparably cheap.

Table 1 presents the data requirements for the encoded video sequence, consisting of the six frames shown in Figure 4. After the Non-Video Bits (NVB) the IDR frame requires almost twice as many bits as the following P frames. The main benefit of model based coding can be expected for the IDR frames, if they can be predicted by model information. Table 2 gives the required data rates if the model as described is used as an additional predictor with our enhanced implementation. Table 3 repeats this experiment with the assumption, that not only the model but also background information is available in front of which the car can be rendered. While for this example this is unlikely, the general case that background information is available by models shall not be neglected.

The rendering of the car model fills roughly 6930 pixels of the 25344 frame pixels, which is roughly  $\frac{1}{4}$  of the frame size.

For a better understanding of the coding choices the H.264 reference implementation takes we visualize the encoded difference when the first frame is predicted from a model. Figure 6 shows a visualization of the difference when predicted from the rendered car model only (a) and when predicted from a model in front of the original background (b).

The computed quality of the video frames was kept constant for all video sequences at a PSNR of  $\sim 40$ dB. The distribution of the noise in the individual frames was found to be comparable in all encoding runs. Therefore the data requirements for the frames are directly comparable.

**Table 1: MPEG-Encoded Frames of Car Sequence**

| Frame    | Bit/Pic |
|----------|---------|
| 00 (NVB) | 176     |
| 00 (IDR) | 26456   |
| 01 (P)   | 14064   |
| 02 (P)   | 14768   |
| 03 (P)   | 15616   |
| 04 (P)   | 14584   |
| 05 (P)   | 15484   |

**Table 2: Model-Based - Encoded Frames of Car Sequence with simple car model**

| Frame    | Bit/Pic |
|----------|---------|
| 00 (NVB) | 176     |
| 00 (IDR) | 20536   |
| 01 (P)   | 14384   |
| 02 (P)   | 14920   |
| 03 (P)   | 15424   |
| 04 (P)   | 14896   |
| 05 (P)   | 15448   |

**Table 3: Model-Based - Encoded Frames of Car Sequence with car model and background**

| Frame    | Bit/Pic |
|----------|---------|
| 00 (NVB) | 176     |
| 00 (IDR) | 4712    |
| 01 (P)   | 12984   |
| 02 (P)   | 14696   |
| 03 (P)   | 15520   |
| 04 (P)   | 15024   |
| 05 (P)   | 15136   |

## 5 EVALUATION AND CONCLUSION

When comparing the Tables 1, 2 and 3 the effect of model based prediction becomes clearly visible, and exactly relates to the expected data reduction. The accumulated data rates (see Table 4) clearly reflect the amount of image information that can be retrieved using the model source.

Encoded with the MPEG reference implementation the IDR frame requires 3.3MB. If we predict the car from a car model with our enhanced encoder, the required data for the IDR frame is reduced to 2.6MB. As given in Section 4 the car makes up roughly  $\frac{1}{4}$  of the frame information, and as can be seen in Figure 6 (a) not the model information still needs to be corrected to a certain degree, especially where the windows are located and



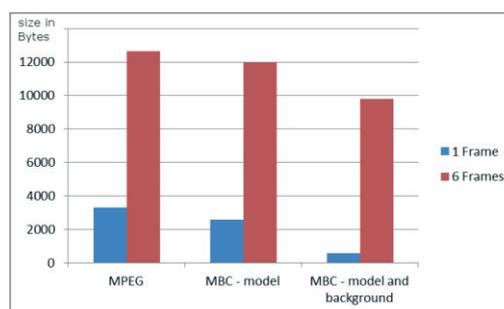
(a) from model only (b) from model and background  
**Figure 6: Visualized difference of predicted IDR frame**

where light reflectance changes the appearance. Therefore a reduction by 0.7MB which is a bit less than  $\frac{1}{4}$  of the frame information is in line with expectations.

Furthermore, if we assume the background of the scene to be known, a reduction of the required data to 0.6MB can be achieved. As visualized in Figure 6 (b) now only  $\frac{1}{4}$  of the video frame needs to be partially corrected.

At the same time it can be noticed that the model knowledge has no meaningful effect on the successive video frames. Therefore the proportional compression gain decreases as the GOP-length increases, means as more pictures are predicted from previously decoded video frames without IDR frames or scene changes in between. A graphical representation of these results is given in Figure 7. The reason for this observation is simple: The realism of rendered images suffices to help predicting I frames, but motion compensation from a previous frame is cheaper than also predicting a P frame from a model.

The results show that a hybrid extension for Model Based Coding of MPEG can lead to very promising results. Essential for the compression are the amount of scene content that can be predicted from models as well as the quality of the models. Fully implemented the model based coding scheme has the ability to compress beyond the minimal entropy content of a video.



**Figure 7: Video size in Bytes**

**Table 4: Accumulated Data Requirements**

| Prediction                            | Bytes/Video |
|---------------------------------------|-------------|
| MPEG                                  | 12646       |
| Model Based (from car model)          | 11973       |
| Model Based (from car and background) | 9781        |

## 6 FUTURE WORK

Right now the encoder employs model data which is rendered externally. A next step in the model based codec development is to integrate the renderer into the encoder. Apart from being tidier to have all tools included in a single application this is also a necessary step to ensure a working codec. Different renderers tend to interpret model data differently, but correct decoding depends on the same understanding of model data.

Combining the model based prediction presented here with the model aided prediction suggested in [3] and [4] can further increase the coding efficiency. As the 3D model is made available for the I frame already, it is a small step to reuse it for model aided motion compensation without additional costs.

Prediction from models employing the same error measures as for prediction from previous frames is suboptimal. This is due to a large gap between perceived error and calculated error. Rendered models tend to have a high perceived quality (good resolution, realistic appearance) but a low SNR compared to the original image, as object edges, material colour gradients or other details irrelevant to a human observer can differ slightly. Currently perceived quality can only be comprehensively measured by subjective studies [10]. However, including such a measurement into the encoder is essential to achieve better compression rates in model based coding.

Image and scene analysis are important for placing models in a scene. Apart from few research scenes where model data is already available this process requires manual input. Currently, a European research project aims at enhancing the video acquisition process, with one goal to detect objects in scenes [11]. Further development of

such methods will be important for successful application of model based coding. Along with model detection the availability of models needs to be enhanced, such that encoder and decoders can refer to the same model knowledge without the need to actually exchange this data.

## References

- [1] G. J. Sullivan, J.-R. Ohm, W.-J. Han and T. Wiegand. Overview of the high efficiency video coding (hevc) standard. *IEEE Transactions on Circuit and Systems for Video Technology*, 22, 2012.
- [2] P. Eisert, T. Wiegand and B. Girod. Model-Aided Coding: A New Approach to Incorporate Facial Animation into Motion-Compensated Video Coding. *IEEE Transaction on Circuits and Systems for Video Technology*, Vol. 10, No. 3. Apr. 2000
- [3] F. Galpin and L. Morin. Computed 3D Models for very low bitrate Video Coding. *Proc. SPIE Conference Visual Communications and Image Processing*. Vol. 4310, 2001
- [4] C.-L. Chang, P. Eisert and B. Girod. Using a 3D Model for Video Coding. *Proc. Vision, Modeling and Visualization*. Nov. 2002
- [5] R. Schäfer, T. Wiegand and H. Schwarz "The emerging H.264/AVC Standard" EBU, Technical Review – Jan. 2003
- [6] Iain E. Richardson, "H.264 and MPEG-4 Video Compression: Video Coding for Next-generation Multimedia"
- [7] C. Haccius and T. Herfet. Model based coding revisited. In *Picture Coding Symposium (PCS)*, 2012, pages 313–316. IEEE, 2012.
- [8] C. Haccius, T. Herfet, V. Matvienko, P. Eisert, I. Feldmann, A. Hilton, J. Guillemaut, M. Kludiny, J. Jachalsky, S. Rogmans. A Novel Scene Representation for Digital Media. *Proc. of Networked and Electronic Media (NEM) Summit*. Oct. 2013
- [9] Mazda 3M MPS 3D Model, The Free 3D models, retrieved from [http://thefree3dmodels.com/stuff/vehicles/mazda\\_3\\_mps/13-1-0-5352](http://thefree3dmodels.com/stuff/vehicles/mazda_3_mps/13-1-0-5352), May 2013
- [10] N. Abukhodair. A user perceived quality assessment of lossy compressed images. *International Journal of Computer Graphics*, 2(2), 2011.
- [11] V. López, E. Fuenmayor, and A. Hilton, *Novel scene representations for richer networked media*. <http://3d-scene.eu/>, Jan. 2013.

# Entropy Constrained Scalar Quantization for Laplacian Distribution: Application to HEVC

Michaël Ropert<sup>1</sup>, Marine Sorin<sup>2</sup>, François Ropert<sup>3</sup>

<sup>1,2</sup>Envivio, Saint Jacques de la Lande, France, <sup>3</sup>Inouco, Caen, France

E-mail: <sup>1</sup>[mropert@envivio.com](mailto:mropert@envivio.com), <sup>2</sup>[msorin@envivio.com](mailto:msorin@envivio.com), <sup>3</sup>[soframarp@wanadoo.fr](mailto:soframarp@wanadoo.fr)

**Abstract:** This paper presents a Rate-Distortion Optimization (RDO) of the dead-zone in the case of a uniform quantization for a random variable modeled by a Laplace distribution. It follows many studies [1], [2], [3] and references therein, addressing the problem of the optimal quantization. This study is restricted to the combination of uniform threshold scalar quantization with dead-zone (UTSQ+DZ), and the uniform reconstruction quantization (URQ). The method of Lagrange multiplier is applied for the entropy constrained minimization of the distortion. The explicit formulas for the  $\lambda$  multiplier, the dead-zone threshold  $T$ , both the rate  $R$  and the distortion  $D$  are briefly presented. Results are compared with other solutions obtained in a more generalized context [12]. Then a bound for the Rate-Distortion  $R(D)$  function is derived, and compared with already known approximations toward low or high bitrates. These new results provide a more accurate description of the quantization behavior, to improve MPEG-like encoder compression performances. A direct application to HEVC improved quantization is presented using the optimal threshold quantization parameter  $T$ .

**Keywords:** Quantization, RDO, compression, Lagrange multiplier optimization.

## 1 INTRODUCTION

In lossy coding, the reconstructed signal is not identical to the original signal. The measure of this difference is referred to as distortion. The approximated signal requires a certain amount of bits to be completely described. The issue of lossy compression is to minimize this number of bits while preserving the distortion as small as possible. Digital video is the application area where the compression is widely utilized. Video compression efficiency is growing continuously since the beginning of the block based JPEG [4] standard. With MPEG, motion has been introduced for video sequences to take advantage of the redundancy between frames. And successive standards H.262/MPEG2 [5], H.264/MPEG-4 AVC [6], MPEG HEVC [7] always improve video compression by the introduction of refinements, or more recently by changing entropy coding mechanism. Basically, the idea is to perform the best causal prediction in such a way the decoder is able to redo the same without anything to be transmitted to the decoder. Close to this ideal configuration, transformed residue distributions (to

be transmitted) become sharper and sharper with new compression standards [8]. Sharp statistical models like Laplace or Cauchy distributions [9] are good candidates to statistically describe those data.

In the initial part of this paper, we present the RDO problem according to the uniform quantization model, the optimal dead-zone and the optimal Lagrange Multiplier. Then, the distortion and the rate are presented and discussed. To evaluate the consistency of the equations, the  $R(D)$  curve for a single coefficient is asymptotically compared with known bounds [14] from the rate distortion theory. Afterwards, the application to adaptive threshold quantization is presented in the context of HEVC compression. To achieve this R-D optimized quantisation, a proposal to combine several Lagrange multipliers into a single one is considered for simplification of the operational R-D model. In the final part, experimental compression gains results are presented using this technique. And finally, the conclusion summarizes the different steps, emphasizes the interesting results and proposes future works.

## 2 OPTIMAL RDO QUANTIZATION

In the first part of this paper, we consider each transformed coefficient separately. The position of each coefficient is not mentioned. So each coefficient has its own distribution and its own standard deviation. The distortion and the bitrate are then attached to the considered coefficient. It is the most commonly used model [2] used for  $R(D)$  curves computations.

The goal of the rate distortion-optimization is to minimize the distortion  $D$  for a given rate  $R_M$ , i.e.,

$$\text{MIN}\{D\} \text{ subject to } R = R_M. \quad (1)$$

Thanks to the Lagrange multiplier method, this classical problem can be re-written as a single minimization:

$$\text{MIN}\{J\} \text{ where } J = D + \lambda.R. \quad (2)$$

$J$  is called Lagrangian, and  $\lambda$  the Lagrange multiplier. Both  $R$  and  $D$  depend on the quantization method.  $R$  also depends of the entropy coder used to transformed quantized values into successive bits. To avoid taking into account the entropy coder efficiency,  $R$  is replaced by  $H$  the entropy, which is the reachable efficiency limit.



## 2.4 Lagrange Multiplier Solution

Minimizing  $J$  in (2) with respect to  $T$  produces two sets of equations [11]:

$$\begin{cases} i > 1 & 2 \cdot \Delta \cdot \left(T - \frac{\Delta}{2}\right) + \lambda \cdot \log_2(e^{-\alpha \Delta}) = 0 \\ i = 1 & 2 \cdot \Delta \cdot \left(T - \frac{\Delta}{2}\right) + \lambda \cdot \log_2\left(\frac{e^{-\alpha T} - e^{-\alpha(T+\Delta)}}{2 \cdot (1 - e^{-\alpha T})}\right) = 0 \end{cases} \quad (9)$$

By identification of the second term in (9), the optimal  $T$  is obtained, then the optimal  $\lambda$ . Finally:

$$\frac{\partial}{\partial T}(J) = 0 \Leftrightarrow \begin{cases} T = \frac{1}{\alpha} \ln\left(\frac{1 + e^{\alpha \Delta}}{2}\right) \\ \lambda = \frac{2 \cdot \ln(2)}{\alpha^2} \cdot \ln\left(\cosh\left(\alpha \frac{\Delta}{2}\right)\right) \end{cases} \quad (10)$$

Clearly, the quantization threshold is linear with  $\lambda$ .

$$T = \frac{\Delta}{2} + \frac{\lambda \cdot \alpha}{2 \cdot \ln(2)}. \quad (11)$$

## 2.5 Optimal $\lambda$ and $T$ values

Given the standard deviation of the Laplacian variable and the quantization step, when the optimal  $\lambda$  is known, then the threshold is fixed. This property can be used to adapt the quantization threshold dynamically. Moreover, from (10), we get the following configuration:

Table 1:  $T$  and  $\lambda$  with respect to  $\alpha$

| distribution | flat (uniforme)                   | sharp (dirac) |
|--------------|-----------------------------------|---------------|
| $\alpha$     | 0                                 | $+\infty$     |
| $T$          | $\frac{\Delta}{2}$                | $\Delta$      |
| $\lambda$    | $\ln(2) \cdot \frac{\Delta^2}{4}$ | 0             |

For flat distribution, i.e.  $\alpha \rightarrow 0$ , as shown table 1, there is no dead-zone, and  $\alpha \rightarrow 0$  the thresholds must be centred between two reconstruction levels. At the opposite, for sharp distributions, i.e.  $\alpha \rightarrow +\infty$ , the dead-zone is maximum, and optimal thresholds mapped on the right for reconstruction levels for positive indexes (resp. left for negative indexes).

The quadratic form [13] for  $\lambda$  corresponding to flat distribution (i.e.  $\alpha \rightarrow 0$ ) is very often used in reference software. However, when the variance is high, the quantization is adapted to achieve compression, in such a way the combination of a low  $\alpha$  and a low  $\Delta$  is not used for compression. Let's define  $x = \alpha \cdot \Delta$ . The area where lambda is used is then limited to  $x$  far from zero ( $\alpha \cdot \Delta \gg 0$ ). In the area of interest,  $\lambda$  is almost linear with the quantization step:

$$\lim_{x \gg 0}(\lambda) = \frac{\ln(2)}{\alpha} \cdot \Delta - 2 \cdot \left(\frac{\ln(2)}{\alpha}\right)^2. \quad (12)$$

Nevertheless, in practice, this approximation is difficult to utilize. It requires the knowledge of the coefficient distribution variance.

## 3 RATE DISTORTION FUNCTION

The optimization problem gave an explicit solution regarding the Lagrangian minimization. It provides the opportunity to exhibit both  $D$  the distortion, and  $R$  the rate.

### 3.1 A posteriori L2 Distortion and Rate

Knowing  $T$  and  $\lambda$  from (10), (7) and (8) are computed explicitly.

$$\begin{cases} D = \Delta^2 \left( \frac{2}{x^2} + \frac{2}{e^x - e^{-x}} \left( 1 - \frac{2}{x} \left( \ln\left(\frac{1+e^x}{2}\right) + 1 \right) \right) \right) \\ R = \frac{1}{\ln(2)} \left( \ln\left(\frac{1+e^{-x}}{1-e^{-x}}\right) + x \cdot \frac{2}{e^x - e^{-x}} \right) \end{cases} \quad (13)$$

Table 2:  $D$  and  $R$  with respect to  $\alpha$

| distribution | flat (uniforme)       | sharp (dirac) |
|--------------|-----------------------|---------------|
| $\alpha$     | 0                     | $+\infty$     |
| $D$          | $\frac{\Delta^2}{12}$ | 0             |
| $R$          | $+\infty$             | 0             |

When the step size becomes huge, it can be verified from (13) that the distortion is bounded by the variance  $\sigma^2$ :

$$\lim_{\Delta \rightarrow +\infty} D = \frac{2}{\alpha^2} = \sigma^2. \quad (14)$$

In table 2, the infinite value of  $R$ , when  $\alpha \rightarrow 0$ , is only valid because  $N \rightarrow +\infty$ . In practice,  $N$  is a finite number, and the entropy should be less than  $\log_2 N$ .

### 3.2 R(D) function

$R(D)$  is difficult to write as a combination of usual known functions. Nevertheless, making  $x$  varying from 0 to  $+\infty$ ,  $R(D)$  can be drawn as a plane curve of  $x$  as described figure 3.

Approximation [14] for high bitrate Gaussian and Laplacian distributions give  $D(R) \approx \varepsilon^2 \sigma^2 2^{-2R}$ . In the case of a Laplace distribution [11], the Taylor expansion of  $2^{-2R(x)}$  and  $D(x)$  for  $x$  close to zero provides the "high bit rate" approximation  $\varepsilon^2 \approx e^2 / 6$ .

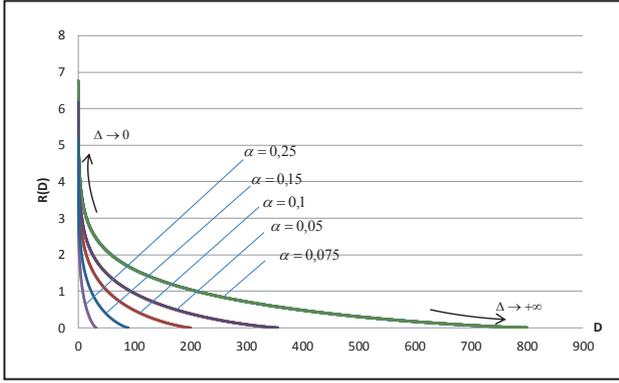


Figure 3:  $R(D)$  plotted as a plane curve.

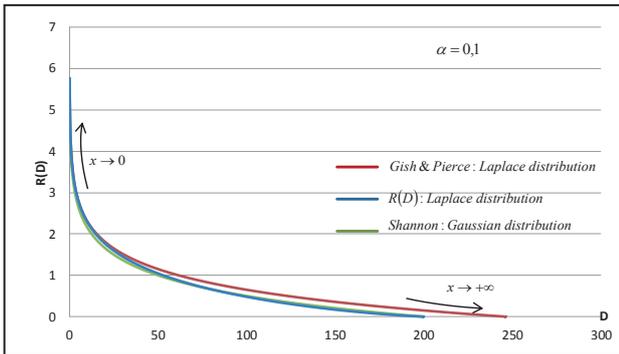


Figure 4:  $R(D)$  and its approximations plotted at  $\alpha = 0.1$ .

The three  $R(D)$  functions are compared in figure 4. As expected, the Gish & Pierce [14] approximation corresponding to  $D(R) \approx e^2 \sigma^2 2^{-2R} / 6$  and the  $R(D)$  function given by (13) are close for high resolution. It also shows that the uniform quantizer with dead-zone behavior is close to the Gaussian case. For a given perfect entropy coder, with independent coefficients (not necessarily i.i.d.), there would be no need for another quantization technique. Unfortunately, transformed coefficients are often dependent, because the decorrelation brought by the transform is not sufficient.

## 4 APPLICATION TO HEVC

The rounding offset is the distance between the threshold  $T$  and the first non-zero reconstruction point. In many existing video encoding systems, the quantization rule has been to use a DZ+UTQ with fixed rounding offset such as 1/2 or 1/3. For example, in the H.265/HEVC reference software encoder (HM 11), the relative rounding offset is set to 1/3 for all Intra modes (small dead-zone) and is set to 1/6 for all Inter modes (large dead-zone). The same rule is used in the H.264/AVC and H.263 reference software encoders (resp. JM, TMN). Obviously, those values are sub-optimal, based on the previous results. Adapted rounding technique should provide improvements with respect to fixed rounding offsets.

### 4.1 Adaptive quantization

The adaptive quantization rounding technique is straightforward. First, the signal statistics of DC and each AC component is collected. The Laplacian nature of AC

coefficients can be observed figure 1. No distinction is performed between DC and AC coefficients for sake of simplicity (a more accurate model for the DC coefficient could be the Gaussian distribution according to the central limit theorem). Coefficients are considered centred.

Three steps are required to perform the adaptive quantization:

1. The variance value of each component is computed taking into account the distribution shape, under the maximum likelihood criterion, leading to the estimation of  $\alpha$  :

$$\alpha = \left( \frac{1}{M} \sum_{j=1}^M |z_j| \right)^{-1}. \quad (15)$$

where  $M$  is the number of transform units (TU) for the considered DCT coefficient. Let's  $p$  be the position of the coefficient in the DCT grid. Then each coefficient has its own  $\alpha_p$  estimation.

2. For the given quantization step  $\Delta$ , and from  $\alpha_p$ , the Lagrange multiplier  $\lambda_p$  and the thresholds  $T_p$  are computed for each coefficient using (10) and (11).
3. The Quantization is applied with the adapted Thresholds.

### 4.2 $\lambda$ adjustment

The global  $\lambda$  is readjusted to limit the estimation bias. Actually, the initial DCT coefficients were obtained after a first encoding pass, via the RDOQ technique which is the following minimisation:

$$\text{MIN} \{ J_{RDOQ} \} \text{ where } J_{RDOQ} = \sum_p D_p + \lambda \cdot \sum_p R_p. \quad (16)$$

This function is to be compared with:

$$\text{MIN} \left\{ \sum_p J_{T_p} \right\} \text{ where } J_{T_p} = D_p + \lambda_p \cdot R_p. \quad (17)$$

$\sum_p J_{T_p}$  and  $J_{RDOQ}$  share the same minimum for a particular value of  $\lambda$ . So  $\lambda$  is adapted to keep the two optimisations coherent:

$$\lambda = \frac{\sum_p \lambda_p \cdot R_p}{\sum_p R_p}. \quad (18)$$

This workaround avoids the global RDO approach as proposed in [10]. No further iteration is needed; it converges to the fixed  $\lambda$  value almost directly. The initial  $\lambda$  provided by the HM is close to this re-estimated value (factor less than 2 observed experimentally, and very close to 1 for small quantization steps).

### 4.3 Experimental results

Several sequences were tested with the adaptive quantization presented above.

A first encoding pass with original HM 11 at fixed quantization step ( $\Delta$ ) allows us to collect the signal statistics:

- 16 frequency components per luma 4x4 block in Inter modes
- Chroma components are not collected
- Components in Intra modes are not collected

From statistics of input data the optimal  $T$  is computed for each component. The final encoding is obtained injecting  $T$  in HM 11 with the new adaptive quantization method (HM 11 + ADZ).

Experiments are conducted using several SD (720x576) format sequences. Only two frames are encoded: one I frame and one P frame: we consider the bitrate and the distortion of the P frame. The original HM 11 is used as the reference encoder with the following configuration:

- 2 frames (IP)
- CABAC
- Maximum CU = 16x16
- Transform size = 4x4
- Search range = 64
- no RDOQ
- no SAO
- Fixed QP

The same configuration is applied in HM 11 + ADZ. R-D curves for HM 11 encoder and HM 11 + ADZ encoder are shown in figure 5-a to 5-d. Improvement (particularly at high bit rates) can be observed by use of the new adaptive quantization technique.

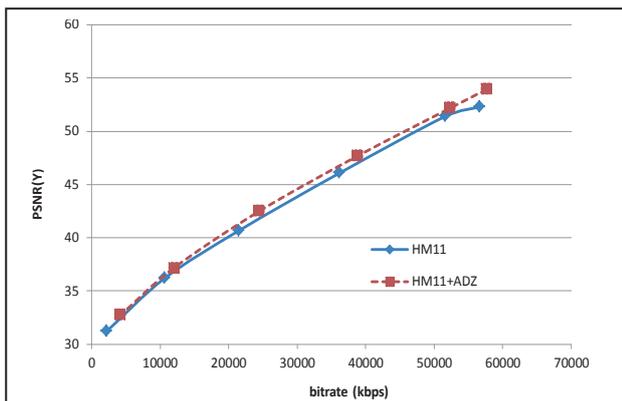


Figure 5-a: PSNR vs bitrate: "City" sequence.

The similar trend of HM 11 and HM 11 + ADZ curves (figure 5-a to 5-d) for low bitrates is explained by the estimation method of DCT coefficients variances which are done once at very high bitrate, and supposed unchanged going down toward lower bitrates. The robustness of this simplification is probably to be evaluated.

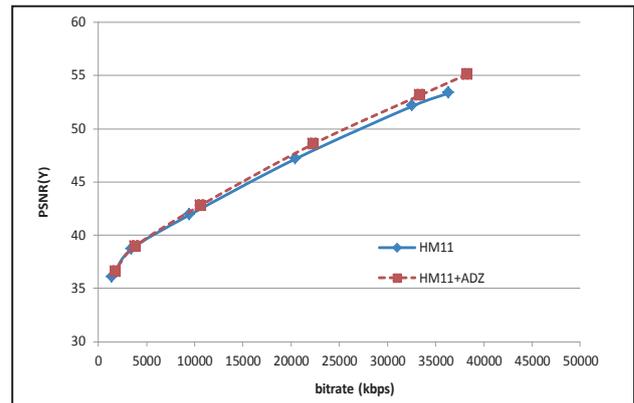


Figure 5-b: PSNR vs bitrate: "Islande" sequence.

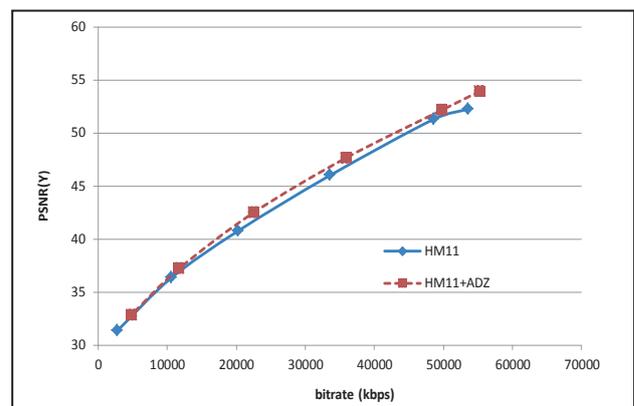


Figure 5-c: PSNR vs bitrate: "Journalist" sequence.

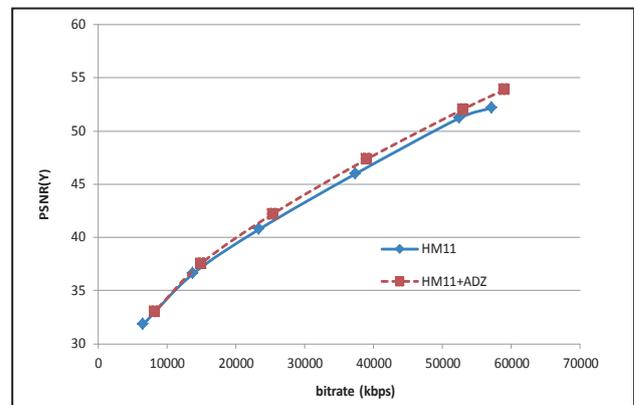


Figure 5-d: PSNR vs bitrate: "Canaries" sequence.

## 5 CONCLUSION

We have presented explicit optimal formulas for Rate and Distortion based on the Lagrangian optimization technique in the case of exponential distributions. It was shown that for each coefficient, the optimal dead-zone is linked to  $\lambda$ : compression gains can be obtained by selecting thresholds  $T$  adapted to  $\lambda$  values. The proposed technique of uniform quantization with optimal dead-zone is applied to the HEVC standard. Compression improvements are observed compared to the software

reference. The experimental results were limited to only one transform size (4x4), the generalisation to several transform size requires further developments. Regarding the applications of the new R(D) results, additional compression efficiency should be achieved by adapting each quantization step to the DCT coefficient distribution statistics (i.e. adapted quantization matrix).

## References

- [1] G.J. Sullivan, "Efficient Scalar Quantization of Exponential and Laplacian Random Variables," *IEEE Trans. on Information Theory*, vol. 42, no. 5, pp. 1365–1374, Sept. 1996.
- [2] X. Li, N. Oertel, A. Hutter, and A. Kaup, "Laplace Distribution Based Lagrangian Rate Distortion Optimization for Hybrid Video Coding," *IEEE Trans. on Circuit and Systems for Video Technology*, vol. 19, no. 2, pp. 193–205, Feb. 2009.
- [3] V.K. Goyal, "Theoretical Foundation of Transform Coding," *IEEE Signal Processing Magazine*, vol. 18, no. 5, pp. 9–21, Sept. 2001.
- [4] ISO/IEC JTC1 10918-1, "Digital Compression and Coding of Continuous Still Images, Part 1, Requirements and Guidelines," Draft International Standard, Nov. 1991.
- [5] ISO/IEC 13818-2 MPEG-2 (ITU-T Rec. H.262) "Generic Coding of Moving Picture and Associated Audio," MPEG-2 International Standard 13818-2, 1994.
- [6] ISO/IEC 14496-10 AVC (ITU-T Rec. H.264) "Draft ITU-T recommendation and final draft international standard of joint video specification," JVT-G050, 2003.
- [7] ISO/IEC 23008-2 HEVC (ITU-T Rec. H.265) "High Efficiency Video Coding," Final Draft International Standard and ITU-T, 2013 .
- [8] E. Lam and J.A. Goodman, "A Mathematical Analysis of the DCT Coefficient Distributions for Images," *IEEE Trans. on Image Processing*, vol. 9, no. 10, pp. 1661-1666, Oct. 2000.
- [9] Y. Altunbasak and N. Kamaci, "An analysis of the DCT coefficient distribution with the H.264 video coder," *IEEE Int. Conf. on Acoustics, Speech, and Signal Process. (ICASSP)*, Montreal, Canada, pp. III-177–80, May 2004.
- [10] C. Parisot, M. Antonini and M. Barlaud, "Optimal Nearly Uniform Scalar Quantizer Design for Wavelet Coding," *Proc. SPIE Visual Commun. Image Process.* vol. 4671, p. 1185-1193, Jan. 2002.
- [11] M. Ropert, and F. Ropert, "RD Optimisation of Uniform Threshold Scalar Quantization for Laplacian Distributions," accepted 30<sup>th</sup> Picture Coding Symp., San Jose, California, Dec.2013
- [12] G.J. Sullivan and S. Sun, "On Dead-Zone Plus Uniform Threshold Scalar Quantization," *Proc. SPIE Visual Commun. Image Process.*, 2005, pp. 1041-1052.
- [13] T. Wiegand and B. Girod, "Lagrange Multiplier Selection in Hybrid Video Coder Control," *Proc. IEEE ICIP*, p. 542-545, Cairo, Egypt, Oct. 2001.
- [14] H. Gish and J. N. Pierce, "Asymptotically efficient quantizing," *IEEE Trans. Inform. Theory*, vol. IT-14, pp. 676-683, Sept. 1968.



# **Enhanced Media Content Generation, Transmission and Consumption II**

# From Raw Data to Semantically Enriched Hyperlinking: Recent Advances in the LinkedTV Analysis Workflow

Daniel Stein<sup>1</sup>, Alp Öktem<sup>1</sup>, Evlampios Apostolidis<sup>2</sup>, Vasileios Mezaris<sup>2</sup>, José Luis Redondo García<sup>3</sup>, Raphaël Troncy<sup>3</sup>, Mathilde Sahuguet<sup>3</sup>, Benoit Huet<sup>3</sup>

Fraunhofer Institute IAIS, Sankt Augustin, Germany<sup>1</sup> Information Technologies Institute CERTH, Thessaloniki, Greece<sup>2</sup> Eurecom, Sophia Antipolis, France<sup>3</sup>

**Abstract:** Enriching linear videos by offering continuous and related information via, e.g., audio streams, web pages, as well as other videos, is typically hampered by its demand for massive editorial work. While a large number of analysis techniques that extract knowledge automatically from video content exists, their produced raw data are typically not of interest to the end user. In this paper, we review our analysis efforts as defined within the LinkedTV project and present the recent advances in core technologies for automatic speech recognition and object-redetection. Furthermore, we introduce our approach for an automatically generated localized person identification database. Finally, the processing of the raw data into a linked resource available in a web compliant format is described.

**Keywords:** Automatic Speech Recognition, Object Redetection, Person Identification, NERD Ontology

## 1 Introduction

Enriching videos (semi-)automatically with hyperlinks for a sophisticated viewing experience requires analysis techniques on many multi-modal levels. In [13], we presented the overall architecture decision for video analysis in the “Television linked to the Web” (LinkedTV)<sup>1</sup> project.

This paper focuses on the issues that we identified as most pressing (and there were quite a few): Local Berlin interviews featured a lot of interviews with the local residents, whose spontaneous speech produced only moderate automatic speech recognition (ASR) results. Speaker identification, while working properly on German parliament speeches, proved to be of little help since we had no localized database of Berlin speakers, a challenge that is shared with face recognition techniques. Object re-detection, for semi-automatically recognizing and tracking important objects in a show such as a local church or a painting, was too slow to be realistically employed in the architecture. Finally, the actual process of hyperlinking was left open in the last paper. In this follow-up paper, we present the new methods and the advances made, and explain our efforts in transforming raw data to semantically enriched and linked content.

This paper is organized as follows. After a brief description of the LinkedTV project (Section 2), we re-visit the ASR performance, which clearly showed deficiencies in spontaneous speech [13]. It has now been adopted to the seed content domain using a huge amount of new training material and a gradient-free optimization of the free decoding parameters (Section 3). Then, we present a stronger and faster solution for object re-detection (Section 4). Next, by interweaving several technologies such as face detec-

tion, video OCR and speaker identification, we can come up with a strong localized database for person identification (Section 5). Last, we elaborate on the actual hyperlinking stage, where the raw data is further processed (Section 6). Finally, we give a conclusion in Section 7.

## 2 LinkedTV

The vision of LinkedTV is of a ubiquitously online cloud of Networked Audio-Visual Content decoupled from place, device or source. The aim is to provide an interactive multimedia service for non-professional end-users, with focus on television broadcast content as seed videos. The project work-flow can be described as follows: starting from the demands of the use case scenarios, coupled with a description of the targeted multimedia content, the videos are analyzed by various (semi-)automatic ways. The raw data obtained from the single approaches is gathered and further enriched in a second step, by assigning media fragment descriptions and interlinking these with other multimedia information, using knowledge acquired from, e.g., web mining. The enriched videos are then shown in a suitably tailored presentation engine which allows the end-user to interact with a formerly linear video, and a recommendation/personalization engine which further gives the possibility to customize this experience.

In [13] we focused on the first two steps in this workflow, namely use case scenario and intelligent video analysis. There, we identified Berlin local news shows as seed content for the *news* use case, and the show “Tussen Kunst en Kitsch”<sup>2</sup> (similar to the Antiques Roadshow of the BBC), shown by Dutch public broadcaster AVRO,<sup>3</sup> as seed content for the *documentary* use case. This paper elaborates on the intelligent video analysis and the linking step as well as their interaction with each other.

## 3 ASR on Spontaneous Speech

Spoken content is one of the main sources for information extraction on all our relevant seed data sets. In [13], we performed a manual ASR transcript evaluation which performed good on planned speech segments, but rather poor on spontaneous parts which were quite common in interview situations in the news show scenarios. We thus decided to extend our training material with new data and adopt the settings of our decoder.

Recently, we collected and manually transcribed a huge new training corpus of broadcast video material, with a volume of approx. 400 h and containing roughly 225 h of clean speech. The new corpus is segmented into utterances

<sup>1</sup><http://www.linkedtv.eu>

<sup>2</sup><http://www.tussenkunstenskitsch.nl>

<sup>3</sup><http://www.avro.nl>

**Table 1:** WER results on the test corpora, for the SPSA iterations and their respective loss functions. Each optimization on a given loss function has been executed two times from scratch with 18 iterations to check for convergence.

| parameter set        | WER<br>planned | WER<br>spontaneous |
|----------------------|----------------|--------------------|
| baseline             | 27.0           | 52.5               |
| larger training data | 26.4           | 50.0               |
| SPSA 1st run         | 24.6           | 45.7               |
| SPSA 2nd run         | 24.5           | 45.6               |

with a mean duration of 10 seconds and is transcribed manually on word level. The recorded data covers a broad selection of news, interviews, talk shows and documentaries, both from television and radio content across several stations. Special care has been taken that the material contains large parts of spontaneous speech. As the effort for acquiring new training data is still ongoing, the final size of the corpus will eventually reach 900 h, making this one of the largest corpora of German TV and radio broadcast material known to us.

This new training material made a revisit of the free speech decoder parameters necessary, to guarantee optimality. In the literature, these parameters are often either set empirically using cross-validation on a test set, which is a rather tedious task, or the default values of toolkits are retained. Few publications analyze the parameter adaption with automatic methods; among them are [3], using gradient descent, [7], using large-margin iterative linear programming, or [5], using evolutionary strategies. Since we aim at facilitating the optimization process by employing a fast approach and therefore enable this step for a wide range of applications, we employ Simultaneous Perturbation Stochastic Approximation (SPSA) [12] for optimizing the free decoding parameters and show in [14] that it leads to stable and fast results.

The algorithm works as follows. For a tuple of free parameters in each iteration, SPSA perturbs the given values simultaneously, both adding and subtracting a random perturbation vector for a total of two new tuples. The gradient at the current iteration is estimated by the difference of the performance (here measured as word error rate, WER) between these two new tuples, and a new tuple is then computed by adapting the old tuple towards the gradient using a steadily decreasing step function. We refer to [14] for further implementation details.

For developing and optimizing the free parameters, we use a corpus from German broadcast shows, which contains a mix of planned (i.e., read news) and spontaneous (i.e., interviews) speech, for a total of 2,348 utterances (33,744 words).

For evaluation, we test the decoding performance on the news show content, separated into a planned set (1:08h, 787 utterances) and a spontaneous set (0:44h, 596 utterances). The results are listed in Figure 1. Here, it can be seen that while the performance for planned speech improved by 2.5% absolute (9.3% relative) in terms of WER, spontaneous speech segments now have a WER of almost 7% lower (13.3% relative) than the original baseline, which is quite a nice advance in the ASR quality.

## 4 Fast Object Re-detection

Since the videos in the presentation engine shall contain interactive (i.e. clickable) objects of interest, we need to associate visual content with appropriate labels. These labels can be automatically generated at the object-class level via high-level concept detection (by detecting concepts such as “car”, “person”, “building”, etc.), where we follow the approach of [10] using a sub-set of the base detectors described there. Moreover, a semi-automatic instance-based annotation of the video can be performed via the re-detection of specific objects of interest selected by the video editor so that, e.g., instances of the same painting in the antique road-show can be identified and tracked throughout the movie, allowing the viewer to click on them for further information or related videos.

We detect instances of a manually pre-defined object of interest  $O$  in a video  $V$  by evaluating its similarity against the frames of this video, based on the extraction and matching of SURF (Speeded UP Robust Features) descriptors [2]. The time performance of our method is a crucial requirement, since the object-based video annotation will be handled by the editor. A faster than real-time processing is achieved by combining two different strategies: (a) exploit the processing power of the modern Graphic Processing Units (GPUs) and (b) introduce a video-structure-based frame sampling strategy that aims to reduce the number of frames that have to be checked.

Regarding the first strategy, GPU undertakes the initial decompression of the video into frames, the extraction and description of the image’s features and the matching of the calculated descriptors for a pair of images. Specifically, for the detection and description of the salient parts of the image a GPU-based implementation of the SURF algorithm is used, while the following matching step is performed in a brute force manner (i.e. each extracted descriptor from the object  $O$  is matched against all the extracted descriptors from the  $i$ -th frame  $F_i$ ) looking each time for the 2-best matches via a k-Nearest Neighbor search for  $k = 2$ . This means that, for each detected interest point of  $O$ , the algorithm searches for the two best matches in  $F_i$  that correspond to the two nearest neighbors  $N_1$  and  $N_2$ .<sup>4</sup>

The next steps aim to filter out any erroneous matches and minimize the incorrect (mis-)detections. Since they have lower computational complexity, they are handled by the Central Processing Unit (CPU). After matching descriptors between a pair of images, erroneous matches are discarded by applying the following rule: keep an interest point in  $O$  and its corresponding best match in  $F_i$  iff:

$$\|DistN_1\|_1 / \|DistN_2\|_1 \leq 0.8,$$

where  $\| \cdot \|_1$  is the Manhattan distance between the interest point in  $O$  and each of the calculated nearest neighbors. Additional outliers are then filtered-out by estimating the homography between  $O$  and  $F_i$  using the RANSAC algorithm [4]. If a sufficient number of pairs of descriptors remains after this geometric validation step, then the object is said to be detected in  $F_i$  and an appropriate bounding box is calculated and stored (i.e. the coordinates of the upper-left corner  $(x,y)$  and its width and height) for this frame, while otherwise the algorithm stores a bounding box of the form  $[0\ 0\ 0\ 0]$ . When the processing of the video frames is completed, a final filtering step is applied on the overall

<sup>4</sup>These GPU-based processes are realized using code included in version 2.4.3. of the OPENCV library, <http://www.opencv.org>

detection results aiming to the minimization of false positives (i.e. erroneous detections) and false negatives (i.e. erroneous misses). The latter is based on a sliding window of 21 frames and a set of temporal rules that decide on the existence or absence of the object  $O$  in the middle frame of this window.

Regarding the second strategy towards faster than real-time processing, further degradation of the needed processing time is achieved by designing and applying an efficient sampling strategy, which reduces the number of frames that have to be matched against the object of interest. The algorithm utilizes the analysis results of the shot segmentation method of [15], which can be interpreted as a matrix  $S$  where its  $i$ -th row  $S_{i,j}, j = 1, \dots, 5$  contains the information about the  $i$ -th shot of the video. Specifically,  $S_{i,1}$  and  $S_{i,2}$  are the shot boundaries, i.e. the indices of the starting and ending frames of the shot and  $S_{i,3}, S_{i,4}, S_{i,5}$  are the indices of three representative key-frames of this shot. By using this data, the algorithm initially tries to match the object  $O$  with the 5 frames of the  $i$ -th shot that are identified in matrix  $S$  (i.e.  $S_{i,j}, j = 1, \dots, 5$ ), and only if the matching is successful for at least one of these frames it proceeds with comparing  $O$  against all the frames of that shot. It then continues with the key-frames of the next shot, until all shots have been checked. Following this approach the algorithm analyses in full only the parts (i.e. the shots) of the video where the object appears (being visible in at least one of the key-frames of these shots) and quickly rejects all remaining parts by performing a small number of comparisons, thus leading to a remarkable acceleration of the overall procedure.

More details on our object re-detection approach can be found in [1].

Our experiments on the object re-detection technique, using objects and videos from the LinkedTV dataset, show that the algorithm achieves 99.9% Precision and 87.2% Recall scores, identifying successfully the object for a range of different scales and orientations and when it is partially visible or partially occluded (see for example Fig. 1), while the needed processing time using a modest modern PC (e.g. having an Intel i7 processor, 8GB RAM memory and a CUDA-enabled GPU) is about 10% of the video's actual duration, thus making the implemented technique an efficient tool for fast and accurate instance-based annotation of videos within the LinkedTV analysis pipeline.

## 5 Towards Localized Person Identification

In the LinkedTV scenarios, object re-detection is one of the most important techniques in the documentary scenario, while person identification is far more crucial for the news show scenario. In [13], we described the challenge of obtaining a reasonable person identification database for local context. To overcome this, we exploit the fact that for most news show, banner information is shown whenever a specific person is interviewed. Manually checking videos of one show over the course of two months, it seems reasonable to assume that (a) the banner is only shown when the person is speaking, and (b) mostly – but not always – only this single person is seen in these shots. We can thus use this information for speaker identification and face recognition (cf. Figure 2 for a graphical representation of this work flow).



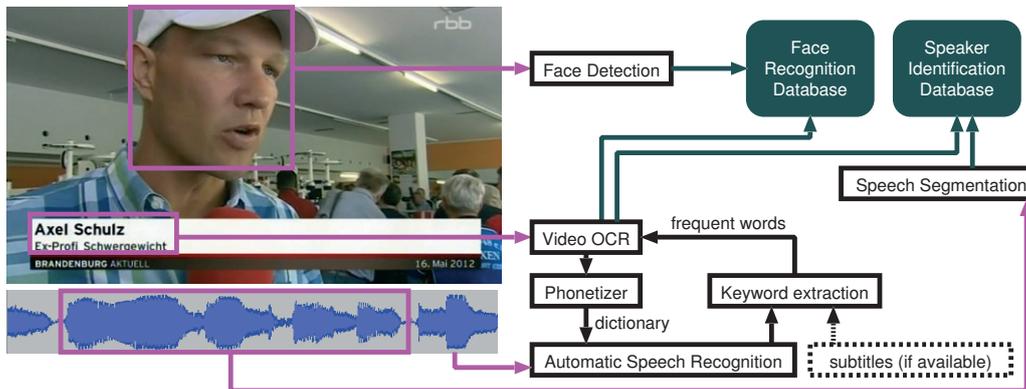
**Figure 1:** Object of interest (top row) and in green bounding boxes the detected appearances of it, after zoom in/out (middle row) and occlusion-rotation (bottom row).

For the show “Brandenburg aktuell”<sup>5</sup>, we downloaded 50 videos over the course of two month, with each of 30 minutes length. Each show contains on average around seven interviewed persons with their name contained in the banner. Since the banner will be always at a certain position, we employ a simple yet effective Optical Character Recognition (OCR) heuristic using tesseract [11]: we check each screen-shot made every half second and decide that a name is found whenever the Levenshtein distance over three consecutive screen-shots is below 2. On manually annotated 137 screen-shots, the character accuracy is at convenient 97.4%, which further improves to 98.4% when optimizing tesseract on the shows font, using a distinct training set of 120 screen-shots.

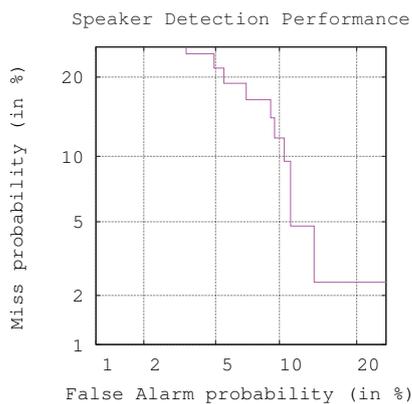
This was used as a first reasonable basis for a speaker identification (SID) database. To obtain the audio portions of a speaker in a news excerpt, the banner is time-aligned to the speaker clustering segment, and other segments which have been assigned to be the same speaker via un-supervised clustering are also aligned to the same data collection. 269 instances with banner information were detected. The length of the spoken parts for a speaker in one show varied between 6 and 112 seconds, for an average of 31 seconds. 32 speakers appeared in more than one video.

For SID, we follow the approach of [9], i.e., we make use of Gaussian Mixture Models (GMMs) using spectral energies over mel-filters, cepstral coefficients and delta cepstra of range 2. An overall universal background model (UBM) is merged from gender-dependent UBMs and forms the basis for the adaptation of person-dependent SID models. For evaluation of the speaker identification, we took every speaker that appeared more than once (32 speakers total) and divided videos of the two months of video material into a 2:1 ratio for training and testing. See Figure 3 for a Detection error tradeoff (DET) curve. The Equal Error Rate (EER) at 10.0% is reasonably close to the performance of German parliament speaker recognition (at 8.5% EER) as presented in our previous paper [13], but with the benefit that it is now on in-domain speakers.

<sup>5</sup><http://www.rbb-online.de/brandenburgaktuell/>



**Figure 2:** Workflow for an automatically crawled person identification database, using news show banner information

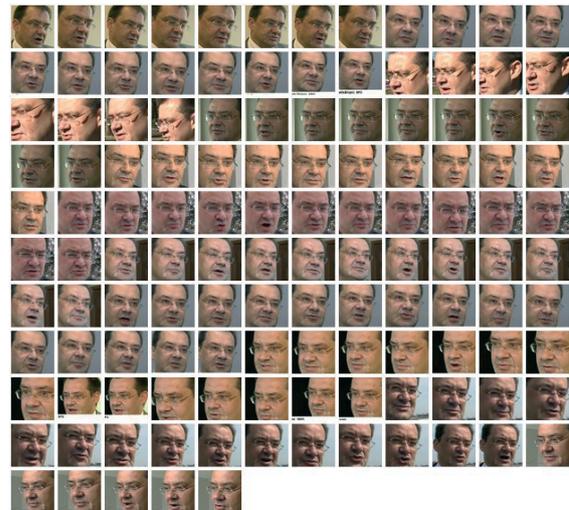


**Figure 3:** DET curve for the speaker identification experiment on RBB material.

In order to build a first database for face recognition, we applied face detection on the relevant screen-shots, using the widely used Viola-Jones detector [16], or more precisely its implementation in the OPENCV library as improved by Lienhart and Maydt [6]. Detection is combined with a skin color detector [8] for filtering out candidate regions that are not likely to be faces. Then, we link detected faces through shots using a spatio-temporal matching of faces: if two faces in adjacent frames are in a similar position, we assume we can match them. Interpolation of missing faces also relies on matching similar bounding boxes in close but none adjacent frames through a shot. This process enables to smooth the tracking results and to reject some false positive (when a track is too short, it is considered as a false alarm). See Figure 4 for the face detection results of one local politician that has been automatically harvested from the videos (he appeared in 11 different instances). These entries will serve as a database for face recognition in future work.

## 6 Hyperlinking

While in the previous sections we have focused on raw information extraction, this sections explains how the outcome from the visual and audio analysis performed over the video resources is transformed into a semantic graph representation, which enhances the way the information is exploited in a television scenario. The resultant Resource Description Framework (RDF) can be easier completed with other descriptions in external resources, better link-able with other content, and becomes available in a



**Figure 4:** Crawled face shots from a local German politician, Jörg Vogelsänger.

Web compliant format that makes possible to bring hypermedia experience to the TV field.

**RDF conversion** In a first step, the aggregated information is converted into RDF and represented according to the LinkedTV Ontology<sup>6</sup>. The REST API service *tv2rdf*<sup>7</sup> performs this operation. The video content is structured in parts with different degrees of granularity, by using the Media Fragments URI 1.0 specification. Those instances of the *MediaFragment* class are the anchors where the entities will be attached in the following serialization step. The media fragment generation introduces a very important level of abstraction that opens many possibilities when annotating certain parts of the analyzed videos and makes possible to associate to fragments with other metadata with temporal references. The underlying model also relies on other established and well known ontologies like the The Open Annotation Core Data Model<sup>8</sup>, the Ontology for Media Resources<sup>9</sup> or the NERD ontology. Table 2 shows some statistics about the number of MediaFragment's created for a 55 minutes chapter of the show *Tussen Kunst* in which five spatial object have been detected.

Below is the Turtle serialization of a spatial object detected in the same *Tussen Kunst* en Kitsch video, accord-

<sup>6</sup><http://semantics.eurecom.fr/linkedtv>

<sup>7</sup><http://linkedtv.eurecom.fr/tv2rdf>

<sup>8</sup><http://www.openannotation.org/spec/core>

<sup>9</sup><http://www.w3.org/ns/ma-ont>

**Table 2:** Number of MediaFragment's generated during the RDF serialization process of a Tussen Kunst en Kitsch episode.

| Serialized Item | N MediaFragment's |
|-----------------|-------------------|
| Shots&Concepts  | 448               |
| Subtitles       | 801               |
| Bounding Boxes  | 4260              |
| Spatial Objects | 5                 |

ing to the LinkedTV ontology. As every object can appear various times during the show, a different MediaFragment instance is created for each appearance. The temporal references are encoded using the NinSuna Ontology<sup>10</sup>.

```
<http://data.linkedtv.eu/spatial_object/faedb8be-8de4-4e33-8d8c-26b35629785e>
  a          linkedtv:SpatialObject ;
  rdfs:label "CERTH_Object-5" .

<http://data.linkedtv.eu/media/e2899e7f-67c1-4a08-9146-5a205f6de457#t=1492.64,1504.88>
  a          nsa:TemporalFragment , ma:MediaFragment
  ;
  nsa:temporalEnd "1504.88"^^xsd:float ;
  nsa:temporalStart "1492.64"^^xsd:float ;
  nsa:temporalUnit "npt" ;
  ma:isFragmentOf <http://data.linkedtv.eu/media/e2899e7f-67c1-4a08-9146-5a205f6de457> .
```

At the same time every appearance is composed of a sequence of square bounding boxes that demarcate the object position, which are also represented as a set of MediaFragments of lower duration. The spatial references are directly encoded in the URL following the Media Fragments URI specification. The fact that one spatial MediaFragment belongs to the entire scope of a particular object is specified through the property MA:ISFRAGMENTOF.

Finally, broadcasters normally make available meta-data related to their TV content, which is also included in the RDF graph during the serialization process. This data normally contains general information about the video such as: title, description, tags, channel, category, duration, language, creation date, publication date, view, comment, and subtitles. The service tv2rdf implements the serialization of TVAnytime<sup>11</sup> files into RDF by using the Programmes Ontology.<sup>12</sup>

**Name Entity Extraction** After the RDF graph is built, certain nodes are populated with extra anchors to the Link of Data Cloud. Named entity extraction processes are performed over the transcripts of the TV content that are available in the subtitle files from the providers or in the ASR results. The tv2rdf REST service launches this task by relying on the *NERD Client*, which is part of the NERD<sup>13</sup> framework. A multilingual entity extraction is performed over the video transcript and the output result is a collection of entities related to each video. Hence, the entities are classified using the core NERD Ontology v0.5<sup>14</sup> and serialized in JSON format, so they have to be translated by tv2rdf into a RDF representation and attached to the right MediaFragment.

During serialization, both Dublin Core<sup>15</sup> and LinkedTV properties are used in order to specify the entity label, con-

<sup>10</sup><http://multimedialab.elis.ugent.be/organon/ontologies/ninsuna>

<sup>11</sup><http://tech.ebu.ch/tvanytime>

<sup>12</sup><http://purl.org/ontology/po>

<sup>13</sup><http://nerd.eurecom.fr/>

<sup>14</sup><http://nerd.eurecom.fr/ontology/nerd-v0.5.n3>

<sup>15</sup><http://dublincore.org/documents/2012/06/14/dces>

**Table 3:** Number of entities per type extracted from the Tussen Kunst en Kitsch video.

| NERD type    | Entities |
|--------------|----------|
| Person       | 37       |
| Location     | 46       |
| Product      | 3        |
| Organization | 30       |
| Thing        | 22       |

fidence and relevance scores, name of the extractor used in the named entity recognition process, entity type and disambiguation URI (in this case, a resource in DBpedia). Below there is an example of the Turtle serialization for the entity Jan Sluijters spotted in the same episode of Tussen Kunst.

```
<http://data.linkedtv.eu/entity/9f5f6bc5-fa3a-4de1-b298-2ef364eab29e>
  a          nerd:Person , linkedtv:Entity ;
  rdfs:label "Jan Sluijters" ;
  linkedtv:hasConfidence "0.5"^^xsd:float ;
  linkedtv:hasRelevance "0.5"^^xsd:float ;
  dc:identifier "77929" ;
  dc:source "semitags" ;
  dc:type "artist" ;
  owl:sameAs <dbpedia.org/resource/Jan_Sluijters> .
```

For having a better understanding of the number of entities extracted in the example video, the Table 3 presents some statistics about the extracted entities per NERD type.

**Enrichment** In a third step, the named entities already incorporated into the data graph are used for triggering processes to retrieve additional media content in the Web. The logic for accessing the external datasets where this information can be collected is implemented inside the LinkedTV REST service MediaCollector.<sup>16</sup> It is here where the original RDF graph is enriched with extra content that illustrates and completes what is shown in the seed video.

MediaCollector gets as input the label of the entities spotted by NERD over the transcript, and provides as result a list of media resources (photos and videos) grouped by source. For this research work the considered sources are selected from a white list defined by the content providers, due to the editorially controlled nature of the scenario. Those sources include mainly corporative Web Sites and some particular video channels in Youtube that have been previously checked by experts. When serializing the information, every item returned by MediaCollector is represented as a new MediaResource instance according to the Ontology for Media Resources. The entity used as input in the media discovery process is linked to the retrieved items through an OA:ANNOTATION instance, as proposed in the Open Annotation Ontology.

**Data Exploitation** Once the metadata about a particular content has been gathered, serialized into RDF, and inter-linked with other resources in the Web, it is ready to be used in the subsequent consumption phases like the editorial review or data display. The creation of a MediaFragments hierarchy with different levels of granularity provides a very flexible model for (1) easily incorporate new data describing the media resource and (2) allowing different interpretations of the available information depending on the final user and the particular context.

For example, the spatial objects detected and named entities can be aligned for obtaining new insights about

<sup>16</sup><http://linkedtv.eurecom.fr/api/mediacollector/>



**Figure 5:** List of media items retrieved from MediaCollector service for the search term "Jan Sluijters".

what is happening in the video. The upper left image in Figure 5 illustrates a painting, detected by the object re-detection algorithm and highlighted with a green bounding box, that appears in the Tussen Kunst en Kitsch show, between the 1492nd and 1504th second. Looking for information attached to temporarily similar MediaFragments in the model, there is an entity about the artist "Jan Sluijters" that is mentioned from the second 1495 to 1502. So it is possible to conclude that this person is the author of the painting or at least is strongly related with it. Similar deductions can be done by relying in other items in the model like keywords and LSCOM concepts. The remaining images in Figure 5 correspond to some of the media items retrieved for the entity "Jan Sluijters". Most of them are about the relevant paintings created by this author.

Finally, as the resulting RDF graph is stored in an standard and Web compliant way, it can be used not only to be visualized in the LinkedTV platform but also for being referenced and consumed by other similar systems consuming television information. This way it is possible to implement solutions that bring innovative hyper-media experiences to the TV scenario.

## 7 Conclusion

In this paper, we presented recent improvements and strategies in the LinkedTV work-flow.

Generally speaking, the main challenge for harvesting semantically rich information from raw video input in sufficient quality is a matter of domain adaptation. We have shown ways to adopt the free decoder parameters to the new domain, requiring only a little amount of training data. Further, we presented improvements in the object re-detection algorithm which allows a fast and reliable detection and tracking of interesting objects. In order to obtain knowledge about the faces and voices of local people, we opted to crawl local news shows which usually contain banner information. We have shown that it is possible to build up a reasonable database fast, using well-established technology. Last, we showed how all this data is incorporated into the LinkedTV hyperlinking layer.

While there are many challenges up ahead, a first breakthrough from a collection of raw analysis data towards a semantically enriched linking has been established. As a next step, we focus on (1) multi-modal topic segmentation for link expiry estimation, and (2) multi-modal person identification, combining the knowledge from face recognition and speaker identification.

**Acknowledgments** This work has been funded by the European Community's Seventh Framework Programme (FP7-ICT) under grant agreement n° 287911 LinkedTV. LinkedTV would

like to thank the AVRO for allowing us to re-use Tussen Kunst & Kitsch for our research.

## References

- [1] Apostolidis, E., Mezaris, V., and Kompatsiaris, I. (2013). Fast object re-detection and localization in video for spatio-temporal fragment creation. In *Proc. MMIX Workshop at IEEE Int. Conf. on Multimedia and Expo (ICME)*, San Jose, CA, USA.
- [2] Bay, H., Ess, A., Tuytelaars, T., and Van Gool, L. (2008). Speeded-up robust features (surf). *Comput. Vis. Image Underst.*, 110(3):346–359.
- [3] El Hannani, A. and Hain, T. (2010). Automatic optimization of speech decoder parameters. *Signal Processing Letters, IEEE*, 17(1):95–98.
- [4] Fischler, M. A. and Bolles, R. C. (1981). Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395.
- [5] Kacur, J. and Korosi, J. (2007). An accuracy optimization of a dialog asr system utilizing evolutionary strategies. In *Proc. Image and Signal Processing and Analysis*, pages 180–184. IEEE.
- [6] Lienhart, R. and Maydt, J. (2002). An extended set of Haar-like features for rapid object detection. In *Proc. Image Processing*, volume 1, pages I–900 – I–903 vol.1.
- [7] Mak, B. and Ko, T. (2009). Automatic estimation of decoding parameters using large-margin iterative linear programming. In *Proc. Interspeech*, pages 1219–1222.
- [8] Rahim, N. A. A., Kit, C. W., and See, J. (2006). Rgb-h-cbcr skin colour model for human face detection. In *MMU International Symposium on Information and Communications Technologies (M2USIC)*, Petaling Jaya, Malaysia.
- [9] Reynolds, D., Quatieri, T., and Dunn, R. (2000). Speaker Verification Using Adapted Gaussian Mixture Models. *Digital Signal Processing*, 10:19–41.
- [10] Sidiropoulos, P., Mezaris, V., and Kompatsiaris, I. (2013). Enhancing video concept detection with the use of tomographs. In *Proc. of the IEEE International Conference on Image Processing (ICIP)*, Melbourne, Australia.
- [11] Smith, R. (2007). An Overview of the Tesseract OCR Engine. In *ICDAR '07: Proceedings of the Ninth International Conference on Document Analysis and Recognition (ICDAR 2007) Vol 2*, pages 629–633, Washington, DC, USA. IEEE Computer Society.
- [12] Spall, J. C. (1992). Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Transactions on Automatic Control*, 37:3.
- [13] Stein, D., Apostolidis, E., Mezaris, V., de Abreu Pereira, N., Müller, J., Sahuguet, M., Huet, B., and Lašek, I. (2012). Enrichment of News Show Videos with Multimodal Semi-Automatic Analysis. In *Proc. NEM-Summit 2012*, pages 1–6, Istanbul, Turkey.
- [14] Stein, D., Schwenninger, J., and Stadtschnitzer, M. (2013). Improved speed and quality for automatic speech recognition using simultaneous perturbation stochastic approximation. In *Proc. Interspeech*, pages 1–4, Lyon, France. to appear.
- [15] Tsamoura, E., Mezaris, V., and Kompatsiaris, I. (2008). Gradual transition detection using color coherence and other criteria in a video shot meta-segmentation framework. In *Proc. Image Processing*, pages 45–48.
- [16] Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Proc. Computer Vision and Pattern Recognition, CVPR*, volume 1, pages I–511 – I–518 vol.1.

# Think Before You Link — Meeting Content Constraints when Linking Television to the Web

Daniel Stein<sup>1</sup>, Stefan Eickeler<sup>1</sup>, Rolf Bardeli<sup>1</sup>, Evlampios Apostolidis<sup>2</sup>, Vasileios Mezaris<sup>2</sup>, Meinard Müller<sup>3</sup>

<sup>1</sup>Fraunhofer Institute IAIS, Sankt Augustin, Germany <sup>2</sup>Information Technologies Institute CERTH, Thessaloniki, Greece <sup>3</sup>International Audio Laboratories Erlangen, Germany

**Abstract:** With a constantly rising demand for interactive videos, automatic enrichment of static videos with web links offers seemingly endless possibilities. Given the content that can be found on the web, however, this can be a rather mixed blessing. In this paper, we investigate web sites with contents of extreme physical violence, political extremism and self-harm and argue that their topic can be quite close to seed videos such as, e.g., news shows. We discuss ways to detect such problematic content and present first solutions to specific challenges.

**Keywords:** LinkedTV, OCR, Cover Song Identification, Concept Detection

## 1 Introduction

Many recent projects focus on automatic enrichment of linear (i.e., static) videos with links to other text resources, images and videos. While offering seemingly endless possibilities, multimedia content that is automatically interlinked with various kinds of data can at best be partially checked for its new content. Thus, special care has to be taken that this data meets legal and moral constraints. This problem is not only inherited by video interlinking but also applies to other fields like, e.g., advertisement placement, which is why some research in this area already exists. However, the main focus of existing applications is the identification of pornographic content, a topic where simple keywords can already identify a large percentage of possible web sites.

The aim of this paper is two-fold. First, we want to give an overview of types of problematic web sites featuring other content than pornography, namely (a) depiction of extreme physical violence, (b) political extremism, and (c) self-harm, e.g., self-cutting and anorexia. Second, we identify challenges for classification algorithms arising from the nature of this material and conduct a series of proof-of-concept and large-scale experiments to evaluate their performance.

This paper is structured as follows. First, we describe the “Linking Television to the Web” (LinkedTV)<sup>1</sup> project and describe why each of the three topics inherit relevant arguments against unconstrained linking (Section 2). We proceed by reviewing related work with a focus on content constraint identification (Section 3). Then, we describe relevant content and indicate challenges for automatic classification techniques (Section 4). Then, we offer proof-of-concepts and thorough experiments on the following classification tasks that have been identified to be potential challenges: audio fingerprints on violent movies (Section 5), song identification on right-extremistic rock records (Section 6), colored text localization on self-harm photos (Section 7), and concept detection on self-injury

images (Section 8). Finally, we conclude this paper with a summary (Section 9).

## 2 LinkedTV

The aim of LinkedTV is to provide an interactive multimedia service for non-professional end-users. To achieve this, linear videos are analyzed by various (semi-)automatic methods, both on the acoustic and the visual level. The raw data obtained is then used to interlink the videos with other multimedia information, using knowledge acquired from, e.g., web mining. The enriched videos are finally shown to the end-user as an interactive video with many links to further web content. Especially for news, which forms the basis of one of LinkedTV’s scenarios, a fully automated process can lead to undesirable content, as the following examples from our seed data illustrate:

- A critical report about the former lawyer Horst Mahler, who has been imprisoned several times for right-wing utterances. The Wikipedia page of Horst Mahler links to uncommented speeches given by Mahler, the first 15 hits in one of the largest online video sites feature uncommented interviews where he states his beliefs – among them the denial of the Holocaust, a statement forbidden in many countries, including Germany.
- A report on anorexia nervosa, which follows the story of a young girl who almost died from undernourishment. In the interviews, she mentions several terms specific to the Pro-Ana movement, which glorifies emaciation. Searching these terms results in few educational web sites but for the most part Pro-Ana blogs and web sites which are full of “trigger” images, i.e., depicting extremely anorectic women (in the Pro-Ana movement, these are often called “thinspos”, a portmanteau of “thin” and “inspiration”).
- After incidents where a man was beaten to death close to a Berlin railway station, a state secretary discusses with the moderator whether there is a trend towards brutalization in our society, the role of computer games, and the phenomenon of “happy slapping” (where somebody is battered, sometimes even worse, in front of a camera). Again, these search terms in an online video portal produce only few educational and critical reports (and never as first results) but mostly actual instances of battering or extremely violent video games which have an age limit “R”<sup>2</sup> or higher.

## 3 Related Work

While there is a vast scientific community focused on knowledge extraction from multimedia objects, little scientific attention is spent towards recognizing inappropriate material on the web, especially for non-textual data.

<sup>1</sup> <http://www.linkedtv.eu>

<sup>2</sup> cf. [www.filmratings.com](http://www.filmratings.com)

There has been specific interest in detecting pornography (e.g., [4, 9, 11]), mainly for filtering search engine results. A special case is given by child pornography, where there is some effort to develop automatic means for detecting and thus support the fight against this crime [1]. Apart from that, the main investigations into detecting material problematic content have been seen in the context of the TRECVID and MediaEval evaluation campaigns, namely the category *physical violence* in TRECVID's feature extraction challenge [17] and the Violent Scenes Detection Task in MediaEval [6]. Beyond these specific domain, there is of course a lot of work in the multimedia analysis community giving a basis for supporting it.

## 4 Problematic content: overview and possible techniques

It is impossible to generalize over all relevant content to be found on such heterogeneous topics as violence, extremism and self-harm. In order to identify promising analysis techniques, we have asked a German bureau responsible for child protection in the Internet to compile lists on each of them, with special care for representative samples.<sup>3</sup> As base data, we manually scanned 3 000 web sites (1 000 per topic) with material not suitable for children of age 17 or below. Some web pages were extremely disturbing even for adults, e.g., featuring real suicides or decapitation, which required security measures such as psychological supervision and restricted data access when handling the data.

In this section, we look for common patterns that can be captured with classification algorithms with the goal of raising warning flags to an editor who manually checks outgoing links for their appropriateness. We make two assumptions: first, the interlinking is targeted for multimedia objects, i.e., audio, image or video, and second, we have no prior knowledge of its content via the surrounding web site or incoming links.

### 4.1 Physical Violence

Web sites with extreme violence typically feature videos or images. The two largest categories are filmed/fotographed content and violent computer games.

**Filmed/Fotographed Content.** Web sites containing violence images and videos are either (a) commercial web sites of horror movies, (b) shock sites that have the sole purpose of disturbing the viewer or (c) sites that want to provoke emotional reaction for political purposes (e.g., news sites against torture, or anti-abortion organizations). The nature of the violence depicted is very heterogeneous and ranges from small quarries to decapitation or other forms of murder (both authentic and fictitious). Fictitious movies are mostly from the genre of horror movies, which range from monster hunts to sexually-sadistic revenge movies. "Best-of" compilations, trailers, single scenes or self-made commentaries are frequent. These often share a logo/url/watermark of a dedicated collection forum/blog.

**Violence Computer Games.** All surveyed material containing violence games are shooters, with either first-person view or third-person view over the shoulder. For

first-person games, the lower part of the screen typically shows the selected weapon, either military (e.g., machine gun), futuristic (e.g., laser pistol) or rough-and-ready (e.g. a simple stick). Some game videos are official trailer, but most are self-made fan videos ("let's play").

**Conclusion.** Among the three topics violence, extremism and self-harm, violence seems to be the topic with the highest heterogeneity of its material, which poses a huge challenge for analysis techniques. Extremely brutal videos such as decapitation are not as frequent but produce a huge number of re-mixes and samples. In our first approach trying to address the problem of violence detection in videos, we tested an algorithm that performs audio fingerprinting on a horror movie in Section 5.

### 4.2 Extremism

While there are many forms of political and religious extremism, the sites given to us were mostly from right-wing extremism, which is why we will focus on this content in the following – by no means a belittlement of other forms of dangerous extremism. Note that some videos from religious extremism for example were rather aligned to the category "violence" because they featured decapitation.

The majority of the web content in right-wing web sites falls into the categories shop sites, propaganda videos and music videos.

**Shop Sites.** The web pages in this category offer right-wing souvenir articles such as apparels and media articles. To attract attention, these sites feature a lot of obvious keywords and palpable images.

**Propaganda Videos.** The videos shown in the resort of propaganda were mostly focused on history revisionist point of views. The language was either in German or English, sometimes both languages were spoken simultaneously. The quality of the videos is, especially for older looking videos presumably digitized from old video cassettes or for war footage, rather mixed. Filming at historic sites, e.g., concentration camps in Germany, appear to have taken place without proper filming equipment, and the speaking person has no dedicated microphone. The name of the people speaking is sometimes shown via banners. Sometimes, extremism symbols like the German Swastika are shown.

**Music Videos.** The extremism music videos that we watched were seldom professional videos like those shown in television, but mostly just the audio plus some images of the CD cover or some band photography. A notable exception were live videos from concerts. Most images contained extremism symbols, such as the triskele, the swastika, or logos of organizations or groups associated with extremism. While some of these objects are depicted on banners, others are sprayed in graffiti, painted on tissue or tattooed. The music itself can be practically any genre such as singer-songwriter or instrumental, but has a strong tendency towards hard rock or metal, with the interpreters mostly shout-singing.

**Conclusion.** Based on the impression of the 1,000 right-extremism web sites surveyed, the detection of this content seems to be the most easiest in direct comparison to violence and self-harm. Many individuals and groups in

<sup>3</sup>jugendschutz.net

this area use re-occurring keywords, named entities and symbols. For music identification, however, fingerprinting alone will only cover a fraction of the material, since fan-based covers and samplers appear frequent (cf. Section 6).

### 4.3 Self-Harm

The given web pages about self-harm included the topics self-injury, eating disorders, substance abuse and suicide. The nature of the web pages ranged from distress calls to tutorials (sometimes disguised as preventional education). Substance abuse was a very broad topic and contained virtually no clues for analysis techniques, while the topic suicide only had a few examples, as most content such as forum discussions is supposedly well-hid in private areas of the web pages (which, of course, holds true for the other topics as well, but to a lesser extend). In the following we will investigate the topics Pro-Ana and Pro-Mia as well as self-injury.

**Pro-Ana/Pro-Mia.** In web sites that glorify eating disorders such as anorexia and bulimia, there are often catch phrases (e.g., “angels have no hunger”) or longer text collections (e.g., “Ana’s letter”, a fictional conversation of a young girl with an incarnate anorexia). While as text, they can be easily spotted, in videos they are often depicted in heavily stylized fonts. Moreover, the letters are often hard to read even for a human eye since often colored letters are put on top of a photo.

According to a survey in [3], 85% of Pro-Ana web sites contain thinspo images. We thus manually analyzed images with thinspos made available on a freely accessible server. In total, the server contained 84 163 images (5.3 GB). Based on this manual survey, we draw the following conclusions: the server mostly contains images that are in relation with fashion shows and catwalk models. Roughly as frequent are images from celebrities. The third largest collection of images features thin women at the beach, unknown people as well as celebrities. Most people depicted seem to have pathological body-mass-indexes, but the images without their context are not obvious eating disorder glorifications.

A second, albeit far smaller group of images, contains material that clearly is related to anorectic context: (a) images that show extreme signs of anorexia, with bones clearly visible all over the body, (b) before-after compilations, both fake and real, of people that starved, and (c) depictions of extremely adipose people (“anti-thinspos”), often accompanied by sarcastic text included in the image.

**Self-Injury.** Web sites which aim at attracting attention to persons who regularly inflict damages to themselves are, in principle, similar to Pro-Ana/Pro-Mia sites in the sense that they contain many pictures and catch-phrases on the topic of self-injury. Often, the difference to eating disorder websites include: (a) self-shots, which do not appear as much in Pro-Ana/Pro-Mia sites where disdain of the own body is often part of the disorder, (b) more general cries for help (“Why won’t anybody listen?”) that do not exclusively circle around the disorder itself. However, the images are mostly self-explanatory, making a specific search a helpful feature in the identification of this topic.

**Conclusion.** The web sites on self-harm are centered around few topics, but the high amount of user generated

content, often from under-age persons, creates a lot of individualized content where fingerprinting is of little help. In the following, we will take a closer look at two challenges: colored fonts on colored ground (Section 7) and high-level concept detection on a proportion of the imagery (Section 8).

## 5 Audio Fingerprints on Violence

Given a small audio fragment as query, the task of audio identification consists in identifying the particular audio recording that is the source of the fragment. In this section, we investigate whether audio fingerprints can be employed on horror movies, where spoken dialogue is only a fraction of the audio signal and noisy fight scenes are more common. For a proof-of-concept, we use the horror movie “Braindead” (1992, director: Peter Jackson) as seed video. The story follows a suburban town drifting into annihilation by some mutant virus which turns the infected into blood-thirsty monsters. The un-cut version is confiscated by law in Germany, a cut version (by 16 minutes) exists which has an age restriction of “16”.

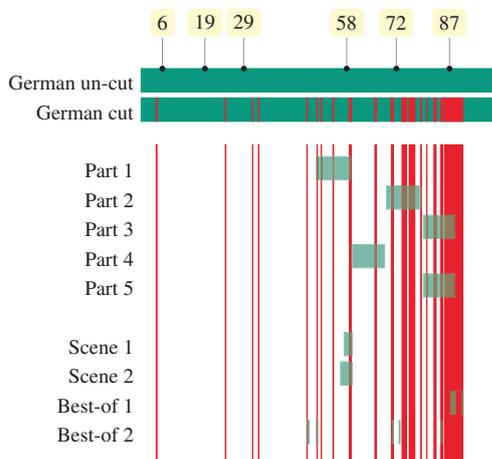
Given the un-cut version of this movie as reference and a possibly modified version of the same movie (e.g., cut version, samples only collecting the most violent scenes, different language version, ...), we want to find out which portion of the un-cut version these scenes relate to, only based on the audio signal. We employ an audio fingerprint method as described in [2].

On the whole movie, we created a fingerprint every ten seconds, resulting in 589 fingerprints total, of which 13 are discarded since they contained no audio. Of the remaining 576 fingerprints, 78.4% are detected in the German cut version, while the remaining sections most probably have been cut due to their brutal nature. Figure 1 illustrates the results: The minutes marked yellow above the top bar indicate gory scenes that are cited in the German confiscation enactment. Red sections are scenes where no fingerprints could be found in the cut version, i.e., these are deleted with respect to the un-cut version because of their violent content. All analyzed video snippets from an online video portal feature scenes of the un-cut version.

For this example, the algorithm also works on versions in other language (possibly due to the high number of fighting scenes where no-one is speaking and the audio signal remains unchanged). The English version featured 18.4% fingerprint matches, and the Spanish version even 22.6% matches.

## 6 Identification on Extremistic Songs

In this section, we look at fingerprinting performance and cover song identification of right-wing music recordings, where a large proportion is executed under bad recording conditions (bootlegs are quite frequent) and quite a few singers are not very melodic. On a data set of 523 albums and 6,631 records with a total length of  $\approx 400$  hours material, we used Echoprint [7] for a indexing and retrieval experiment. Of 6,000 songs, we indexed 3,000 of them in a Echoprint database and checked whether they were found correctly and whether the remaining 3,000 were correctly marked as unknowns. We thus obtained 86% true positives compared to 90% true negatives. These numbers are misleading, however, meaning that the performance is actually much higher, as the albums contain duplicates (one third



**Figure 1:** Fingerprints taken from the un-cut German version of “Braindead” that are identified in video snippets found in an online video portal.

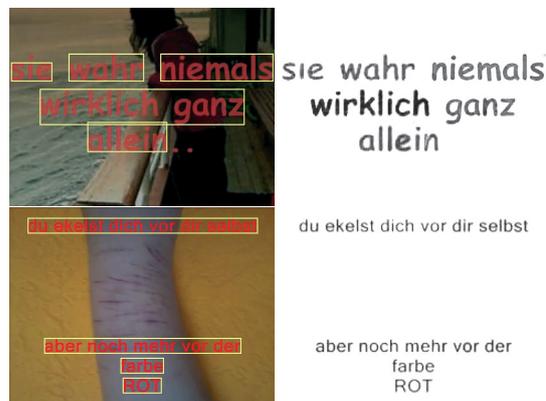
of the false positives feature the words “tribute” and “sampler” in either song title or album title, which means that they should be evaluated as true positives since the song is probably indeed already known).

Another peculiarity of this data collection was the high amount of cover songs, especially for notorious right-wing bands like “Skrewdriver”. Since fingerprinting will reject cover songs as completely different music, we proceed The goal of cover song identification is to identify different versions of the same piece of music within a database (opposed to an audio recording of a specific version as in audio identification). In the cover song scenario, one has to deal with changes in instrumentation, tempo, and tonality, as well as with more extreme variations concerning the musical structure, key, or melody [10]. This requires document-level similarity measures to globally compare entire documents.

The overall procedure crucially depends on the used feature variant, the type of score matrix, and a number of other parameters. In our implementation, we use a chroma variant supplied by [13] and apply various enhancement strategies to improve structural properties of the score matrices, see also [10, 12, 15].

For a proof-of-concept, we used the album version of the song “Hail the new dawn” from “Skrewdriver”, plus seven covers, where two cover songs are from the same band and five songs are from different bands. Then, we compiled data sets containing other songs plus their cover versions, containing (a) 112, (b) 340 and (c) roughly 2,000 songs and measured for each Skrewdriver song how close it was assigned to the songs of its own group. For each query, the result is ranked with respect to decreasing similarity (see Table 1).

Overall, one may say that current systems for cover song identification yield reasonable results as long as the versions to be detected roughly follow the harmonic progression of the original song—at least in passages that have a duration of at least 40 to 60 seconds. This is, for example, clearly the case for the “Skrewdriver” cover group. When there is no dominating harmonic content in the song and if there are lots of variations in the melody (as may be the case for punk music or hard rock), however, the usage of chroma-based audio features becomes problematic and identification systems are likely to fail.



**Figure 2:** Text localization and text extraction from screen-shots of self-cutting videos

## 7 Colored Text Localization

OCR algorithms typically require that the text portion within a video is automatically detected and separated from the background. We noted that in self-harm videos the text is often highly stylized and colored, and also placed on top of colored background pictures. State of the art text detection methods [5, 8] use the gray-scale image and cannot cope with this situation. In this section, we introduce a new text detection algorithm.

The text detection is based on color segmentation using the statistical region merging (SRM) [14]. The algorithm determines uniform colored connected components (CC), which represent the characters and other objects/parts in the image. In a subsequent step the extension of SMR for handling occlusions is used to merge the characters to words. Additionally to the color information of the original SRM, the difference of the height of the CC and the distance between the CC is used.

In order to refine the text separation for the OCR, a Gaussian model is calculated for the CC and for the background of the words. An up-scaled version of the image is used to create an up-scale gray-scale image. Each pixel is assigned a gray-scale value which is proportional to the probability of the text model given the RGB value. The last step is to use Tesseract for the recognition based on the gray-scale image and the information about the found text regions.

Figure 2 shows two example screen-shots where conventional text localization and text separation failed for conventional models, and where our method produces the correct results.

## 8 Concept Detection on Self-Injury

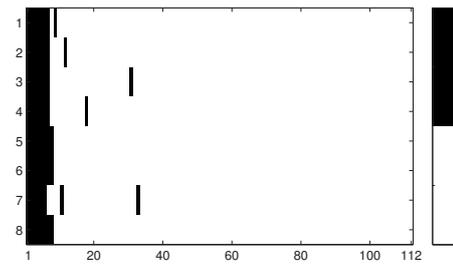
In this section, we want to investigate whether a visual concept detection algorithm can identify images of self-injury. On a non-protected server with the topic self-injury we had access to a collection of 5 383 images (625 MB), which mainly features self-cutting with sharp items, e.g., razors or scissors. The cuts appear all over the body, but mainly on the lower arm. The majority shows bleeding wounds with or without visible fat tissue, others are in the state of healing and thus show coagulated blood tissue or scars. Other images contained self amputation or blunt injuries.

As a comparison class, we reduced these images to cuts on the forearm, and used material from a web server on fore arm tattoos as counter class. See Figure 3 for some

(a) Average precision (AP, s. [15] for definition) values for the 8 queries and the three datasets

| Idx | Band          | Remark                  | AP <sub>112</sub> | AP <sub>340</sub> | AP <sub>2000</sub> |
|-----|---------------|-------------------------|-------------------|-------------------|--------------------|
| 1   | Skrewdriver   | original                | 0.982             | 0.924             | 0.883              |
| 2   | Skrewdriver   | demo tape               | 0.948             | 0.929             | 0.878              |
| 3   | Skrewdriver   | live version            | 0.891             | 0.875             | 0.860              |
| 4   | English band  | very “bawly”            | 0.916             | 0.849             | 0.777              |
| 5   | American band |                         | 1                 | 1                 | 1                  |
| 6   | Swedish band  | female singer           | 1                 | 1                 | 1                  |
| 7   | Swedish band  | melodic female singer   | 0.831             | 0.789             | 0.620              |
| 8   | British band  | quite close to original | 1                 | 1                 | 1                  |

(b) Ranking matrix for each song, showing pairwise document-level similarities between the query and all documents contained in the dataset 112



**Table 1:** Cover song identification on eight selected cover songs for the band Skrewdriver

examples from the set and the similarity of the classes. In order to increase the challenge for the classification algorithm, we further discarded tattoos in any other color than black or red. In total, 400 images for each class remained for training, and we used 67 (cutting) and 86 (tattoo) images for testing.

For classification, we make use of high-level concept detection algorithms, where we follow the approach of [16] with a large sub-set of the base detectors described there. The results for tattoo images are shown in Figure 4(b), the results for self-cuts are shown in Figure 4(a). With an assumed threshold of 0.5, 8 out of 151 images are misclassified.

## 9 Conclusion

In this paper, we elaborated on the needs to identify problematic web content in applications that rely on automatic interlinking such as the ones foreseen in the LinkedTV project. We gave a brief overview of possible contents in the areas of physical violence, extremism and self-harm, by manually scanning through thousands of web sites deemed inappropriate for children and adolescents. On the four selected challenges — (a) audio fingerprints in a horror movie, (b) right-wing cover song identification, (c) colored font OCR and (d) self-cutting concept detection — we presented our approaches and gave both proof-of-concepts and evaluations regarding their performance.

Overall, whenever interlinked content is not constrained via white-lists, we believe that outgoing links should be checked by human editors whenever the seed videos feature vast topics such as news, and that automatic analysis techniques are capable of producing warning flags which will support this task to a large degree.

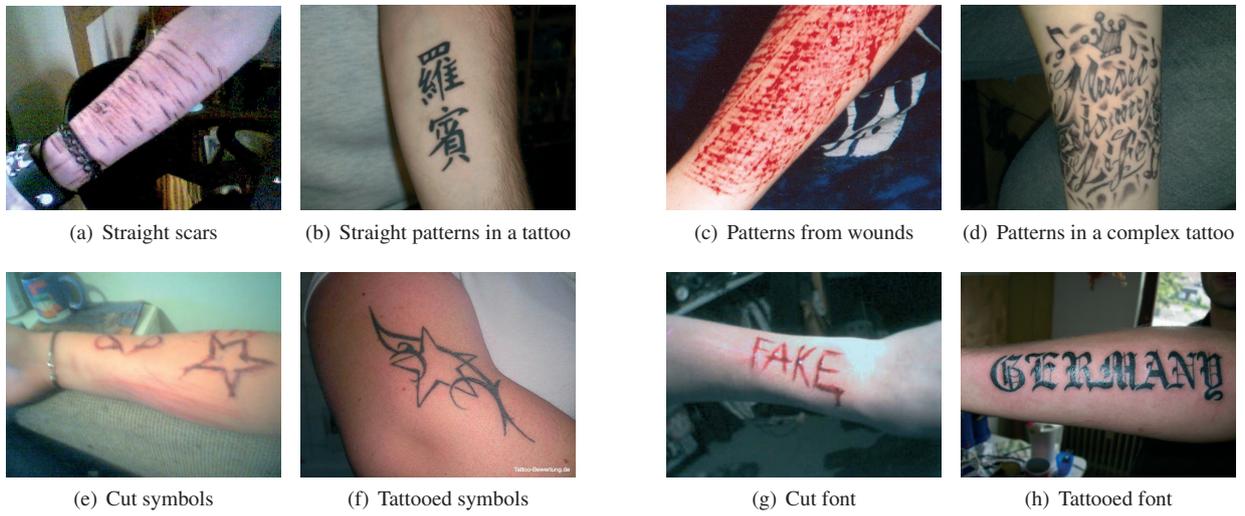
## Acknowledgments

This work has been partly funded by the European Community’s Seventh Framework Programme (FP7-ICT) under grant agreement n° 287911 LinkedTV.

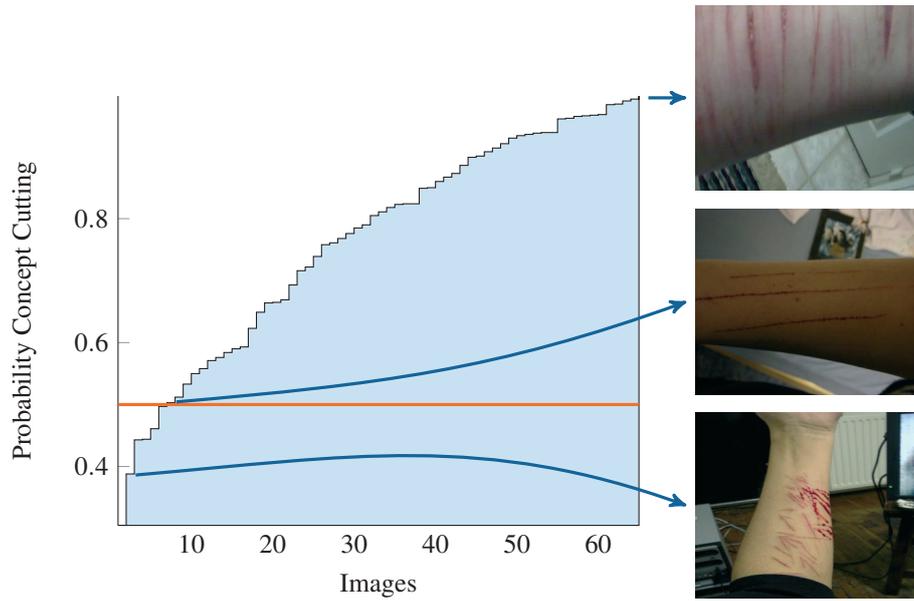
## References

[1] Aldhous, P. (2011). The digital search for victims of child pornography. *New Scientist*, 210(2807):23–24.  
 [2] Bardeli, R., Schwenninger, J., and Stein, D. (2012). Audio fingerprinting for media synchronisation and duplicate detection. In *Proc. MediaSync*, pages 1–4, Berlin, Germany.  
 [3] Borzekowski, D., Schenk, S., Wilson, J., and Peebles, R. (2012). e-Ana and e-Mia: A Content Analysis of Pro-Eating

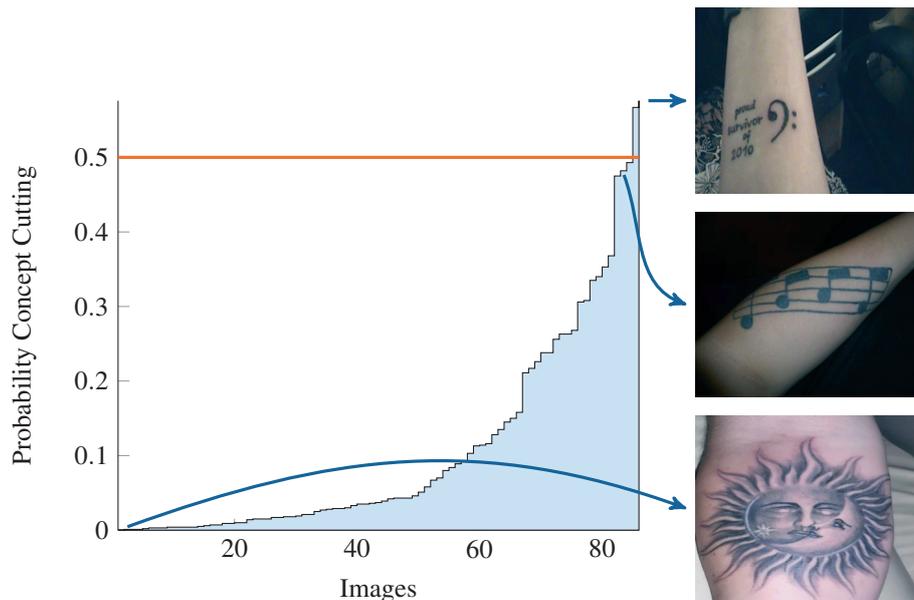
Disorder Web Sites. *American Journal of Public Health*, 100 (8):1526–1534.  
 [4] Bosson, A., Cawley, G. C., Chan, Y., and Harvey, R. (2002). Non-retrieval: Blocking pornographic images. In Lew, M. S., Sebe, N., and Eakins, J. P., editors, *Image and Video Retrieval*, volume 2383 of *Lecture Notes in Computer Science*, pages 50–60. Springer Berlin Heidelberg.  
 [5] Chen, H., Tsai, S. S., Schroth, G., Chen, D. M., Grzeszczuk, R., and Girod, B. (2011). Robust text detection in natural images with edge-enhanced maximally stable extremal regions. In *2011 IEEE International Conference on Image Processing*, Brussels.  
 [6] Demarty, C. H., Penet, C., Gravier, G., and Soleymani, M. (2012). The MediaEval 2012 Affect Task : Violent Scenes Detection. In *MediaEval 2012 Workshop*, volume 927.  
 [7] Ellis, D. P., Whitman, B., and Porter, A. (2011). Echoprint: An open music identification service. In *Proc. ISMIR*.  
 [8] Epshtein, B., Ofek, E., and Wexler, Y. (2010). Detecting text in natural scenes with stroke width transform. In *CVPR*, pages 2963–2970. IEEE.  
 [9] Fleck, M. M., Forsyth, D. A., and Bregler, C. (1996). Finding naked people. In Buxton, B. F. and Cipolla, R., editors, *ECCV (2)*, volume 1065 of *Lecture Notes in Computer Science*, pages 593–602. Springer.  
 [10] Grosche, P., Müller, M., and Serrà, J. (2012). Audio content-based music retrieval. In *Multimodal Music Processing*, volume 3, pages 157–174. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, Dagstuhl, Germany.  
 [11] Kim, M. J. and Kim, H. (2012). Audio-based objectionable content detection using discriminative transforms of time-frequency dynamics. *IEEE Transactions on Multimedia*, 14(5):1390–1400.  
 [12] Müller, M. and Clausen, M. (2007). Transposition-invariant self-similarity matrices. In *Proc. ISMIR*, pages 47–50, Vienna, Austria.  
 [13] Müller, M. and Ewert, S. (2011). Chroma Toolbox: MATLAB implementations for extracting variants of chroma-based audio features. In *Proc. ISMIR*, pages 215–220, Miami, FL, USA.  
 [14] Nock, R. and Nielsen, F. (2004). Statistical region merging. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 26(11):1452–1458.  
 [15] Serrà, J., Serra, X., and Andrzejak, R. G. (2009). Cross recurrence quantification for cover song identification. *New Journal of Physics*, 11(9):093017.  
 [16] Sidiropoulos, P., Mezaris, V., and Kompatsiaris, I. (2013). Enhancing video concept detection with the use of tomographs. In *Proc. ICIP*.  
 [17] Smeaton, A. F., Over, P., and Kraaij, W. (2009). High-Level Feature Detection from Video in TRECVID: a 5-Year Retrospective of Achievements. In *Multimedia Content Analysis, Theory and Applications*, pages 151–174. Springer Verlag, Berlin.



**Figure 3:** Comparison of pictures from the category self-injury and tattoo, both on the forearm



(a) Results of concept detection "Self-cuts", for actual cuttings on the lower arm



(b) Results of concept detection "Self-cuts", for tattoos on the lower arm

**Figure 4:** Overview for the probabilities of the concept detection "cutting vs. tattoo", with an assumed threshold of 50%.

# Social Backup and Sharing of Video using HTTP Adaptive Streaming

Hans Stokking<sup>1</sup>, Victor Klos<sup>2</sup>, Jin Jiang<sup>3</sup>, Claudio Casetti<sup>4</sup>

<sup>1,2</sup>TNO, Delft, The Netherlands, <sup>3,4</sup>Politecnico di Torino, Torino, Italy

E-mail: <sup>1</sup>hans.stokking@tno.nl, <sup>2</sup>victor.klos@tno.nl, <sup>3</sup>jin.jiang@polito.it, <sup>4</sup>claudio.casetti@polito.it

**Abstract:** This paper is on social backup, sharing and remote access of video using HTTP Adaptive Streaming. A social backup is a backup at the location, and thus on the equipment, of (close) friends and family. Backups are created at friends' locations, matching the hosting user's interest with the content and taking into account the available bandwidth between the respective locations. A backup of high-quality video content can be done in segments. These segments allow, once distributed across various locations, remote access in a streaming fashion. This allows the bundling of upload speeds of various locations, making available enough bandwidth for streaming the content. Even then, bandwidth bottlenecks may still exist. Our prototype implementation shows two ways of dealing with this limitation. One is to request additional bandwidth from your friends, assuming they are willing to give you priority over other bandwidth usage. The other is to stream in lower quality, making use of the adaptive part of HTTP adaptive streaming.

**Keywords:** backup, sharing, caching, social, P2P, HTTP Adaptive Streaming, federation, home gateway

## 1 INTRODUCTION

Nowadays, people create, share and consume content in a multitude of ways. Using their mobile devices, people take photos and shoot videos, and post them directly online to share with their friends, family and often with larger groups in their social network. Some people share everything with everyone, while others share specific items with specific friends or family members.

The FP7 FIGARO project focusses, among other things, on secure distributed backup, sharing and remote access to content, using people's social network. To enable this, the project has created an architecture of federated home gateways of the users. This federation allows a user's home gateway to make backups at the gateways of friends and family. These backups can be 'social' backups, to be used for the purposes of sharing the content with friends and enabling access to content while on the move. This allows for backups to be uploaded to the network once, and re-use these uploads for sharing purposes with others.

If video content is of high quality, the bandwidth demands for streaming such content will be (much) higher than the average uplink of a single home gateway. Therefore, for the purposes of sharing of and remote access to content, the backup of the (video) content is segmented into small parts. This allows for

streaming of high-quality (HD) video, using the uplink bandwidth of multiple friends' gateways at the same time. Each gateway can then stream certain segments, to the extent the uplink bandwidth permits.

As a setting for this paper, FIGARO proposes an evolvable future Internet architecture based on gateway-oriented federation of residential networks. The residential gateway has a central role in the FIGARO vision of the future Internet. It interconnects the residential network with the Internet and is responsible for aggregating a multitude of devices and services within the residential network. In FIGARO, residential gateways undertake the federator role, internally as well as externally. Figure 1 shows residential networks connected at the edge of the Internet and illustrates a simplified view including the two types of residential network federations. The upper part illustrates external federation interconnecting multiple gateways to form a cooperative overlay across residential networks. This federation enables further collaboration to offer added value in terms of, for example, access and sharing of content, storage and network capacity. The right-part of the figure shows the internal federation within a residential network.

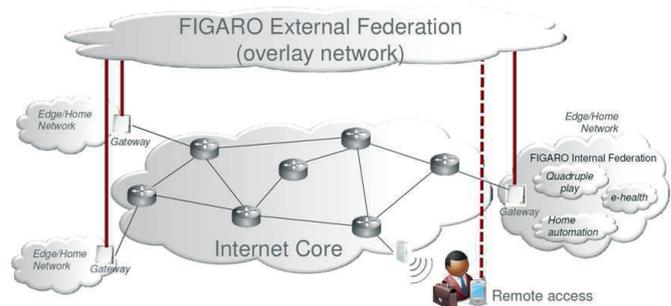


Figure 1 Overview of the FIGARO environment

This paper will start with describing related work. After that, a use case is introduced upon which the work is based. Next, the FIGARO architecture is described, and our implementation of the use case is introduced. The paper will end with our conclusions, including our lessons learned from our validation through implementation. The main focus of our work is in combining segmented backup and HTTP adaptive streaming.

## 2 RELATED WORK

Related work is in the area of P2P with helpers, of P2P backup and of social network based systems for such services, and of streaming applications using the social network.

Tribler [1] first introduced the concept of 'helpers' in a peer-to-peer (P2P) environment. Using BitTorrent as a P2P

network, they created a new Tribler client. This new client has as one of its features the concept of cooperative downloading. Using this concept, the Tribler client can request the help of friends of the user for improving a download. Normally, BitTorrent download speed is limited by the upload speed of a user, because of the tit-for-tat mechanism. With Tribler, friends can also initiate a download of segments, and deliver those segments to you without needing segments in return. This bypasses the tit-for-tat mechanism, and thus can increase your download speed.

Friendstore [2] describes a backup system using the social graph. Earlier P2P backup solutions do make use of other peoples storage, but only in an anonymous fashion. This does provide only little confidence in the reliability of the backup. With Friendstore, users can make backups at trusted nodes owned by friends or colleagues. The Friendstore system deals with matters like freeriding through the social network. Disk usage is monitored and users are informed if others are unwilling to offer storage in return, allowing users to act on this.

Earlier work in Figaro [3][4] also investigated the possibilities of combining backup with sharing, focussing on the optimal placement in terms of user interest and bandwidth availability. A user's files can be backed up to other users' gateways based on shared interest or available bandwidth, or both. It is shown that placement according to both is optimal. Also, when users' quota available for backup fill up, a replacement algorithm is run to regularly optimize the placement of content, increasing the average benefit of the backups.

Others work on using social relationships to improve P2P-based video streaming. [5] uses the social relationships between users to improve the P2P performance of video streaming. The main insight offered is that users mostly watch video's from their direct friends. This insight is used to construct a very effective prefetching algorithm, using the knowledge of the social network to push first segments of a video to direct friends. [6] creates a trust model based on social network workings. The trust model is based on the ratio between upload and bandwidth, with high ratio's gaining a user trust while low ratio's losing trust. This trust model deals mainly with the freeriding problem, using a trust threshold to exclude users from the P2P swarm.

Finally, our earlier work is described in [7]. This paper focusses on some security aspects of combing social backup and sharing. Friends are divided in inner-circle friends, not needing encryption when backing up content segments at their location, and outer-circle friends, which do require encryption when backing up content segments at their location. The paper offers a novel way of combining content segments in the encrypted domain, enabling the offloading of processing to outer-circle friends.

The solution presented in this paper is different, because it combines segmented backup capabilities with real-time, multi-source streaming. The notion of sharing in this paper is not the sharing in the traditional P2P sense of offering some file to all users of the P2P network. Instead, here sharing is an

active deed by a user to share some content with some specific friend. Further, real-time multi-source streaming is accomplished by using HTTP adaptive streaming (HAS), see e.g. [8].

### 3 USE CASE AND APPROACH FOR STREAMING

Meet Alice, fond of photography and filming. She has a large collection of home media. Photos and home movies occupy some 700 GB on her hard disk, and the collection grows each week. This data is very valuable to her, so she has made backups. Using her Figaro-enabled home gateway, this backup also finds its way to the homes of other people in her social vicinity where it is stored in an encrypted form.

One evening, Alice has dinner with Bob and his family. They have a pleasant evening, and talk about the holiday season. Alice tells them about her last holiday in which she went parasailing and how cool that was. Also, she had a 1080p helmet actioncam at the time and she finished a movie of that brilliant active trip. She grabs her phone and shares the movie with Bob. He is notified and sends the movie to his TV. There it is played—in full HD quality—and they all love it. Afterwards they all compliment her on her obvious talent for movie directing and on her courage.

The available upload bitrate from Alice's own home gateway is insufficient to provide the complete movie in real time. By using HAS streaming from multiple sources – being the Figaro-enabled home gateways of Alice's friends – the total amount of bandwidth suddenly is sufficient to provide a maximum quality movie experience on the big screen in Bob's living room.

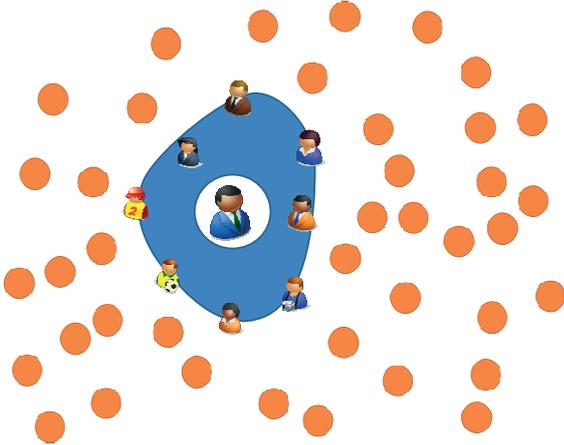
This use case contains a number of aspects, on which the approach for our implementation is based.

Backup provides durability through duplication, while caching facilitates the availability of content. The mechanisms of these functions can be combined to establish network efficiency when both doing backups and sharing content in a group of close friends. E.g. when backing up your family pictures, you can create a backup at your family's gateways and at the same time share the content with your family. The content is then transmitted over the network only once, while serving the two purposes of both backup and sharing, including remote access. This backup is comparable to work presented in [2]. Social backup is different in the sense that it is based on close relationships, e.g. family and good friends, and that the knowledge of friends' interests is taken into account when selecting the remote gateway to cache a specific content type. We envision having offline contact with these relations, and agreeing on the use of the others' bandwidth and storage. This usage does not have to be reciprocal. A main insight here is that people you trust and know, are willing to help you, and this circumvents the idea of freeriding.

Real-time streaming, such as described in the use case, in a P2P like fashion always has to deal with bandwidth bottlenecks. A single user's upload speed is limited, which is why we present a multi-source approach to combine multiple

upload speeds in a P2P like fashion. Even then, because users are also using the same bandwidth for other purposes, bandwidth may still be limited even though multiple friends are streaming different segments. To solve this bandwidth bottleneck, we identify two approaches.

The first approach is to ask inner-circle friends for more bandwidth. The rationale behind this, is that really good friends and family are willing to go out of their way to help you. If only a few peers are willing to deliver more bandwidth, this may be enough to provide the real-time streaming. We have implemented this through the use of a ‘request bandwidth’ button. Alternatively, this could be automated if some kind of priority scheme is implemented. Peers could then decide which services may be delayed to help close friends. This concept makes use of the distinction between what we call inner-circle friends (in marked area) and outer-circle friends (faceless circles), as shown in Figure 2. While outer-circle friends would allow you to use their resources, only inner-circle friends would really go out of their way and give you priority in using their resources.



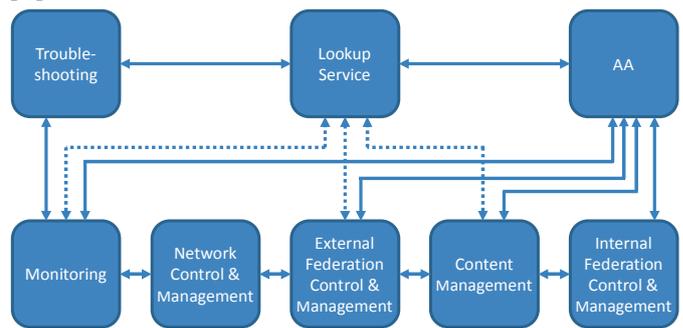
**Figure 2 Inner-circle and outer-circle friends**

The second approach is to stream at a lower bandwidth, using the adaptation part inherent in HTTP adaptive streaming. The most obvious way to achieve this is to have multiple bitrate segments available, so a lower bitrate may be chosen when needed. But, this does not make sense in our backup & sharing scenario. Having multiple copies at various peers will, even at a lower bitrate, take up more storage space. This will reduce the amount of files which can be backed up, given a certain amount of storage space. And, either the user doing the backup or the user receiving the backup will have to transcode the original content in various bitrates. Given that it is uncertain that a lower bitrate will ever be used, this can be very inefficient. Alternatively, we have come up with a just-in-time transcoding method using again the inner-circle friends. These inner-circle friends are willing and able to provide the necessary processing to perform transcoding at times lower bitrate segments are needed.

## 4 ARCHITECTURE AND COMPONENT INTERACTION

The FIGARO architecture is shown in Figure 3. The lower five functions typically run on a home gateway, whereas the upper three functions typically reside in the network.

Home gateways typically contain both internal federation and external federation. Internal federation is about the interaction with various devices in the home environment, both end user devices and infrastructure devices. External federation is about the interaction between gateways of various users. Content management contains functionality for content backup, sharing and remote access. The monitoring function monitors a.o. the available bandwidth, and the network control & management performs advanced networking functionality. Centrally, the lookup service is the entity keeping track of where everything is. It keeps track of which friends are using which home gateways, but also keeps track of where content is stored and backed up. The AA (Authentication, Authorisation) functionality provides the security needed for performing the inter-gateway federation. The troubleshooting function covers trouble-shooting, not further covered in this paper.



**Figure 3 FIGARO architecture**

The social-aware backup protocol in the FIGARO environment is shown in Figure 4. In this figure, a client device is shown making a backup of some (piece of) content at a gateway  $x$  ( $GW_x$ ) of some friend. The identification of friends amounts to a matching of which user’s friends on a specific social network (e.g., Facebook or Google+) are also FIGARO users, while also collecting their interests (a task which is outside the scope of this paper and is addressed in [4]). In a first step, the client device requests a list of available friends’ gateways from the Lookup Service. The Lookup Service is the part of the centrally located Federation Manager which amongst other things keeps track of all available gateways and of the users of those gateways. After receiving a list of available gateways the client device has the objective of finding the selection of friends on whose gateways to back up its items.

The matching of item and remote gateway should benefit the hosting users, i.e., by closely matching their interests with the content type to be backed-up. Also, it should transfer data effectively, i.e., by maximizing the utilization of available bandwidth between the respective gateways. In our previous

work [3] we have shown that joint optimization between bandwidth and interest is a clear winner compared to optimization on one or the other. See [3] for more detail on the optimal content placement solution. The result of the matching is a list, called “*query list*”, ranking, for each content type, the best choices of friends’ gateways where such content can be cached. The procedure behind the identification of the “best” list of remote gateways is the following: every time the procedure is scheduled, a user sends a backup request to the remote friend gateway (e.g. GWx) whose ID is in the element that tops the query list. Such element is then removed from the list if the request is accepted. After sending backup requests on behalf of a user for a specific total item size (chosen as tuneable parameter), the procedure stops and it is rescheduled randomly at a later time.

After selecting gateways, the backup process itself is started. First a backup request is sent, the target GWx authenticates the client device as belonging to a friend and the content is encrypted. Encryption could be done beforehand, but doing encryption after selecting the target GW allows for personalized keys for each target GW. Then, the backup itself is performed. This backup is performed in segments to enable remote access.

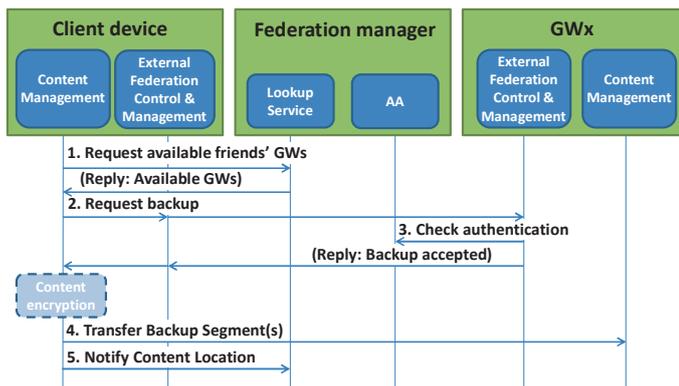


Figure 4 Social-aware backup

To enable remote access to content, either by the owner of the content or by users with whom the content is shared, content can be segmented. Reason for this is the limited upload network bandwidth most connections have. E.g. streaming of a high-quality video may require a 4 Mbps connection, while many upload speeds are limited to e.g. 512 kbps or 1 Mbps. By segmenting the content in small segments, and distributing these segments across various locations, even high-quality video becomes remotely accessible for streaming. Thus, by creating a segmented backup with a good distribution of segments across various gateways, multiple gateways together can provide real-time streaming access to high-quality content. For granting other users such access when sharing content, the FIGARO Federation Manager keeps track of these authorizations in the central AA function.

After backing up the content, the user can now share this content as well. If the content is shared with a user from a location at which a backup is placed, for segments already located at this location only an authorization step is necessary,

as the content itself is already distributed. This is shown in Figure 5. The client device first requests the friend’s GW location from the Lookup Service, and then sets the access rights for the other user in the AA module at the Federation Manager. These access rights are used in the mobile access scenario in the next subsection. Finally, a notification of the shared content is sent to the other GW, and optionally a key for access to the content.

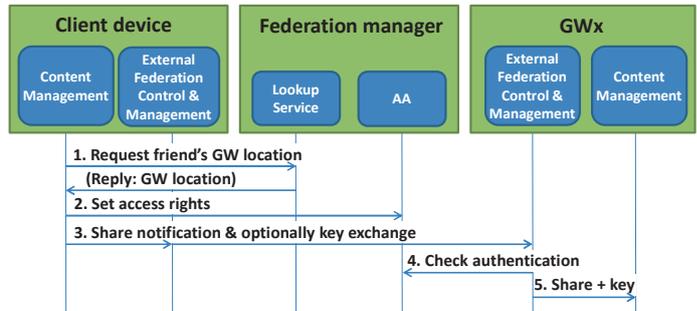


Figure 5 Sharing of previously backed-up content

Figure 6 shows this remote access scenario. The client device first has to discover the location of the backed-up content segments, and can then request these segments. Gateways containing these segments first have to check if the requesting device has authorization, and can then deliver the segments. For streaming, here is where some of the functionality for dealing with limited bandwidth fits in. Client devices can either explicitly request bandwidth from gateways of inner-circle friends, or can select lower-bitrate versions of segments to request.

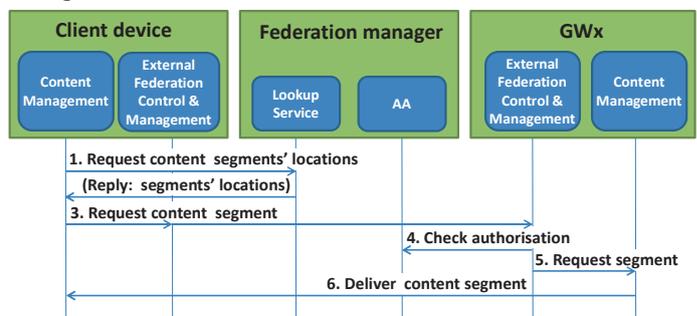
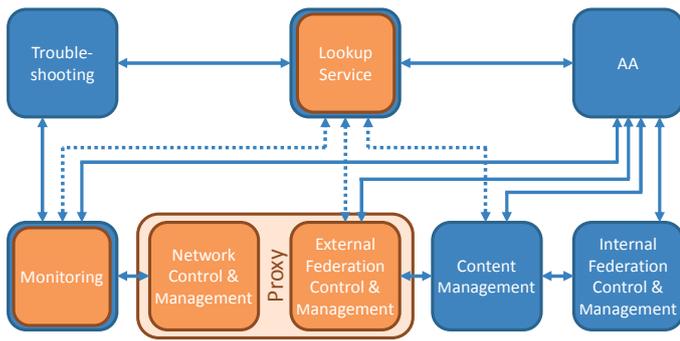


Figure 6 Remote access to backed-up content

## 5 IMPLEMENTATION

This section describes the prototype demonstrator we have implemented, with the goal to prove our theory and to emphasize the unique and novel parts of our solution. The implementation of our backup-and-share HAS streaming application therefore includes less elements than the architecture above describes. Figure 7 depicts the main differences.

As can be seen from the figure, there is no AA and content management. On each gateway, the directory that is backed-up is a fixed one. This simplification frees us from the need to implement AA and greatly simplifies content management.



**Figure 7 Demonstrator architecture marked in orange**

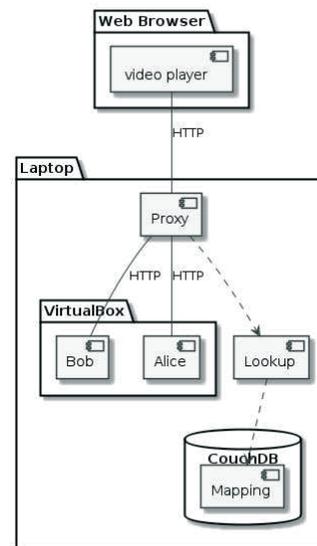
Note that this decision does not mean the system is insecure: each fragment is backed up only after it has been encrypted. This encryption is performed using GPG with symmetric AES, where each video segment is coupled to its own, uniquely generated key. While it does make sense to use a single key for a complete video, it is deemed unsafe to use a single key for all videos of a single user. In addition, our key-per-segment approach may make it more convenient to implement some kind of revocable privacy scheme later on. The filename of each encrypted segment is changed to a string generated using a UUID. Per segment, the original filename, the UUID-based backup file name, the key and the MIME type is stored in the lookup service. That service may reside on the gateway itself or on some kind of central (e.g. Servicer Provider specific) location. In our prototype, network distribution of backed-up content is not taken into account: all content is immediately available to all gateways over a Storage Area Network (SAN) that is connected over Gigabit Ethernet.

Now that the backup is created, it can be shared. Sharing content is a matter of sharing the invitation, which has the form of a metadata file. The metadata consists of a description to render an interface (e.g. movie title, duration, genre, etc.) plus an embedded HTTP Live Streaming (HLS) playlist. HLS is the Apple version of HTTP Adaptive Streaming. Inside the playlist references to file names contain the obfuscated UUID-based string, which can be retrieved from friends as they are part of the backup. This is kept transparent for the player: a regular HTML5 video tag is used and the HLS playlist is provided to it. Each entry in the playlist is rewritten to start with 'http://localhost:port/', so that when a segment is requested by the video player, the request finds its way to an HTTP proxy which runs at the specified port number.

We have introduced an HTTP proxy in the request chain, to be able to divert requests for segments to various gateways. The proxy has access to the monitoring information, which includes current bandwidth usage of the various other gateways. Using this information it then selects the gateway with the highest amount of available bandwidth. The proxy then fetches the segment, decrypts it and forwards it to the player. Using a proxy in this manner allows us to request segments from various locations in a flexible manner, and perform on-the-fly decryption, while using an off-the-shelf HAS player.

In the current version of our prototype, the gateways are equipped with the possibility to transcode content. Apart from the differently generated HLS playlist, the invocation scheme slightly changed. Most importantly, the gateway becomes an active participant in the content delivery process. If the HAS player requests a segment in a different quality, this request is forwarded as usual by the proxy to a selected gateway. But, to enable the selected gateway to transcode the content, that gateway needs to have the key for decryption as well. In this case, only gateways belonging to inner-circle friends are used for transcoding, since you trust them with your content. When the selected gateway agrees to cooperate, it indicates so by returning a JSON message that encompasses a temporary URI on the gateway itself. It then proceeds to decrypt and transcode the segment. The proxy uses the contents of the JSON message to construct a HTTP 307 'temporary redirect' response. When the player receives that response, it follows the redirect and ends up downloading the transcoded (and decrypted) video segment. Thus, using the redirect, the proxy steps out of the loop thereby increasing efficiency.

The invitation file itself could either be sent directly or by reference. Either way, additional signing, client certificates and other security measures could be used to further manage access.



**Figure 8 Implementation setup**

Figure 8 shows the implementation setup of our prototype. We have used a single laptop with multiple virtual machines as gateways. Gateway upload limitations are simulated using Traffic Shaping based on tc, available in the Linux kernel since version 2.4. This allows us to simulate current capacity and bandwidth limitations of various gateways, and also simulate gateways providing additional bandwidth upon request.

## 6 LESSONS LEARNED

We have only carried out our implementation as a proof-of-concept of our approach, and have not carried out a proper

evaluation as such. Still, during our design phase and testing of our implementation, a number of insights emerged.

Performing a backup and sharing content are two completely distinct activities, both from a user perspective as well as from a technical perspective. From a user perspective people expect their backups to be reliable, but not necessarily real-time available. Of course people will want to be able to restore content when needed, but given a free-of-charge P2P backup people will be happy with having to wait a bit for their content to be restored. On the other hand when someone streams a movie over the Internet, e.g. from Vimeo or from another video website, they prefer it to start immediately. During playback a temporary reduction in quality, e.g. due to networking circumstances, is easily forgiven in exchange for the real-time character of the playback. Of course people will want the highest quality available for playback, but they don't expect to wait for this either. These differing demands are schematically represented in Table 1.

**Table 1 Differing user demands**

|                | Speed | Content reproducibility |
|----------------|-------|-------------------------|
| Backup         | -     | +                       |
| Video play out | +     | -                       |

Implementing a HTTP proxy to do lookup and decryption of video segments has proven quite profitable. As remarked, a standard off-the-shelf HTML5 video player can be used to enjoy the benefits of optimized multipath streaming. Also, measuring the available bandwidth does not always yield correct results. Even though bandwidth measurements themselves may be accurate, they often require some time to give good results. When network circumstances change suddenly, it takes a while for bandwidth measurements to reflect this. But, an off-the-shelf HAS player can also switch quality itself. For example other devices may generate unexpected traffic that can significantly change the available bandwidth on the network. There can thus be cases where the player suddenly has a drop in available bandwidth, not detected in time by the bandwidth monitoring tool. In these cases, the general mechanism behind HAS dictates the client to go and try to download the slow or failed segment in another (lower) bitrate. During play out, the segments are played sequentially without any glitches. Only the quality degradation itself – due to the lower bitrate – may be noticeable.

## 7 CONCLUSIONS

In this paper we have described the concept of social backup and sharing of content. Content such as video content is backed up in segments across various friends' locations. This allows for streaming the content using multiple upload bandwidths, thereby increasing the resolution and quality that can be streamed in this P2P fashion. In our implementation of this concept, we have introduced a proxy between content player and backup network. This proxy allows us to

dynamically select locations from which to stream, while still making use of an off-the-shelf HTTP Adaptive Streaming client.

## 8 FUTURE WORK

From a technical perspective, there are also differences between backup data and shared content. In the case of a backup, bytes of data must be secured elsewhere and possibly be restored at a later time. For many years, solutions are on the market that reliably do this. The data itself has no meaning to the backup system, the system only has to be able to reliably transfer bytes from one place to the other, store them and keep track of what data is where. Technically, there are differences between data and content. When dealing with content backups in a social manner as in our work, the data itself becomes meaningful. This meaning is captured in the form of required additional metadata, as we have implemented in this study. Also, the ability to transform content, as in transcoding to a lower-quality version to provide a lower-bitrate version, can only be performed because the actual content is accessible. Therefore, it would be interesting to pursue a more general study into the 'holistic' differences between data and content. This may lead to insights to be exploited in future work in the area of social aware backup and sharing of content.

Additionally, we have so far only built a prototype as a proof-of-concept. We have not performed any structural experiments to evaluate the performance of our system, nor have compared it to existing alternatives. This we have left for future work also.

## Acknowledgment

This work has been financed partly by European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. ICT-2009-5- 258378 (FIGARO project).

## References

- [1] Pouwelse, Johan A., et al. TRIBLER: a social-based peer-to-peer system. *Concurrency and Computation: Practice and Experience* 20.2 (2008): 127-138.
- [2] D. N. Tran, F. Chiang, and J. Li. Friendstore: Cooperative online backup using trusted nodes. In *SocialNets '08: Proceedings of the 1st workshop on Social network systems*, 2008
- [3] J. Jiang, C. Casetti. Distributed Content Backup and Sharing using Social Information. *IFIP Networking 2012*, Prague, Czech Republic, May 2012.
- [4] J. Jiang, C. Casetti. Socially-aware Gateway-based Content Sharing and Backup. *HomeNets 2011- SIGCOMM workshop*, Toronto, Canada, August 15, 2011
- [5] Ze Li; Haiying Shen; Hailang Wang; Guoxin Liu; Jin Li, "SocialTube: P2P-assisted video sharing in online social networks," *INFOCOM, 2012 Proceedings IEEE*, vol., no., pp.2886,2890, 25-30 March 2012
- [6] Ho, Donghyeok; Song, Hwangjun, "Resource allocation algorithm based on social relation for video streaming services over P2P network," *Networks (ICON), 2012 18th IEEE International Conference on*, vol., no., pp.185,190, 12-14 Dec. 2012
- [7] Thijs Veugen, Hans Stokking, Secure Processing Offload in Recombining Media Segments for Mobile Access, 34<sup>th</sup> WIC Symposium on information Theory in the Benelux, 30-31 May 2013
- [8] Stockhammer, Thomas. "Dynamic adaptive streaming over HTTP--: standards and design principles." *Proceedings of the second annual ACM conference on Multimedia systems*. ACM, 2011.

# A NOVEL SCENE REPRESENTATION FOR DIGITAL MEDIA

C. Haccius<sup>1</sup>, T. Herfet<sup>1</sup>, V. Matvienko<sup>1</sup>, P. Eisert<sup>2</sup>, I. Feldmann<sup>2</sup>  
A. Hilton<sup>3</sup>, J. Guillemaut<sup>3</sup>, M. Kludiny<sup>3</sup>, J. Jachalsky<sup>4</sup>, S. Rogmans<sup>5</sup>

<sup>1</sup>Intel VCI, Saarbrücken, DE; <sup>2</sup>Fraunhofer HHI, Berlin, DE; <sup>3</sup>University of Surrey, Surrey, UK; <sup>4</sup>Technicolor, Hannover, DE; <sup>5</sup>Minds, Hasselt, BE

**Abstract:** This document presents a novel multidimensional scene representation architecture which bridges the gap between classical model based approaches, such as meshes, and vision based approaches, such as video plus depth. The architecture is described conceptually and a proposed implementation is presented. The layered architecture and its implementation present a tidy way of conceptualizing the interactions of data up the production chain. Beyond that this architecture enables innovative computational videography processing of multidimensional material. High quality storage of computer generated and captured video data as well as support for intermediate processing steps and novel content representation and interaction complete the architecture to provide a means for future developments for enhanced scene visualization.

**Keywords:** Multidimensional Scene Representation, Computational Videography, Content Interaction

## 1 INTRODUCTION

*SCENE* is an on-going research project dedicated to create and deliver richer media experiences [1]. A consortium of international research and industry partners aim to enhance the whole chain of multidimensional media production. These enhancements include new capturing devices, scene content processing tools, renderers dedicated to render *SCENE* data. At the core of this project is a novel representation architecture. This novel architecture is results from a change of paradigm the *SCENE* project introduces to cinematic movie production processes.

This paper is structured as follows. In the next section the change of paradigm introduced by the *SCENE* project and the historical motivation for this change are explained. Section 3 contains the conceptual description of the scene representation architecture, highlighting the features and advantages of such a layout. The paper continues with the actual implementation of the envisioned scene representation. The final section draws a conclusion and points to research conducted by different partners in the *SCENE* consortium. It also outlines future work which will be done on the Scene Representation.

## 2 SCENE – A PARADIGM CHANGE

Throughout history the bottleneck of image or movie capturing devices has been the film; in recent times the image sensor. As the sensitivity of the film or image

sensor was comparably low, this bottleneck enforced constraints on the optical system and the capturing process. For low light conditions long exposure times or large lenses had to be chosen; the first resulting in motion blur of moving objects and the second limiting the depth of field. These artefacts have coined movie productions throughout the last century; they even became desired artistic elements and stylistic devices in movie productions.

During the last years new chip technologies have enhanced available image sensor to a level where this physical bottleneck is removed. The amount of light necessary to create an image does usually not dictate camera parameters any more. Nevertheless, motion blurs and limited depth of field are still applied for artistic means.

Computational Photography alters image content by computational means to create visually appealing and artistically interesting results [2]. Successful implementation of ideas from computational photography requires high quality data and information on the scene content. The same holds for computational videography, which transfers the ideas of computational photography to motion pictures.

Data distortion introduced for artistic means as described above limit the application of computational videography and therefore limit the artistic freedom in post processing steps. *SCENE* changes the way data is acquired by striving to capture as much undisturbed information as possible by maintaining artistic freedom and directors decisions. Thus *SCENE* enables the full spectrum of computational videography without limiting neither director nor camera man in his creative freedom.

## 3 THE SCENE REPRESENTATION

The SRA is a key innovation to enable the paradigm change described above. Major achievements are

**Single Format:** When processing multidimensional video data on a computer a multitude of information sources are required: Video from several sources, camera calibration data, lighting information and spatial knowledge are just naming a few. Our proposed architecture unites all this information necessary for movie production in a single format.

**Undistorted data:** When introducing artistic elements like motion blurs, depth of field or colour offsets these effects traditionally modify the captured data. Post-processing such data is time consuming and difficult. The

**Corresponding author:** Christopher Haccius, Intel VCI, Saarland University, Campus C6.3, 9.05, 66123 Saarbrücken, +49 681 302 6544, haccius@intel-vci.uni-saarland.de

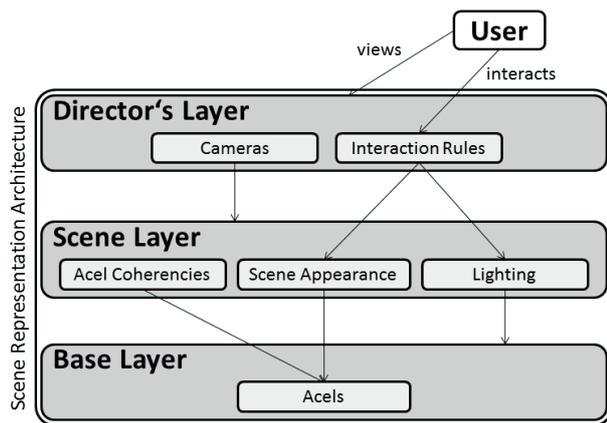


Figure 1: Scene Representation Architecture Layout

scene representation stores all data in the best available quality and introduces altering effects in a higher layer, thus preserving all available data for facilitated image and video processing steps.

**Content Interaction:** Image or video content is usually frame based. The scene representation is object based and therefore allows segmented content. Knowledge about objects in a scene allows interaction such as updated product placement, object modification or camera interaction.

**Unified Representation:** Computer Generated (CG) content and Captured Video (CV) stem from two very different worlds and are processed largely independent in movie productions. The scene representation allows a unified representation of both, CG and CV data as well as any intermediate processing steps, thus merging both worlds in an early stage and facilitating post production.

These achievements are enabled by a layer-based architecture (see Figure 1). Details of the different layers are given in the following subsections.

### 3.1 The Base Layer

The base layer of the SRA contains elements which are either CG or CV data. The architecture suggests that these elements are the smallest meaningful units that a capturing device can detect. We therefore name those units atomic scene elements, abbreviated ‘acels’. Each acel is coherent in itself, but independent from other acels. The number of dimensions an acel uses is conceptually unlimited. Possible dimensions are spatial and temporal dimensions, colours or reflectance. All common data types like images, meshes or videos are supported as acels, but any intermediate representation or additional dimensions on top of existing data types can easily be represented as well.

Many ideas for acel representations from Captured Video can be transferred from research on patches. Patches represent solid (sub-) surfaces for one animation/time instance of a scene. They evolve over time in a way which is plausible for human assumption, i.e. their position and shape are altered according to temporal and physical coherence. Patches represent physical entities and where introduced in the context of real-time reconstruction of

human faces, for example in [3]. Directly mapping the patch properties of acels shows that acels are well suited to represent solid and non-solid objects physical coherences and supports the use of acels for numerous CG effects like relighting or shadows. Multiple more features can be easily added.

### 3.2 The Scene Layer

Multiple acels have to be registered in a global scene context. This registration is done in the scene layer of the SRA. The dimensions of a scene are the superset of all acel dimensions contained in a scene. Registration is not only done in space and time, but colour offsets and other measurement differences between acels can be corrected during registration. This component of the scene layer has therefore a structure comparable to a scene graph comprising multidimensional offset information in its branches. Sowizral and Nadeu propose methods to present multidimensional scene volume information in graph structures [4, 5]. In addition to placing acels in a global scene, the scene layer provides the lighting information for the scene. Lighting can be adjusted according to the scene lighting conditions independent of where acels were captured initially [6, 7]. Relations among acels are also expressed in the scene layer [8]. A coherency table expresses coherencies among the individual acel dimensions and features.

### 3.3 The Director’s Layer

The director’s layer defines the usage of scene content. The most important form of scene usage is scene perception. Cameras describe the traditional way of perception by defining intrinsic and extrinsic camera parameters. A novelty is that these virtual cameras are not limited to physical plausibility, but can feature several depth planes or shaped focal depth, introduce motion blurs, which are contradicting physical motion, or change the light sensitivity over one frame. Moreover, by defining user interaction rules, users further down the processing chain may be allowed to modify director’s decisions.

## 4 ALGORITHMS

While storing traditional image and video formats without any further knowledge is allowed in the SRA, novel scene analysis and modification techniques can provide numerous benefits. As the SRA imposes hardly any constraints, future development is limited by the researchers’ creativity only. Exemplarily, we present in the following novel algorithms developed in the context of multidimensional scenes, which contribute to and largely benefit from the proposed SRA.

### 4.1 Superpixels

In [9] Ren and Malik introduced the idea of utilizing superpixels as primitives for image analysis and processing tasks. These superpixels are groups of pixels sharing similar features like colour and texture (see Figure 2). They can be utilized as auxiliary information that is stored with the content in order to allow interactivity



(a) Original (© Technicolor) (b) Segmented Image

Figure 2: Superpixels

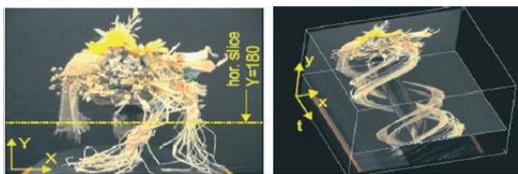
especially for video-based content. Basic functions that can take advantage of such information are selection and tracking of objects. For a good and robust tracking, key criteria for superpixels are temporal consistency and the ability to adapt to structural scene changes. Superpixel information can be added to acels as a further dimension or acels can be segmented according to superpixel information already. Temporal consistency of the superpixels is maintained in the scene layer of the proposed SRA.

#### 4.2 Image Cube Trajectories

A second algorithmic approach to create spatio-temporal consistent acels is the analysis of image cube trajectories. The main idea of this method is to represent each 3D point by a related trajectory in a so called image cube (see Figure 3). It has been shown in [12] that it is possible to reconstruct the 3D scene from the parameters of the trajectories in the image cube. A key component for this process is the trajectory detection within the cube. It is based on image cube parameterization as well as on robust estimation of the trajectory colour. The main advantages coming with the proposed SRA are on the one hand to be able to store trajectory parameterizations, such as shape, colour, reflectance properties or detection confidence. On the other hand the parameters of the image cube, such as dimensions, related camera calibration information, camera path or the original image data can be kept and stored directly.

#### 4.3 Spatio-Temporal Point Correspondence

For the analysis of dynamic multi-view sequences point correspondences in spatial and temporal direction are of often required. Common methods either produce too many faulty or too few corresponding points. We have therefore developed a method for key point matching that reliably establishes correspondences in both spatial and temporal direction and that returns more matches than standard approaches [13]. Instead of considering individual key points independently and removing those who might have ambiguities, we look at the spatial configuration of neighbouring points. Figure 4 illustrates matching points found by an incrementally constructed



(a) Sample Image (b) Image Cube Representation

Figure 3: Flower Sequence with circular camera path



(a) Input Data (b) Basic Matching (c) Our approach

Figure 4: Point Correspondences for Spatio-Temporal Scenes

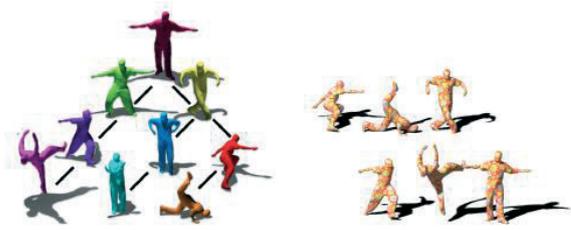
Delaunay mesh over key point candidates assuming locally affine displacements. While in given example standard SIFT matching obtains 3100 correspondences with a significant amount of false pairs, our approach provides 3500 matches with less outliers. Detected correspondencies and their reliability measures can be stored in the scene layer of the SRA.

#### 4.4 Temporally consistent meshes

Since the pioneering work by Kanade and Rander [14], who with their virtualized reality system introduced surface capture for human motion, significant work has been done on reconstructing human characters from image and depth data. More recently it has become essential to produce not only accurate models of each independent time frame but full 4D temporally consistent character reconstructions. These models can be used for automatic propagation of mesh and texture edits [15] saving significant artistic effort. This style of data can be stored efficiently within the acel representation with spatial dimensions representing the 3D position of mesh vertices and temporal dimensions their motion over time. Of two possible surface reconstruction and tracking approaches the first involves building up a series of 3D models based on the work of Stark et al. [16]. Visual hull and multi-view stereo information is combined within a graph cut framework to build accurate frame by frame models of a character. Subsequently these models are tracked using non-sequential geometric tracking algorithms presented by Budd et al. [17]. The second makes use of appearance information to track open surfaces. Rough initial tracking of sparse points with a standard KLT tracker yields a set of point clouds whose similarity in Euclidean space gives a metric to build the shape tree. The tracking is refined with a dense patch based tracking approach defined by Kludiny et al. [18].

### 5 IMPLEMENTATION

In order to meet the requirements to the SRA defined by the intended advances and the algorithms presented above a flexible and extendable implementation is required. This



(a) Unaligned Mesh Tree (b) Aligned Mesh Database

Figure 5: Temporally Consistent Meshes

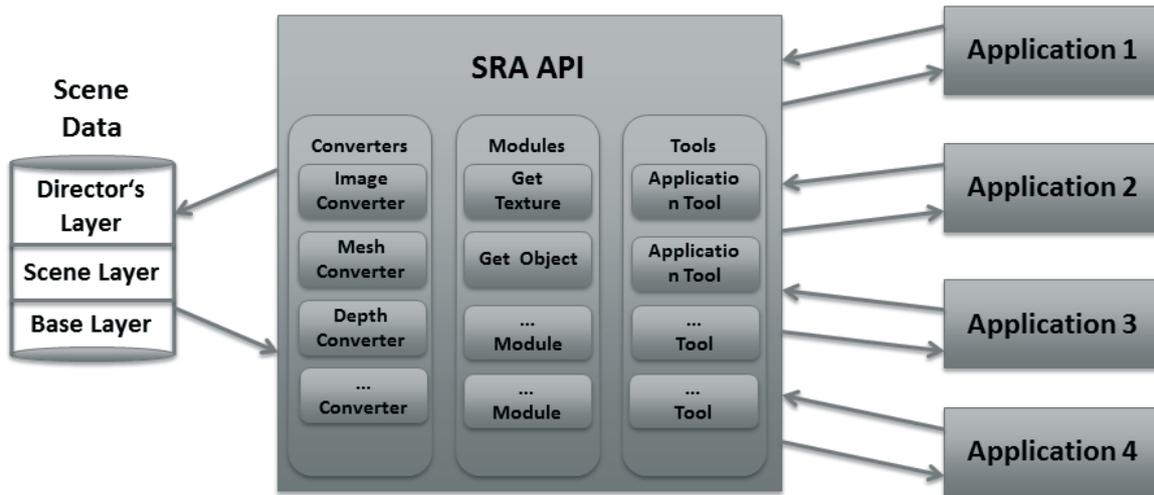


Figure 6: Structure of SRA Implementation

becomes especially important as all components of the video processing chain are constantly enhanced and further developed. A possible implementation enabling these demands is an API to an underlying data structure. The SRA API is an implementation of the concepts presented in Section 3 enabling the algorithms described in Section 4.

The underlying scene data can be any structured data that the SRA API supports. The structure represents the different layers of the SRA as well as scene elements such as acels, configuration data or interaction rules. The file readers which understand the underlying format are not part of the SRA API, but can be exchanged with the format of choice.

Support for a certain type of data is enabled in the SRA by the necessary converters in the API. These converters need to understand the data and be able to provide it to different processes in the API in the required representation. Exemplarily a mesh converter can be asked to return a mesh representation of an arbitrary input acel. If the data type of the input acel is supported as a mesh the request can be processed and an internal mesh representation can be provided. The set of converters can be arbitrarily extended to meet the data types of different data sources as well as the input requirements of further processing tools and applications.

Scene Modules in the API are used to initialize and execute computational processes. These modules can make use of converters and additionally implement further algorithms that process and enhance scene data. A module requires scene data in a certain format as input and can be triggered to be executed on demand. Exemplarily for such modules are getter-modules for textures or objects. These modules apply converters to transform acel information into a desired representation and present them to the next higher level as an object or a texture.

The third part of components contained in the SRA API contains interfaces for tools. Different applications have different demands to the Scene Representation. A certain tool interface can fulfil these demands by providing scene

content application specific. Exemplarily a video rendering tool assures that acel data is presented frame based to a video renderer, which can then render the content of the scene per frame. Alternatively, a free-view interface can present the full content of a 3D scene and be rendered as a static scene to navigate in.

Neither the number of converters nor computational modules or application interfaces is limited. All of these can be extended with the growing demands from users, applications and algorithms. As such the implementation of an API for SRA access presents currently the perfect solution to have an extendable and flexible interface which does not limit the creativity of its users.

The SRA API is a C++ library which can be included in applications in order to make use of the scene features. It can then be accessed by scripting languages (Python) or through the header functions exposed by the API. Thus it provides an easy to use interface to the scene developments.

## 6 VERIFICATION

A prototype to prove the conceptual ideas presented above was created. 100 frames of a billiard scene are represented in the *SCENE* layers and rendered. Figure 6 shows one of the video-clip frames, which exemplarily presents novel features and the paradigm change enabled by the *SCENE* format.

The prototype contains five acels: the static background, two independent players, the colored balls, the white ball

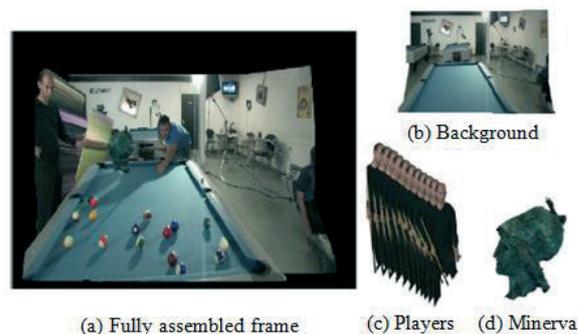
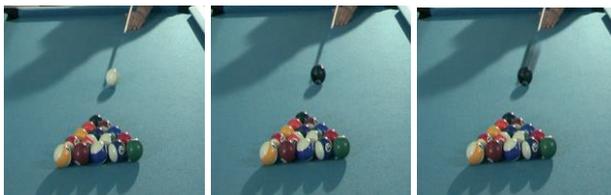


Figure 6: Proof-of-Concept SRA Implementation



(a) Captured Data (b) Replaced Object (c) Motion Blur  
Figure 7: Artistic Effects

and the Minerva head. The white ball is stored as an individual ace for the artistic effect shown in Fig. 7. While the background is a single color bitmap plus depth [19], the players and balls have a temporal dimension as well. The Minerva head is a mesh with material properties to show the seamless integration of bitmaps and meshes in one single layout. Lighting conditions were captured with a 360° environment camera and an environment map was created. This information was used to relighten the Minerva head according to the lighting conditions of the scene [20, 21]. Lighting information and scene composition are stored in the scene layer.

The director's layer describes a camera, which renders the scene off-angle to the original capturing device to make depth visible. Fig. 7 presents the feature of adding artistic effects in the director's layer. The captured data is unblurred and can be easily segmented and tracked (see Sections 4.1 and 4.3) and replaced by a black ball. Adding a motion blur is another simple algorithmic step. While our blur perfectly shows the ability to insert artificial blurs, photorealistic algorithmic blurs exist and can be included in the renderer [22].

## 7 CONCLUSION AND FUTURE WORK

This paper presents a paradigm change in the way future video content can be produced to enable computational videography. Furthermore, a representation design allowing this change from an architectural viewpoint as well as state-of-the-art algorithms to create and process content for this architecture are introduced.

To our knowledge this is the first approach to redesign the full movie production process with the goal of enabling computational videography on multidimensional video content. Scientific interchange and future research will surely be able to enhance the ideas presented here, yet we are sure that the paradigm change introduced is imminent and work described in this paper represents a valid foundation for further research.

Future work will need to further specify the SRA to meet the quality and algorithmic demands posed by content consumers and developers. Existing and novel ideas of computational videography can be designed to make use of the extra information provided through the SRA. Acquisition hardware will be designed to capture an ever increasing amount of multidimensional data for advanced video processing.

Some of this work is currently covered by SCENE project partners. Next to the five institutes mentioned in the list of authors the companies ARRI, Barcelona Media, Brainstorm and 3Dlized are collaborating partners in the SCENE project. A full project description and the latest developments can be found online [1].

This work has been supported by the EC within the 7<sup>th</sup> framework programme under grant agreement no. FP7-IST-287639.

## References

- [1] V. López, E. Fuenmayor, and A. Hilton, "Novel scene representations for richer networked media", <http://3d-scene.eu/>, Jan. 2013.
- [2] R. Raskar and J. Tumblin, Computational Photography: Mastering New Techniques for Lenses, Lighting, and Sensors, AK Peters, Ltd., 2009.
- [3] W. Waizenegger, N. Atzpadin, O. Schreer, and I. Feldmann, "Patch-sweeping with robust prior for high precision depth estimation in real-time systems," in *Image Processing (ICIP), 2011 18th IEEE International Conference on*. IEEE, 2011, pp. 881–884.
- [4] D.R. Nadeau, "Volume scene graphs," in *Proceedings of the 2000 IEEE symposium on Volume visualization*. ACM, 2000, pp.49–56.
- [5] H. Sowizral, "Scene graphs in the new millennium," *Computer Graphics and Applications, IEEE*, vol. 20, no. 1, pp. 56 –57, jan/feb 2000.
- [6] R. Ng, R. Ramamoorthi, and P. Hanrahan, "Triple product wavelet integrals for all-frequency relighting," in *ACM Transactions on Graphics (TOG)*. ACM, 2004, vol. 23, pp. 477–487.
- [7] R. Ramamoorthi and P. Hanrahan, "A signal-processing framework for inverse rendering," in *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*. ACM, 2001, pp. 117–128.
- [8] P. Huang, C. Budd, and A. Hilton, "Global temporal registration of multiple non-rigid surface sequences," in *Computer Vision and Pattern Recognition (CVPR), 2011, IEEE Conference on*, June 2011, pp. 3473 –3480.
- [9] X. Ren and J. Malik, "Learning a classification model for segmentation," in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*. IEEE, 2003, pp. 10–17.
- [10] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," 2012.
- [11] C.L. Zitnick and S.B. Kang, "Stereo for image-based rendering using image over-segmentation," *International Journal of Computer Vision*, vol. 75, no. 1, pp. 49–65, 2007.
- [12] I. Feldmann, P. Eisert, and P. Kauff, "Extension of epipolar image analysis to circular camera movements," in *Image Processing (ICIP), 2003. Proceedings of International Conference on*. IEEE, 2003, vol. 3, pp. III–697.
- [13] J. Furch and P. Eisert, "Robust key point matching for dynamic scenes," in *Proc. European Conference on Visual Media Production (CVMP)*. IEEE, Dec 2012.
- [14] T. Kanade and Rander. P., "Virtualized reality: Constructing virtual worlds from real scenes," 1997.
- [15] M. Tejera and A. Hilton, "Space-time editing of 3d video sequences," in *Proceedings of the 2011 Conference for Visual Media Production*, Washington, DC, USA, 2011, CVMP '11, pp. 148–157, IEEE Computer Society.
- [16] J. Starck, "Surface capture for performance-based animation," *IEEE Computer Graphics and Applications*, vol. 27, 2007.
- [17] C. Budd, P. Huang, M. Kludiny, and A. Hilton, "Global non-rigid alignment of surface sequences," *IJCV*, 2012.
- [18] M. Kludiny, C. Budd, and A. Hilton, "Towards optimal non-rigid surface tracking," in *ECCV*, 2012, pp. 743–756.
- [19] C. Richardt, C. Stoll, N.A. Dodgson, H.-P. Seidel, and C. Theobalt, "Coherent spatiotemporal filtering, upsampling and rendering of RGBZ videos," May 2012, vol. 31.
- [20] T. Haber, C. Fuchs, P. Bekaer, H.P. Seidel, M. Goesele, and H.P.A. Lensch, "Relighting objects from image collections," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 627–634.
- [21] T. Yu, H. Wang, N. Ahuja, and W.C. Chen, "Sparse lumigraph relighting by illumination and reflectance estimation from multi-view images" in *ACM SIGGRAPH Sketches*. ACM, 2006, p. 175.
- [22] S. Lee, E. Eisemann, and H.P. Seidel, "Real-time lens blur effects and focus control," *ACM Transactions on Graphics (TOG)*, vol. 29, no. 4, pp. 65, 2010.

# HbbTV: a powerful asset for alerting the population during a crisis

Ralf Pfeffer<sup>1</sup>, Sebastian Siepe<sup>2</sup>, Benedikt Vogel<sup>3</sup>, Roberta Campo<sup>4</sup>, Cristina Párraga Niebla<sup>5</sup>

<sup>1,2,3</sup>Institut fuer Rundfunktechnik GmbH, Muenchen, Germany; <sup>4</sup>Eutelsat SA, Paris, France; <sup>5</sup>German Aerospace Center (DLR), Wessling, Germany

E-mail: <sup>1</sup>pfeffer@irt.de, <sup>2</sup>siepe@irt.de, <sup>3</sup>vogel@irt.de, <sup>4</sup>rcampo@eutelsat.fr, <sup>5</sup>Cristina.Parraga@dlr.de

**Abstract:** Alerting the population during a disaster is key to limiting the damage, personal losses and injuries that are sustained. To this end, an alerting system that can utilize diverse distribution channels to reach as many people as possible is essential. The aim of this paper is to present a powerful and effective system which is capable of broadcasting alerts to the population via their televisions, using existing technologies on the receiver side, so that the procurement of specialized devices is unnecessary. The Alert4All project is developing an innovative multi-channel public alert system that uses HbbTV - an already available worldwide standard on the mass market - to disseminate alerts.

**Keywords:** Alerting system, crisis, HbbTV, broadcast, A4A, CAP

## 1 INTRODUCTION

In the context of natural and man-made disasters, public alert systems have the potential to significantly reduce the impact in terms of victims and losses to property and infrastructure, if the relevant information arrives on time to the people at risk. Consequently, several countries are improving their public alert systems, moving from a “single-channel” approach to a “multi-channel” approach that exploits the complementarities in terms of warn effect vs. information content that the different alert channels can offer. Some examples are the USA’s IPAWS [1] (Integrated Public Alert and Warning System), Germany’s MoWaS (Modular Warning System) [2], Israel’s eVigilo [3] and the NL Alert System [4] in the Netherlands.

The dissemination of alert messages using these multi-channel systems over TV devices is mainly based on the delivery of alert messages to the TV broadcasters’ news redaction, which then provides the information to the public through their news programs. Only eVigilo uses a direct-to-TV alert interface and this requires either a specific set-top box or a dedicated application at the receiver TV, based on subscription service [3].

The EU FP7 Alert4All project [8], in short A4A, is developing a multi-channel alert system that utilizes, amongst others, digital satellite and terrestrial TV as alert dissemination channel, via two different approaches:

- Applying HbbTV [7] to deliver the alert message with multi-media content and interactivity;
- Applying a dedicated application at the receiver using a novel transport protocol - the A4A protocol [6] - that minimizes the required capacity to transmit the alert.

This paper focuses on the HbbTV solution and describes the architecture and interfaces required to implement the HbbTV-based alert service for satellite (DVB-S/S2) and terrestrial (DVB-T/T2) broadcast systems.

Following this introduction, Section 2 provides an overview of the A4A system, Section 3 introduces the HbbTV protocol and explains how it is used to disseminate alerts to the population in case of need and Section 4 presents some conclusions on the alerting system.

## 2 THE ALERT4ALL SYSTEM – AN OVERVIEW

The multi-channel approach adopted in the A4A communications system is based on an alert message dispatcher, called the Global Alerting Gateway (GAG), that allows authorities and responders to send alert messages over different communication technologies, as shown in Figure 1. The GAG applies the de-facto standard for alert messages, the Common Alerting Protocol (CAP) [5], which is XML based, which ensures interoperability with other public warning systems that also use this standard. In addition, it incorporates a novel protocol – the A4A protocol - that significantly reduces the capacity requirements for transport imposed by CAP and ensures best practices in the formulation of alert messages [6]. Both interfaces are developed for digital TV broadcasting based on DVB. In particular, DVB-S/S2, DVB-T/T2 and DVB-SH have been selected to showcase the solutions.

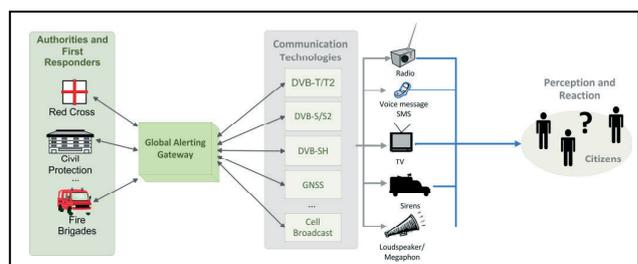


Figure 1: Alert4All communications system overview

**Corresponding author:** Ralf Pfeffer, Institut fuer Rundfunktechnik, Floriansmuehlstrasse 60, 80939 Muenchen, 0049 – (0)89 – 323 99 – 345, pfeffer@irt.de

The A4A protocol has been developed as a dedicated application suitable for a DVB-SH prototype receiver and a USB DVB-T2 receiver connected to a computer. This solution requires a software update for DVB-T2 set top boxes with enhanced services, such as digital recorders.

However, given that CAP is XML based, a CAP alert message, even enriched with multimedia content, can be relatively easily fed into HbbTV content. For this reason, it is very attractive to distribute alert messages to legacy HbbTV-enabled TVs (or set top boxes).

### 3 THE HBBTV-BASED ALERT SERVICE

Hybrid Broadcast Broadband TV [7], or HbbTV, is a major new pan-European initiative aimed at harmonising the broadcast and broadband delivery of entertainment to the end consumer through connected TVs and set-top boxes.

HbbTV started from joint initiatives in France and Germany that were then followed by Spain and the Netherlands, and other EU countries. It has been published as ETSI standard in its first version in 2010 and it is now acquiring a general consensus worldwide, including in the US, South America and Asia.

HbbTV combines broadcasting and internet services in order to create brand new services and usage experiences for TV, including catch-up TV, video on demand (VoD), interactive advertising, personalisation, voting, games and social networking, as well as programme-related services such as digital teletext and Electronic Programme Guides and second screen applications. In a nutshell, the receiver uses a browser to display the HbbTV page and from there the user can find information and navigate to other webpages on the Internet.

In this hybrid system each of the two contributing networks is used for specific purposes:

- Broadcast (one-to-area) is used to deliver traditional digital TV (i.e. linear audio and video) together with application data and signalling information to all the users;
- Broadband Internet is used to ensure bi-directional communication and interactivity between the user and the application provider.

However, the HbbTV applications do not need to make use of both links: while interactivity through broadband networks is fundamental for voting and video on demand, broadcast only is suitable for digital teletext and information services.

HbbTV is independent of the physical communication link and is used by terrestrial and satellite broadcasters (DVB-T/T2/S/S2) and all network operators (ADSL, cable, DVB-C, fibre and satellite).

### 3.1 The Alert4All Solution for Delivering Alert Messages over HbbTV

The GAG issues alert messages and distributes them to the population via several communications systems. For this purpose, an interface based on web services has been developed in the A4A project. This interface has a data part and a signalling part. The data part is used to deliver alert messages to the access points of each communication technology, or the “Alert Channel Access Point”, in short ACAP. The signalling part is used to interact with the ACAP when for example, checking a system’s availability, which allows the authority that sends the alert messages to monitor the status of the alert channels. Hence, the HbbTV ACAP at the broadcaster site contains a web server that connects to a web client in the GAG.

The GAG can push different request types, e.g. to check the availability of the system chain, or test a request, and in these instances no alerts need be forwarded from the broadcasters.

With the request type 1, the CAP-message will be searched for the specific XML-objects that are needed to create the alert-message. The plaintext of the XML-objects is transmitted within multiple Stream Events towards a multiplexer. The multiplexer will then include them, together with the static HbbTV-application, inside the DSM-CC (Digital Storage Media Command And Control) carousel. The carousel is part of the DVB transport stream (TS) that also includes the video and audio information (V/A). The interface chain used is illustrated in Figure 2.

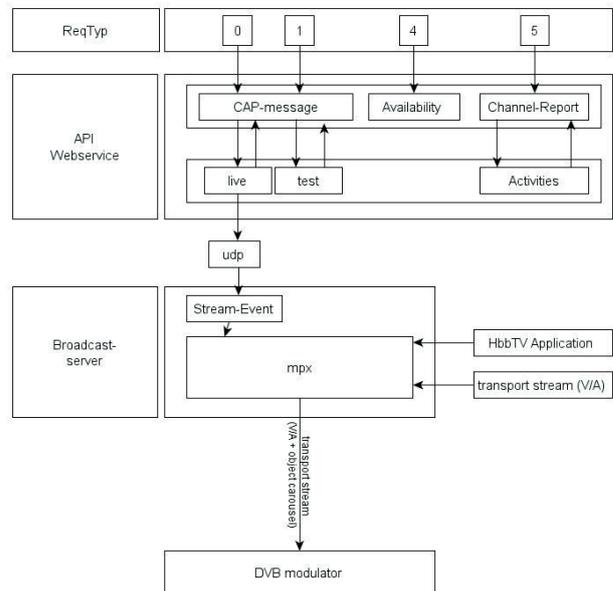


Figure 2: Interface at a broadcaster

The Stream Events contain the actual alert. They fill the static HbbTV application which then builds the frame of the alert messages shown on the TV.

The signalling interface includes two types of acknowledgements that allow it to inform the GAG that

the broadcaster has received an alert message and also that the alert message has been forwarded to the viewers.

All the specified functions used in the A4A solution are standardized to conform with DVB and HbbTV, so there is no need to enhance any standard.

### 3.2 Hybrid Broadcast Broadband TV

When using HbbTV as an alerting-system with terrestrial or satellite transmission it is important that the system is independent from broadband communication channels so that the risk of unavailability of the service in case of crisis is minimized. Therefore the implementation for the A4A HbbTV alerting channel has been structured on two levels:

- all the essential information is delivered within the transport stream, i.e. the transport container for video, audio and additional data via Broadcast;
- complementary information, e.g. detailed maps, or accessory instructions which are not mandatory to handling the crisis can be retrieved via broadband when such a connection is available.

HbbTV is based on XHMTL. Supporting receivers embed a browser to handle this special HbbTV-HTML. In the solution being developed by the A4A project, an A4A-HbbTV-Application (i.e. the HbbTV-XHMTL) is transmitted constantly within the DSM-CC carousel in the transport stream. This application is static and is automatically running when a service has been chosen. The application is invisible as long as there are no new or updated alerts. It is only if an alert is sent with a Stream Event that the application will be shown on the television screen. Normally the sound of the running program is on, although optionally it can be muted while an alert is shown; this increases the "wake-up" effect of the displayed alert, and trigger potentially higher interest in the displayed information.

The HbbTV-application under development in the A4A project is shown in Figure 3.

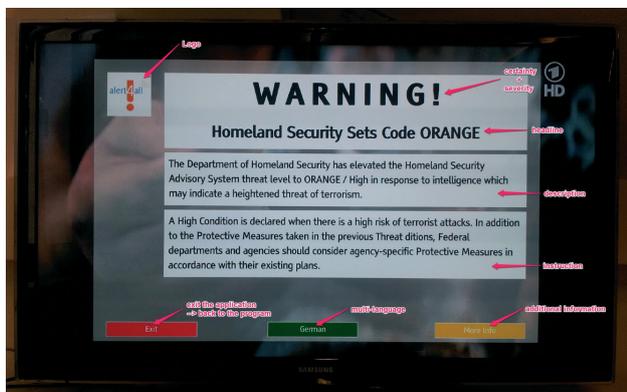


Figure 3: An alert displayed on a common television set

A space is reserved on the top left of the display for the logo of the authority that issued the alert message (this space is currently occupied by the A4A logo).

Depending on what is defined within the received CAP-file, one of the following three levels of severity is shown: *Alarm*, *Warning* or *Information*.

The headline highlights the core information of the alert, while the body of the message gives the details in two paragraphs:

- the description of the alert event;
- the instructions on what to do in this special case (e.g. leave the house, stay inside and close the windows, evacuate to a specific point and similar).

Three coloured buttons appear at the bottom of the alert window: viewers can either close the application (red), view the alert in a different language (green) or access additional information on the event (yellow).

When a submenu is selected, e.g. *More Info*, a fourth *back*-button will appear. The colours used for these options correspond to the coloured buttons available on most of the standard remote controls for easy access and use of the application.

Alert4All supports multi-language alerts. The alert is by default displayed in the same language as the on-going TV service. Other languages are accessed via the green button.

All the data seen up to this point is considered as essential information and thus it is delivered via broadcast. No Internet connection is required for this part. This allows for disseminating the alerts to a larger number of receivers: TVs that are HbbTV-enabled but which have not been connected to the internet or those that have been connected to internet but for which, as a consequence of the disaster event, the broadband link is out of service.

The HbbTV application is associated with a specific TV channel: this means that the alert is displayed on the screen when the user is watching that specific TV channel. In a real implementation, the HbbTV application will be associated with all the TV channels in order to reach all the users watching, regardless of the TV channel to which they are tuned.

The data composing the essential information from the basic A4A-HbbTV-Application has purposely been kept very small - on the order of 100 KB per alert. The data rate required for disseminating such alerts in less than 8 seconds is in the order of 100 kbps. Therefore there is a small overhead to be considered for a high definition (HD) TV channel that is transmitted on average at 8 Mbps.

Besides the essential information broadcast, there is the possibility to access complementary information via a broadband network when such a connection is active. For example, there is a URL-link in the *More Info* window, as well as a QR-code to open the URL on a smartphone that uses the Internet connection so that further information like pictures, maps, etc. can be gathered. However, in emergency situations, Internet networks could be

congested or the infrastructure affected and not every household is equipped with an Internet connection. Additionally, sometimes all the features of HbbTV-enabled devices are not exploited by their users.

### 3.3 Transmission

The HbbTV application containing the alert is located in a DVB transport stream; therefore the transmission system is standard and interoperable. Any transmission system that handles a transport stream at the input of a modulator can be used, while on the receiving side, any receiver supporting DVB and HbbTV can be used.

Within the Alert4All Project, two transmission systems are in use: DVB-T/T2 and DVB-S/S2.

Both systems need a modulator to transfer the transport stream along with its video, audio and alert (i.e. HbbTV application and Stream Events) to a radio frequency (RF)-signal being transmitted over the air.

DVB-T/T2 is the standard for terrestrial transmission: the RF signal is transmitted from the broadcasters' towers and antennas. The viewer uses the traditional terrestrial receiving antenna at home.

DVB-S/S2 is the standard for digital Television via satellite. In this instance the RF output signal of the modulator is sent to a satellite transponder and then broadcast over a very large geographical area (spot). The viewer receives the signal with a receiving satellite antenna, typically a parabolic antenna.

Together with DVB-C - the standard for cable-bound transmission - a broadcast can reach most people. As this is broadcasting, it doesn't matter how many people are watching the service, the network won't collapse due to too many users, as could and would presumably happen in a mobile communications network.

## 4 CONCLUSION

With HbbTV, a wide spread standard is used to transmit alert messages to citizens using existing broadcasting systems. Broadcast is favoured for essential information, as it represents a very powerful and well approved

communication system and it can reach millions of households with a single transmission. Broadband (Internet) is used as a complement to provide additional information only, as the reliability of this link might be reduced during a disaster.

The HbbTV solution introduced here is for broadcasting alert messages and the Alert4All project itself will prove to be a considerable advantage when handling and coordinating crises. This will contribute significantly to saving lives and reducing damage.

As part of the activities of the project, the A4A Consortium is carefully considering what the most suitable model for deploying and exploiting the multi-channel alerting system on a large scale is. Any entitled organization would enter into service level agreements with television and radio broadcasters at both an institutional level as well as with individual service providers in each country adopting the warning system. This will ensure effective delivery of the informing and warning messages.

## Acknowledgement

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n° [261732].

## References

- [1] Federal Emergency Management Agency, "IPAWS Open Platform for Emergency Networks", March 2012.
- [2] [http://www.bbk.bund.de/DE/TopThema/TT\\_2011/Warnung\\_gross\\_e\\_Gefahren.html](http://www.bbk.bund.de/DE/TopThema/TT_2011/Warnung_gross_e_Gefahren.html)
- [3] [http://www.evigilo.net/?page\\_id=143](http://www.evigilo.net/?page_id=143)
- [4] <http://www.newswire.ca/en/story/841143/acision-and-one2many-to-deploy-cell-broadcast-at-kpn-for-the-netherlands-nl-alert-programme>
- [5] OASIS (Organization for the Advancement of Structured Information Standards), "Common Alerting Protocol (CAP) Version 1.2", March 2010.
- [6] T. De Cola, J. Mulero Chaves, C. Párraga Niebla, "Designing an Efficient Communications Protocol to Deliver Alert Messages to the Population During Crisis Through GNSS", in Proc. of the 6<sup>th</sup> ASMS and 12<sup>th</sup> SPSC Conference 2012, September 5-7, Baiona, Spain.
- [7] ETSI TS 102 796 (V1.1.1): "Hybrid Broadcast Broadband TV"
- [8] <http://www.alert4all.eu>



**Experience, Inclusion and  
Environmental Responsibility and  
Networked Media Analytics**

# Communicating deictic gestures through handheld multi-touch devices

Clinton Jorge<sup>1</sup>, Jos P. van Leeuwen<sup>2</sup>, Dennis Dams<sup>3</sup>, Jan Bouwen<sup>4</sup>

<sup>1</sup>University of Madeira, Madeira-ITI, Funchal, Portugal;

<sup>2</sup>The Hague University of Applied Sciences, The Hague, Netherlands; <sup>3,4</sup>Bell Labs, Alcatel-Lucent, Antwerp, Belgium

<sup>1</sup>clinton.jorge@m-iti.org, <sup>2</sup>j.p.vleeuwen@hhs.nl, <sup>3</sup>dennis@research.bell-labs.com

<sup>4</sup>jan.bouwen@alcatel-lucent.com

**Abstract:** Deictic gestures are gestures we make during communication to point at objects or persons. Indicative acts of *directing-to* guide the addressee to an object, while *placing-for* acts place an object for the addressee's attention. Commonly used presentation software tools, such as PowerPoint and Keynote, offer ample support for *placing-for* gestures, e.g. slide transitions, progressive disclosure of list items and animations. Such presentation tools, however, do not generally offer adequate support for the *directing-to* indicative act (i.e. pointing gestures). In this paper we argue the value of presenting deictic gestures to a remote audience. Our research approach is threefold: identify indicative acts that are naturally produced by presenters; design tangible gestures for multi-touch surfaces that replicate the intent of those indicative acts; and design a set of graphical effects for remote viewing that best represent these indicative acts for the audience.

**Keywords:** Multi-touch device, deictic gestures, remote presentations, slide presentations, grounding.

## 1 INTRODUCTION

Increasingly, knowledge workers work outside the traditional office, and more and more teams are distributed over multiple physical locations. Teams often communicate their work through PowerPoint (and other) slide presentations. These presentations tend to follow a standardized path: single slides display information with very little interaction from the presenter or audience [4]. In presentations that are attended or viewed online, the remote presentation generally is displayed as one of three situations: solely the slides being presented; the slides and the presenter's voice; and in some occasions with the addition of a video feed of the presenter. Experimental studies have indicated that merely linking spaces through audio-video links does not improve performance to the levels observed between side-by-side collaborators [8].

Communication is a collective activity of the first order. Studies performed by Hindmarsh et al [7], have demonstrated how communication and collaboration depend upon the ability of individuals to invoke and refer to features of their immediate environment. Many

activities within collocated working environments rely upon the participants talking with each other and monitoring each other's conduct. When A speaks to B, A must do more than merely plan and issue utterances while B must do more than just listen and understand. A, must speak only when A acknowledges B is attending, hearing and trying to understand what A is saying, and B must guide A by giving A evidence that B is doing just this [5]. This mutual acknowledgment of understanding between A and B is called *Grounding in Communication*. During a conversation people tend to utter back-channel responders such as "uh huh", "yeah". In Grounding, these confirmations or negations of understanding are named Evidence. Positive evidences become more noticeable while conversing over a telephone or during teleconferencing activities where there is a deficiency of visual cues, such as facial expressions.

Pointing is one of the mechanisms for grounding in communication that require least collaborative effort between the communicating parties. Clark and Brennan [5] argue that deictic gestures combined with communicative statements help establish common understanding and that appropriate gestures that are easily interpreted are preferable over complex sentence constructions. Pointing is a deictic gesture used to reorient the attention of another person so that an object becomes the shared focus of attention. There are four important stages for performing a successful pointing gesture: Mutual orientation; Preparation and staging; Production of the gesture; and Holding (until confirmation) [3].

*Directing-to* and *placing-for* are two basic techniques for indicating [6], *Directing-to* produces a signal that directs the addressee's attention to an object; *placing-for* places an object for the addressee's attention. Graphical user interfaces in computers demonstrate the extended notions of these basic indicating techniques. A click is a virtual form of *directing-to*, and dragging is a virtual form of *placing-for*.

Baecker et al [2] performed studies on a moving point such as a screen cursor and laser pointer that defines the remote person's reference space. Baecker described the results as "giv[ing] them the gestural and referential capability of a fruit fly." Similarly, Kirk et al [8] argue "laser pointers have lower bandwidth for the expression

**Corresponding author:** Clinton Jorge, University of Madeira, Madeira-ITI, Caminho da Penteadá, Funchal 9020-105. Portugal. +351964452305. clinton.jorge@m-iti.org

of gestural information than the direct presentation of hand gestures or sketches.”

In this paper we approach the issue of limited information bandwidth of deictic gestures in remote slide presentations (e.g., pointing with laser pointers) that hinder natural (deictic) communication in these settings. We argue that handheld multi-touch devices are capable of enhancing the representation of a presenter’s deictic gestures without introducing a steep learning curve or high cognitive load. We describe our theoretical framework and present the results of our experiments. We conclude by discussing our design guidelines and future work suggestions.

## 2 STATE-OF-THE-ART

Commonly, the mouse cursor or physical laser pointer are the tools used within collocated presentations as an extension of the performer’s gestures. Figure 1 shows a collocated presentation that was recorded and then broadcast online. They recognized and approached two issues for recording the local presentation for online visualization: how to capture the presenter and the slide projection within the same frame with enough quality to perceive both; and how to capture the presenter’s indicative gestures towards the slide projection. The cameraman positions the presenter to one of the sides of the video frame while the content being discoursed is augmented in the remaining portion of the frame. In this specific scenario the presenter uses a laser pointer to point to referents on the slide. Since the slides are augmented on the video (Figure 1, left), there is no visual feedback to where the presenter is pointing. To repair this detachment between verbal utterances and gestures, the cameraman pans the camera to capture the projected slide presentation (Figure 1, right), thus showing where the presenter’s laser pointer is located. At this point the audience can link verbal utterances to the laser pointer but at the cost of removing the presenter from the frame and viewing the content (slides) at a much lower quality.

Pointing gestures made towards a display (e.g. slide projection) are in general not retrievable at remote sites and participants are unable to tell what object has been pointed at. Lucero et al [9], describe an interactive wall-mounted display named Funky Wall, to support designers in easily conveying messages or ideas in the form of an asynchronous visual presentation. The authors designed four different proximity regions to act as individual interactive triggers. The closest region allowed users to record their gestures by augmenting them onto the content as white translucent streaks. Cheng and Pulo [3] proposed extending the reach of the performer of the gesture with a physical laser pointer, not only for indicative purposes but also as a direct interaction device. The authors argued the form of interaction would thereby reduce the cognitive load of the user and improve users’ mobility while interacting and performing actions. In [11], Tan et al presented a system capable of visually detecting pointing gestures and estimating the 3D pointing direction in real-time. The system offered at best an 88% detection rate and a 75% precision.

Keynote, FuzeMeeting, and other web conferencing tools currently support virtual laser pointers on their tablet applications. This resolves the interaction issues encountered with physical laser pointers but does not address the lack of gestural expressiveness or capture the larger array of presenter gestural intentions.

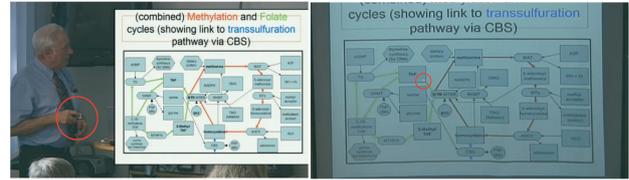


Figure 1: Two distinct repairs performed to enhance audience’s perception and understanding  
source: [http://iaomt.media.fnf.nu/2/skovde\\_2011\\_me\\_kroniskt\\_trotthetsyndrom](http://iaomt.media.fnf.nu/2/skovde_2011_me_kroniskt_trotthetsyndrom)

## 3 RESEARCH FRAMEWORK

We propose a theoretical framework (see Figure 2) for developing support for deictic gestures, which involves two entities: the presenter speaking and using the slides as a visual aid; and the audience to whom the presenter is speaking. The framework represents the presenter’s intention, which is to transmit a message to the audience. The gestures he performs are intentional, for example: directing the audience’s attention to a particular section of the slides. These intentions are exteriorized through gestures (in addition to utterances). The system recognizes the gestures and creates visual representations thereof as an effect for the audience to perceive. The audience then interprets their perception of the effects and creates their own mental model of what the presenter’s intention could be.

The framework is described in further detail during the subsequent subsections and guided our research methodology in this project.

### 3.1 Intent, Gesture, Effect, Perception

This project’s research activities were designed around the four key nodes of the theoretical framework: intent, gesture, effect, and perception.

#### 3.1.1 Intent: Presenter

The *intent* node defines the high-level meaning for the performed gesture. The presenter has an intention and externalizes this by performing a gesture in order to, e.g., direct the audience’s attention to a specific part of a slide. Ideally, the addressees should easily understand the presenter’s intent and act accordingly.

#### 3.1.2 Gesture: Presenter/Computer

The *gesture* node describes interactions gestured by the presenter based on his intentions and captured by the system—the handheld multi-touch device. These gestures may be triggers for events (navigation) or to communicate deictic gestures. In section 4.1 we present an experiment designed to understand what interaction can transform the presenter’s intent into gestures.

### 3.1.3 Effect: Computer

The *effect* node is the result of recognizing the gestures performed by the presenter and translating them into graphical effects displayed to the audience. Different effects are associated to different gestures (and therefore intentions), influencing the audience's interpretation of the effects and thereby of the presenter's intents.

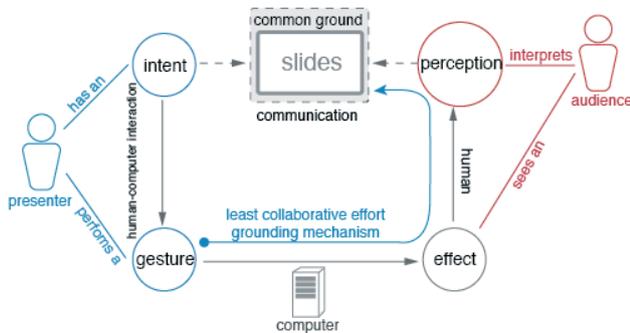


Figure 2. Proposed theoretical framework representing how the presenter through a multi-touch device transmits his intentions to the audience.

### 3.1.4 Perception: Audience

The *perception* node is the result of the effect (the remote representation of the gesture) being perceived and interpreted by the audience whose members create their own mental model of the presenter's intention. It is at this node that the effectiveness and value of our research is evaluated (see sections 4.2 and 4.3).

## 4 USER NEEDS STUDY

In our project we performed small-scale studies on each of the four nodes of the framework, focusing mostly on the *perception* node—in order to understand whether pointing effects added any meaningful information to a remote slide presentation. The latter study (see section 4.2) was designed involving thirteen subjects and performed in a lab setting at The Hague University of Applied Sciences. Subsequent refinements to the user experiment led to an online experiment (section 4.3) that involved nineteen participants.

An initial experiment was carried out that was related to the *intent* and *gesture* nodes of our framework, this experiment is described next.

### 4.1 Mapping Intentions to Gestures

The objective for this user experiment was to understand the connection between some common gestural intentions identified through observations and literature reviewing. The experiment, required subjects to perform the first gesture that came to mind when the researcher read out a pre defined "intent" (e.g., "point out the second bullet point"). A list of intents was created for each of four slides shown, where each intent required the subject to perform a gesture. The intents are categorized as being *pointing*, *indicating*, *highlighting*, or *grouping*. Subjects were seated in front of an iPad displaying a single slide in full screen running on the drawing application Adobe Ideas. Interactions were recorded, overlaying the

displayed slide with a pen tool (50 pixel (similar size to finger tip) 50% transparency and red in color).

### 4.1.1 Findings

Twelve subjects participated in the experiment held in a lab environment. Subjects worked at Bell Labs in technology related positions and were over 35 years old. Four were novices and never used an iPad or multi-touch device, eight owned iPhones or were familiar with the technology. The results were analyzed individually and then compared to identify similarities or patterns.

A total of 134 gestures were recorded and observed. 31.34% of all recorded gestures were 1-finger pointing gestures (e.g., tap or touch on the device). 17.91% of all recorded gestures were grouping 1-finger gestures (such as circular gestures). For *pointing*, 11/12 subjects performed an index-finger indicative gesture equivalent, a tap or touch. For *indicating*, 9/12 subjects perceived this intent to be similar to pointing and performed an equivalent tap or touch gesture. For *highlighting*, often interpreted as a persistence technique using a semi-transparent coloring tool, 8/12 performed 1-finger dragged gestures to highlight text and 7/12 subjects performed a circular gesture to highlight individual artifacts. For *grouping*, 9/12 subjects grouped objects with a circular gesture.

Similar results (gestures) were found in Lucero et al's experiment [9] and can be categorized as "standardized multi-touch gestures".

We found that experienced users tend to simplify gestures, while novice users perform more personal, embodied gestures and techniques—especially for highlighting and relating content on paper.

### 4.2 Personal Perceptions of Pointing

The designed experiment required test subjects to view three video presentations on a laptop. The Repertory Grid Technique (Kelly 1955) [1] was used to elicit subjects' personal constructs (perceptions) and scoring without researcher bias, and was followed-up with a semi-structured interview. Each experiment required around 45 minutes to complete (depending on the interview).

Subjects viewed three videos subsequently, each a part of the same presentation. Each video was shown in a different visualization style, randomly ordered: slides and audio (A); slides and audio with an additional video feed of the presenter (V); and slides and audio combined with a virtual laser pointer – representing gestures (P).

Having viewed the three videos, subjects were asked to choose two presentation styles and compare these to the third, writing down the similarities or differences in their experiences, in the form of constructs. This was repeated for all possible combinations. Subjects then scored the three styles for each of these constructs, on a 7-point Likert scale.

The experiment was held at The Hague University of Applied Sciences over the course of a day. Thirteen subjects participated in the experiment, including students in design and engineering as well as professors.

### 4.2.1 Findings

Eight male and five female, subjects participated in the Repertory Grid (RGT) experiment and generated 96 construct pairs (e.g., “helps concentrate versus distracting”). These constructs were analyzed and subject preference (for a single presentation variant) was obtained based on the sum of scores: highest as the preferred presentation variant.

These participants scored the three variants as follows:

- 5/13 scored slides, audio and pointer (P) highest
- 5/13 scored slides, audio and video (V) highest
- 3/13 scored slides and audio (A) highest

Key differences were found between male subjects, who preferred the pointing (five-out-of-eight, 5/8) and disliked the video, and female subjects, who preferred the opposite (4/5). These results were consistent with the outcome of the semi-structured interviews that followed with each subject (see following subsections).

### 4.2.2 Male Subjects

Eight male subjects took part of the semi-structured interviews. During the interviews subjects were not bound to the three presentation variants thus 4/8 subject commented on preferring the combination of pointing and video (VP) (a style not included in the study). The interviews confirmed the disliking of the slide and audio (A) 1/8 and video (V) 1/8 variant. The pointing (P) variant received highest score of the displayed variants in the experiment with 2/8. Comments (6/8) about the pointing (P) included how pointing helped them “think like the presenter,” because their “eyes are guided through the constructions” and “pointing directs you to important stuff on the slides.” Two-out-of-eight subjects did not see the immediate benefit of pointing in remote presentations.

### 4.2.3 Female Subjects

Five female subjects took part of the semi-structured interviews. Female interview results were consistent with the RGT experiment. Four-out-of-five (4/5) female subjects preferred the video (V) variant while 1/5 preferred audio (A). Similar to the male interviews, 2/5 expressed preference for pointing and video integrated (this style was not included in the experiment). One subject commented on how pointing (P) was useful while three found pointing useful only for complicated or complex presentations, when guidance is needed.

During the semi-structured interview 4/5 of females preferred pointing for complicated presentations. They commented on the visual and kinetic aspect of the pointing cursor, that the drag effect was distracting and the motion erratic.

### 4.2.4 Discussion

Three-out-of-thirteen subjects that disliked the pointing (P) variant were professors. They commented on not liking to be guided and how they preferred to think for themselves.

During the interviews these similar comments arose on how pointing helped better understand the content and the

thought process of the presenter in more complex scenarios such as graphs.

## 4.3 Significance of Pointing in Presentations

Another, online, experiment was performed with the objective of expanding on the findings of the previous studies. By refining the videos (shorter duration) and the pointing effect (improved effect and movement) we aimed to find further evidence of the benefits of pointing in remote presentation scenarios.

In this experiment, subjects accessed a webpage to view the three video presentation variants, again in randomly ordered styles. Skipping videos or parts of the video was disabled. Subjects were then asked to score each presentation style based on constructs resulting from the previous user study. The audio variant (A) was replaced by video and pointing (VP). None of the previous subjects participated in the online experiment.

### 4.3.1 Findings

Nineteen subjects completed the online experiment; eight females and eleven males, aged between 21 and 51. Subjects were recruited from three universities: University of Madeira, Eindhoven University of Technology and The Hague University of Applied Sciences.

No male subject preferred the video and slides style (V), while 4/11 preferred the pointing style (P). The combination of pointing and video (VP) scored the highest with 7/11. Interestingly, only one female subject preferred video and slides (V), while 4/8 (50%) of female subjects scored the pointing style (P) the highest. 3/8 preferred the combination of pointing and video. The contradiction in the female results with the previous study is remarkable. The female subjects seem very susceptible to the pleasantness of the effect and movement of the pointing cursor. These two attributes were refined for this study and the pointing was used only when required with the deictic utterances. Video was clearly less scored with only 1 out of 19 subjects preferring it. Video and Pointing scored the highest with 10 out of 19, while Pointing appeared second best, with 8 out of 19.

From our analysis of the scores on constructs, it appears that pointing (P) helped subjects to concentrate (high scores on the construct *concentrate*), while video (V) did not. Also, the combination of pointing and video scored low on the aspect of concentration, meaning that the added video is experienced as distracting. The same negative effect of adding video to pointing leads to reduced scores for *helpful* and *better understanding*.

*Emotional, personal and presence* constructs were scored lowest for pointing (P) with some exceptions of individual high scores. When analyzing the combination of video and pointing, these constructs—that scored highest in the video style (V)—suffer little to no reduction in their scoring. This led us to conclude that, while pointing does not add as much social presence, personal information and emotion as the video feed of the presenter does, it also does not negatively affect the

qualities in the presentation as adding video does for the *concentration* construct.

## 5 DESIGN GUIDELINES

The *persistence* of the visual effects augmented to slides during presentations influences how the audience perceives gestures. Regarding the persistence as a spectrum, running from transient to persistent, we identified artifacts for both extremities of this spectrum. At the most transient extreme, the mouse cursor and laser pointer are located, which convey very little information (current location only). At the other extreme, we find persistent graphics, e.g. notes, highlights and annotations. These artifacts convey increased information but their persistency may not at all times be useful. Our contribution is to the intermediate spectrum that has not been fully explored: between transient, user cancelled events and slide exposure duration.

Indicative gestures are related in time to utterances and to referents (objects), thus no pointing cursor should exceed the duration of a slide exposure or be too transient to be missed due to late glances by attendees. We propose the following gestures and effects (pointing cursors, see Figure 3) for some of the most common gestural intents identified in slide presentations.

The *touch* cursor is similar to the laser pointer. It allows for referencing a single referent easily by moving around or by tapping at a location. The ripple effect provides an “epicenter-like” event, and provides a brief persistency, enough for late glancing addressees to view.

The *drag* cursor leaves behind a trail similar to a heat surface concept. This should allow for late glances to get enough feedback to follow the presenter’s chain of thought throughout the slide and easily identify past referents and present ones.

The *sticky* cursor derives from the notion of the fourth stage of deictic pointing: holding. A little wiggle gesture places a cursor (a fingerprint) remaining there until the user cancels it or until the end of the slide exposure; no continuous interaction is needed. Multiple objects can be referenced through multiple sticky cursors with different colors or shapes.

The *region* cursor surrounds a group of objects or an area of the slide, whereas the *shape* cursor (a repetition of the same gesture) highlights that area and is more persistent (during slide exposure). The *highlight* cursor is a two-finger gesture for highlighting text.

We argue that these effects should represent the majority of presenters’ deictic gesturing needs and subsequently aid addressees’ focus attention and follow the presenter especially in more complex (visual) slides.

| CURSOR NAME         | GESTURE   | DESCRIPTION   |
|---------------------|---|---|
| touch cursor        |  | <br>ripple effect                |
| drag cursor         |  | <br>beginning touch      present |
| multi-sticky cursor |  | no time lapse effect<br>persistent until user cancels   |
| shape cursor        |  | highlight tool<br>persistent until user cancels   |
| region cursor       |  | transient region focus grouping   |
| highlight cursor    |  |  text highlight                  |

Figure 3. Gestures and effects for the most common gestural intentions in slide presentations.

## 6 DISCUSSION

Through the work presented here we argue that handheld multi-touch devices are a low-cost solution capable of facilitating deictic gestures. Our designs support more gestural intentions than the common laser pointer and mouse cursor thus increasing the expressiveness of the gestures in these communicative activities. While the cursors presented here have not yet been subject to user testing, we expect to find that they are representative of user intentions. The drag cursor could work in conjunction with face or eye gaze tracking software used for single person audience, allowing the system to recognize when the remote user is not looking at the presentation, triggering the drag effect.

Our studies show that pointing is considered a helpful tool for addressees in concentrating and understanding a presentation – in particular remote, distributed presentations. A combination of pointing with a video feed of the presenter provides the best of both worlds for some individuals. Our studies also indicate a substantial variation in the appreciation of these tools; a result that is not unusual when analyzing sex difference data from experiments [10]. This suggests that options to disable and show each one of these modal communication tools would be required.

## 7 FUTURE WORK

Our first prototype multi-touch app for presentations to remote audiences does not support all designed cursors. Future work would involve implementation of our cursor designs and further user studies for confirming and refining the gestures and the related effects. Deployment in real work environments would provide invaluable feedback from the presenters’ perspective.

## 8 CONCLUSION

This paper focused on remote slide presentations and the lack of gestural expressiveness perceived by remote audiences. Deictic gestures are part of our natural language and are not fully supported in these scenarios.

Our objective was to explore this issue and present some guidelines to aid future research in the area. We present six cursor designs (deictic gesture representations) that we argue are representative of most deictic gestures and can be captured on a handheld multi-touch device.

## 9 ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. ICT-2011-287760. This project received financial and operational support from Alcatel-Lucent Bell Labs, Belgium and Madeira Interactive Technologies Institute, Portugal, and operational support from The Hague University of Applied Sciences. We thank all participating subjects.

## 10 REFERENCES

- [1] Alexander, P. and Loggarenberg, J. Van. The repertory grid: discovering a 50-year-old research technique. ... *of the 2005 annual research ...*, (2005), 192 – 199.
- [2] Baecker, R., Harrison, S., and Buxton, B. Media spaces: past visions, current realities, future promise. *CHI&#39;08 extended*, (2008), 2245–2248.
- [3] Cheng, K. and Pulo, K. Direct interaction with large-scale display systems using infrared laser tracking devices. *In Proceedings of the Asia-Pacific symposium on Information visualisation - Volume 24 (APVis '03)*, Tim Pattison and Bruce Thomas (Eds.), Vol. 24. Australian Computer Society, Inc., Darlinghurst, Australia, Australia, 67-74.,
- [4] Chiu, P., Liu, Q., Boreczky, J., et al. Manipulating and annotating slides in a multi-display environment. *Proceedings of INTERACT*, Citeseer (2003), 2.
- [5] Clark, H. and Brennan, S. grounding in communication. *Perspectives on socially shared cognition*, (1991).
- [6] Clark, H. 2003. "Pointing and placing." In: Kita S. Pointing: A foundational building block of human communication. *Pointing: Where Language, Culture, and Cognition Meet*, Lawrence Erlbaum Associates, Inc., Mahwah, New Jersey 1 (2003): 1-8., (2003).
- [7] Hindmarsh, J.O.N., Benford, S., and Greenhalgh, C. Object-Focused Interaction in Collaborative Virtual Environments University of Nottingham. *Design 7*, 4 (2001), 477–509.
- [8] Kirk, D.S. and Fraser, D.S. The effects of remote gesturing on distance instruction. *Proceedings of th 2005 conference on Computer support for collaborative learning: learning 2005: the next 10 years!*, International Society of the Learning Sciences (2005), 301–310.
- [9] Lucero, A., Aliakseyeu, D., Overbeeke, K., and Martens, J.-B. An interactive support tool to convey the intended message in asynchronous presentations. *Proceedings of the International Conference on Advances in Computer Entertainment Technology - ACE '09*, (2009), 11.
- [10] Passig, D. and Levin, H. Gender preferences for multimedia interfaces. *Journal of Computer Assisted Learning 16*, 1 (2001), 64–71.
- [11] Tan, K., Gelb, D., Samadani, R., Robinson, I., Culbertson, B., and Apostolopoulos, J. Gaze Awareness and Interaction Support in Presentations. *ACM MUltimedia*, (2010), 643–646.

# Automatic 3DTV Quality Assessment Based On Depth Perception Analysis

J.A. Rodrigo<sup>1</sup>, J. P. López<sup>2</sup>, D. Jiménez<sup>3</sup>, J.M. Menéndez<sup>4</sup>

<sup>1,2,3,4</sup>Escuela Técnica Superior de Ingenieros de Telecomunicación, Universidad Politécnica de Madrid, Madrid, Spain

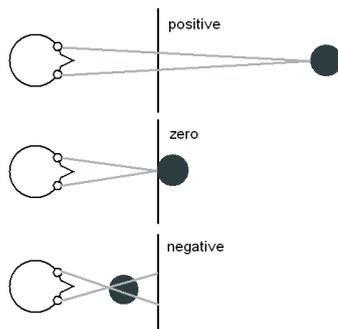
<sup>1</sup>jrf@gatv.ssr.upm.es, <sup>2</sup>jlv@gatv.ssr.upm.es, <sup>3</sup>djb@gatv.ssr.upm.es, <sup>4</sup>jmm@gatv.ssr.upm.es

**Abstract:** Quality assessment is a key factor for stereoscopic 3D video content as some observers are affected by visual discomfort in the eye when viewing 3D video, especially when combining positive and negative parallax with fast motion. In this paper, we propose techniques to assess objective quality related to motion and depth maps, which facilitate depth perception analysis. Subjective tests were carried out in order to understand the source of the problem. Motion is an important feature affecting 3D experience but also often the cause of visual discomfort. The automatic algorithm developed tries to quantify the impact on viewer experience when common cases of discomfort occur, such as high-motion sequences, scene changes with abrupt parallax changes, or complete absence of stereoscopy, with a goal of preventing the viewer from having a bad stereoscopic experience.

**Keywords:** 3DTV, depth maps, zone of comfort, VQA, motion estimation, parallax.

## 1 INTRODUCTION

3D content is claiming importance in media environment. The success of new 3D services is a reality due to the improvement in technology, but visual comfort analysis is demanded. Quality issues are currently a bigger concern in 3D media than they were in traditional media. Although there are some impact factors and initial measurement methods in this field, there is still no common way and procedure to compare 3D video content and integrated solutions and obtain an evaluation of quality.



**Figure 1. Parallax comparison**

Stereoscopic 3D video perception is based on the fact that two different video signals are captured in order to feed to each of the viewer's eyes; recreating the experience of watching a real world scene, where two different images are captured by each eye and the difference between them depends on the position of the elements in the world related to the viewer's position. This means that the system is feeding the observer with a

disparity depth cue. Parallax created by disparity is determined by the virtual perceived location of the objects in a scene, as shown in Figure 1.

Watching 3DTV is significantly different from a natural view, as the point of view is prefixed by the fixed point of view of the camera lenses that have captured the scene, and is therefore the focus. Furthermore, in natural viewing, the eyes focus (accommodate) and converge to the same distance, but when looking at a 3D object displayed on a screen, a viewer's eyes must focus on the screen for a while, and at the same time, they converge on a point in space that may be located beyond the screen, on the screen, or in front of the screen. This is known as the vergence-accommodation conflict. This conflict limits the amount of parallax that a viewer can tolerate without suffering visual discomfort, also known as the Zone of Comfort [13].

This paper aims to study the effects of stereoscopic disparity in quality assessment through the analysis of depth maps of a sequence and its temporal evolution. We try to quantify objectively the effects of parallax, depth and motion, exporting the common situations in which discomfort is substantial, from opinions of observers derived from empirical and subjective tests.

Following, related studies are presented in Section 1 and the description of subjective assessment developed is compiled in Section 3. The proposed method is defined in Section 4. The test results are given in Section 5 and conclusions in Section 6.

## 2 RELATED WORK

Video quality assessment is a difficult process which plays a major role in various processing applications [1]. A lot of work has been developed in this field, defining metrics and algorithms to predict the quality of a video sequence. An overview of the extensive and most interesting work in quality assessment is collected in [2], [3] and [4]. This work is always related to subjective quality assessment and most of the published adhere to the procedures contained in the Recommendation ITU-R BT.500 [5].

In 3D media, new factors related to the optic effect of stereoscopy are concerned in order to assess quality; such as visual discomfort or perceptual inconsistencies between depth cues, as stated in [12]. Much work has also been developed in this field relating depth and motion, such as [6], where filtering is used to reduce visual discomfort on screens.

In [7], an overview describing the main topics relevant to comfort in viewing stereoscopic television is developed, analyzed after subjective tests, related to accommodation-vergence conflict, parallax distribution, binocular mismatches,

**Corresponding author:** J. A. Rodrigo, Universidad Politécnica de Madrid, Madrid, Spain, +34 913 367 344, jrf@gatv.ssr.upm.es

depth, and cognitive inconsistencies. In [8], it is reported that depth and motion are closely related in terms of calculating visual discomfort. And [9] offers a visual comfort model for detecting a salient object's motion features in depth of field.

An interesting subjective evaluation of visual discomfort is developed in [10], where parallax limits and regions of comfort, dependent on the screen size, disparity and viewing time, are obtained. Other artifacts such as stereo window violation (SWV) and temporal continuity of the disparity (TCD) have been studied in [11] where guidelines to create comfortable and faster stereoscopic films are included.

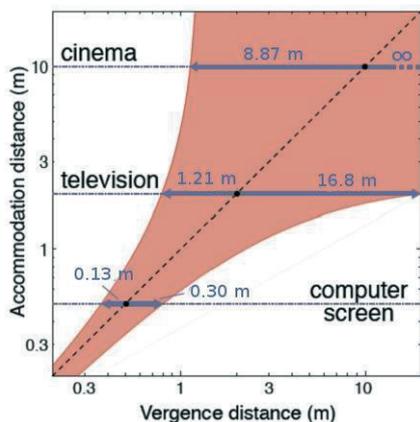


Figure 2. Shibata's Zone of Comfort

The Zone of Comfort (ZoC) was first introduced by Percival [14]. He suggested the limits to vergence-accommodation postures that could be achieved without causing discomfort. More recent studies such as Shibata et al. [15] concludes that the ZoC may differ from Percival's, when the experiments are based on stereoscopic vision rather than on vision through spectacles. In stereo vision the vergence-accommodation conflict constantly changes, while in a lens or spectacles system it is maintained fixed [13]. Figure 2 shows Shibata's ZoC for different accommodation distances in stereo vision. According to this diagram, images with positive parallax have little-to-no capability to induce discomfort, while negative parallax is most likely to cause discomfort if not controlled.

In order to adapt a stereoscopic ZoC to 3D video, it is necessary to take into account motion and time of exposure in a stereo scene. The ZoC will be further reduced when these elements appear. The time to converge and accommodate in this case is relevant, thus there is a need to adapt the concept of ZoC. In [10] the variation of time of exposure is studied in order to determine its effects on visual discomfort.

Determining an image parallax range uses an associated depth map. There are several ways to obtain it from a stereo image, depending on computing complexity and accuracy restrictions. As a rule of thumb it can be stated that complexity is proportional to accuracy, thus, low-complexity algorithms such as Sum of Absolute Differences (SAD) can typically perform well, as stated in [16]. SAD-based algorithms are between the least-complex and more often-used. Census-based algorithms are common in real-time hardware-based systems and may work better in homogeneous zones of the

image. Its complexity increases when used in software-based systems because of its bit-based nature. Some systems mix both algorithms in order to obtain the best results from each one of them.

### 3 SUBJECTIVE 3D EVALUATION

Tests have been run over a set of 3D video sequences to understand and analyze different features which generate visual discomfort or quality reduction. A group of 16 observers were asked to rank the sequences taking into account their 3D quality. Results were compared to the objective data obtained through our developed tools to decide which features would be a possible cause of visual discomfort and how to modify them to obtain good 3D experiences. All tests were carried out on a 46" screen with passive glasses at the recommended distance.



Figure 3. Example of sequence "Itaca 3D"

Sequences used for the assessment included a 3-minute sequence called "Modernism" (Figure 4), created by Mediapro, in which different scenes appeared with different levels of motion and depth, sequences "Rain Fruits" and "Fountain" from EBU were also used [17], as well as synthetic test sequences "Palco HD" and "Itaca 3D" (Figure 3), which include parallax and object distances variation experiments created by us. All these sequences are high-definition resolution (1920x1080) with no compression, available in side-by-side formats.



Figure 4. Examples of frames in sequence "Modernism"

#### 3.1 Cases of Study

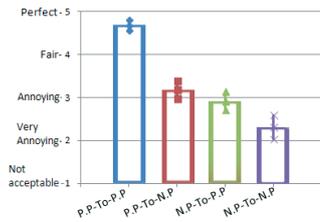
As a conclusion of opinions obtained after subjective assessment with sequences with variations of parallax (positive and negative), motion and scene changes; different cases of study can be isolated. Different experiments were developed:

- Pairs of sequences with transitions from different types of parallax, negative and positive, to detect the impact over abrupt stereoscopic changes. A "Positive parallax" sequence (P.P) is considered when it has not remarkable negative parallax and pixels with positive parallax represent more than 25% of an image. On the other side, "Negative Parallax" sequences (N.P.) are sequences whose images posses more than 15% of pixels in negative parallax (assuming an environment of positive parallax). See Figure 5 for results.

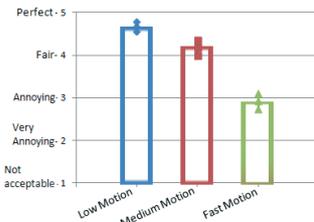
- Negative parallax sequence with different levels of motion: low, medium and high motion. Test statistics are collected in Figure 6.

- Sequences with window violation (W.V) produced in different sides of the image, in lateral or top/bottom regions of the image. See subjective results in Figure 7.

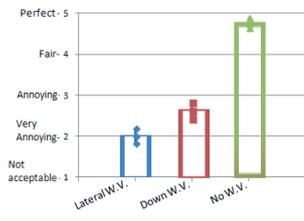
- Long sequence with soft variation of parallax, at the end the sequence starts from the beginning producing an abrupt parallax change. Results from this experiment are in Figure 8.



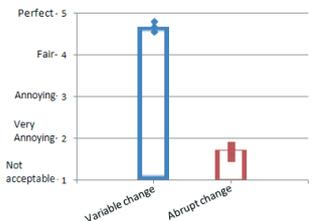
**Figure 5. Transitions between types of parallax**



**Figure 6. Results in impact related to motion**



**Figure 7. Sequences with Window Violation (W.V)**



**Figure 8. Progressive or abrupt parallax variations**

## 4 WORK IMPLEMENTATION

In this section, the work developed is described in two subsections. First (Figure 9.a), the tools implemented in order to obtain quality through depth map histograms, calculating degradations related to each individual frame, are described in detail. In the second part (Figure 9.b), the work for static images is extended to sequences, analyzing video motion and the effect of depth when there are variations in parallax, derived from depth maps. Also some cases of study are analyzed when combining depth and motion.

### 4.1 Quality in Static Images Using Depth Map Histograms

To resolve the validity of a stereoscopic image it is required to determine whether it delivers visual discomfort or annoyance. The developed algorithm obtains parallax information through the computation of a depth map.

First of all, the depth map histogram is compared to the suited ZoC, in order to check if it doesn't fall out of its boundaries. Vergence-Accommodation conflict needs to be confined between ZoC limits to prevent visual discomfort.

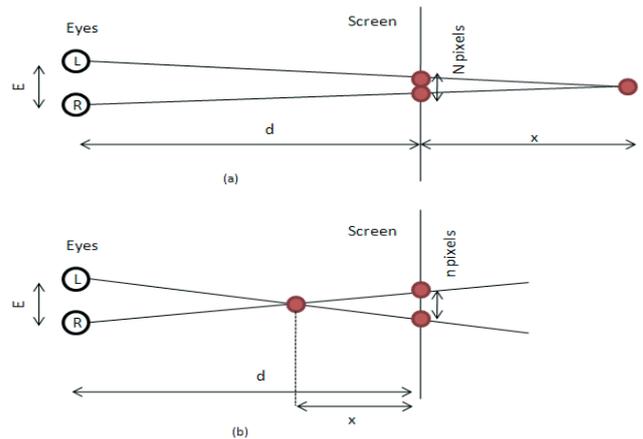
In order to evaluate the disparity results it is necessary to understand the relation between disparity in pixels and virtual perception of depth. Figure 5 shows the trigonometric relations between the observer's location, an object's depth perception and its disparity measured in pixels. The relation

between  $d$  and  $x$  is not a fixed value, because the distance between the eyes doesn't change with the screen's size, thus it has to be assumed a size for the screen.

Assuming a display of diagonal size  $D = 46''$  and with an aspect ratio (AR) of 16/9 and 1920x1080 resolution, screen width ( $W$ ) would be 101 cm, therefore  $R=101\text{cm}/1920\text{pixels}=0.053 \text{ cm/pixel}$ . Assuming  $E=6.5 \text{ cm}$  and  $d=2.4 \text{ m}$ , that would leave us with a parallax range limited to  $[-125, 107]$  for Shibata's ZoC, measured in pixels. Anything relevant that the algorithm finds out of those bounds will be considered as a cause of visual discomfort.

The other feature measured the window violation which is another suspect of causing visual annoyance. Window violation occurs when an object with negative parallax doesn't fit the screen and, therefore, is cut by the screen edges. Having negative parallax, it is supposed to be out of the screen, which means that screen edges shouldn't be able to hide its view. This generates an incoherent depth cue situation.

In order to measure this feature, the algorithm will examine the depth map's limits looking for negative parallaxes, which will be computed as a factor of visual annoyance.



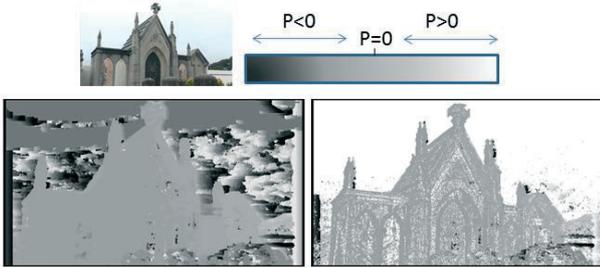
**Figure 9. Positive (a) and negative (b) parallax estimation**

#### 4.1.1 Depth Map Calculation

To compute depth maps from stereoscopic images, the system performs a SAD-based algorithm. We need to obtain general depth characteristics of a scene and its evolution, though pixel depth accuracy in the whole image is not necessary. SAD-based algorithms work well enough to fulfill our goal and are less computationally demanding.

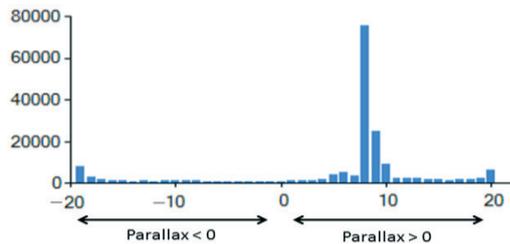
The weakest detections with SAD algorithms occur in homogeneous zones, where the capability to discern between possible pair candidates is low. In order to alleviate these probable errors, the system creates a difference between both views in order to calculate depths only over those pixels that will differ from one image to another, reducing homogeneous zones and, therefore, noise in resulting depth maps. Discarded pixels won't be taken into account. Figure 10 shows the original depth map (left) and the filtered depth map (right). In the original depth map there are several errors in the background zone, where the sky is homogeneous. In the left image this zone isn't calculated and therefore not taken into

account to classify the image's general depth, improving the absence of errors in the histogram.



**Figure 10. Advanced Depth Map**

Figure 11 shows the histogram calculated for the previous depth map. All the elements in the scene have positive parallax. There is a very small amount of negative parallax pixels which represent noise (bad information) that result from the depth map algorithm calculation.



**Figure 11. Example of parallax histogram**

## 4.2 Depth and Motion QoE Decision Algorithm

In 3D stereoscopic video, motion is a basic element to take into account when assessing quality, as it is a primal reason for visual discomfort and is related to high depth levels which combine areas with negative and positive parallax.

The steps to follow for formulating a QoE decision are the ones which follow. Firstly, the complete sequence is processed to obtain the motion vectors in order to find the scene changes. In the exact moment where scene change occurs, the motion vectors are calculated between consequent frames to obtain the level of motion in that specific scene.

Depending on the level of motion, the scene is classified as slow, medium or fast motion. This is necessary to decide the necessity of calculating new depth maps for various frames if it is fast motion, or assuming same disparity for a collection of similar frames, saving time of computation.

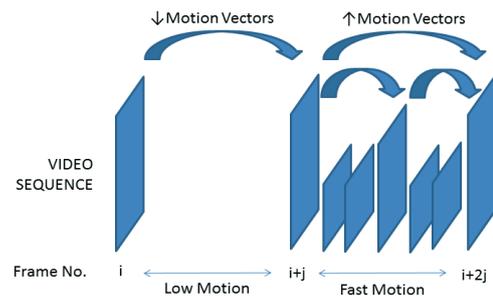
After deciding the key-frames (one or more), depth maps are obtained for each of these frames by making use of the difference in images between left and right view, which is used as a mask to simplify the process. Depth map is calculated as explained in the previous section about static images. The comparison between parallax histograms, derived from each key-frame, allow us to make a statistic about the variations in objects depth and, consequently, quantify the probability of visual discomfort appearance.

### 4.2.1 Motion Vectors and Motion Estimation

The work derived from the static image process is related to motion. It is necessary to evaluate the motion level in a video

sequence, to conclude how much this motion affects the perception of the third dimension in stereoscopic video. For this purpose, the motion vectors calculation is obtained.

The whole sequence is processed in order to detect frames where motion is produced. In consecutive static frames or areas with low motion, the depth map is assumed to be the same for that sequence of frames. When medium or fast motion happens, more depth map information is necessary to compare results. Motion is calculated through average motion vectors in sequence (Figure 12). Motion is calculated as the average valid motion vectors (without discarding incoherent ones), always related to the variance of motion lengths.



**Figure 12. Video sequence motion analysis**

For motion vectors calculation, only the left stereoscopic image is selected, and a grid is created to detect the block motion, in the case of the example a grid with 3 lines and 5 columns allow us to obtain 15 different motion vectors. The blocks between 9x9 and 15x15 pixels are searched in the next frame left image, homogeneous blocks are discarded to avoid false detections. The motion must be coherent in distance, so vectors with length values over two times the variance are also discarded. The final average motion vector length either reveals if the image is static or the corresponding motion level (low, medium or fast) related to the objects.

The last case is when a scene change occurs. Then depth maps from both previous and next are processed. This is a concrete case of motion vector abrupt variation, in which the variance of the vectors length is higher than when fast motion happens. As manifested from observers, the abrupt changes of negative parallax in a positive/negative parallax environment provoke a high visual discomfort in the observer's eye. Discomfort is usually produced in environments with significant-variance of negative parallax and motion, even with low and medium motion, and especially in fast motion sequences.

## 5 TEST RESULTS

With the results obtained from subjective assessment, studies were developed in static images and motion video sequences.

### 5.1 Results in Static Images

Tests were run over still images to classify stereo features without dealing with motion effects. Tests were focused on ZoC measurements, window violations and depth distribution.

To evaluate the effects of parallax out of ZoC we have rendered virtual images such as the one showed in Figure 3. When disparity was forced to be near Shibata's ZoC the

perception of the observers was negative, even when disparity fell below 70% of ZoC range. In order to secure a good comprehension of the scene, the threshold was fixed at 2/3 of the total of Shibata's ZoC. Further away the vergence-accommodation conflict was found to be nearly unsolvable or, at least, it took a lot of time to be solved. This effect of time will be dealt with in subsequent sections. Outside of the ZoC violations were easily detected by the algorithm analyzing the resultant parallax histogram. Other still images from the sequence (Figure 4) were used to quantify window violation cue conflict. During that sequence, the text is turning and, from time to time, some of that text crosses the screen's limits. As the text has a negative parallax, it should never touch the borders. From the tests results, it was determined that window violation cue conflict became difficult to overcome when at least 20% of the screen edge was filled with negative parallax pixels. Again, we were able to detect window violations measuring positive parallax pixels over the edges from computed depth maps results.

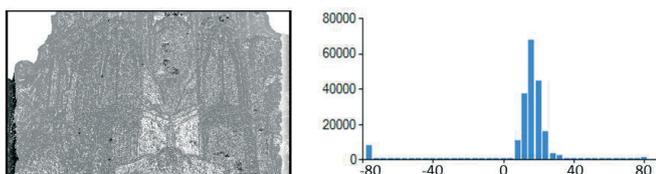
The last still image test was related to QoE rather than annoyance or discomfort. In this case, a set of images were ranked for their 3D effectiveness. Results were compared to their depth map histogram distribution. Figure 13 to Figure 16 show depth maps and histograms of the images submitted to test. Table 1 holds variance statistics for all the images tested. Note that histogram value for -80 pixels always shows a peak. This peak is considered as noise related to depth map calculation techniques and will not be taken into account when statistics are calculated.

The 3D perceiving the church and the cemetery images, observers usually prefer the second image because there's a wider range of depth. This is statistically measured as a bigger positive parallax variance. The "Library" and the "table" were found to be the preferred images due to its variety of depths, from positive to negative parallax.

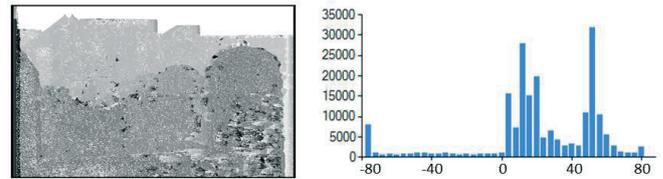
**Table 1: Histogram variances**

| Image | Positive Variance (pix) | Negative Variance (pix) |
|-------|-------------------------|-------------------------|
| 1     | 10                      | -                       |
| 2     | 19,6                    | -                       |
| 3     | 12,8                    | 23,2                    |
| 4     | 15,6                    | 16,4                    |

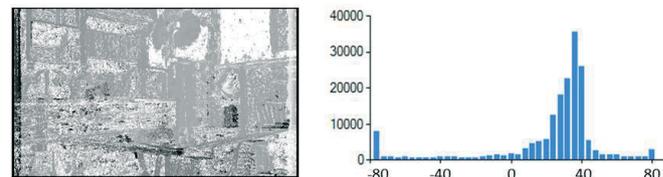
All these tests revealed an interest in depth variance and negative parallax because of their higher immersive capabilities. Our developed tools confirmed these features in each one of the images through statistical depth analysis, which led us to believe the system is well-suited to detect possible indications of 3D quality of experience through objective analysis of still images.



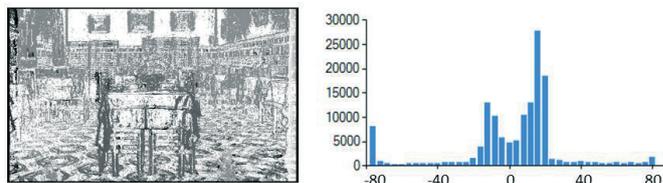
**Figure 13. Church depth map and histogram**



**Figure 14. Cemetery depth map and histogram**



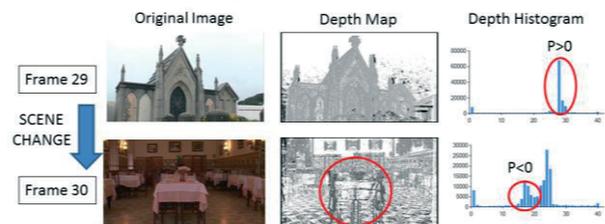
**Figure 15. Library depth map and histogram**



**Figure 16. Table depth map and histogram**

## 5.2 Depth and Motion QoE Decision Algorithm

First of all, the scene changes were analyzed with different variations of negative parallax in an environment of positive high-variance. As seen in Figure 17 the depth map and histogram are calculated, and their related statistics evaluated, to detect the scene changes.



**Figure 17. Example of scene change detection**

**Table 2: Positive and negative parallax from histogram analysis**

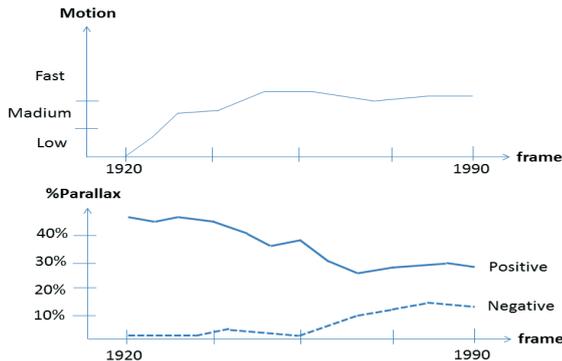
| Frame | Positive Parallax (% pixels) | Negative Parallax (% pixels) |
|-------|------------------------------|------------------------------|
| 29    | 44,26%                       | 2,60%-                       |
| 30    | 31,58%                       | 18,15%                       |

The variation of negative parallax from frame 29 to the next one is more than 15%, taking into account that, although the negative parallax in frame 29 nears zero, the positive parallax is very significant, with a score of more than 25%, which means that there is a high probability of detecting visual discomfort, as observers manifested in subjective tests, which need time to focus the objects in negative areas. Similar results have been obtained with scene changes with variations of negative parallax higher than 10%. Tests related to motion with high parallax variation offers similar results.



**Figure 18. Frames with high negative parallax**

Fast motion is detected in some sequences when the motion vectors reveal a movement higher than 2 pixels per frame, as happens in Figure 18, which shows the camera making a “travelling” fast movement.



**Figure 19. Evolution of motion and parallax percentage between frames 1920 and 1990**

Figure 19 shows both negative and positive parallax percentage in parallel to motion description. It is remarkable that an abrupt increase in negative parallax is not enough for visual discomfort to be detected. It is necessary to create an environment with parallax variance and motion. The probability of discomfort is higher with faster motion.

## 6 CONCLUSIONS

Depth and motion are main factors in perceived quality of experience. Information provided by depth maps and estimated motion vectors is useful to avoid effects that can cause visual discomfort and fatigue in observers when contemplating 3d stereoscopic contents.

Subjective assessment allowed us to isolate the main features to be detected, in order to perform an algorithm which could translate user’s opinions into an automatic objective system.

The presence of objects with negative parallax on a static image, and especially when motion is detected in the video sequence which contains that image, requires quantifying the probability of the observer’s annoyance. This information can be obtained through depth maps, motion vectors and parallax histograms. In graphics comparing parallax and motion evolution the relation between both parameters in the final experience of users is remarkable. Previous Zone of Comfort (Zoc) studies have been found to be greatly affected by motion and time of viewing, diminishing its range significantly. Parallax getting near the ZoC edges (especially negative) has been proven to be undesirable when fast motion or high parallax variance appeared.

Tests that have been developed showed good results when applying the techniques to video sequences that contain effects which could be considered annoying for the human eye. Results obtained offer guidelines for stereoscopic video creation, extracting probabilities of visual discomfort and fatigue and reaching consensus between 3D sensationalism and annoyance to the observer’s eye. Nevertheless, the user has the final decision to accept or reject a determined content.

## Acknowledgement

This paper is based on work performed in the framework of the project 3D-Contournet with research in techniques to assess quality in stereoscopic video. The work is also related to Immersive TV public funding project, headed by Indra Company and in collaboration with Mediapro, with the objective of developing an immersive environment with the use of CAVE and stereoscopic screens, and by the project TEC2012-38402-C04-01 HORFI, as well. We would like to acknowledge Jordi Alonso and people from Mediapro for lending 3D stereoscopic video contents with variations of parallax, available for the test development.

## References

- [1] Z.Wang, A.C.Bovik, and L.Lu, “Why is image quality assessment so difficult?,” IEEE International Conference on Acoustics, Speech, & Signal Processing, vol. 4, pp. 3313-3316, May 2002.
- [2] S. Winkler. “Digital Video Quality: Vision Models and Metrics”. Ed. Wiley. March 2013. ISBN-13: 978-0470024041
- [3] H.R. Wu, K.R. Rao, “Digital Video Image Quality and Perceptual Coding (Signal Processing and Communications)”. CRC Press. November 2005. ISBN-13: 978-0824727772.
- [4] Z. Wang, A. Bovik. “Modern Image Quality Assessment (Synthesis Lectures on Image, Video, & Multimedia Processing)”. Morgan & Claypool Publishers. February, 2006. ISBN-13: 978-1598290226
- [5] “Methodology for the Subjective Assessment of the Quality of Television Pictures,” Recommendation ITU-R BT.500-11, ITU Telecom. Standardization Sector of ITU, 2002.
- [6] Jung, Y.J.; Sohn, H.; Lee, Seong-il; Speranza, F.; Ro, Y.M., "Visual importance- and discomfort region-selective low-pass filtering for reducing visual discomfort in stereoscopic displays," Circuits and Systems for Video Technology, IEEE Transactions on , vol.PP, no.99.
- [7] Wa James Tam; Speranza, F.; Yano, S.; Shimono, K.; Ono, H., "Stereoscopic 3D-TV: Visual Comfort," Broadcasting, IEEE Transactions on , vol.57, no.2, pp.335,346, June 2011
- [8] F. Speranza, W.J. Tam, R. Renaud, and N. Hur, “Effect of disparity and motion on visual comfort of stereoscopic images” in Proc. of SPIE, vol 6055, pp. 94-103, 2006.
- [9] Y.J. Jung, S. Lee., H. Sohn, H.W. Park and Y.M. Ro. “Visual comfort assessment metric based on salient object motion information in stereoscopic video”. Journal of Electron Imaging, vol. 21, Issue 1, Feb. 2012.
- [10] Sang-Hyun Cho; Hang-Bong Kang, "Subjective evaluation of visual discomfort caused from stereoscopic 3D video using perceptual importance map," TENCON 2012 - 2012 IEEE Region 10 Conference , vol., no., pp.1,6, 19-22 Nov. 2012
- [11] Kun-Lung Tseng; Wei-Jia Huang; An-Chun Luo; Wei-Hao Huang; Yin-Chun Yeh; Wen-Chao Chen, "Automatically optimizing stereo camera system based on 3D cinematography principles," 3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D.
- [12] J.A. Rodrigo, D. Jiménez and J.M. Menéndez, “Real-Time 3-D HDTV Depth Cue Conflict Optimization” ICCE Berlin, September 2011, p. 8.
- [13] M.S. Banks, J.C.A. Read, R. S. Allison, and S. J. Watt “Stereoscopy and the Human Visual System”. Motion Imaging Journal (May/June 2012).
- [14] A. S. Percival, “The Relation of Convergence to Accommodation and its Practical Bearing”. Ophthalmol. Rev. 11: 313-328, 1892.
- [15] T. Shibata, J. Kim, D. M. Hoffman, and M.S. Banks, “The Zone of Comfort: Predicting Visual Discomfort with Stereo Displays”, J. Vision, 11: 1-28, 2011.
- [16] P. Leclercq, J. Morris. “Assessing Stereo Algorithm Accuracy”, IVCNZ ’02: Proceedings of Image and Vision Computing, 2002.
- [17] EBU Test Sequences. <http://tech.ebu.ch/testsequences>

# Hybrid TV services for work integration of people with disabilities

Carlos Alberto Martín<sup>1</sup>, José Manuel Menéndez<sup>2</sup>, Guillermo Cisneros<sup>3</sup>

<sup>123</sup>Grupo de Aplicación de Telecomunicaciones Visuales (G@TV) – Universidad Politécnica de Madrid, Madrid, Spain

E-mail: <sup>1</sup>cam@gatv.ssr.upm.es, <sup>2</sup>jmm@gatv.ssr.upm.es, <sup>3</sup>gcp@gatv.ssr.upm.es

**Abstract:** During the last years, a new, powerful paradigm has arisen among the networked electronic media: the Hybrid TV or Connected TV, characterized by receivers able to play both broadcast content and Internet content. In this paper, the authors explain how this technology can be used to provide specific services for people with disabilities. A set of tools, based on the HbbTV standard, has been implemented to make easier the work integration for people with disabilities in the *INLADIS* project. Accessibility has been a sine qua non condition for this set of tools,

**Keywords:** HbbTV, Hybrid TV, Connected TV, work integration, people with disabilities, accessibility

## 1 INTRODUCTION

A new technological paradigm has recently arisen in the world of the networked electronic media: the Connected TV or Hybrid TV. This paradigm is characterized by receivers or TV sets that are able to receive and to play content coming from a broadcast network (for example, the digital terrestrial television, DTT) and Internet (i.e., a broadband network). This paradigm is also named HBB, Hybrid Broadcast Broadband.

The main set manufacturers have specified their own proprietary platforms to develop their own broadband content portals. These specifications do not require interoperability thus the TV set must just play the multimedia content coming from the manufacturer servers. The manufacturer keeps the control on the content consumed by users. This model is sometimes called “walled gardens”.

On the other hand, a standard specification called HbbTV (Hybrid Broadcast Broadband TV) [1] has been created to ensure interoperability between TV sets (regardless the manufacturer) and the broadband content portals (regardless the content provider). HbbTV allows both manufacturers and TV operators to exploit their media content through the broadband network. This standard solution was chosen to develop the tools explained in this paper.

Hybrid TV avoids the problems that the interactive TV systems have traditionally suffered since this paradigm implies the existence of an always-on, broadband return channel, used to provide personalized audio-visual content, according to the user preferences and interactions.

The question we tried to answer in the beginning of our project was how we could take advantage of this powerful technological paradigm to offer new services for people with disabilities. After this introduction, section 2 describes the technological features of HbbTV. Section 3 explains the project and how HbbTV features and other web technologies have been used in it. Section 4 describes how accessibility has been taken into account. Section 5 briefly drafts some conclusions and other future work.

## 2 HBBTV

HbbTV (Hybrid Broadcast Broadband TV) [1] is an international initiative to provide a standardized interoperability solution for the Connected TV. The specifications created by this initiative have been published by the ETSI, European Telecommunications Standards Institute. The newest specification is named HbbTV 1.5 and integrates some additional features with regard to HbbTV 1.0.

One main characteristic of HbbTV is the use of previous technologies to achieve an actual, fast deployment of the standard. HbbTV-compliant devices can be easily found in consumer electronic stores and some European TV operators (like ARD, the main German broadcaster) are exploiting HbbTV applications since 2010.

In figure 1, extracted from the HbbTV norm [1], the technologies referenced by HbbTV can be observed. The main idea when considering HbbTV content is that applications are CE-HTML documents, which are interpreted by an Internet browser embedded in the TV set. CE-HTML, formally named “*Web-based Protocol and Framework for Remote User Interface on UPnP Networks and the Internet (Web4CE)*” [2], is a specification created by the CEA (Consumer Electronic Association, an American industrial alliance), with the code CEA-2014. CE-HTML can be described as an HTML language for consumer electronic devices.

HbbTV references CE-HTML to specify the languages for the above mentioned applications: XHTML, CSS (the well-known language for style sheets) and JavaScript. HbbTV also includes AJAX (Asynchronous JavaScript and XML), which is a set of web technologies that allows the browser to communicate with the server in background meanwhile web content is depicted. Due to the mentioned reasons, application technologies are related to the embedded browser.

**Corresponding author:** Carlos Alberto Martín Edo, Grupo de Aplicación de Telecomunicaciones Visuales (G@TV) – Universidad Politécnica de Madrid. ETSI de Telecomunicación. Ciudad Universitaria s/n 28040 Madrid (Spain); +34 91 336 73 44; [cam@gatv.ssr.upm.es](mailto:cam@gatv.ssr.upm.es)

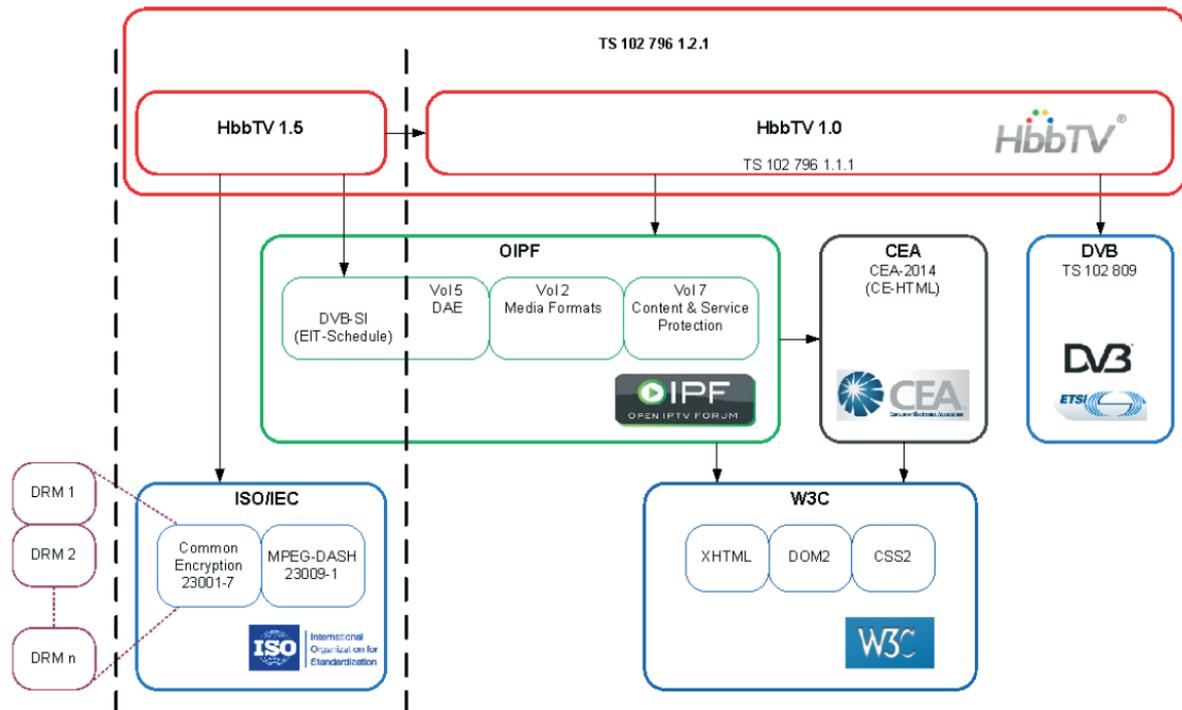


Figure 1. HbbTV normative dependencies, extracted from the specification [1]

Moreover, HbbTV integrates the DOM through CE-HTML. DOM (Document Object Model) is a W3C specification to dynamically access and update the content, structure and style of documents.

Precisely, HbbTV references in a direct or indirect manner numerous specifications from the Web environment (like the ones created by the W3C, World Wide Consortium). DOM 2 allows the TV set to handle the key events, i.e. the user interactions by means of the remote control.

The OIPF (Open IPTV Forum) is another large source of referenced technologies for HbbTV. The OIPF specifies the DAE, Declarative Application Environment [3], which includes the JavaScript APIs needed in an audio-visual environment to carry out functions on the TV set. For example, these APIs allow an HbbTV EPG application to tune another TV service if it is selected by the user in the application graphical interface.

Another key specification proposed by the OIPF defines the Media Formats for audio-visual content delivered through Internet in the HbbTV applications [4]. Although the specification identifies more alternatives, HbbTV suggests a more limited set of formats. AVC is specified for high definition and standard definition video; HE-AAC and E-AC3 are the preferred audio formats; and two formats are identified for the system layer: MPEG-2 transport stream and the MP4 container. In this way, HbbTV reminds us that audio-visual and file formats are independent. Moreover, HbbTV applications are able to insert broadcast video inside its graphical interface and scale it, too.

HbbTV also references an OIPF norm to define the protocols allowed in the standard (OIPF Protocols Specification) [5]. The minimum requirement for an HbbTV receiver is to be able to accept file streaming over HTTP. Besides, HbbTV 1.5 specifies MPEG-Dash for adaptive streaming, characterized by providing a specific bitrate according to the instantaneous throughput of the channel.

HbbTV references a DVB specification for two critical aspects to provide the service: application carriage and signalling. The mentioned specification [6] was elaborated by DVB for any generic HBB system and not only for HbbTV. The transport of the applications can be carried out by means of two main mechanisms: a DSM-CC object carousel (in the broadcast channel) and HTTP (in the broadband channel). DSM-CC is the part 6 of MPEG-2 [7] and it specifies a way to provide data in a cyclic manner inside MPEG-2 transport streams. This carousel may be used to provide a first user interface, such a launcher or red-button application, without using the broadband connection. In interactive TV, the launcher is a simple graphical interface that is automatically depicted when the user tunes a channel. It works as a gate to know the other interactive services provided by the broadcaster. This launcher usually consists in an invitation to push the red button in the remote control to access to the rest of applications. On the other hand, some TV operators have deployed their HbbTV services without multiplexing this carousel, but providing all the needed information through the broadband channel since the moment the user tunes the TV service, including the red-button application.

The signalling of an HbbTV application requires two different actions: a table called Application Information Table (AIT) and a specific descriptor in the PMT. This AIT is very similar to other signalling tables defined by DVB in its Service Information norm (DVB-SI). Moreover, this AIT table is fully compatible with the AIT table that was created by DVB for MHP, the old standardized middleware for interactive TV. This table contains all the required information for running the application in the receiver and, particularly, the URL where the application data can be found. The AIT is multiplexed with a specific PID (packet identifier) in the broadcast transport stream. This PID is referenced in the PMT of the service, close to a *stream\_type* that means “private table” and to a specific descriptor named *application\_signalling\_descriptor*, which indicates that the application is an HbbTV one.

As shown in the figure 1, HbbTV 1.5 adds support for the scheduled version of the Event Information Table (EIT), which is the DVB-SI table that carries the TV programme information. This feature is very useful to build EPG HbbTV applications. Finally HbbTV 1.5 adds support for MPEG common encryption in the broadband content, in order to integrate DRM systems that allow to monetize the broadband assets and to build business models in the future.

The different kinds of information that the TV receiver exchanges (including signalling, applications and linear–broadcast – and non-linear multimedia – broadband) are well represented in Figure 2, extracted from the HbbTV standard [1]. It must be remarked that the HbbTV applications are located on the browser, thus this module is in charge of interpreting them.

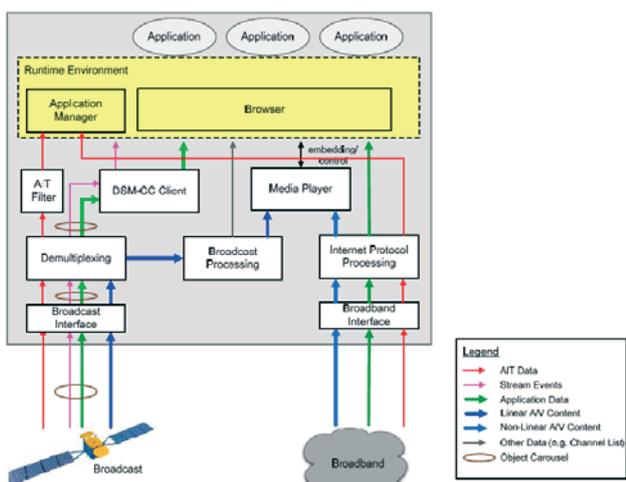


Figure 2. Functional block diagram and kinds of information in an HbbTV receiver [1]

The presence of the broadband, always-on channel and the Internet technologies integrated in the standard bring huge opportunities for the deployment of new media services. Many main European broadcasters have developed catch-up applications, where hundreds of previously emitted programmes can be found. Some

operators, like the German ARD, are offering these pieces of content with a very good quality (peaks of 4 Mbps in AVC video). Some broadcasters have developed applications such as the so called “digital teletext”, which integrates an improved graphical interface, images and video clips. The variety of possible applications is enormous, including EPG with trailers, sport information, etc. In our project, we tried to take advantage of this technology to provide work integration services for people with disabilities. On the other hand, there have been some proposals to provide access services for this collective by means of HbbTV [8]. However, the originality of the work described in this paper consists in the implementation of specific services. Moreover, the project has used HbbTV features to achieve the accessibility for people with disabilities in the developed applications.

### 3 INLADIS PROJECT

Our project, called *INLADIS – Multiscreen Platform for the Work Integration of People with Disabilities*, has been sponsored by the Indra-Fundación Adecco Chair in Universidad Politécnica de Madrid for the Accessible Technologies. Precisely, the two main objectives of the Chair are promoting the work integration of people with disabilities and improving the accessibility in media. As in other previous projects, the authors have tried to use the latest TV technologies to develop services for people with disabilities, although such technologies have not been designed or deployed for this purpose.

We enlaced this project as a way to develop innovative services using a new networked media technology. From the beginning, our aim was to achieve a multiscreen platform, available through a connected TV set and also through the PC screen. Since the limited browser embedded in the TV set is able to interpret HbbTV content, a conventional PC browser can be used to play the content, too.



Figure 3. Home page of the work integration platform

Four tools have been developed during the project, as depicted in the initial screen (whose snapshot is shown in Figure 3):

- T-learning, offering multimedia courses to improve the employability
- Actual job offers for people with disabilities

- Forum about work integration for people with disabilities
- General information about the work market for the collective of people with disabilities

Different strategies have been implemented to make accessible this set of tools, as explained in the next section.

With regard to the T-learning tool, the objective in the project was not to implement the courses themselves, but to provide a way to easily author these courses. For this purpose, a graphical interface has been created to let an author introduce the learning content (including text, images and video clips) by means of a PC. Each course consists of several chapters and each chapter, of several pages. In this way, the user can more easily follow the course. The learning content introduced by the author through the web interface is saved in XML files, which will be interpreted by the HbbTV application by means of the standard features to handle this kind of files. In this way, a modular approach is reached, distinguishing the learning content (XML file) and the graphical interface to present it (HbbTV application).

The second tool shows actual job offers for people with disabilities. A certain grade of disability is required to apply for those jobs. The offers are supplied in real-time by Fundación Adecco authors by means of a web interface. Then, an XML file including all the fields of the offer is sent to the server and it is parsed to automatically insert the offer in the platform. Users can search for offers according to two criteria: the geographical region and the industrial sector. Figure 4 shows the offer list after a user search.

| OFERTAS    |   |   |                    |
|------------|---|---|--------------------|
| Fecha      | Oferta  | Categoría                                 | Localidad          |
| 09-10-2012 | Cajero/a-Atención al cliente                    | Atención al cliente                       | Zaragoza           |
| 09-10-2012 | Jardinero/a                                     | Construcción, Instalación y Mantenimiento | Isla de Canela     |
| 08-10-2012 | Responsable de reprografía                      | Otros                                     | Madrid             |
| 08-10-2012 | COMERCIAL, FRANCES BILINGÜE RESIDENTE BARCELONA | Comercial / Ventas / Telemarketing        | Barcelona          |
| 03-10-2012 | Técnico/a de Laboratorio                        | Química y Farmacia                        | Madrid             |
| 03-10-2012 | Recepcionista con inglés bilingüe               | Atención al cliente                       | Madrid             |
| 03-10-2012 | Ayuda capacitación administrativos/as con di    | Atención al cliente                       | Madrid             |
| 02-10-2012 | Envasador/a                                     | Agricultura, Silvicultura y Pesca         | Valencia           |
| 01-10-2012 | Mozo/a Almacén                                  | Transporte / Logística                    | Alcalá de Guadaíra |

Figure 4. List of job offers according to user criteria

The forum was implemented after a suggestion from Fundación Adecco to provide a meeting point for people with disabilities and trained in ICT, who do not have other forums.

The fourth and last tool offers static, generic information about work integration for people with disabilities. The content has been conceived to be useful for both employers and employees.

This set of tools has been deployed in a conventional Apache2 web server, taking advantage of the application languages specified in HbbTV: XHTML, XML, JavaScript and CSS. Moreover, PHP has been used on the

server side for the dynamic generation of HTML content, to be depicted in the browser embedded in the TV set. Thus, PHP allows the portal to combine the flexibility for offering personalized content and the fulfilment of the HbbTV language requirements on the client side.

Images are included in the formats specified in HbbTV: jpeg, png and gif. On the other hand, the video clips included in the t-learning tool are coded according to the formats recommended in HbbTV: AVC for the video and MP4 container for the file. Audio is coded in the AAC format, to reach full compatibility to HE-AAC, which is included in the standard.

The purpose of the project was not only to develop the set of tools, but make them accessible, in the way explained in the next section.

## 4 ACCESSIBILITY

Since the set of tools for the work integration is focused on people with disabilities, the accessibility was a primordial objective of the project. This issue was taken into account from the beginning of the project, according to the *design for all* principle.

Different strategies have been followed, according to the requirements of the people with disabilities collectives. Moreover, the content had to be accessible in the two screens considered in the project: on an HbbTV TV set and on a computer. Especially for this last screen, the W3C accessibility guidelines have been a source of information and ideas to get accessible interfaces. The Web Accessibility Initiative is led by the W3C to promote the accessibility in Internet. For the purposes of this project, the guidelines included in WCAG (Web Content Accessibility Guidelines) [9] have been of particular interest.

With regard to the people with a visual impairment, two different strategies have been implemented. On the one hand, an improved graphical interface, characterized by the use of a larger font and a colour scheme that maximizes the contrast (black text in the foreground over a white background). Moreover, images are removed in the course pages in this accessible mode to gain space in the screen and present the larger font.

On the other hand, audio clips have been included to make accessible all the tools and screens, except the forum. Not only is the content synthesized but also the navigation instructions (this kind of access services is known as audio navigation), a welcome message, etc. The audio clips have been generated by means of the web demos of quality text-to-speech software. The clips for the static content (fourth tool and menus) have been asynchronously created, whereas the clips for dynamic content (courses and job offers) are automatically generated when the author submits the data. After the speech-synthesis phase, clips are coded in MPEG-1 layer 2. Moreover, tests have been successfully carried out to integrate ReadSpeaker, a text-to-speech web tool that generates and delivers the audio clips on the fly.

With regard to users with a hearing impairment, the course authoring interface enables the submission of

video clips for the sign language service. If this clip is present in a page, it is automatically played in a scaled window when the user visits it. Moreover, a special icon is used in the menus to indicate the presence of sign language video clips.

Finally, some design decisions have been made to achieve an easier navigation in the portal, which will benefit any user, including people with a physical or intellectual disability. The navigation among pages can be carried out by means of both the numeric keyboard and the colour buttons and the menu options are presenting in carousel. The audio navigation increases the usability for these users, too.

Moreover, the platform offers a personalized configuration. For this reason, username and password are required in the first screen. However, the objective is not to limit the access, but automatically take into account the user preferences, particularly with regard to the access services enabling (audio navigation, audio clips and improved interface or sign language).

## 5 ON-GOING AND FUTURE WORK

Currently, the prototype is fully functional in the laboratory. The means of the research group enable an actual broadcasting inside the laboratory, including the signalling specified in the norms. In this way, it has been possible to test the application in commercial HbbTV sets, simulating real operation conditions. In order to exploit the platform with real users, a conventional TV broadcaster should include the HbbTV signalling to point the *INLADIS* server. In this case, the mentioned text-to-speech web tool would be necessary since the demo versions used in the prototype do not allow any kind of exploitation. Some important broadcasters have been contacted. Although they have shown their interest, a more visual interface would be required for the actual exploitation. On the other hand, the t-learning courses could be improved by means of e-learning models like SCORM.

In other words, the prototype is currently functional and some minor changes would be necessary for a real use in commercial TV operators.

With regard to user trials, informal tests with people with disability have been carried out in the laboratory. Currently, a validation phase is being designed with the support of Fundación Adecco. The objective is to perform a qualitative survey with real users, including interviews and questionnaires.

Finally, the work explained in this paper is going to continue in the HBB4ALL European Project, which is going to deploy lab tests and large pilots of access services based on HbbTV in different European countries, with the support of the European Commission by means of CIP-ICT-PSP (621014).

## ACKNOWLEDGMENT

The authors gratefully acknowledge Indra and Fundación Adecco for sponsoring the *INLADIS* project. Special thanks to the rest of our colleagues that have worked in this project: Juan Pedro Fernández and Paola Cano.

## References

- [1] European Telecommunications Standards Institute (ETSI). TS 102 796 "Hybrid Broadcast Broadband TV". V1.2.1. Nov, 2012
- [2] Consumer Electronic Association. CEA2014 "Web-based Protocol and Framework for Remote User Interface on UPnP™ Networks and the Internet (Web4CE)". Jul, 2007.
- [3] Open IPTV Forum (OIPF). Release 2 Specification. Volume 5 "Declarative Application Environment" [V2.1]. Jun, 2011.
- [4] Open IPTV Forum (OIPF). Release 1 Specification. Volume 2 "Media Formats" [V1.2]
- [5] Open IPTV Forum (OIPF). Release 2 Specification. Volume 4 "Protocols" [V2.2]. May, 2013.
- [6] European Telecommunications Standards Institute (ETSI). TS 102 809 "Digital Video Broadcasting (DVB); Signalling and carriage of interactive applications and services in Hybrid broadcast/broadband environments". V1.1.1. Jan, 2010.
- [7] ISO/IEC International Standard 13818-6:1998, "Information technology -- Generic coding of moving pictures and associated audio information -- Part 6: Extensions for DSM-CC (Digital Storage Media Command and Control)"
- [8] Peter Olaf Looms – Danish Broadcasting Corporation (DR) "The future of DTV Access services". EBU Technical Review. 2010 Q4.
- [9] W3 Consortium. "Web Content Accessibility Guidelines (WCAG) 2.0". W3C Recommendation. Dic, 2008. Available in: <http://www.w3.org/TR/WCAG/>

# A Self-organizing Isolated Anomaly Detection Architecture for Large Scale Systems

Emmanuelle Anceaume<sup>1</sup>, Erwan Le Merrer<sup>2</sup>, Romaric Ludinard<sup>3</sup>, Bruno Sericola<sup>3</sup>, Gilles Straub<sup>2</sup>

<sup>1</sup> IRISA / CNRS (France), <sup>2</sup> Technicolor Rennes (France), <sup>3</sup> Inria Rennes - Bretagne Atlantique (France)

<sup>1</sup> firstname.name@irisa.fr, <sup>2</sup> firstname.name@technicolor.com, <sup>3</sup> firstname.name@inria.fr

**Abstract**—Monitoring a system is the ability of collecting and analyzing relevant information provided by the monitored devices so as to be continuously aware of the system state. However, the ever growing complexity and scale of systems makes both real time monitoring and fault detection a quite tedious task. Thus the usually adopted option is to focus solely on a subset of information states, so as to provide coarse-grained indicators. As a consequence, detecting isolated failures or anomalies is challenging. In this work, we push the monitoring task at the edge of the network. We present a peer-to-peer based architecture, which enables nodes to adaptively and efficiently self-organize according to their “health” indicators. By exploiting both temporal and spatial correlations that exist between a device and its vicinity, our approach guarantees that only isolated anomalies (an anomaly is isolated if it impacts solely a monitored device) are reported on the fly to the network operator.

**Keywords**—Monitoring, Anomalies, Reporting, Peer-to-Peer.

## I. INTRODUCTION

The number of IP-enabled devices keeps on growing in a steady manner, often reaching millions of units managed by a single operator. If those devices are able to provide a service to the user in their intended running state, deviations in behavior or hardware/software problems are generally detected offline by human intervention. The technical barrier for efficient online monitoring and analysis is the size of the devices set to operate, together with the huge amount of parameters and states to consider. Network operators deploy helpdesk in order to support their customers when they are facing problems. In the last years the cable and telecom industry have developed different remote management standards [3] to better support the helpdesk operator via dedicated protocols and tools. As a consequence, the helpdesk operation represents an important part of the overall operating cost of a network provider. Reducing the number of calls as well as their duration is an important key for every network operator to sustain profitability and reduce the total cost of ownership. Nevertheless both telecom and cable industries came up with client-server architectures where a single server (or a farm of servers) is in charge of managing a set of devices. Such architectures are typically used for management tasks (*e.g.*, service provisioning, device firmware upgrading) rather than for real time monitoring activities, essentially because of scaling issues. Indeed, the massive scale we are considering calls for efficient monitoring algorithms. A first option is to gather all the devices logs in a single place, and to analyze collected data using for instance

the MapReduce paradigm [11] to detect the causes of the anomalies. This nevertheless implies a significant detection latency and processing costs in the cloud.

The second option is to push monitoring procedures on devices. Actually, standardized procedures exist at devices level to autonomously trigger asynchronous alarms in presence of anomalies. However, these procedures are never used for practical reasons. Indeed if the cause of the anomaly lies in the network itself (*e.g.*, at routers, links or data center outages) this may impact a very large number of devices, and thus letting thousands of impacted devices reporting the problem to the helpdesk operator, which may quickly become a disaster due to the volume of generated messages. On the other hand, it is of utmost importance to minimize the overall network footprint by giving each device the capability to self distinguish network-based anomalies from *isolated* ones – anomalies that only impact the device itself – so that only isolated anomalies are reported on the fly to the helpdesk. This is the problem we address in this paper. Specifically, we propose a novel distributed monitoring tool, called FixMe, that enjoys the following properties.

- FixMe is self-managing: all the monitored devices self-organize according to their “health” indicators so that they can detect any correlation between their state and the one of their neighbors,
- FixMe is dynamic: (*i*) monitored devices may join the system or may be removed from it at any time, and (*ii*) there is no assumption regarding the QoS repartition of the monitored nodes (*i.e.*, we do not assume that the repartition is uniform),
- FixMe does not rely on any complex bootstrap procedure. In contrast to most of the monitoring tools, devices do not need to be prearranged into a predefined number of clusters (as required for instance in k-means based solutions),
- FixMe is scalable: the end-to-end detection process, *i.e.* from the local detection to the management operator reporting, requires a logarithmic number of messages.

Analysis, formal proofs and pseudo-code are deferred in [2].

The remaining of the paper is organized as follows. Section II provides an overview of existing monitoring approaches. Section III presents the model of the system, and defines the addressed problem. Section IV describes the FixMe overlay and its associated operations. Section V describes the algorithm that solves the problem. Section VI concludes and presents future works.

## II. RELATED WORK

This Section provides an overview of the existing techniques used in large scale systems to continuously and automatically monitor time-varying metrics. The authors in [17] exploit temporal and spatial correlations [4], [8], [14] among groups of monitored nodes to decrease monitoring communication costs, *i.e.*, the cost incurred by the periodic reporting of the updated metrics values from the monitored nodes to the management node. The idea is to prevent any reporting message from occurring when such a reporting would contain metrics values that could be directly inferred by the management node. This is achieved by giving each monitored node the capability to locally detect whether the current values of its monitored metrics are in accordance with predicted ones (through Kalman filters tools [6] installed at both monitored nodes and the management node), and by gathering nodes into clusters (such that, for each monitored metric, a set of clusters group together nodes that share correlated values of the considered metric according to the Pearson correlation coefficient). At clusters level, an elected leader is in charge of communicating with the management system when the current metric values of its group members differ from each others. Although close to our objectives, the main drawback of this solution lies on the centralized clustering process. All the nodes of the system are continuously organized into clusters computed through the k-means algorithm exclusively run by the management node, which is a clear impediment to the scalability of their approach. Other works aim at minimizing the processing cost for continuous monitoring [15], [10], [16] in the light of the theoretical results of [5], however similarly to [17], all these approaches suffer from a centralized handling of the clustering process.

In contrast, our objective is a fine-grain detection tool capable of accurately and efficiently detecting isolated events. As will be described in the remaining of the paper, we combine clustering and structured peer-to-peer architectures to reach this objective.

## III. MODEL OF THE SYSTEM

We consider a set of  $N$  nodes that communicate among each other through the standard synchronous message-passing model. Each node in the system is assigned a unique random identifier derived from a standard hash function (*e.g.* MD5, SHA-1). Each node has access to  $D$  services numbered  $1, \dots, D$ . At any time  $t$ , the QoS of each service is locally measured with an end-to-end performance measurement function

$$\begin{aligned} Q_i : \{1, \dots, N\} \times \mathbb{N} &\longrightarrow [a_i, b_i] \\ (p, t) &\longmapsto \text{QoS } i \text{ at node } p \text{ at time } t \end{aligned}$$

Without loss of generality we suppose that the QoS range  $[a_i, b_i]$  of service  $i$  is equal to  $[0, 1]$ . We define the *position* of a node  $p$  at time  $t$  by the vector  $Q(p, t)$  defined as

$$Q(p, t) = (Q_1(p, t), \dots, Q_D(p, t)). \quad (1)$$

For each monitored service  $i = 1, \dots, D$ , we split interval  $[0, 1]$  into  $n_i$  disjoint intervals  $[x_i^{(j-1)}, x_i^{(j)})$ ,  $1 \leq j \leq n_i$ ,

with  $x_i^{(0)} = 0$  and  $x_i^{(n_i)} = 1$ , the last interval being closed. Integer  $n_i$  is a parameter of the system. These  $n_i$  intervals can be thought as  $n_i$  QoS classes of service  $i$ . For instance, one can consider a division of  $[0, 1]$  for service  $i$  such that  $|x_i^{(j)} - x_i^{(j-1)}| \geq |x_i^{(j+1)} - x_i^{(j)}|$ . Such a division could be used to reflect the increasing sensitivity of users regarding QoS variations. A user is more sensitive to a very small variation of a high QoS than to a large variation of a low QoS. Without loss of generality, we suppose a regular division into identical length intervals and we define  $\rho_i = |x_i^{(j)} - x_i^{(j-1)}| = 1/n_i$ . In the following these intervals are named *buckets* (a more precise definition is given in Section IV).

In addition to the functions  $Q_1, \dots, Q_D$ , each node has access to  $D$  anomaly detection functions  $A_1, \dots, A_D$ . At each time  $t$ , each function  $A_i$  is fed with the sequence of the  $\ell_i \geq 1$  last QoS values  $Q_i(p, t - \ell_i + 1), \dots, Q_i(p, t)$  and provides some meaningful prediction of what should be the next QoS value. Note that  $\ell_i$  is a parameter of  $A_i$ . These functions are implemented to cope with the specific variations of their input values, and thus different kinds of anomaly detection functions exist, ranging from a simple threshold based functions, to more sophisticated ones like the Holt-Winters forecasting or Cusum method. In this paper, we suppose that the output of these anomaly detections are boolean. At time  $t$ ,  $A_i(p, t) = \text{true}$  if the sequence  $Q_i(p, t - \ell_i + 1), \dots, Q_i(p, t)$  is considered as an anomaly, it is *false* otherwise. Implementation of both  $Q_i$  and  $A_i$  functions are out of the scope of the paper.

Finally, suppose that a node locally detects an anomaly whose origin comes from a network/service dysfunction or failure. Then this anomaly will have an impact on the QoS of other nodes, and thus these nodes will locally detect it. On the other hand, we suppose that if a node locally detects an anomaly whose origin is local (hardware or software), then this anomaly will only impact its QoS, and thus no other nodes will be impacted by this specific anomaly.

Prior to defining the addressed problem, let us consider the following simple scenario presented in Fig. 1. The QoS of a single service monitored by two nodes  $a$  and  $b$  is represented by interval  $[0, 1]$ . At time  $t$  the quality positions  $Q(a, t)$  and  $Q(b, t)$  of both nodes lie in bucket  $j$ , while at time  $t+1$ , at least one of the two nodes experience a QoS change. Five situations can be observed. In situations (1), (2) and (4) node  $a$  is the only node that observes a QoS change. In situation (1), this change does not push  $a$  position outside bucket  $j$ , while in situation (2) and (4) it does. However in both situations (1) and (2), the anomaly detection function  $A_1(a, t+1) = \text{false}$ , thus  $a$  does not consider this move as an anomaly, therefore does not do any more investigation. In the other hand, in situation (4),  $A_1(a, t+1) = \text{true}$ , and thus node  $a$  triggers a FixMe message. Now observe the two last situations (3) and (5). Both nodes observe a QoS change considered as an anomaly by their function  $A$  (*i.e.*,  $A_1(a, t+1) = A_1(b, t+1) = \text{true}$ ). However in situation (3) the QoS degradation is the same for both nodes ( $Q(a, t+1)$  and  $Q(b, t+1)$  lie in bucket  $k$ ) and thus neither  $a$  nor  $b$  consider this anomaly as isolated, while in situation (5)  $Q(a, t+1)$  and  $Q(b, t+1)$  respectively lie in buckets  $k$  and  $\ell$ . Thus both nodes trigger a FixMe message. We now formally

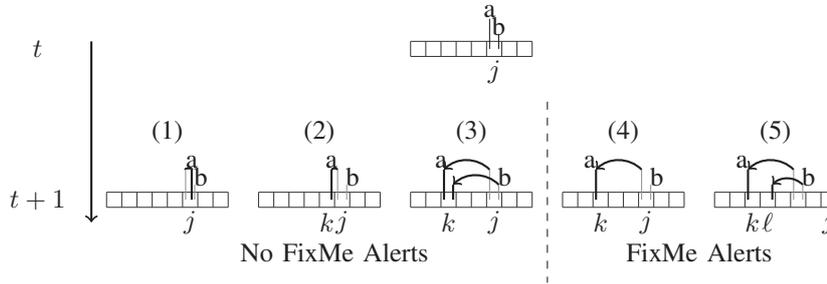


Fig. 1. Isolated anomaly detection of one monitored service. Node  $a$  triggers FixMe message in both cases (4) and (5), while node  $b$  triggers it only in case (5).

define the problem we address in this work.

**Definition 1** (The Isolated Anomaly Detection Problem). *Let  $\mathcal{S} = \{1, \dots, N\}$  be the set of monitoring nodes, and an additional node named the management operator with which any of the  $N$  nodes communicate. Let  $\mathcal{S}_{j,k}^t \subseteq \mathcal{S}$  be such that  $\forall p \in \mathcal{S}_{j,k}^t$ ,  $p$  has moved from bucket  $j$  to bucket  $k$  from time  $t - 1$  to time  $t$  and there exists a service  $i$  such that  $A_i(p, t) = \text{true}$ . Then at time  $t + 1$ , an alert is raised at the management operator if and only if  $|\mathcal{S}_{j,k}^t| \leq \tau$ , with  $\tau$  a parameter of the system. In Fig. 1,  $\tau = 1$ .*

#### IV. FIXME FRAMEWORK

##### A. Rationale

In this Section, we describe how we address the Isolated Anomaly Detection problem in a distributed system composed of  $N$  monitored nodes. FixMe framework orchestrates the monitored nodes into an overlay network, named in the following FixMe overlay. An overlay network is actually a virtual network built on top of the physical network within which nodes communicate among each other along the edges of the overlay by using the communication primitives provided by the underlying network (e.g. IP network service). The algorithms nodes use to choose their neighbors and to route their messages define the overlay topology. The topology of unstructured overlays conforms with random graphs (i.e., relationship among nodes are mostly set according to a random process which reveals to be inefficient to find a particular node or set of nodes in the overlay). On the other hand, structured overlays build their topology according to structured graphs (e.g., tree, torus, hypercube). Most of the structured overlays are based on Distributed Hash Tables (e.g., [12], [13]). The efficiency and scalability of all these proposed DHTs rely on the uniform distribution of the nodes in the identifiers space at the expense of breaking the application logic. This is why, for specific applications such as streaming applications, broadcast spanning trees structures, that support the application-level broadcast, have been proposed [9]. Our concern is to exploit the QoS relationship among monitored nodes, which make all the aforementioned solutions non adapted. As a consequence, we propose to organize nodes so that at any time  $t$  the neighbors of any node  $p$  are the nodes  $q$  whose QoS (i.e.  $Q(q, t)$ ) are

closer to the QoS of  $p$  (i.e.  $Q(p, t)$ ). The description of such an organization is done in Section IV-B. From the application point of view, three operations are provided by the system: the lookup, the join, and leave operations that allow nodes to respectively find a position in the overlay, join the overlay or leave it. From the topological structure point of view, two operations are provided: the split and merge operations that guarantee the scalability of FixMe overlay when some regions of the overlay become too dense or too sparse. All these operations are described in Section IV-C. Finally, when too many monitored nodes share exactly the same QoS (or equivalently sit at the same position in the overlay), nodes within the bucket self-organize into an hypercube as described in Section IV-D.

##### B. Overview of FixMe Overlay

The FixMe overlay is a virtual multi-dimensional cartesian coordinate space on a multi-torus. The entire coordinate space is tessellated into a collection of *buckets*. A bucket is the cartesian product of  $D$  intervals of respective length  $\rho_1, \dots, \rho_D$  (cf. Fig. 2, where FixMe overlay is made of  $16^2$  buckets). When a node  $p$  joins FixMe at time  $t$ ,  $p$  joins the bucket which corresponds to its quality position (or simply its position)  $Q(p, t)$ . When a bucket is populated by more than  $S_{\min}$  nodes this bucket is called a *seed*. The entire coordinate space is dynamically partitioned into distinct zones, named *cells*, such that a cell contains at most one seed (cf. Fig. 2, where FixMe overlay on the left is made of four cells, and the one on the right is made of five cells). More formally,

**Definition 1** (Cell). *A cell is defined as an hyper-rectangle of buckets, among which at most one is a seed. A cell is fully and uniquely characterized by a set of  $2^D$  buckets called the corners of the cell, sorted using the lexicographic order.*

Figure 2 shows these different elements for  $D = 2$  and  $\rho_1 = \rho_2 = 1/16$ . The buckets are elementary squares, the seeds are represented by the black squares, and the cells are depicted by the coloured rectangles. Note that neither cell 4 (on the figure on the left) nor cell 5 (on the figure on the right) have a seed. The reasons will be detailed in the following.

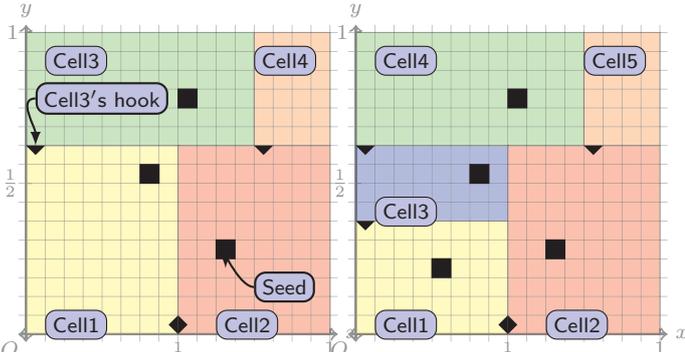


Fig. 2. FixMe overlay before (on the left) and after (on the right) a split operation

### C. FixMe Operations

a) *Lookup operation.*: We describe how a node locates the seed that is in charge of a given bucket  $b$  through the lookup operation. In FixMe, routing is exclusively handled by seeds. Each seed maintains a *routing table* that contains an entry for each of its  $2D$  neighboring seeds in the coordinate space. An entry contains the IP address and the virtual coordinate of the seed. A *lookup message* contains the destination coordinates. Using the neighbor coordinate, a seed routes a lookup message toward its destination using a simple greedy forwarding to the neighbor seed that is closest to the destination address. CAN [12] uses this routing to cross its zones. However, as such the lookup operation needs to cross in average  $\mathcal{O}(DN^{1/D})$  zones. We combine the multidimensional routing of CAN with Chord-fashioned long-range neighbors [13], [7] to improve the lookup operation cost. Specifically, in addition to its  $2D$  neighboring seeds, each seed associates a location key to each neighbor seed of its routing table. Hence, if the seed coordinates are  $(x_1, \dots, x_d, \dots, x_D)$ , then the  $+i$ th (respectively the  $-i$ th) key location for the  $d$ th axis is defined by  $(x_1, \dots, x_{(d,+i)}, \dots, x_D)$ , where  $x_{(d,+i)} = x_d + 2^i \rho_d$  (respectively  $x_{(d,-i)} = x_d - 2^i \rho_d$ ). In addition, the distance between the seed and the location key is bounded by  $R_d$  where  $R_d$  is a system parameter corresponding to the absolute farthest location to be accessed in one hop in the  $d$ th axis. Each seed  $s$  also maintains a predecessors table that contains couples  $(s', l)$ , where  $s'$  is a seed pointing on location  $l$  in  $s$  cell. The predecessors table is used when a split or merge operation are triggered to update the predecessors routing table.

b) *Join operation.*: When some new node  $p$  wants to join the system at time  $t$ , it contacts some node  $q$  already in the system. This bootstrap node  $q$  sends a lookup request for the incoming node position  $Q(p, t)$  to find the seed  $s$  responsible for the cell in which  $p$  must be inserted. Once  $p$  gets  $s$  address, it asks  $s$  to join the bucket that matches its position  $Q(p, t)$ . If that bucket is the seed  $s$  itself, then the procedure described in Section IV-D is run. Otherwise,  $s$  updates its *cell routing table* by inserting  $p$  address and its position  $Q(p, t)$ . Similarly,  $p$  keeps a pointer to  $s$  (as described above, routing is handled by seeds, thus  $p$  only needs to point to  $s$ ). Now, if the number of nodes that sit in  $p$  bucket exceeds

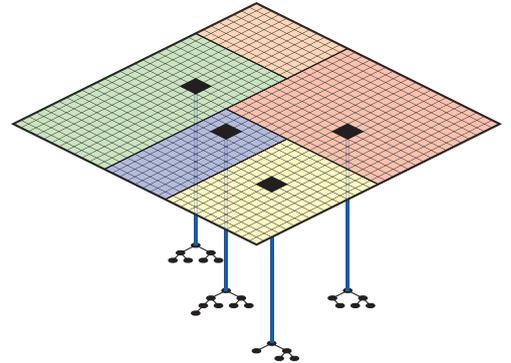


Fig. 3. FixMe cell-layer overlay and the embedded clustered overlays.

$S_{\min}$  then this bucket becomes a seed, and a split operation is triggered by  $s$  (see below).

c) *Split operation.*: A cell splits into two smaller cells when the population of one of its buckets exceeds  $S_{\min}$  nodes and the cell has already one seed. The cell splits along the dimension that corresponds to the largest distance between the two seeds. More precisely, let  $s_1$  and  $s_2$  be the two seeds whose coordinates are  $s_1 = (x_1^{(1)}, \dots, x_D^{(1)})$  and  $s_2 = (x_1^{(2)}, \dots, x_D^{(2)})$ . Let  $i_0 = \operatorname{argmax}_{1 \leq i \leq D} |x_i^{(1)} - x_i^{(2)}|$ . Then the cell is split along the hyperplane orthogonal to  $i_0$  axis and passing through the point  $\lfloor (x_{i_0}^{(1)} + x_{i_0}^{(2)}) / 2 \rho_{i_0} \rfloor \rho_{i_0} e_{i_0}$  where  $e_{i_0}$  is the  $D$  dimensional vector with  $e_{i_0}(i) = 1_{\{i=i_0\}}$ . Both seeds  $s_1$  and  $s_2$  update their respective cell routing tables to point to the nodes whose bucket falls in respectively  $s_1$  and  $s_2$  cells, as well as their routing table to point to their respective neighboring seeds. Figure 2 depicts the split operation of cell 1.

d) *Leave operation.*: Let  $p$  be a node,  $c$  be the cell node  $p$  sits in, and  $s$  be the seed in charge of  $c$ . When node  $p$  leaves the overlay (either voluntarily or not) then seed  $s$  simply discards  $p$  from its cell routing table. If  $p$  was sitting in  $s$  and the population of  $s$  undershoots  $S_{\min}$  nodes, then  $p$  departure provokes the merging of cell  $c$  with another cell  $c'$  as described in the sequel.

Prior to describing the merge operation, we introduce the notion of *cell hook* represented in Fig. 2 by black triangles.

**Definition 2** (Cell hook). *Let  $c$  be a cell in a  $D$ -dimensional FixMe overlay. Each corner of  $c$  has  $2D$  neighbors buckets. The hook of  $c$  is the first bucket (in the lexicographic order) of these neighboring buckets that does not belong to  $c$ .*

### D. Self-organizing Nodes in Dense Seeds

In the context of QoS monitoring, it is not unusual to observe that a very large number of nodes perceive a quite similar QoS for a set of services. In such cases, FixMe would show cells with very dense seeds, that is seeds with a quite large number of nodes. Thus to keep the scalability property of FixMe, we propose to self-organize these nodes into a structured graph so that the routing cost among them remains logarithmic in their

population size. Any structured graph proposed in the literature can be chosen. In this work we use PeerCube [1] essentially because each vertex of the hypercube gather from  $S_{\min}$  to  $S_{\max}$  nodes, which makes this cluster-based DHT highly robust to churn. Thus, in FixMe overlay as shown in Fig. 3, all the seeds are organized as follows. The first  $S_{\min}$  nodes that are in a seed form the *root* of the hypercube, and upon new nodes arrivals, the dimension of the hypercube increases [1]. From the point of view of the neighboring seeds of any other seed  $s$ , only the root of the hypercube is visible.

## V. SOLVING THE ISOLATED ANOMALY DETECTION PROBLEM

We now propose an algorithm that solves the isolated anomaly detection problem. The algorithm, whose pseudo code is presented in Fig. 4, is cyclically run by any node  $p$ , and is made of the following three tasks. Briefly, in Task 1, node  $p$  changes its position in FixMe overlay according to the QoS change of its monitored services (if necessary). If this QoS change is diagnosed as an anomaly by its function  $A$ , then  $p$  determines whether this anomaly is isolated or not (Task 2), and in the affirmative sends a FixMe message to the management operator (Task 3).

Let  $r$  be the current round of the algorithm. In Task 1 node  $p$  computes its current position  $Q(p, r)$ . Let  $b_r$  be the bucket that corresponds to this position,  $c_r$  be the cell that contains  $b_r$ , and  $s_r$  be the seed in charge of cell  $c_r$ . If  $Q(p, r)$  differs from  $p$  position at time  $r - 1$  (we note  $b_{r-1}$  the bucket that corresponds to this position), then  $p$  leaves bucket  $b_{r-1}$  and joins bucket  $b_r$ . If there exists a service  $i$  for which  $A_i(p, r) = \text{true}$  then  $p$  runs Task 2. The goal of Task 2 is to enable node  $p$  to determine whether there are other nodes in the overlay that have experienced the same QoS change as  $p$ , that is, nodes that left bucket  $b_{r-1}$  at the beginning of round  $r - 1$  and join bucket  $b_r$  at the beginning of round  $r$ . This is achieved as follows. By construction of FixMe, an hypercube is embedded in the seed  $s$  of each cell (see Section IV-D), and all the nodes in that cell point to the cluster root of seed  $s$  (see Section IV-C). Let  $H_r$  be the hypercube embedded in seed  $s_r$ . Then  $p$  computes a random key  $h$  that depends on both round  $r - 1$  and its previous position  $b_{r-1}$ , and asks the node in  $H_r$  that is in charge of key  $h$  (by construction of any DHT, such a node always exists) to increment a counter  $v$  (initially set to 0 at the beginning of round  $r$ ). After  $T$  time units, Task 3 starts. Node  $p$  reads counter  $v$ , and if it strictly less than  $\tau$  (i.e., no more than  $\tau$  nodes have jump from bucket  $b_{r-1}$  to bucket  $b_r$ ) then  $p$  sends a FixMe message to the management node, which ends Task 3.

## VI. CONCLUSION

In this work, we have formalized the isolated anomaly detection problem. Such a problem is recurrent in various large scale monitoring applications, and in particular in the cable and telecom industry where it is of utmost importance to make the difference between isolated anomalies and network based anomalies. One of the reasons being a financial one. In this context we have proposed the FixMe tool that pushes

| <b>Algorithm 1: p.updatePosition(r:round)</b> |   |
|---|---|
| <b>Data:</b>                                  | T: delay such that all nodes have moved to their new bucket (if necessary).   |
| <b>Output :</b>                               | The positioning of $p$ in the appropriate bucket, and the sending of a FixMe message if $p$ detects an isolated failure |
| 1   | begin   |
| 2   | Task 1  |
| 3   | $r \leftarrow r + 1;$   |
| 4   | oldposition $\leftarrow$ p.bucket ;   |
| 5   | newposition $\leftarrow$ Q(p,r);  |
| 6   | newbucket $\leftarrow$ p.lookup(newposition);   |
| 7   | if newbucket $\neq$ p.bucket then   |
| 8   | p.leave();  |
| 9   | p.join(r,p);  |
| 10  | end   |
| 11  | EndTask   |
| 12  | if $\exists i, 1 \leq i \leq D, A_i(p, r) = \text{true}$ then   |
| 13  | Task 2  |
| 14  | $h \leftarrow \mathcal{H}(\text{oldposition}, r - 1);$  |
| 15  | p.incrementValue(h);  |
| 16  | EndTask   |
| 17  | Wait Until T;   |
| 18  | Task 3  |
| 19  | $n \leftarrow$ p.cell.seed.get(h);  |
| 20  | if $n \leq \tau$ then   |
| 21  | send FixMe msg to Management Operator;  |
| 22  | end   |
| 23  | EndTask   |
| 24  | end   |
| 25  | end   |

Fig. 4. Isolated Anomaly Detection algorithm run by any node  $p$

monitoring to end devices, and by combining local algorithms to detection functions provides a scalable and efficient solution to the isolated anomaly detection problem. As a future work, we first plan to analyze the evolution of FixMe in a stochastic model to study, in particular, the influence of the distributions on the cells repartition and their sizes. The long term objective is the implementation, and deployment of FixMe.

## REFERENCES

- [1] E. Anceaume, R. Ludinard, A. Ravoaja, and F. Vilar Brasileiro. Peer-cube: A hypercube-based p2p overlay robust against collusion and churn. In *Proceedings of the IEEE International Conference on Self-Adaptive and Self-Organizing Systems (SASO)*, 2008.
- [2] Emmanuelle Anceaume, Erwan Le Merrer, Romaric Ludinard, Bruno Sericola, and Gilles Straub. Fixme: A self-organizing isolated anomaly detection architecture for large scale distributed systems. In *Principles of Distributed Systems*, 2012.
- [3] Broadband Forum. TR-069 CPE WAN Management Protocol Issue 1, Amend.4, 2011.
- [4] A. Desphand, E.C. Guestrin, and S.R. Madden. Model-driven data acquisition in sensor networks. In *Proceedings of the International Conference on Very Large Databases (VLDB)*, 2002.
- [5] S. Har-Peled and B. Sadri. How fast is the k-means method? *Algorithmica*, 41(3):185–202, 2005.
- [6] R. E. Kalman. A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering*, 82(1):35–45, 1960.
- [7] B. Kovacs and R. Vida. An adaptive approach to enhance the performance of content-addressable networks. In *Proceedings of the International Conference on Network and Computer Science (ICNS)*, 2007.
- [8] S. Krishnamurthy, T. He, G. Zhou, J. A. Stankovic, and S. H. Son. RESTORE: A Real-time Event Correlation and Storage Service for Sensor Networks. In *Proceedings of the International Conference on Network Sensing Systems (INSS)*, 2006.

- [9] J.W. Lin. Broadcast scheduling for a p2p spanning tree. In *Proceedings of the IEEE International Conference on Communications*, 2008.
- [10] K. Mouratidis, D. Papadias, S. Bakiras, and Y. Tao. A Threshold-Based Algorithm for Continuous Monitoring of K Nearest Neighbors. *IEEE Transactions on Knowledge and Data Engineering*, 17(11):1451–1464, 2005.
- [11] A. Rabkin and R. Katz. Chukwa: a system for reliable large-scale log collection. In *Proceedings of the International Conference on Large Installation System Administration (LISLA)*, 2010.
- [12] S. Ratnasamy, P. Francis, M. Handley, R. M. Karp, and S. Shenker. A scalable content-addressable network. In *Proceedings of the SIGCOMM Conference*, 2001.
- [13] I. Stoica, R. Morris, D. R. Karger, M. Frans Kaashoek, and H. Balakrishnan. Chord: A scalable peer-to-peer lookup service for internet applications. In *Proceedings of the SIGCOMM Conference*, 2001.
- [14] M. C. Vuran and I. F. Akyildiz. Spatial correlation-based collaborative medium access control in wireless sensor networks. *IEEE/ACM Transactions on Networking (TON)*, 14(2):316–329, 2006.
- [15] X. Xiong, M. Mokbel, and W. Aref. SEA-CNN: Scalable Processing of Continuous K-Nearest Neighbor Queries in Spatio-Temporal Databases. In *Proceedings of the IEEE International Conference on Data Engineering (ICDE)*, 2005.
- [16] Z. Zhang, Y. Yang, A. K. H. Tung, and D. Papadias. Continuous k-means monitoring over moving objects. *IEEE Transactions on Knowledge and Data Engineering*, 20(9):1205–1216, 2008.
- [17] Y. Zhao, Y. Tan, Z. Gong, X. Gu, and M. Wamboldt. Self-correlating predictive information tracking for large-scale production systems. In *Proceedings of the International Conference on Autonomic Computing (ICAC)*, 2009.

# Technology Enablers for a Future Media Internet Testing Facility

Michael Boniface<sup>1</sup>, Stephen Phillips<sup>1</sup>, Athanasios Voulodimos<sup>2</sup>, David Salama Osborne<sup>3</sup>,  
Sandra Murg<sup>4</sup>

<sup>1</sup>IT Innovation, Southampton, UK; <sup>2</sup>NTUA, Athens, Greece; <sup>3</sup>Atos, Madrid, Spain; <sup>4</sup>Joanneum Research, Graz, Austria

E-mail: <sup>1</sup>mjb@it-innovation.soton.ac.uk, scp@it-innovation.soton.ac.uk,  
<sup>2</sup>thanosv@mail.ntua.gr, <sup>3</sup>david.salama@atosresearch.eu, <sup>4</sup>sandra.murg@joanneum.at

**Abstract:** Creating innovative Future Media Internet (FMI) products and services is a complex endeavour requiring consideration of socio-technical factors and an increasingly diverse technology landscape. Accelerating time to market requires the availability of technology enablers adapted to local contexts and integrated together to create added value patterns of use. In this paper we present a set of such technology enablers used within a FMI testing facility. We describe how each enabler has been constructed to support the lifecycle of different classes of content and integrated to provide coherent and representative aggregations of content expected in FMI applications and services. A set of lessons learnt are derived from experiments conducted using the enablers at venues of the facility.

**Keywords:** future media internet, augmented content, technology enablers, trials and experiments

## 1 INTRODUCTION

Offering collective experiences to consumers is at the heart of many new Internet business models. Traditional business models that place digital information as the primary asset (through, for example, content download), have largely failed to deliver due to the difficulty in protecting value in digital goods. Providers of digital services are now looking to create value by linking people to each other and to locations (both real and virtual) in such a way as to capture the popular imagination. These new interactions exploit the needs of consumers to share their experiences and thus create new channels for revenue creation and advertising.

Systems that deliver new forms of social interaction and experience are complex socio-technical structures requiring a broad range of capabilities to acquire, process, aggregate and present different types of multi-stakeholder content. In this paper we present a set of technology enablers developed to provide such capabilities and how together they deliver key aspects of the Future Media Internet (FMI). We describe how the enablers are used as the foundation for experimentation and trials at the EXPERIMEDIA Facility [1]. The EXPERIMEDIA Facility is a Future Internet Research and Experimentation (FIRE) Facility aiming to accelerate research, development and exploitation of innovative

products and services that explore new forms of social interaction and experience in mixed online and real-world communities. EXPERIMEDIA provides three complementary Smart Venues offering communities and live events for trials conducted in real world locations. The technology enablers form the essential element of EXPERIMEDIA providing a baseline for experimentation and enhancing infrastructure so that each venue shifts towards the FMI.

## 2 MOTIVATION

Experience is at the heart of new economy [6]. For the FMI to contribute, innovative applications and services must be developed that support capabilities such as enhanced personalised experiences, real-time social interaction, non-linear story-telling, and greater levels of immersion. Communication must be enhanced by rich and augmented audio-visual, sensor and 3D content delivered to virtual and real locations. Of course, these innovative applications place significant demands on network and content management infrastructures as providers attempt to deliver guaranteed Quality of Service (QoS), enhanced Quality of Experience (QoE) and Quality of Community (QoC) to communities that dynamically organise themselves around socially distributed, fixed and mobile content.

Significant multidisciplinary challenges exist when considering the interactions between social and technical parameters, some of which are:

- Developing the diverse technical capabilities (e.g. connectivity, access technologies) that are localised to different types of venues, from large public arenas, home entertainment systems to online virtual venues;
- Diverse content creation (e.g. 3D Internet, augmented reality, social media, broadcast media) and adaptive delivery processes that operate together across combinations of environments; e.g., synchronisation of hybrid broadband broadcasts with on-going live events at a venue, where both spectators at the venue and those with whom they are networked (locally or remotely) may wish to share or experience the best camera view out of hundreds available;
- Flexible up- and down-scaling of infrastructure capacity to support ad-hoc and short-lived communities during certain live media events;

**Corresponding author:** Michael Boniface, IT Innovation Centre, Gamma House, Enterprise Road, Southampton, S017 7NS, + 44 (0) 23 8059 8866, mjb@it-innovation.soton.ac.uk

- Discovering ways in which media preferences can be shared in real-time and synchronised between diverse social groups no longer defined by presence at a specific location but defined by network relationships between social groups at a range of live or virtual venues; and
- Large-scale management of mixed content environments (user generated, digital TV, mobile broadcasting) where everyone is simultaneously playing the role of content consumer, content producer and content mediator to provide virtually enriched and personalised media services that can be shared between members of ad-hoc social groups.

Addressing these challenges and developing successful media services is a complex endeavour. It is essential to provide a suitably featured testing and experimentation environment that can be used to explore a full range of system properties required by networked media systems. Such systems build on a range of technology enablers providing capabilities to support diverse content types and lifecycle management approaches. Capabilities need to be designed so that they can not only be composed and integrated into larger systems supporting aggregated information flows, but also so that they are generically applicable to different applications and in consideration of the social, cultural, ethical and environmental constraints found in different locations. It is likely to be prohibitively expensive for any individual organisation (especially Small and Medium Enterprises) to develop and support all the capabilities necessary. In any event, for such a broadly based testbed to be effective and acceptable to those that use it, it is essential that it is seen as being developed and managed openly and independently of any specific proprietary interest.

### 3 TECHNOLOGY ENABLERS

Technology enablers are software or service components whose functionality allows users to achieve added value through use, either by design (i.e. the purpose is known in advance) or more frequently by openness (i.e. the purpose is opportunistically established by the user). Technology enablers are a key part of future innovation in programmes such as Future Internet Research and Experimentation (FIRE) [3] and the Future Internet Public-Private Partnership (FI-PPP) [4]. Technology enablers of the FMI must support a range of social, audio-visual, pervasive and 3D content. Each class of content has distinct characteristics, content lifecycles (authoring, management and delivery) and platforms to support it. Developing a new platform supporting all content types is unrealistic and the approach must focus on developing open interfaces to existing platforms that allow for greater levels of interaction between information and control flows.

We define a component model that focuses on different content aspects within the FMI and have developed implementation technologies supporting the lifecycle of the specific content. Tools and services are provided that support the mixing of different content types in the

delivery of user experience, where the content lifecycles are managed within separate systems. Figure 1 shows the component model for Social, Audio Visual, Pervasive and 3D content to which we add Experiment content supporting all data related to the setup, execution, monitoring, analysis and security of trials. A key element of the components is that they are designed on the principle of openness and transparency in terms of observability, configuration and security policy. Each component includes a structural (i.e. entities) and metric model (i.e. QoE, QoS, QoC), and is instrumented to allow deep measurements. Deep measurements are observable system properties not normally exposed collectively. These could include response times and interaction logging in client applications and server load or frame rates in infrastructure and services. The disclosure of such information is essential for understanding the interplay between different system components, along with the observation of behaviours in larger composed Internet ecosystems including communities.

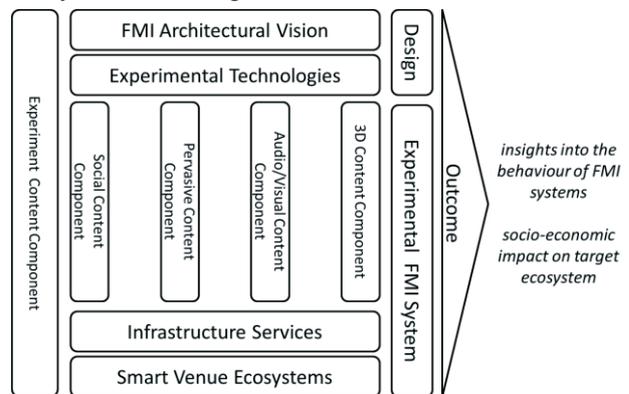


Figure 1: Technology Enablers of an FMI Testing Facility

The metric modelling framework is generic and supports a range of potential measurements. The objects of experimental observation (referred to as 'Entities') are loosely coupled with the agent making the observations. Entities themselves contain one or more Attributes that are the subject of actual instrumentation and measurement activity. Measurement data itself is logically structured within Metric Generators (typically used to represent metrics linked to a particular component or user). Further organisation is offered through the grouping of sets of measurements using one or more named Metric Groups. A Measurement Set contains zero or more measurements that are specific to a particular attribute; Metric Groups may contain one or more Measurement Sets. The semantics of each Measurement Set is defined by its Metric, which in turn has a Metric Type and Unit of measure. An agent interface supporting the metric model and measurement is based on Advanced Message Queuing Protocol (AMQP) and is available through Java, Java Android, C#, C++ and Ruby. A configuration interface is provided that supports set up and runtime adaptation of some QoS parameters. A security model is provided that describes authentication, access control and how personal data is processed. The latter element is necessary to assure compliance with ethical

experimentation and associated data protection legislation. The capabilities of each component are described in more detail in the following sections.

### **3.1 Social Content Component (SCC)**

Social content is characterised by user generated content produced by and consumed within online communities. Photos, videos, comments and opinion is disseminated by individuals to related friends using social networking platforms. The SCC offers the capability to access social content, explore social graphs, extract general social knowledge (e.g. sentiment and controversy) and media specific QoS/QoE for adaptive, efficient and personalised delivery of experiences [5][6]. Using an open social API experiments can navigate a user's social profile in a range of social networking platforms. The virtualisation of social network APIs is important as although the predominant network is Facebook, other online platforms are used by target participant communities. A pluggable social analytics dashboard is offered allowing different algorithms to be incorporated with default algorithms provided to detect individual and group preferences based on attitudes, selections and beliefs. The dominant attitudes, beliefs and communication ways of social groups (rather than individuals) can be used to optimise streamed, delivered or even transmitted media content. In addition, the detection of the proximity of consumers to content, similar behaviours and searching for popular user generated content can improve media delivery, enhance live streams, or augment information that is aligned with preferences of consumers.

### **3.2 Audio Visual Content Component (AVCC)**

Audio visual content is primarily characterised by video and metadata that's streamed and consumed by applications (i.e. players). AV content is produced by professionals and users using content production, management and content distribution networks. The AVCC offers capabilities for all aspects of the content lifecycle (acquisition, production, transcoding, distribution, etc) and advanced capabilities for acquisition and synchronisation between camera feeds, audio and metadata, including matching exact frames from different cameras; integrated automatic data collection and management systems; and metadata annotation and generation tools based on domain standards.

The AVCC is a general purpose component whose use is applicable in a wide range of applications with high levels of technical readiness. The system was recently deployed for the 2012 Paralympics. The HD recording system supports different audio-visual sources, a huge storage capability and fast distribution to different endpoints. Metadata synchronisation allows social content and pervasive content to be annotated on audio-visual streams and presented to consumers depending on personal preferences. The metadata can be introduced manually or automatically by pluggable video analytics or from other sources such as the SCC or PCC. In some cases, like music that is annotated on video, synchronisation is

essential. An important characteristic of the AVCC is that it operates in real-time and provides augmented content live to consumers rather than processing offline and return results later. This capability is critical for applications supporting live social and networked media interaction and experience.

### **3.3 Pervasive Content Component (PCC)**

Pervasive content is produced by mobile users and sensors located in real-world environments. Human sensing (e.g. biomechanics, physiology, etc), human location tracking (indoors and outdoors), location-based content, real-world community interaction models, environment sensing, points of interest all characterise pervasive content.

The PCC offers capabilities that collectively gather data about a user's physical location, QoE, points of interest and interactions. Physical location is used in both the context of tracking a user's location and also as a means by which Augmented Reality (AR)-based content can be selected for delivery and user generated data can be mapped to a spatial location. A pervasive gaming platform is provided supporting the gamification of activities and allowing for adaptive narratives and content that's customised for different experiences. The platform allows professionals and users to co-create content, such as a locative game integrated with the structure, narrative, and content of the event itself. Users attending the event can consume and produce content in real time using Smart mobile devices. The unfolding events, as experienced by users, can be adapted and orchestrated in real time. Users primarily participate locally at the event but can also contribute via the Internet, and synchronised but distributed live events can be joined to provide a common experience. The platform allows access to content and services both before and after the event, thus supporting community building and operation. Metadata generated is published and can be used to annotate audio-visual streams so that other participants can search for and retrieve available content.

### **3.4 3D Content Component**

3D Content is characterised by reconstructions of humans and real-world objects as geometrics models. Many other approaches exist such as RBG plus depth, stereo cameras and multiview that claim to be 3D but actually only deliver the illusion of 3D. These other techniques are concerned with presentation

The 3DCC supports experimenters in acquiring and manipulating 3D information from depth sensing devices (e.g. the Kinect). Low, medium and high level capabilities are supported. Low level functions offer depth images, human skeleton and RGB acquisition. Several filtering algorithms are provided to enhance depth measurements, improve skeleton tracking and to calculate various biomechanical measurements (angles between bones, human joints, body part surface areas, etc). Finally, a tool is offered for avatar editing and avatar interactive motion is provided.

### 3.5 Experiment Content Component (ECC)

Experiment content is produced and consumed by developers performing tests on FMI systems to understand and gain insight into structure, behaviour and performance. System configuration, system dependency graphs, input/out data sets, testing procedures and monitoring data all characterise experiment content.

The ECC allows a developer to set up, execute and tear down tests on FMI systems deployed at different locations. Mechanisms are provided to deploy services in virtual machines and connect them together according to a desired experimental configuration. Dependency handling allows relations between sets of services to be specified using service recipes and deployed on target hosting platforms (e.g. targeting VM Ware machines or OpenStack clouds, depending on the systems available at the venues). The ECC monitors, derives experimental data from, and manages the system under test through integration with the ECC API. The ECC elicits QoS, QoE and QoC data from the other components and delivers it to the experimenters so that they can analyse the behaviour of technical systems in relation to user experience. The ECC manages the delivery of monitoring metrics that are stored and available for both live and post/batch analytics. Monitoring agents are available for services, mobile clients and web applications. A dashboard is provided leading developers through an experiment lifecycle that includes setup, live monitoring, analysis and tear down.

## 4 COMPOSITION PATTERNS

Composition Patterns are standard ways that technology enablers can be used together to investigate new forms of social interaction and experience. We define a set of important patterns to help experimenters understand how the components of the facility can best support their experimental objectives and to provide stimulus for new ideas.

The 1<sup>st</sup> pattern, “Instrumentation and Observation”, focuses on instrumentation of technology enablers. Each component is described in terms of metrics (QoS, QoE and QoC) associated with their specific content domains (social, audio-visual, pervasive, and 3D) and is required to generate measurements of these metrics during the runtime. Additional infrastructure metrics regarding infrastructure performance are generated by hosting components such as the Cloud Manager (e.g. compute, storage and networking). All metrics assist experimenters in understanding the behaviour of the system in terms of both technical performance and user experience. For example, the AVCC generates metrics related to AV streaming such as frame rate, frames dropped, video quality, etc. When combined with networking metrics (e.g. bandwidth, latency, etc) an experimenter can study the network characteristics necessary to deliver a certain QoS (e.g. 25 fps, HD with a 1/1000 frames dropped) to a group of consumers. This is standard, although not simple, and undertaken initiatives such as ITU QoE study areas (e.g.

ITU-R Rec. BT.500-11 provides methodology for the subjective assessment of the video quality)

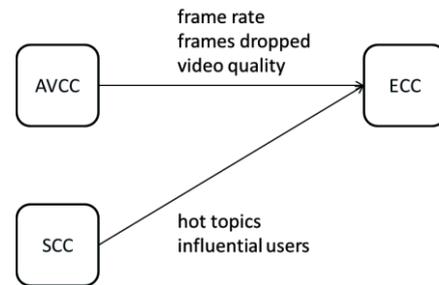


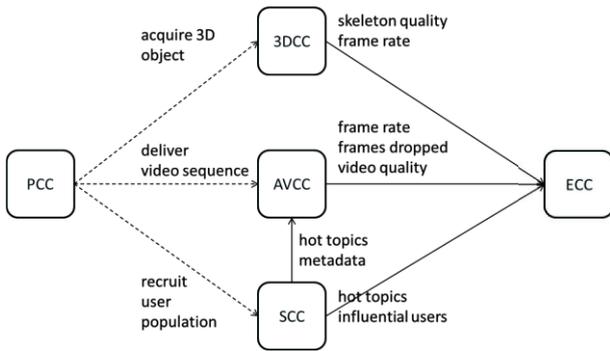
Figure 2: Correlating between discussion topics and delivered content

FMI technology enablers must focus more on how different content, aggregations of content and social interaction affect experience. With each Content Component generating metrics, experimenters can begin to correlate monitoring data between components. For example, navigating to a certain location in virtual world may create a popular discussion in a social networking group. By identifying popular discussions and looking at which point in the story/presentation (e.g. seeking a specific time point a recorded video stream) these occurred, the experimenter can begin to understand why specific events cause specific outcomes in the target community, and if necessary initiate a deeper analysis (e.g. direct user evaluation) with the community on these target areas. Changing the narrative after the production would be considered a “design” phase adaptation. However, increasingly we envisage adapting the narrative during the production based on emerging profiles and interests of social groups and how they react to the content being delivered. In this case rather than undertaking a post analysis of the metrics we could automatically annotate a video stream with metadata indicating points of interest/questions associated with the content. The Content Author could then adaptive the narrative based on discussions, questions, or votes for more information by reviewing an annotated stream timeline.

The 2<sup>nd</sup> pattern “Mixed Information Flows” focuses on how content from each component can be orchestrated in information flows as part of a new experience. Examples include:

- Annotating video streams (AVCC) with metadata from social networking trends (SCC)
- Annotating video streams (AVCC) with metadata derived from sensors (PCC)
- Adapting the narrative of a pervasive game (PCC-Creator) based on social networking trends
- Reconstructing people in physically different locations (3DCC) in a single virtual location

An interesting element is how by mixing the content between different platforms, the experience and technical performance changes in each component. For example, does changing the narrative as a consequence of the social networking topic reduces the discussion on the social network because the focus of attention is has changed?



**Figure 3: An integrated view of EXPERIMEDIA based on Experiment Composition Patterns**

It is critical that technology enablers be offered with an integrated view based added value composition patterns (See Figure 3). Here we show how all components can be used together in an FMI system. The PCC orchestrate the narrative (control flow is the dotted lines, data flow is the solid lines) for the gamification of activities. As such, the PCC can initiate controlling actions such as recruiting user populations through information dissemination in social networks, delivering popular video sequences to specific communities and acquiring 3D representations of objects and people. With all components instrumented using a behavioural model resulting metrics generated are acquired by the ECC and available for real-time and post analysis.

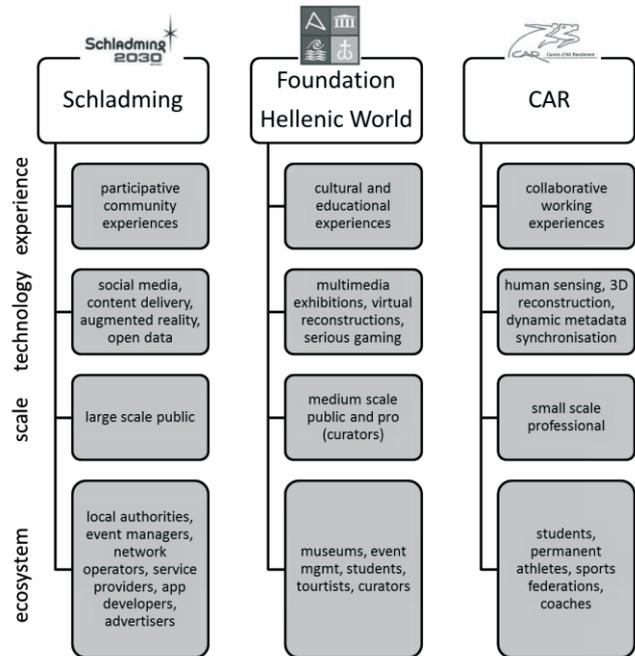
## 5 RESULTS

The technology enablers have been developed as the baseline technology of the EXPERIMEDIA facility. EXPERIMEDIA offers a new approach to European testbed provision by providing a facility with the four foundation elements necessary for socio-technical experimentation of the FMI conducted in the real world:

- Smart venues: attractive locations where people go to experience events and where experiments can be conducted using smart networks and online devices;
- Smart communities: online and real-world communities of people who are connected over the Internet and available for participation in experiments;
- Live events: exciting real-world events that provide the incentives for individuals and smart communities to visit the smart venues and to become participants in experiments;
- Baseline technology enablers: state-of-the-art Future Internet testbed infrastructure for social and networked media experiments supporting large-scale experimentation of user-generated content, 3D internet, augmented reality, integration of online communities and full experiment lifecycle management.

EXPERIMEDIA has three complementary venues with each offering different experiences, live events, stakeholder ecosystems and scale [5]. The venues include Schladming Ski Resort, Austria; CAR High Performance

Training Centre, Spain; and Foundation for the Hellenic World, Greece (See Figure 4)



**Figure 4: EXPERIMEDIA Smart Venues**

The technology enablers have been verified and validated in nine experiments, three at each venue. Topics for experiments included:

- Visitor planning using interactive video and situated augmented displays [8]
- Content sharing in hyper local and temporal communities [9]
- Improving athletes' performance through real-time 3D motion capture using non-invasive and wearable sensor technologies [10]
- Remote sports training using augmented reality 3D video conferencing tool [10][11]
- Participative battles and wars in human history using gamification and immersive technologies [12]
- Personalising users' experiences inside a museum based on cognitive styles [13]

Each experiment was designed to deliver value impact to both participants and venue stakeholders through the use of the enabling FMI technologies. Not all enablers are relevant to every experiment but two or more enablers have been used by each experiment. Important lessons have been learnt from the first phase of experiments. Firstly, technology maturity assessment is an essential element of enabler development and use. A balance must be struck between reliability and innovative features. Putting technology into realistic contexts to be evaluated by participants at scale requires Technology Readiness Level 7 [14]. In addition, due to the openness of use by experimenters it's essential that during experiment design System Readiness Level be assessed to determine maturity of a specific integration of components [15]. Where maturity levels have not been achieved, it is not

cost effective to scale up until deficiencies have been remedied. Secondly, a clear model of capability deployment and delivery must be established early, so that experimenters are clear on the scope of their responsibilities for systems integration and operation. Where possible, capabilities should be delivered as services rather than software distributions installed by experimenters to reduce the cost of experimentation. This decision has consequences for the providers of technology enablers in terms of responsibilities. Providers of technology enablers may become service providers active within trials and must take operational (e.g. delivery of QoS) and legal (e.g. protection of personal data) responsibilities. Finally, sustainability is probably the most critical factor, and access rights to the technology enablers are essential in realising the value impact from experiments. Experimenters depending on technology enablers must be presented with a route to market that either incorporates commercialisation of the enablers or offers alternative paths through standards compliant implementations. Failure to address these concerns creates barriers to adoption and reduces the potential benefits of the technology enablers.

## 6 CONCLUSIONS

In this paper we have presented a set of technology enablers for the FMI that can accelerate research, development and innovation in products and services targeting new forms of social interaction and experience. The enablers have been structured so as to support different classes of content being acquired, managed and delivered by platforms. Patterns of integration between enablers have been established so that applications and services can create advanced information flows that mix classes of content to create augmented presentations and experiences. The definition of composition patterns helps experimenters understand how the components of the infrastructure can best support their experimental objectives. The enablers have been verified and validated through a series of trials conducted at the EXPERIMEDIA facility demonstrating generic applicability in different application and localisation of the enablers to environment, social, legal and ethical contexts. The availability of advanced technology enablers with well-defined composition patterns, deep instrumentation of QoS, QoE and QoC linked to value impact, and the ability to be localised to different real-world contexts reduces the complexity of developing and evaluating innovative FMI systems.

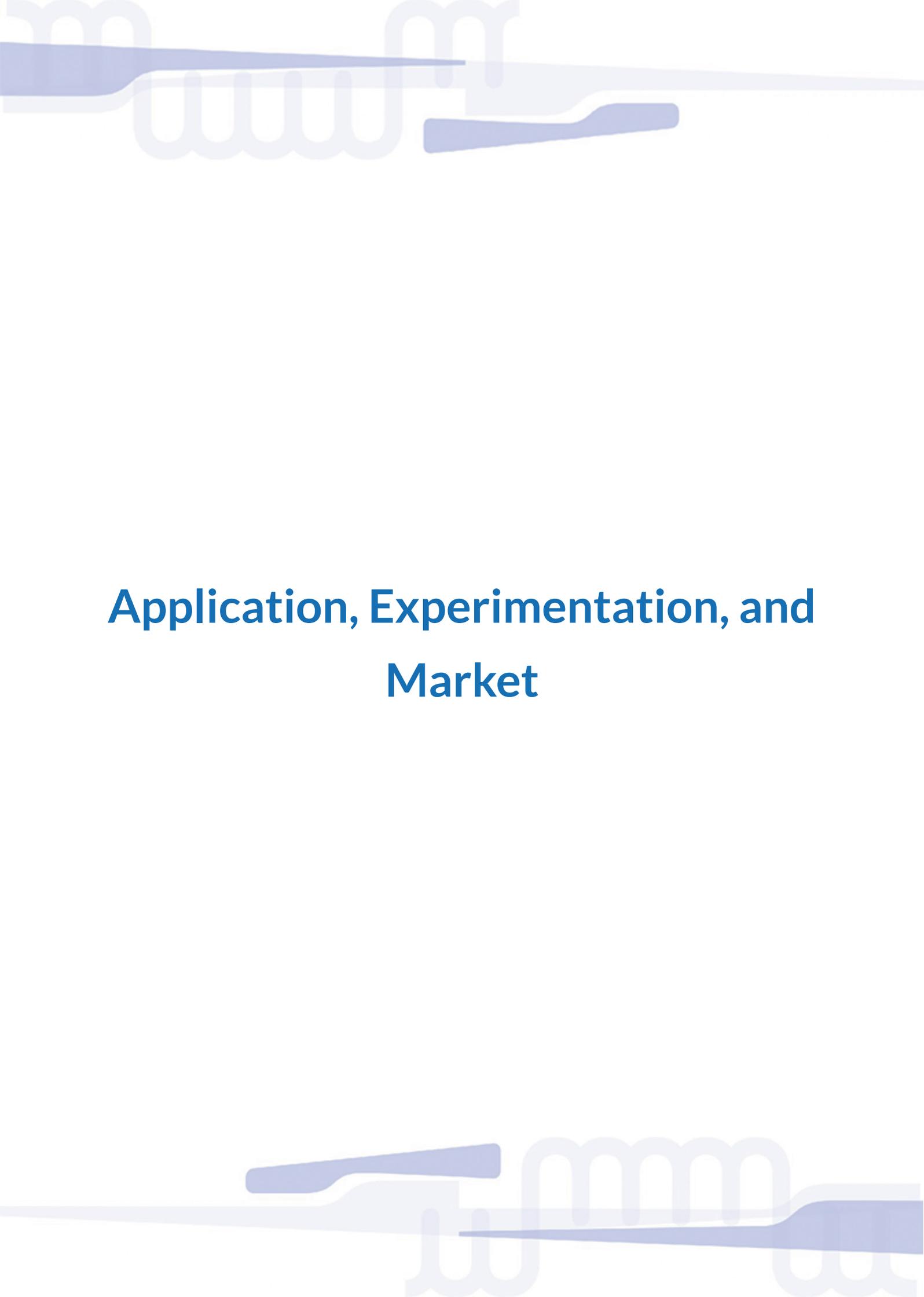
## 7 ACKNOWLEDGEMENTS

The EXPERIMEDIA project has received research funding from the European Commission under the

Information Communication Technologies Programme (ICT), contract number 287966. The project has a consortium of 20 partners from industry and academia as well as non-profit organizations. The authors wish to thank their collaborators in the EXPERIMEDIA consortium for their help and support.

## References

- [1] EC EXPERIMEDIA Project, [www.experimedia.eu](http://www.experimedia.eu)
- [2] Joseph, P. I. N. E., & Gilmore, J. H. (2011). *The experience economy*. Harvard Business School Press.
- [3] Gavras, A., Karila, A., Fdida, S., May, M., & Potts, M. (2007). Future internet research and experimentation: the FIRE initiative. *ACM SIGCOMM Computer Communication Review*, 37(3), 89-92.
- [4] SurrIDGE, M., Alvarez, F., Carrillo, M., Salvadori, E., Hierro, J., & Bohnert, T. (2012). Trade-offs and responsibilities in Phases 2 and 3 of the FI-PPP Program. White paper, <http://www.fi-ppp.eu/wp-content/uploads/2012/09/FI-PPP-Phases-2-3-White-Paper-Draft-Final.pdf>
- [5] Naveed, Nasir and Gottron, Thomas and Sizov, Sergej and Staab, Steffen (2012), "FREuD: Feature-Centric Sentiment Diversification of Online Discussions". In: *Proceedings of the 4th International Conference on Web Science (WebSci'12)*, June 22 – 24, 2012.
- [6] Naveed, Nasir and Sizov, Sergej and Staab, Steffen (2011) ATT: Analyzing Temporal Dynamics of Topics and Authors in Social Media. pp. 1-7. In: *Proceedings of the ACM WebSci'11*, June 14-17, Koblenz, Germany
- [7] Salama, D. Infrastructure and Software Assets Inventory <http://www.experimedia.eu/wp-content/uploads/sites/4/2013/04/D3.1.1-First-Assets-Inventory-v1.02.pdf>
- [8] Nixon, L., Bauer, M., & Bara, C. Connected Media Experiences: interactive video using Linked Data on the Web. WWW '13 Companion Proceedings of the 22nd international conference on World Wide Web companion
- [9] DigitalSchladming website, <http://onmeedia.com/apps/schladming/>
- [10] Anargyros Chatzitofis, Nicholas Vretos, Dimitrios Zarpalas, Petros Daras, Three-dimensional Monitoring of Weightlifting for Computer Assisted Training(, 15th International Conference on Virtual Reality and Converging Technologies, Laval 2013
- [11] Sergiusz Zieliński et al. (2013) Stereoscopic videoconferencing with augmented reality in technology enhanced sports training", *eChallenges 2013*
- [12] Martín López-Nores, Yolanda Blanco-Fernandez, Jose J. Pazos-Arias, Alberto Gil-Solla, Jorge García-Duque, Manuel Ramos-Cabrer, and Manolis Wallace, (2013) REENACT: Learning about Historical Battles and Wars through Augmented Reality and Role Playing - An EXPERIMEDIA Experiment, 5th International Conference on Computer-Supported Education (CSEDU)
- [13] Antoniou, A., Lepouras, G., Lykourantzou, I., Naudet, Y. (2013) Connecting physical space, human personalities, and social networks: the Experimedia Blue project. *Proceedings of the International Biennial Conference Hybrid City, Subtle Revolutions*. D. Charitos, I. Theona, D. Gragona, H. Rizopoulos, M. Meimaris (Eds). University Research Institute of Applied Communication, Athens, 23-25 May, p. 197-200.
- [14] Mankins, J. C. (1995). Technology readiness levels. White Paper, April, 6. <http://www.hq.nasa.gov/office/codeq/trl/trl.pdf>
- [15] Sauser, B., Verma, D., Ramirez-Marquez, J., & Gove, R. (2006, April). From TRL to SRL: The concept of systems readiness levels. In *Conference on Systems Engineering Research*, Los Angeles, CA.



# **Application, Experimentation, and Market**

# OnEye – Producing and broadcasting generalized interactive videos

Alain Pagani<sup>1</sup>, Christian Bailer<sup>2</sup>, Didier Stricker<sup>3</sup>

<sup>1,2,3</sup> German Research Center for Artificial Intelligence DFKI GmbH, Kaiserslautern, Germany

E-mail: <sup>1</sup>alain.pagani@dfki.de, <sup>2</sup>christian.bailer@dfki.de, <sup>3</sup>didier.stricker@dfki.de

**Abstract:** Interactive videos where objects are enriched with additional information have several important applications including e-commerce, education and gaming. However, the production of such videos is difficult and costly due to the lack of tools to automatize the necessary tasks. In addition broadcasting such videos still remains an issue as current video players do not incorporate the possibility to add supplementary media content. In this paper, we present OnEye, a framework that allows video producers to make objects clickable in their videos and to easily incorporate additional content to the video. The framework consists of different tools that support the creation of such enriched media along the production chain up to the visualization by the end-user. The technologies involve state of the art tracking methods and intelligent user interface, as well as web-based player capabilities. We present an application scenario based on online shopping.

**Keywords:** Interactive videos, clickable videos, embedded advertising.

## 1 INTRODUCTION

Digital video is the driving force behind the expansion of the webTV, IPTV and the new "Generation Mobile". To cope with actual trends, non-promotional and affordable collections of digital videos should be made attractive to the user. The new way for advertising arises rather from the placement of products within a scene, leading to a replacement of the classical TV commercials. This activity is already used in television and cinema and is known as "product placement". Thus, the watching experience is not interrupted, and still the products are presented to the audience. The next development step is called "embedded advertising": an object within a scene "contains" the advertisement, and the viewer can select the object in order to view additional information (such as product photos, specifications, etc.), and may also order the article. This combination of embedded advertising and e-commerce is referred to as t-commerce. But the successful application of this business model presupposes a good use of the technology of object tracking within the digital video to track over longer sequences and to allow selection by the viewer.

In this paper, we introduce tools to create such enriched video content and to present them to the audience in a specific way. We show that a vision based object tracking can help in the generation process, but also that an



Figure 1: The three components of the OnEye system: OnEye Creator, OnEye Videos and OnEye Player.

interactive process is necessary. Furthermore, we present a new video player capable of reading enriched video content over the web. Our technology called "OnEye" is composed of three elements (see Figure 1): OnEye Creator is a web-based software that allows for editing standard videos, tracking objects and creating hyperlinks for tracked objects. The outputs of OnEye Creator are enriched videos that we call OnEye Videos. They contain encrypted supplementary information in form of an XML file. These videos can be read by a specific player called OnEye Player, which is available for Desktop, Tablets and Smartphones.

The remainder of the paper is organized as follows: In Section 2, we review existing systems and discuss the requirements of a production tool for interactive videos. We then present the software OnEye Creator in the light of the provided tracking methods and the intelligent user interface in Section 3. Section 4 presents the Videos and the Player. We present the results of our evaluation in Section 5 before concluding and addressing future work.

## 2 RELATED WORK

In the recent years a lot of effort was put into creating full automatic object tracking approaches. An overview can be found in overview works like [1, 2] or tracking evaluation works like [3, 4]. By contrast, only very few works addressed the problem of semi-automatic tracking, although full automatic tracking is still not reliable enough for many practical applications. One semi-automatic framework is presented by Bertolino et al. [5]. Their tracking algorithm is segmentation based, and it can create very accurate results with exact object borders, as long as it tracks correctly. The user's task is to initialize the segmentation and to correct it if it gets erroneous over time. To fulfil the task the application provides the user

**Corresponding author:** Alain Pagani, German Research Center for Artificial Intelligence, Trippstadter Straße 122, 67663 Kaiserslautern, Germany, +49 631 205 75 3530, alain.pagani@dfki.de

several frame based editing tools. Further semi-automatic segmentation based tracking approaches that work similar



Figure 2: The Graphical User Interface

can be found in [6] and [7]. For these approaches however, a complete segmentation of the object has to be provided, which is often not necessary for clickable videos, and needs a time-consuming human supervision. In our approach in contrast, we optimize the time consumption by providing intelligent tools in order to minimize the user interaction while guaranteeing verifiable correct results. Actually, the requirements for producing reliable clickable videos are quite different from the requirements of classical object tracking. In the object tracking literature, researchers implicitly aim at solving the full-automatic tracking problem, which can be defined as follows: given one single view of an object of interest (usually in the first frame of a video sequence), follow this object throughout the sequence despite possible appearance changes. Results of such papers usually show that in some cases it is possible to follow an object for a time frame ranging from a few seconds to a few minutes. Benchmarks [4] have shown that no tracker is able to track reasonably every sequence, and the best trackers can count on a success rate not exceeding 80% (and less, depending on the sequence). In the production of clickable videos in contrast, the requirements are reversed: the result must be 100% correct, whatever it costs. This means that at least a last verification by a human user is indispensable before validating the results. In our approach, we make use of this human intervention, but recognizing that human resource is costly, we explicitly aim at minimizing user interaction while guaranteeing perfect results. Existing applications like Klikthrough [8] or VideoClix [9] use manual intervention and sell video processing as a service (pay-per-video). WireWax [10] is the only currently available video edition software that allows for interactive object tagging. In this software, faces are automatically detected and tracked, but the tracking of other objects is difficult due to the lack of user intervention and validation. Our system aims at bridging the gap between automatic tracking and full manual tracking by providing the right tools to the user.

### 3 ONEYE CREATOR

#### 3.1 Object multiple selection and tracking

In this section we describe our video edition software for object tracking, OnEye Creator. The software consists in a server-side application and a web-based client. The client



Figure 3: The sequences used in the evaluation. From top to bottom: "Lemming", "Liquor", "Board", "Faceocc" and "David"

implements the graphical user interface, and the tracking algorithms are running on a distant server. Figure 2 shows the graphical user interface for video edition: In the upper half, the video is shown as in a standard player. A timeline allows the user to seek a given frame or to play /fast forward or rewind the video. The user then has the possibility to provide examples of the considered object in one or several so-called reference frames by simply drawing a bounding box around the object of interest. We call the frames where the object has been selected by the

user **user-specified frames** or **USF**. Note that the user will usually provide several USFs for a given object over the full sequence. This contrasts with classical automatic tracking, where the user usually provides only the position of the first frame of the sequence. This allows us to develop advanced tracking strategies that exploit the multiple user input.

Once a sufficient number of USFs have been entered by the user, automatic tracking can take place. We have implemented 6 of the best state-of-the art object trackers in the system in a modifiable way, so that more and more trackers can be added to the system. The available trackers so far are the following: General methods. The general tracking methods we implemented are an object detection-based methods that builds upon the idea of the P-Channel representation [11], a modified version of the MILTrack algorithm [12] (with HAAR and HOG features), Visual Tracking Decomposition (VTD) [13] and Circulant Structure with Kernels (CSK) [14]. Additionally, we implemented two specialized methods: the first one is a color based tracker that is extremely reliable if background and foreground can be separated, and the second other one is a blob tracker that works only with static background.

### 3.2 Necessity of a multi-tracker approach

We conducted a study with standard sequences in order to characterize the different trackers and to evaluate the feasibility of tracker selection. The results of the evaluation are detailed in Section 5. The outcome of the evaluation showed, that no existing tracking algorithm was able to track successfully an object automatically in all the sequences. Some trackers seem to be specialized for specific cases, and some sequences are too complex to allow for automatic tracking. However, in our application, we are seeking guaranteed 100% correct results. We therefore implemented all trackers in the software and allow the user to try many trackers at the same time for object tracking. It is convenient to use different trackers on the same sequence, and we have developed an intelligent tracker fusion mechanism based on an adapted majority voting that can automatically select the best tracker among the tested ones to ensure better tracking results.

### 3.3 Exploiting multiple user input

Because the user can select the object to track several times over the complete sequence, we have an advantage over standard automatic tracking methods. We currently exploit this advantage as follows: we first split the sequence into subsequences starting at a USF and ending at a USF. For each subsequence, we can apply one of these strategies: (1) Track forward from the starting USF until the middle of the sequence and backward from the ending USF until the middle of the sequence. This proves to add robustness when compared with standard (single-direction) tracking. (2) Track forward until the end of the subsequence and backward from the end to the beginning

and compare the outputs of each direction. We then alert the user with a color-coded timeline whenever the two directions tracks differ, and he/she can add more USF in the differing parts. (3) With multiple USFs, we can interpolate the trajectory of the object between USFs using a 2D B-spline. Here again, we use a color coded timeline for indicating compliance (green) between the interpolation and the track or differences (red). The user can then rapidly go to the timeframe where interpolation and track differ in order to add one or more USFs in the critical timeslots. Thus, the user can iteratively converge to the correct track. Once the user is satisfied with the results of a tracker, he/she can *validate* a single frame or a range of frames. These validated frames are called **user-validated frames** or **UVF**. Here again, a color code on the timeline of the video easily show which parts of the sequence have been successfully processed, and which ones remain to track and validate. The goal is to validate all the frames of a video sequence.

### 3.4 Exporting tracking results

The procedure of interactive tracking can be repeated for as many objects as wanted in a given video sequence. For that, the user simply chooses “new object” in the menu of the software and can start to define the track of the second object. Once all objects of the sequence have been tracked, the results can be exported to an XML file that stores for each frame and each object the position of the bounding box of the object. After exporting, the video file is enriched with extra information about object location that can be used for making clickable video.

## 4 ONEYE VIDEOS AND PLAYER

An OnEye Video is a video file that contains the location of one or several objects over the sequence as supplementary information. In the current version of our software, this information is encoded in a separate XML file, but in future versions we will encode it directly into the video file. In order to use this information, OnEye Videos can be played in a specific Player – the OnEye Player. Our player is implemented in HTML5 and Javascript and can play the video in the same manner as the standard players, while providing the extra possibility to interact with the selected objects by clicking on the object. Once a click has been detected on a pre-defined object, an event is launched – usually the video pauses and an object-specific information is shown on or besides the video. In our prototype, this is implemented by adding an URL to each object in the accompanying XML file. When an object is clicked, the URL of the object is opened besides the video in a mini-browser. This solution is generic and allows for different kinds of content being loaded by clicking.

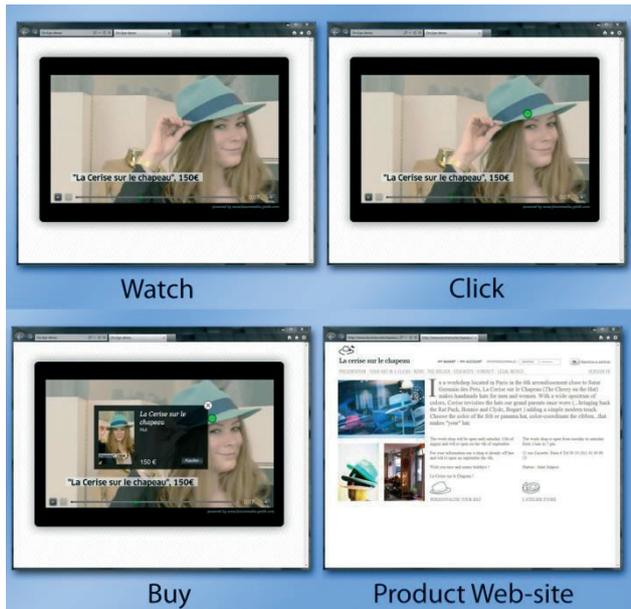


Figure 5: View of the OnEye Player in a commercial scenario

## 5 EVALUATION OF EXISTING TRACKERS

The first idea of the project was to select the best possible tracker for generic object tracking in video sequences. We therefore implemented and evaluated different state-of-the-art tracking methods and compared their output on different representative sequences.

For this experiment, we took 5 sequences usually used for tracker evaluation: “Lemming”, “Liquor”, “Board” from the PROST Dataset [15], “Faceocc” from the FragTrack Dataset [16] and “David” from the IVT Dataset [17] (see Figure 3 for exemplary frames from these sequences). For each sequence, we tracked the object of interest with all the following methods: MILTrack with HAAR features (HAAR), MILTrack with HOG features (HOG), MILTrack with both HAAR and HOG features (HAAR+HOG), MILTrack with Color HOG features (CHOG), MILTrack with HOG features without online learning (only the first frame is taken into account in the appearance model)(HOGffo), P-Channel and VTD. The result of tracking is shown in compliance diagrams: for each frame we measure the compliance between the bounding box found by the tracker  $B_{track}$  and the ground truth  $GT$  by computing the overlap as follows:

$$o = \frac{B_{track} \cap GT}{B_{track} \cup GT}$$

The diagrams in Figure 4 show in the x-axis a threshold of the overlap  $o$  and on the y-axis the percentage of frames of the sequence that attain at least an overlap of value  $o$ . If we take an overlap threshold of 0.5 or 0.6, we see in these experiments that the tracking algorithm that works best for a specific sequence usually performs poor on other sequences, and that no single tracker produces acceptable tracking results for all the sequences. It was therefore necessary to adopt a strategy where many different trackers are used, with an intelligent fusion of tracker results as well as a user-initiated validation.

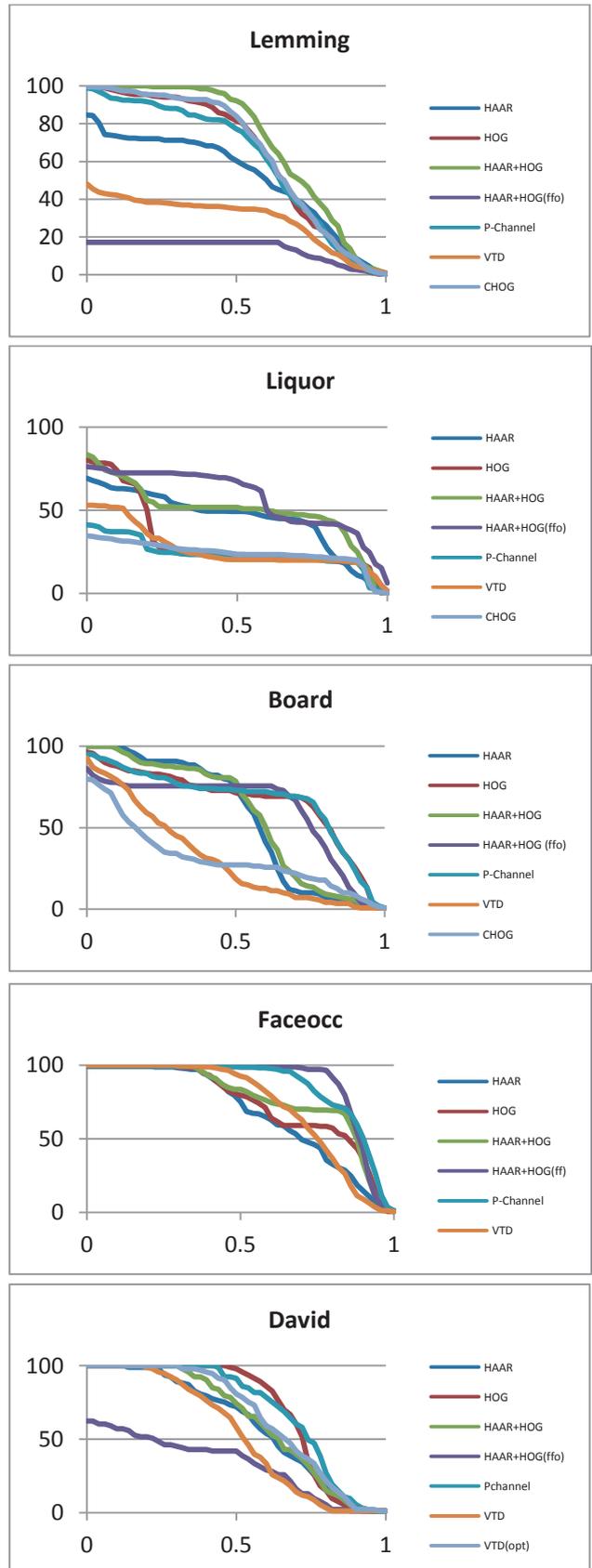


Figure 4: The sequences used in the evaluation. From top to bottom: “Lemming”, “Liquor”, “Board”, “Faceocc” and “David”

## 6 CONCLUSION

In this paper, we presented OnEye, a framework for producing and broadcasting clickable videos. The system comprises a web-based video editor that allows for object tracking with user interaction in order to guarantee correct results. The time required to produce the tracks is optimized thanks to tracker fusion techniques and fast validation process based on interpolation techniques. The outputs of the editor are so-called OnEye Videos that contains the tracks of one or several objects of interest. These videos can be played in a generic player called OnEye Player that reads the tracks and transforms clicks into events. Such an event can be the opening of a mini-browser showing a webpage where the object can be purchased. In future work, we plan to further investigate our fusion strategies in order to validate the automatic choice of the best tracker.

### Acknowledgement

This work has been partially funded by the BMBF VIP project OnEye under contract number 03V0007 and the European project AlterEgo under contract number 600610.

### References

- [1] H. Yang, L. Shao, F. Zheng, L. Wang, and Z. Song. Recent advances and trends in visual tracking: A review. *Neurocomputing*, 74(18):3823–3831, 2011. This is reference No. 2
- [2] M. Chate, S. Amudha, V. Gohokar, et al. Object detection and tracking in video sequences. *Aceee International Journal on signal & Image processing*, 3(1), 2012.
- [3] Q. Wang, F. Chen, W. Xu, and M.-H. Yang. An experimental comparison of online object-tracking algorithms. *SPIE: Image and Signal Processing*, pages 81381A–81, 2011.
- [4] Y. Wu, J. Lim, and M.-H. Yang. Online object tracking: A benchmark. *CVPR 2013*
- [5] P. Bertolino. Sensarea: An authoring tool to create accurate clickable videos. In *Content-Based Multimedia Indexing (CBMI), 2012 10th International Workshop on*, pages 1–4. IEEE, 2012.
- [6] H. Zhong, L. Wenyin, and S. Li. Interactive tracker—a semi-automatic video object tracking and segmentation system. *Microsoft Research China*
- [7] I. Grinias and G. Tziritis. A semi-automatic seeded region growing algorithm for video object localization and tracking. *Signal Processing: Image Communication*, 16(10):977–986, 2001.
- [8] <http://www.clikthrough.com> – June 2013
- [9] <http://www.videoclix.tv> – June 2013
- [10] <http://www.wirewax.com> – June 2013
- [11] A. Pagani, D. Stricker, and M. Felsberg. Integral p-channels for fast and robust region matching. In *Image Processing (ICIP), 2009 16th IEEE International Conference on*, pages 213–216. IEEE, 2009.
- [12] B. Babenko, M.-H. Yang, and S. Belongie. Robust object tracking with online multiple instance learning. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(8):1619–1632, 2011.
- [13] J. Kwon and K. M. Lee. Visual tracking decomposition. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1269–1276. IEEE, 2010.
- [14] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista. Exploiting the circulant structure of tracking-by-detection with kernels. *ECCV 2012*
- [15] PROST: Parallel Robust Online Simple Tracking - Jakob Santner, Christian Leistner, Amir Saffari, Thomas Pock und Horst Bischof – *CVPR 2010*
- [16] Robust fragments based tracking using the integral histogram - Amit Adam, Ehud Rivlin and Ilan Shimshoni – *CVPR 2006*
- [17] Incremental Learning for Robust Visual Tracking - David Ross, Jongwoo Lim, Ruei-Sung Lin, Ming-Hsuan Yang – *IJCV 2007*

# KoKoo (Kontent Kooration)

## Evolving a Content Curation System To a comprehensive Editorial backend platform

Fabio Luciano Mondin – **Telecom Italia** - [fabioluciano.mondin@telecomitalia.it](mailto:fabioluciano.mondin@telecomitalia.it)  
Daniele Merola – **Politecnico di Torino** - [daniele.merola@polito.it](mailto:daniele.merola@polito.it)  
Lucia Longo – **Politecnico di Torino** - [luca.longo@polito.it](mailto:luca.longo@polito.it)  
Maurizio Belluati – **Telecom Italia** - [maurizio.belluati@telecomitalia.it](mailto:maurizio.belluati@telecomitalia.it)

## 1. Introduction

The aim of this paper is to show how a prototypal system, designed as a general purpose stand-alone *content curation* tool could be evolved by following some alpha user's feedbacks to an comprehensive multi-service platform. The widespread diffusion of mobile devices, such as smartphones and tablets as long as the availability of mobile wideband services are increasing day by day the number of players in the ICT Market.

In such a scenario, following the user needs becomes a critical issue, since it is likely for the users. to find products and services quite similar to the one you are offering, better fulfilling their needs. KOKOO (KONtent + KOO(Ü)ration) is a comprehensive platform made by Telecom Italia R&D division.

It is a solution for solving the growing content provider needs to find new and most interesting news to offer to other users on different media, aggregating them in a personal journal with a similar look and feel. Chapter 2 will show the old system (presented also at Nem Summit 2012 showcase), chapter 3 will present user feedbacks and chapter 4 will show the new system and all of its aggregated services, stressing how this was designed by following user feedbacks.

## 2. The Old Content curation Service

The first *content curation* tool, shown at Nem Summit 2012, was developed in order to offer the users a platform to create their personal journals by means of aggregating content coming from different sources. The initial implementation included twitter and *Youtube* as sources. Afterwards the code structure was redeveloped in order to make simple for the developers to add other new sources and as a proof of concept, *Virgilio* and *OkNotizie*, two Italian news providers were added to the *curation* tool.

The idea was rather simple. An editing environment composed of a dashboard and a simple search box.

Typing the keyword on the search starts the search over the enabled sources, each "found" news is "normalized" to a standard format called "nip" and by simply dragging and dropping the nip on the dashboard, the news is added to the journal.

Concerning the view of the journal, the layout is automatically computed basing on the number of articles inserted in the journal.

From a service perspective, the idea was to give the user an environment to build and share their personal journals. Moreover, we were trying to develop dynamic queries, a mechanism of automatic journal generation on a topic base, an interesting use case in which the user sets the system so to build up a journal on a specific topic with the first n news from the x source.

The next figure shows a snapshot of the old interface.



Figure 1 – The old tool's interface

## 3. Alpha User's Feedbacks on Content Curation

As long as the first prototype evolved, we felt the need to check whether the application design performed by IT experts fits well with the final user needs or not.

Following the approach theorized in the User Centered Design paradigm, two focus groups were performed aiming at verifying the acceptability of the application, its easiness of use and its pleasantness.

Both focus groups were organized with users belonging more or less at the same age. The first group was for young people with ages spacing from 16 to 24, while the second group was composed of adults aged 35-45.

The results was that for this kind of users, the platform as it was built, was almost useless. It seems that users did not feel the need to build their own journal to share with friends on a specific platform. The idea to create a newspaper just for himself to read all alone was perceived quite useless and a bit complicated. Users have considered instead that build a newspaper with others collaborators would be very helpful and could meet the needs of organized groups (such as. Companies, associations, etc ...) with the necessity to create content and spread news. Almost all of the users evidenced the strong need to share whatever to online social networks such as Facebook or Twitter.

Very positive on the other hand was assessed the possibility of having an intelligent news aggregator that would allow people to select, automatically and with little effort, relevant news about topic of interest. To be able to have multiple sources in parallel in searching content seemed very useful and enriching, so that our platform, for users interviewed, had a benefit compared to other types of services available.

The perception of this service was quite different from the expected one: users expected the system to make something for them, rather than the contrary. The Service profile described by the focus groups was something standing in the middle between *Pinterest* and *Flipboard*, that is why we decided temporarily to abandon the “general” user target for the *curation* system.

We then tried to understand if an (even modified) version of the tool could suite someone else needs and so we could point out that such a tool could be of interest for users needing to produce content quickly to send to other users. Our new target was then made of bloggers, editors, TV provider aiming at broadcasting additional content quickly and so on as if being a content producer could be the way to overtake the natural user expectation of a system “making things” for the user.

## 4. The new KoKoo tool

Since the focus groups and alfa-user testing had shown that this service seemed to be more suitable for users who were “expert” in aggregating and dealing with content. Our idea was to evolve it into a comprehensive “backend” platform, integrated with other services developed or in development in Telecom Italia, a platform in which the final user is standing on a different level with respect to who is using the platform.

*Kokoo* is a tool that emphasizes the aspect that gives more importance to the content curation, unlike automated services (e.g. Google News), that is the human nature of who is the content curator: in fact, it offers to the user a complete tool to search and find quality contents in line with what user wants, by using multiple sources at the same time, social networks (*Facebook*, *Google+*, *Twitter*) and the web in general, an important aspect for our new target.

Contents are listed for their intrinsic value, calculated according to several social and internal parameters: once selected, and customized if necessary, they are aggregated in social journal, easy to share on various social networks, and automatically saved on the platform. The web platform architecture is shown in the following figure 2 below.

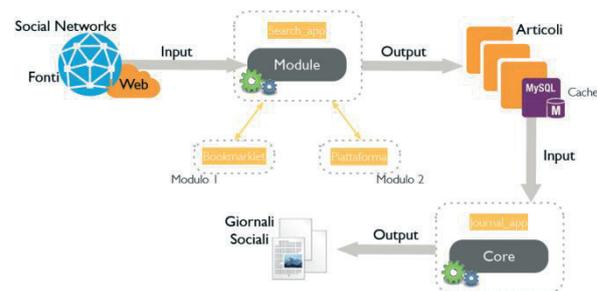


Figure 2 – Kokoo internal architecture

In the current web content's generation and curation scenario, Flipboard seems to be the only competitor, but besides the possibility to create personal journals, belonging to both platforms, Kokoo enhances the user experience by its own set of APIs: in fact, the creation of social newspapers is extended with other specific services.

Users can publish a newspaper with a specific summary, generated by a powerful algorithm carefully designed, They can compose articles using several sources of information; not only those related social network but also TV programs watched in real time and blog articles from around the world, enhance the lessons in a class with interactive content made available by the teachers.

At The time of writing, there is no content curation platform used in the public administration: Kokoo will also be a tool to generate information content made available by the government to the citizens of their communities, with the main aim of making more active the figure of the people within the institutions themselves.

User searches contents by using a module called *Search\_app*, composed of two sub-modules whose activation depending on whether the user is adding news to its journal by using the platform or the *bookmarklet*. The first sub-module uses social network APIs (Facebook, Google+, Twitter) to collect information on a specific topic while the second just allows the user to add to his social journal the content of the web page who is visiting.

Moreover, the platform has been designed to show to the user the latest social contents according to a specific order (using a caching engine and a ranking algorithm specially studied), on the basis of the most influential *hashtag* during day.

Once the contents have been added to the social newspaper, the core module of *Kokoo*, called *Journal\_app*, organizes them, caring on graphics and persistence of the same newspaper within the platform. The final output will be the complete social journal, accessible directly on *Kokoo* and on various social networks.

*Kokoo* has been redesigned with the role of middleware in a more complex system, whose architecture is shown in figure 3.

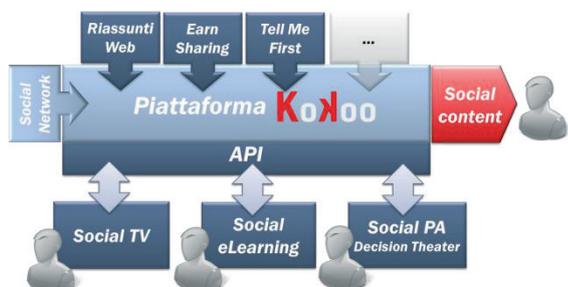


Figure 3 – Overall System Architecture, Involving Kokoo.

On the one hand, the modularity that characterizes the web platform makes easy adding external services (e.g. *EarnSharing*) that can enrich the user experience throughout the process of *content curation*.

Moreover, using a set of proprietary APIs, the ability to interface *Kokoo* with other platforms such as *SocialTV*, the *Social eLearning* and *Social PA* is an advantage in terms of quality of contents produced and in terms of the degree of contents' integration in the social panorama.

As we said previously, the platform is interfaced with many different services that are going to be described in the next paragraphs.

#### 4.1. Kokoo for Social TV

In a social TV scenario, *Kokoo* will be used by the broadcaster to quickly produce content to send on tablets running a second screen application synced with specific programs. *Social Journal* on second screen can give an hint of the most interesting social reactions to a TV program.

#### 4.2. Kokoo for Social eLearning

In this scenario *Kokoo* acts as a content creation platform for school material. Internet and social networks, can be a rich environment for finding teaching content, teachers and students can search on *Google*, *Wikipedia*, *Youtube*, *Twitter*, *Facebook* etc. articles, videos and general contents related to the topic they are studying at school. Using *Kokoo* platform their can create collections of related content that can be exported as e-book and used as studying material.

#### 4.3. Kokoo and EarnSharing

*Earnsharing* is a platform allowing authors of multimedia content to get money from the viral diffusion of content on online social networks. Money earned are shared between both the author and the people who shares the content. *Kokoo* has been integrated. *Kokoo* uses *earnsharing* repository in order to increase the number of sources available in *kokoo*.

#### 4.4. Kokoo and “Riassunti Web”

*Riassunti Web* is a tool developed in a project work financed by Telecom Italia. It is basically a tool to summarize sets of content with a common topic by extracting statistically the most important topics. It is interfaced to *Kokoo* both as a summarization tool for a single journal and a tool to summarize a set of news coming out from a search keyword.

#### 4.5. Kokoo and “Social PA”

*Kokoo* has been presented to subjects belonging to the public administration. It will be a backend tool for the public administration in which someone will

build a journal with specific news to be broadcasted to citizens via a specific apps, especially containing news about events etc.

## 4.6. Kokoo and “TellMeFirst”

We are working to integrate our platform with third party semantic service, called “tellmefirst” [<http://tellmefirst.polito.it>]. This service, which was developed externally with a funded tutoring program by Telecom Italia is now released as open source. The service actually trying to understand the seven main topics in a text by performing queries on semantic databases.

## 5. Conclusions

### 5.1. KoKoo business model

The main KoKoo business model is based on advertising. Collecting user usage data, profiling and clustering users with similar interest (you can get this kind of information analysing content, keyword and main arguments of selected articles or joined journals), you can provide different and user centric advertising. Like traditional and printed journals or more recent social network business model, advertising is one of the most important kind of revenue on content curation system.

The model could be more interesting and innovative if advertising revenue is shared among different players in the content provisioning chain: the platform provider, the content creator that create the content and the curator that select, collect, share and

promote the content. This kind of model encourage users to create and collect articles. Each click on an advertise on a curator’s journal make a gain that is shared among the players.

Thinking about specific context environments, KoKoo could be a good backend platform for producing content at support of other kind of services.

In an educational environment for example, teachers can use KoKoo for selecting and producing content for their students. In a Public Administrations environment KoKoo could be used for collecting and producing content information for citizens and in very similar manner in entertainment marketplace could be used for producing and collecting contents from different sources but related to the same event.

In those scenarios, acting as a back-end platform, KoKoo could be provided as ‘software as a service’ business model requiring a fee for using the service for a defined period of time (i.e per month, year) and numbers of users that can use the platform as content curator.

This experience shows how a prototype must always be kept under discussion, until its very last release to the final users. The first “*content curation*” tool was quite an advanced prototype, advanced both in terms of development and technologies used, but in fact the user always has to be put in the center of the development process. On the other hand, even a negative indication coming from the final user, does not mean necessarily that a project must be trashed and forgotten: usually an innovative, well designed tool naturally has its own space, the challenge is in finding the right customers to turn a prototype into a real service.

# Networked Visualisation in Professional Markets: Prospects & Challenges

Augustin Grillet

Barco nv, President Kennedypark 35, 8500 Kortrijk, Belgium

E-mail: [augustin.grillet@barco.com](mailto:augustin.grillet@barco.com)

## 1 INTRODUCTION

Recent technological advances in computing, internet as well as software technologies are opening new ways to perform visualisation functions in networked environments. This brings exciting opportunities for the development of new products and services, as well as new business models. Yet, professional markets have specific requirements that need to be understood and carefully addressed if one wants to take full advantage of these technological advances.

In this paper, we review recent developments in the industry and highlight a couple of open challenges requiring, to the authors' view, further R&D and ecosystem innovation. Barco is a technology company headquartered in Kortrijk, Belgium and a worldwide leader in professional visualisation markets.

## 2 MARKET & TECHNOLOGY TRENDS

Since the first market introduction of flat panel displays and micro-display light valves, display –and related– technologies have continuously progressed to a point where the possibilities to differentiate based on image quality attributes (e.g. size, brightness, contrast, resolution, frame-rate, colour gamut, etc.) are becoming increasingly tiny. Indeed and for a significant portion of end applications, visualization products are already today being primarily benchmarked based on their ability to intuitively ease end-user interaction as well as collaboration between local or remote users.

Such a trend, already common in consumer markets, is now largely impacting professional applications, beyond the pioneering Broadcast market. Markets such as Healthcare, Security & Control, Digital Signage, V&AR, Entertainment and Education are offering a myriad of relevant use cases, from e.g. telemedicine to IP video-surveillance and collaborative research, design & engineering.

Meanwhile, the Internet and its associated hardware, software & services have approached a level at which it is becoming possible to transport and manipulate very high quality video in real-time, in a more cost effective way. As a result, the full arsenal of IT related technologies can now be applied to high-end multimedia applications and commodity IT components can be used for applications that have traditionally been served with proprietary &

very dedicated hardware solutions, typically running on closed networks only.

Last but not least, regulations on environmental compliance of IT & visualization equipment are becoming stricter, and strategies to reduce power consumption need to be rightfully balanced with workflow requirements of the end-users.

## 3 EXAMPLES OF RECENT ADVANCES FROM THE FIELD

### 3.1 Digital Operating Rooms

The Healthcare sector has to cope with an increased shortage of medical resources, leading to a growing need for collaboration between hospital campuses, but also between hospitals and other remote locations (e.g. universities, doctors' or patients' homes, etc.). At the same time, pressure on reducing total cost of ownership and/or maximizing equipment availability calls for more integrated and remotely manageable image distribution systems. Together with healthcare integrators and hospitals, Barco is currently experimenting the introduction of IP technology in the Operating Room (in place of historical AV equipment), with the recently developed Nexxis solution<sup>1</sup>, a full IP-based system for near zero latency (using raw uncompressed format) distribution of video, audio and computer data within the same or multiple operating rooms in the hospital (this innovation was partly developed within the ITEA2 project MEDIATE<sup>2</sup>).

### 3.2 Collaborative Meeting Rooms

The need to improve meetings productivity is a widely shared requirement over many industries. Efficient & reliable collaboration tools are those that can be hassle-free to the user when being operated, either locally or remotely. The new ClickShare<sup>3</sup> concept of Barco has been developed as part of a regional funded R&D project to address both wireless streaming and simple user interaction aspects of screen sharing within meeting rooms, and since introduction in 2012 has won several market/innovation awards. The challenge is now to bring the same level of experience to more demanding use cases, e.g. multi-sites collaborative design reviews meetings, during which high resolution and stereoscopic content

**Corresponding author:** Augustin Grillet, Barco nv, President Kennedypark 35, B-8500 Kortrijk – [augustin.grillet@barco.com](mailto:augustin.grillet@barco.com)

(e.g. CAD model, geo-seismic 3D maps, etc.) mixing graphics, videos, voice and data has to be seamlessly exchanged over public networks or dedicated optical fibre links, and involving remote users.

### 3.3 Digital Cinema

Multimedia content streaming to cinemas using public IP networks and bandwidth reservation mechanisms was recently experimented within the FI-CONTENT FP7 project<sup>4</sup> (part of FI-PPP), as a possible new feature for extra value creation within the digital cinema eco-system. Real time access to cinema distribution locations, besides reducing the operational costs, will enable new business models for the creative industry, from independent producers to local advertisers and to exhibitors themselves (e.g. interactive cinema, business events, live sport or music event streaming, etc.). One of the biggest technological challenges will however be in enabling cinemas to manage real-time workflows, in place of today only scheduled workflows.

## 4 CHALLENGES AHEAD

### 4.1 System performances vs. cost of bandwidth

The rise in image resolution, dynamic range and frame rate is constantly pushing the limits in terms of bandwidth requirements. Along with cinema & broadcast, many industries are embracing high resolution formats at 4K and beyond, sometimes in combination with high frame rate, e.g. for displaying stereoscopic 3D content. For long limited to 8-bits, new applications such as automotive design (V&AR) are now building on 10 or even 12-bits color depth for highly realistic imaging & visualization. The ultimate demand in bandwidth will however certainly come from the wide deployment in the future of applications making use of full resolution auto-stereoscopic or multi-views 3D formats, a development expected to take place within the next decade.

To compensate for the increase in data, new video codecs such as H.265 are getting these days lots of -duly deserved- attention, yet similar developments need to take place also on the graphics side, targeting in particular RGB codecs with 4:4:4 color mode. Nevertheless, not all applications can tolerate compression, e.g. for certification reasons (notably in safety critical applications), and therefore also for this reason significant decrease of the transmission cost per bit/s should be a prime target of the networking industry.

### 4.2 Interoperability

Just like in any other domains, the lack of interoperability is preventing broad adoption of networked visualization systems. More standardization efforts in the industry are therefore required, both at sub-system (including open standards for codecs) and application levels (over different markets and industry sectors). The AVB standards<sup>5</sup> for Pro-AV applications today appear as the

best option to support applications requiring a guaranteed bandwidth/throughput, low or at least consistent latency, real-time monitoring and clock synchronization.

### 4.3 Information Security Assurance

A cornerstone of many B2B networked visualization systems, information security (confidentiality, integrity, availability) represents both an education challenge (for professionals with an AV background), economical challenge (how to handle the complexity and multiplicity of applicable regulations), and technological challenge (e.g. how to support the move to cloud based and virtualized architectures while maintaining all necessary security attributes, and how to implement HDCP-like features over WAN & LANs for digital content protection).

### 4.4 Useability

Limited useability is still dominating in the professional world, in significant contrast with the mainstream evolution seen on the consumer marketplace. This calls for more intuitive HMIs supporting functions such as e.g. transparent & secure user identification. Ambient computing, social media, BYOD, etc. are a few examples of consumer technologies & trends that have yet to be fully adopted by professional environments. Additionally, the reliability of advanced HMIs needs to be improved for reaching a qualified useability in professional, safety critical environments.

### 4.5 Workflow Integration

Integration within professional workflows constitutes another aspect of the useability challenge, as well as a huge opportunity for new value creation. The European eco-system is recognised for the density and quality of B2B system integrators and end-users; developing a privileged relationship with this community would allow more & targeted innovation to take place in Europe regarding workflow integration.

### 4.6 New Business Models

IP networked & virtualized architectures enable the exploration of new business models, such like Visualization-as-a-Service. This however requires new technological solutions, e.g. for the development of QoS guaranteed highly performing image processing pipelines over dispersed COTS servers. Second, open standards are needed to hook up new networked based applications providers to professional services platforms.

Finally, where new business models can typically be nicely explored & demonstrated by start-ups, durable market transformation will in many cases require more profound & coordinated efforts from incumbent players in the eco-systems.

## 5 CONCLUSIONS

The increasing drive to better serve the B2B industry needs with tools that enable remote and intuitive collaboration between professionals is forcing a shift in the value proposition of visualization products & services. Besides the now rather well established research roadmap regarding improvement in image quality attributes, new challenges regarding IP networked & cloud based system architectures require the attention and joint innovation of the broader eco-system, if one wants to support the competitiveness of the creative & content industry with advanced networked electronic media solutions.

It should be the aim of the NEM ETP and Horizon2020 to facilitate such eco-system innovation in Europe.

## References

- [1] <http://www.barco.com/en/products-solutions/networked-solutions/digital-operating-room>
- [2] <http://www.itea2.org/project/index/view/?project=10041>
- [3] <http://www.barco.com/en/Products-Solutions/Presentation-collaboration/Clickshare-presentation-system/Wireless-presentation-and-collaboration-system.aspx>
- [4] <http://www.barco.com/en//News/Press-releases/barco-fundamental-research-is-pushing-the-technology-frontier-and-shaping-the-cinema-of-the-future.aspx>
- [5] <http://www.avnu.org/>

# Composite Media; A new paradigm for online media

Ingar M. Arntzen, Njål T. Borch, Northern Research Institute & Motion Corporation

## Abstract

*Media has been shaped by inherent limitations of the available distribution mechanisms since the advent of broadcasting. We seek to break free of this heritage by fundamentally reconstructing media. We promote the concept of motion as a fundamental building block in all media. By structuring and executing media according to shared motion, a world of opportunity opens up. In particular, our recent invention of highly scalable, cross-Internet motion-synchronization implies that motion-based media is collaborative and multi-device by design. We envision a shift away from the current paradigm, where fixed pieces of content are produced, transmitted and consumed. Instead, media will be composed again and again from a continuously developing corpus of online content and online motion. In this world, viewing, navigation, interaction and authoring can all be collaborative, multi-device activities. We call this Composite Media.*

## Introduction

Over the last decades broadcast media has grown much in complexity. The origins were simple, at least conceptually; TV providers used a one-way broadcast network to entertain a largely passive audience. These days viewers are increasingly interacting using a wide variety of devices, and live TV shows are coupled with web offerings, SMS services, Twitter streams, Facebook integration, infographics, live analytics, user feedback, custom apps, and more. The life cycle of a single TV program has also become more complex, as it is made available through broadcast networks, web-based live streaming options, and catch-up services for archived content. In addition, programs are time-shifted using personal video recorders (PVRs) and DVB receivers.

This increasing complexity is challenging. In particular, there are issues relating to coupling of media, services, devices and users. For example, Twitter messages and web offerings associated with live TV shows are coupled only as they coincide in real-time. This coupling is weak, and breaks as TV shows are time-shifted. Or, as TV shows are migrated from broadcast networks to archives, the lack of coupling forces viewers to manually navigate services, viewing devices, media applications and media offsets. On the other extreme, time-sensitive infographics merged into a streamed TV show couples these media types too strongly. As the TV show is time-shifted, one might want the infographics to reflect this, for instance by including the effects of other interacting VOD viewers. Instead, overlaid infographics are frozen in time.

We argue that such issues with coupling are pervasive and illustrate fundamental limitations of the current paradigm. Broadcast media is still conceptualized in terms of its simple origins; fixed pieces of media content are produced, transmitted and consumed. In particular, the coupling of media types, services, devices and user navigation has no prominent place in this model. To address these issues, we propose Composite Media (CM) as a new paradigm for online media.

## Composite Media

**Composite Media (CM) is online and multi-device by design. A wide range of networked devices, including smart phones, tablets, laptops and smart TVs may connect to a single, shared instance of CM. As this media presentation is requested to *play* by one device, this affects all devices in synchrony. If another device requests the presentation to *pause*, or perhaps to *skip* or *fast-forward*, all devices behave accordingly. Conceptually, instances of CM execute online so that client devices are free to connect and disconnect at any time. Client devices host independent views into the presentation, suitable for their capabilities and the role they play in the presentation. A single user can use multiple devices covering different aspects of the presentation. Collaboration is supported as multiple users access the same presentation.**

For example, a Tour-de-France CM production may define multiple roles, such as “video”, “infographics” with race statistics and maps, and “interactivity” with status updates and comments. In addition there might be several roles for authoring, intended only for the production team. Viewers may then start up with all roles hosted by one device, say the smart TV. However, as a tablet joins the presentation, responsibility for “infographics” and “interactivity” is dynamically transferred from the TV to the tablet. If the race is time-shifted, or navigated for the highlights, all views and roles are still kept in synchrony. In short, CM is a flexible platform for “secondary screen” and beyond.

Composite Media is derived from two distinct ideas:

1. map media content to time - or other relevant media dimensions - and synthesize media presentation based on the current positions on media dimensions
2. shared navigation along media dimensions

### 1. CM as linear media

Real-time synthesis of media based on user navigation is the foundation of linear media. For example, video presentations are synthesized from video frames and subtitle tracks. Flash[2] animations are produced by mapping vector graphics to a navigable event-timeline. Similarly, slide show applications map individual slides to a dimension of discrete slide numbers. More recently, popcorn.js[4] demonstrates that arbitrary web content can be synthesized in accordance with the progression of HTML5 video elements.

CM takes this familiar idea to its logical conclusion. Linear navigation is recognized as common to all linear media, and promoted as a fundamental component of CM. **In CM, media navigation is universally represented by a single concept; unidimensional motion. All CM presentations are driven by motion.** More formally; CM presentations are synthesized in real-time from media content referring to media dimension(s). This synthesis is at all times directed by motion(s) along those media dimension(s).

To implement motion, CM relies on the technical concept Media State Vector (MSV) [1]. The

MSV is based on the classical equations of unidimensional motion under constant acceleration, where motions are described by initial conditions  $[p,v,a,t]$  (i.e. position, velocity, acceleration, time). The MSV is simplistic, yet expressive. Crucially, the MSV supports common navigational primitives of linear media, including discrete steps (next, prev, goto), continuous motion (play, fast-forward), or even acceleration. With MSV's as a fundament, CM readily supports all these navigational primitives. In addition, CM supports mixed navigation styles as well as multi-dimensional navigation.

A central idea in CM is the decoupling of motion from media, and the promotion of motion as a resource in its own right. This gives CM significant power and flexibility. The decoupling of motion implies that motions can be shared between content sources as well as presentational (e.g. visual) components. A single CM presentation can thus be synthesized from multiple, heterogeneous content sources, while allowing content subsystems to operate independently, yet in a coordinated manner. Similarly, UI components directed by shared motions can appear as tightly interconnected parts of a single media presentation, without requiring UI components to communicate, or even be aware of each other. In short, motion decouples content from presentation, and allows CM presentations to recombine them in flexible ways.

In contrast, consider how rich, linear web-based media presentations are currently made. Frameworks like Flash, SMIL, popcorn.js or Prezi [2,3,4,5] all build linear media around similar concepts of time-based user navigation. Yet all of them rest on implementations of motion that are internal and custom to the framework. This makes it hard to combine such presentations with external resources. The HTML5 video element makes this easier by exposing its internal motion through an API, thereby allowing the popcorn.js JavaScript library to synchronize arbitrary web content to the progression of a video. Still, popcorn.js misses a grand opportunity by insisting that the source of motion must be video (or audio). By making motion an object in its own right, CM makes it easy to include multiple videos in a single presentation. Perhaps more importantly, CM allows the construction of continuous media presentations, without requiring the inclusion of continuous media at all.

In effect, CM implies the possibility of a universal API for motion, to which presentational components and frameworks would interface. For instance, a motion-enabled video element would accept external motion, and operate as a slave to that motion. We have verified that this is possible with current HTML5 video elements, even without optimizing the internals for this purpose.

## **2. CM as Online Media**

Above we indicated the value of sharing motions in the context of single-device media presentations. However, the defining characteristic of CM is that motions can be shared across the Internet. In CM, motions (MSVs) are online objects. This allows clients across the Internet, including regular web browsers, to connect and closely mirror the motions of ongoing media presentations. In fact, the accuracy of MSV synchronization [1] implies that shared motions may be thought of as simultaneous, even if clients are separated by large geographical distances, or

communicating over different networks carriers. This is possible by compensating for network latency and clock skew. Network latency is only evident as motions are actively changed, for instance when play or pause commands are issued.

Equally important, online MSVs are extremely lightweight and can be hosted by highly scalable services. This implies that the scalability of CM is likely to be limited not by motion sharing, but by the scalability of its content services. CM presentations may make use of a large number of MSVs. Some of these MSVs may be used in large-scale broadcast scenarios, where asymmetric media control is required. For instance, in the classical secondary screen scenario a TV provider can own and control a broadcast MSV, whereas the audience is restricted to read-only access on their smart TV and secondary devices. Symmetric media control is relevant in smaller social groups, for instance in social collaborative applications. Additionally, CM presentations may allow individual users to create and use private MSVs, and even switch back and forth between private and shared MSVs. Private MSVs are a natural fit for on-demand media, and since private MSVs are still online they can easily be shared among the devices of a given user. Opening a private session for a friend is simply a matter of sharing access and url of the MSV.

The online nature of CM is clear as both motion and content are online resources. Conceptually, CM presentations execute online, independent of its clients. In particular, a CM presentation may continue to play even without connected clients. Clients may connect and disconnect at any time, without disturbing the execution of the CM presentation or other clients. Joining clients may quickly learn the current state of the presentation, as they synchronize with its MSVs and (subsequently) starts loading the relevant media content. Still, despite its online nature, the cost of synthesizing CM presentations is paid locally by clients.

In short, by making motion sharable across the Internet, all the advantages of motion sharing is extended to the distributed setting. In particular, shared motion provides the intermittent glue required for loosely coupling content types, services, devices, UI components, and users, across Internet. The name “Composite Media” similarly refers to how motion-based media presentations are composed seamlessly (in real time), from online content sources, UI components, users, and devices.

We have conceptualized and implemented shared, online motion, and verified its utility as technical foundation for CM. This work was recently published under its technical name; The Media State Vector (MSV) [1]. The Motion Corporation is the world's first commercial company to offer shared motion as a hosted service.

## Applications

Composite Media is to a large extent a new model for online media rather than a replacement of technology. Though the model is fundamentally new, implementation can still build upon existing technology and services while simultaneously open for extensions and novel applications. The following section discusses a few selected topics relevant for broadcast media.

## **Seamless interaction between providers and consumers**

TV broadcasters often provide content from multiple backends such as the broadcast network, live online streams (possibly with a live-window of several hours), and on demand offerings for archived material. While this improves outreach and functionality, it also adds complexity for the end user. The user must for example often manually select which backend to use. Paused live streams can not be resumed if the live window is surpassed. Instead the user must manually relocate the show within the archived service, as well as relocating the offset where the show was previously paused.

When a Composite Media presentation is created around the various data streams and backends, this complexity can be hidden as the user interacts with the same presentation at all times. By consulting an online representation of user navigation (the MSV) clients can at any time select the appropriate backend as well as the correct media offsets. This way, seamless transitions between different technologies can now be done by the client, ensuring a coherent and pleasing service regardless of backend and choice of device. It also allows the user to move the viewing experience from one device to another by simply opening the presentation on a different device. There is no need to pause the content, communicate locally between devices or even for the devices to know about each other. On the same note, collaborative viewing of content can be enabled by allowing users to invite other users to use the same MSV. No additional support is required.

## **Web-based Content Production**

We argue that CM is a generic, flexible and cost effective approach to online media. In particular, by providing a generic synchronization mechanism for the web, CM immediately allows HTML5 to become the standard authoring framework for time-synchronized second screen applications, as well as multi-device media applications in general. As MSVs synchronize all visual elements (on a single page as well as on multiple devices), visual elements do not need to communicate with each other in order to act coherently. This allows sophisticated UI's to be assembled from a set of simpler, general-purpose, motion-enabled UI components. The MSV's inherent support for time-shifting also implies that live CM productions may be reused without modification for later on-demand viewing. Furthermore, authoring may be live, or come from a mixture of prearranged material and live cues. Authoring may also be collaborative, allowing responsibilities to be shared within a production team.

Transforming existing HTML based visualizations to support MSVs and thus be integrated in CM presentations will often lead to simplifying the components as opposed to adding complexity. For example, a content provider can create a control element for CM presentations with buttons for play, pause and navigation, as well as sharing or storing bookmarks etc, all visualized for various devices. Any presentation (even just a single video element) can have that controller added to it's page, thus allowing the user highly recognizable and unified controls for all services, on all devices. This control element does not need to be integrated directly with any media object as this coupling is mediated by the underlying MSV.

## Multi-device Applications

Composite Media is multi-device by design as it depends on shared motion. CM presentations are built from a set of data sources and a set of visual or interactive elements connected to these data sources. The UI elements visualize and interact according to motion, and will as such present the correct data at the correct time for each user. This also means that different visual components can be used on different devices to provide true multi-device experiences.

While multi-device CM has obvious applications within secondary screen scenarios, it can also expand more traditional services. As Smart TVs are more easily available, infographics can be generated by the TV as an independent HTML overlay. This will enable graphics based on current data even if time shifted, or customized data for the user, for the user's location or even integrating the infographic with social media sites, e.g. showing comments from Facebook friends, playing a quiz with friends on the TV etc. The potential of using CM to deliver more targeted commercials is also easily apparent, opening for time-sensitive game style commercials as well as localized or personalized commercials.

## Summary

In short, CM is a highly flexible and holistic paradigm for online media, where devices, users and content providers can work seamlessly together. It fits perfectly with the webification of media content, the increasing abundance of smart devices and the flexibility and social expectations of end users. We have validated the central technical concept, shared motion, and have implemented a hosted service, designed for scalability and ease of use. With this we have made proof-of-concept demos targeting a variety of multi-device applications, including broadcast, secondary device, online education, online newspapers, distributed music orchestration, collaborative slide-shows, collaborative viewing, collaborative authoring, collaborative documentation, as well as interactive visualization of scientific data. We argue that CM reduces the complexity of online media by providing a flexible and unifying model.

[1]: "The Media State Vector; A unifying concept for multi-device media navigation", Ingar M. Arntzen, Njål T. Borch, Christopher P. Needham, MoVid'13, Proceedings of the 5th Workshop on Mobile Video, ACM, pages 61-66, Feb 2013, Oslo, Norway.

[2]: Adobe Flash. [www.adobe.com/flashplatform/](http://www.adobe.com/flashplatform/)

[3]: SMIL. Synchronized Multimedia Integration Language, [www.w3.org/TR/2008/REC-SMIL3-20081201/](http://www.w3.org/TR/2008/REC-SMIL3-20081201/)

[4]: popcorn.js. The HTML5 Media Framework, [popcornjs.org](http://popcornjs.org)

[5]: Prezi, [prezi.com](http://prezi.com)

# Impact of new technologies and social networks on a secondary education theatre project

J. P. López<sup>1</sup>, P. Ballesteros<sup>2</sup>, D. Jiménez<sup>3</sup>, J. M. Menéndez<sup>4</sup>

<sup>1</sup>Universidad Politécnica de Madrid, Madrid, Spain; <sup>2</sup>I. E. S. Al-Satt, Algete (Madrid), Spain; <sup>3</sup>Universidad Politécnica de Madrid, Madrid, Spain, <sup>4</sup>Universidad Politécnica de Madrid, Madrid, Spain

E-mail: <sup>1</sup>jlv@gatv.ssr.upm.es, <sup>2</sup>mandaladu@hotmail.com, <sup>3</sup>djb@gatv.ssr.upm.es, <sup>4</sup>jmm@gatv.ssr.upm.es

**Abstract:** This paper describes the potential impact of social media and new technologies in secondary education. The case of study has been designed for the drama and theatre subject. A wide set of tools like social networks, blogs, internet, multimedia content, local press and other promotional tools are promoted to increase students' motivation. The experiment was developed at the high-school IES Al-Satt located in Algete in the Comunidad de Madrid. The students included in the theatre group present a low academic level, 80% of them had previously repeated at least one grade, half of them come from programs for students with learning difficulties and were at risk of social exclusion. This action is supported by higher and secondary education professors and teachers who look forward to implanting networked media technologies as new tools to improve the academic results and the degree of involvement of students. The results of the experiment have been excellent, based on satisfactory opinions obtained from a survey answered by students at the end of the course, and also revealed by the analytics taken from different social networks. This project is a pioneer in the introduction and usage of new technologies in secondary high-schools in Spain.

**Keywords:** Social Media, Social Networks, Audiovisual Content, Secondary Education, Analytics

## 1 INTRODUCTION

Networked media technologies have become key elements in many areas. For education projects [1], [2], audiovisual content and social networks constitute a powerful set of tools, that can be used to motivate learning and help the students to widen their knowledge, without a cost increase. Thus, networked media technology should enable a more effective use of resources, and should be used as much as possible to improve the relationship between students and teachers, and facilitate students' access to interesting material, without additional cost.

Social networks and web 2.0-based social media services (e.g., Facebook®, Twitter®, YouTube®, etc.) have recently become very popular, especially among young people [3]. One of the main reasons is, because in social networking sites users can participate intensively in activities and services sharing content and opinions, debating and create different groups depending on their needs and interests. The use of ICT as a learning methodology needs to be firmly incorporated

into the classroom, in order to improve the experience of students and teachers alike..

A sense of community is a very important factor in raising motivation in a drama and theatre project. Feelings of membership, belonging to the groups and sharing emotional connection are basic features for motivation [4], as well as the essence of social networks aims. In [5] the usage of social networks is constructively considered in the learning process. Facebook allows members to participate, both inside and outside of the classroom, and to keep connected outside the classroom.

According to the Nielsen "Social Media Report" [6], the usage of social networks among young people is continuously increasing. The proliferation of smartphones and tablets has facilitated access to social media, complementing the access through PCs or laptops, which remain only slightly ahead in terms of medium of connectivity to social media. The trend of utilizing Social Network Sites for education is widely used as a tool for sharing collaboration, posting comments, which leads to the development of a new interactive teaching and learning platform [7]. Among social networks, Facebook is the most used, followed by Twitter and blogs, as reflected by the final analytics. These three systems formed the core of media usage in the experiment, apart from a website and the diffusion of video and audio through the platforms YouTube and Vimeo.

There have been similar experiences in higher and primary education, but this is a pioneer project in public secondary school in Spain. Students from secondary school are aged from twelve up to eighteen years old, which is commonly the age when they start to strongly interact with new technologies, begin to have their own mobiles, and open their eyes to the possibilities of internet and communications in respect to their education.

A variety of experiences have been developed in higher education for different purposes. In [8], Almadhoun proposes social networks for promotional activities in higher education. E-learning and on-line access to materials, in addition to the materials developed in the classroom, are the basis of some research. In many professions, social networks play an important role in different fields, such as civil engineering education [9] or [10], which employs YouTube videos as an important tool for teaching. Additionally, other experiments for social network integration in education can be found in [11], [12], [13], [14] and [15].

**Corresponding author:** J. P. López, Universidad Politécnica de Madrid, Spain, +34 913 367 344, jlv@gatv.ssr.upm.es

There are several studies in which the impact of ICT's capabilities on children and young people is analysed as well, such as [16] and [17]. The first one concludes that the younger group's interaction with technology is different from that of the adult population. According to the second, social media in schools contributes to the organization of big projects or activities, transmitting news and information, and can also be helpful to orient students toward their future career.

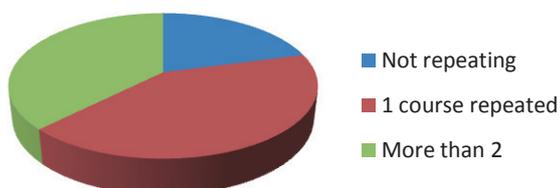
Following, after presenting an introduction and the state of the art in Section 1, we will present an overview of the work environment in Section 2, defining the conditions and main objectives of the project. In Section 3, the work that has been developed is presented, describing the set of tools that has been used for the project. Finally, Section 4 includes the results of the final survey answered by students at the end of the course, while their conclusions are drawn in Section 5.

## 2 ENVIRONMENT

This paper is the result of joining the efforts from university and secondary school, to introduce new networked media technologies in classrooms. For this purpose, the elective subject of drama and theatre from the official educational program has been selected in public secondary high-school I. E. S. Al-Satt, located in Madrid region.

The group was composed of 27 students, girls and boys aged between 14 and 18 years studying 3<sup>rd</sup> course of Secondary Obligatory Education (E.S.O), who had previously selected the optative subject of drama and theatre among other optional subjects offered.

The students included in the theatre group present a low academic level. Analysing their profiles, an 80% of them had previously repeated at least one grade (Figure 1). In Spain, when a student presents low academic achievement during a given school year, they must repeat the year to improve the results obtained in the present course.



**Figure 1. Students' academic profile in number of years repeated**

Also, half of them come from "diversification" programs, which is an educative adaptation for students with learning difficulties and at risk of social exclusion.

All parents or legal tutors of the students signed a consent form at the beginning of the course to allow their children to be photographed or filmed during the project, and the material obtained to be published on the internet.

The project extended from September 2012 until June 2013. In this period, students executed six different performances in front of an audience; the last of them took place on May, the 15<sup>th</sup>.

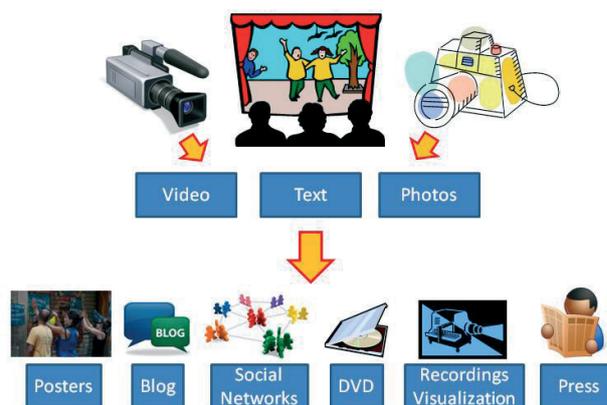
The theatrical production was called "Child Soldier", and was composed of different scenes based on methodology of *avant-garde* and community theatre. This method aims to distribute the spotlight among every individual actor/student. Similar examples of this work methodology are described in [18].

Attracted by the original material generated for this project, "Child Soldier" was selected by La Caixa Foundation for its funded "Caixa Escena" Encuentros, which is one of the most important theatre contests in secondary school.

## 3 IMPLEMENTATION

The main objective of introducing media content and promoting interaction using social networks to improve the relationship between each one of the students and the scene, motivating them to make the best effort and obtain the best result possible on stage. The initial idea was to use already-established social networks (Facebook or Twitter), instead of new platforms, such as Moodle, whose access by students would be less frequent compared to the networks that they normally use.

The evolution of the project was progressive. In the first three months, rehearsals of the first scenes developed were photographed and the material was included in social networks. The best of these photographs were used to produce promotional posters for the project. These posters were used to create a visible image of the theatre company and to raise both the students' and the school community's enthusiasm regarding the project.



**Figure 2. Working schema**

The schema of the project appears in Figure 2. The main purpose was to generate multimedia material in the classroom and on stage to distribute it through different mediums, and to create a brand image of the theatre company.

Three mediums of material were used to document the project: text, video and photos. All these mediums have been mixed to present interesting products to be used for different applications. Photographs were employed for posters and informative papers, which were posted on the walls of the institute to be seen by other students. Also photographs were distributed through social networks and the blog. The content must be attractive enough to increase participants' interest in the networks.

Video recordings were used for visualization in the classroom and a DVD authoring. The students' ability to see themselves on screen was a chance to correct their errors, learn from them, and be able to see what a spectator sees. Video and multimedia were also very innovative in high-schools, a main factor in the learning process.

Local press media played an important role in maximizing project's reach and improving students' motivation.

### 3.1 Social Networks

Social network profiles played a main role from the beginning. Profiles on Facebook®, Twitter®, YouTube® and Vimeo® were created for interaction with students, content sharing and improving the communication and relationship between teacher and student.

Following the studies of [5], Facebook was selected as the first tool in the experiment. The profile was private and limited to the students included in the theatre group and contributors to the project development.

Twitter® is the fastest growing Web 2.0 technology when compared to other micro-blogging platforms [19] and was discovered to be the most popular application amongst the members of the group, as the analytics revealed later. Then, alternative usage of both platforms was performed, in order to motivate students, when posting photos, text relative to the performance and other useful information. Students could interact with this content, posting comments, saying "I like it" or sharing them with other users.

Finally, the video sharing networks, YouTube and Vimeo [10], were presented as the most powerful tools. Videos from rehearsals and performances were uploaded to these two platforms, but also presentation trailers and edited video used as backdrops on stage. Students had the chance to see themselves on the internet, analyse their performances, and be able to improve their features, after watching the content.

### 3.2 Blog

As the social network Facebook had private use, and was limited to the members, another public platform was necessary for promotional and diffusion purposes. A blog developed in Wordpress platform was created to upload general information about next performances, news from different press media or different announcements.

### 3.3 DVD Authoring

The first performance - 20<sup>th</sup> December 2012 - was photographed and filmed during early rehearsals and in front of an audience. Different cameras were used for this purpose, a reflex professional camera for capturing photographs, and high-definition videocameras (with resolution 1280x720) were used from different points of view, including close up and full shots.

All material was classified and ordered, and was used for creating a DVD with the whole performance, which was distributed to the students as learning material. This

constitutes an innovative tool in the learning process, which was considered as the text book of the subject.

The content was edited with a professional edition software tool (Adobe Premiere) to obtain the best result, employing different viewing angles to offer a more impressive viewing. The DVD authoring was developed with another software tool (Adobe Encoder).

Also, the DVD included subtitles in Spanish and English languages in order to ease the memorization of the script. This material received a very good reception among students.

The DVD was used as a promotional tool among other secondary high-schools, for teachers and organization managers in Caixa Escena Encounters, with good reception. Through this DVD, other schools' teachers received an invitation to visit the project's social networks.

### 3.4 Audiovisual Content Viewing

Before distributing the DVD, the first viewing of media recordings was in the classroom. The complete edition of the performance was projected in the first class after the Christmas holidays, and the students' reactions were analysed.

The impact on students was strong and, as revealed in the final survey, crucial in their implication with the project and posterior development. The motivational factor of technology usage was reaffirmed by the fact that all students were moved by seeing the recording of themselves acting on stage.

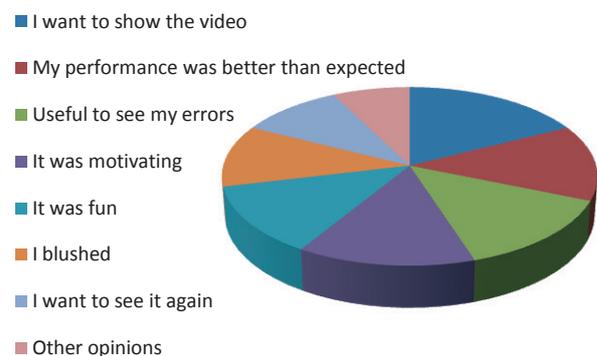


Figure 3. Most-repeated opinions after watching themselves on screen

Among the most common comments and opinions collected after the video viewing, students wanted to show them to relatives or friends, and they judged their own performance as being better than they had expected. Also, video was considered a useful tool in improving their performances. In most cases, students found it motivating or fun to recognize their recurring errors. Only a low percentage admitted to having blushed after watching themselves on screen.

### 3.5 Backdrops

One of the most innovative proposals for the theatre project is the usage of video projections used as backdrops for theatrical purposes. Backdrops substitute the typical stage sets, creating a futuristic effect that fill the stage with light and shocking images.



Figure 4. Samples of backdrops usage

Backdrops were used to complete the stage, but also some videos were edited and projected over the screen with a meaning, not only for artistic purposes. For example, there were cases of “bullying” in the high-school, and one of the videos projected denounced against this type of behaviour. This video received a good reception among the audience and students, as the final survey revealed.

### 3.6 Other promotional tools: posters, website, mailing and local press

But social networks were not the only technology used with motivational purposes in the project. The posters created through photographs captured in rehearsal and performances were important too. The photographs were modified with tools, such as Adobe Photoshop, to offer a more professional result, searching for shocking images and icons, to leave no one indifferent.

The project had an e-mailing address to be used for communication among the members of the group, and also for inquiries from outsiders.

Also, a website was created, with the interface that can be seen in Figure 5.



Figure 5. Website Interface

Finally, local press media, such as “La voz de Algete” or “Crónica Norte” used the material for more extensive promotional ways in their paper edition and also online. Both published the news of the group, photos and information, which students were able to share, motivating them to continue with their effort and helping them to improve. Also, national newspapers Europa Press and La Vanguardia, mentioned the “Caixa Escena” Encounters, highlighting the performance of I. E. S. Al-Satt.

## 4 RESULTS

The project results are collected through different channels. First of all, subjective reactions and reception of the audience and students offered a general qualitative impression. In addition quantitative and measurable data was collected,

coming from the internet analytics about the media content consumption and on the heels of a survey conducted at the end of course, revealing interesting information about the impact of the project.

### 4.1 General Impressions

An improvement in students’ behaviour was detected, which is reflected by a decrease in the number of disciplinary reports among students involved in the project.

The degree of school absenteeism has decreased in theatre class compared to the previous year, computing a percentage of attendance higher than 90% for more than 80% of students, which is a very high level in secondary school, especially compared to the same subject in previous year. The other students presented a percentage of attendance not lower than 70%, most of the cases with not-justified absence, which is another indication of students’ difficult profile.

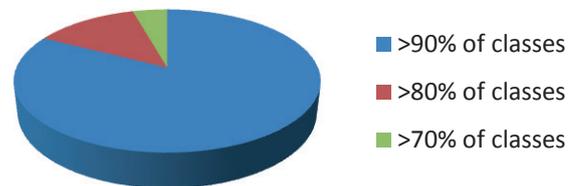


Figure 6. Percentage of students' class attendance

On the other hand, it must be mentioned that their academic performance did not experience an improvement, except in theatre class, in which their effort was rewarded with good grades.

### 4.2 Social Networks Analytics

Analytics from YouTube reveals a growing interest in the evolution of the project. The statistics reveal more than 1800 views of an amount of thirty different videos, and more than 3300 estimated minutes watched, from November 2012 to June 2013.

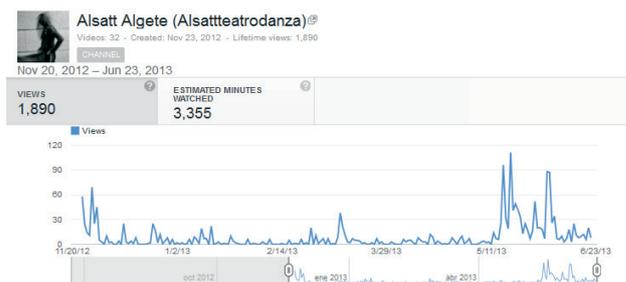


Figure 7. Global analytics from YouTube

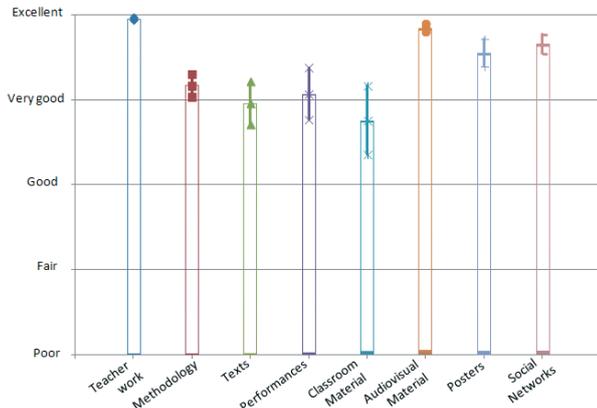
The evolution of interest in other networks, such as Twitter or the blog, also present significant trends, similar to Figure 7. Additionally, videos uploaded to the Vimeo profile must be taken into account, making a total of more than 2000 views in videos from the theatre company.

### 4.3 Final Survey

The results from the final survey are collected in this section. They were asked about different aspects of the experience, the

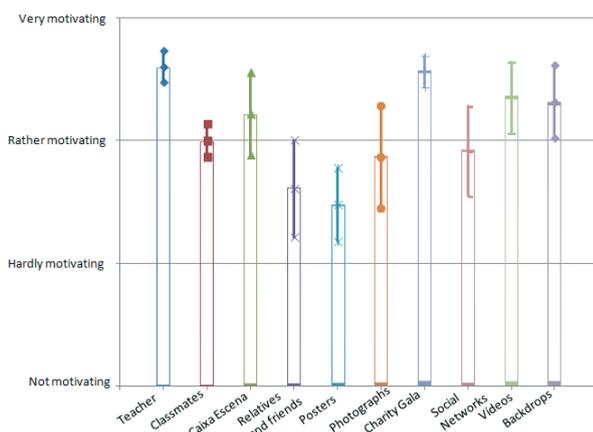
quality of material, motivational factors, social networks, performances, and finally they had some space for expressing free opinions.

Firstly, they were asked to evaluate the quality of different aspects from the project. Results revealed the high quality of the project, especially in reference to the teacher's work on the project, but also to the quality of social networks, methodology and graphic material (Figure 8).



**Figure 8. Students' subjective assessment about different aspects of the project**

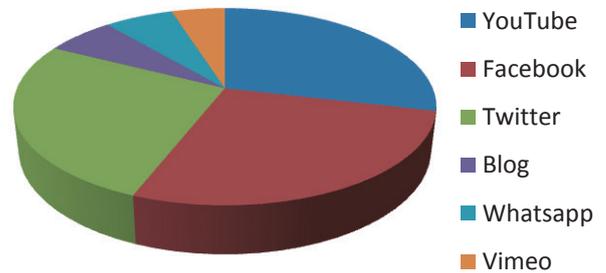
Figure 9 shows the most motivational factors from the project. According to the students, the teacher is the most motivating factor. The videos, backdrops and, of course, the performances (Caixa Escena and Charity Gala), were also highly rated. Then, content in social networks was also highlighted, because it offers the chance to interact with other students, improving their communication. Posters and photographs were less considered in comparison with other factors, but it is true that students received these positively.



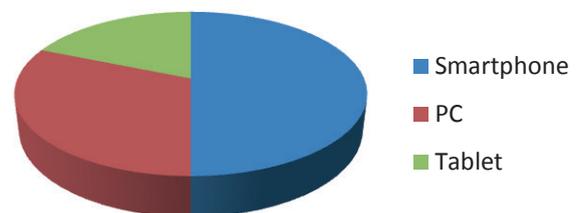
**Figure 9. Motivational factors**

Among the most-used social networks, we find that students prefer the networks they commonly use in their normal environment. YouTube, Facebook and Twitter usage is more frequent than other specific sites, as happens with the blog (Figure 10). Most of the students claimed to connect to these networks daily or weekly, searching for theater content and photos. The most commonly reported method of connection

was through smartphones, in first place, and PC's as the second option, while tablets are less frequent for their age (Figure 11).

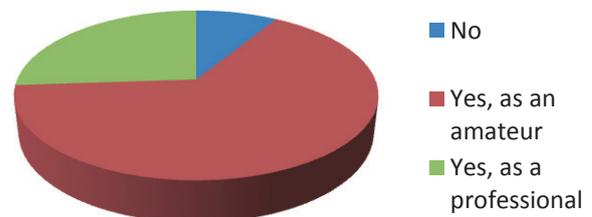


**Figure 10. Most-used social networks**



**Figure 11. Devices used for social network access**

Finally, as an indicator of the student's enthusiasm in theatre following the project, we asked the question: "Would you like to dedicate time to theatre in the future?". Most of them answered "Yes, as an amateur" or "Yes, as a professional", as seen in Figure 12. Also, in free opinions section, students congratulate teachers for their effort and satisfactory results, and they mentioned the great fellowship created among them.



**Figure 12. Opinions about future dedication to theatre**

## CONCLUSIONS

The project could be exportable to other educational centers. Not only for similar drama and theatre experiences, but even for other subjects, in which technology and established social networks could represent a key factor for improving their motivation.

The usage of social networks that they normally use is very important, instead of using a specifically-created platform. As seen in analytics, YouTube, Facebook and Twitter have better reception than the blog, because the information is in their environment and they do not need to make the effort of accessing a new platform or website. Communication has improved between teacher and student without any cost, and photographs and graphical content are accessible a short time

after being produced or generated. Social networks transform a local project in something global.

Audiovisual and graphical content are profitable in the learning process. Students are able to improve after watching their performance and share their experience with relatives and friends, as revealed in the survey, which is also motivating. Posters, photographs and news on press, maximize the project's scope and make them feel satisfied, which will influence their work effort. If they find the material created for them as professional as possible, they will take it seriously. It will be their responsibility.

The youth are subject to multiple stimuli nowadays, and they usually find traditional education system too static. The insertion of new technologies and social networks could make them find this process more dynamic and motivational, especially when they are students with learning difficulties. For this purpose additional effort is needed for taking up the educational approach, including time for preparing the professional material, dissemination and setting up the entire digital environment.

For future work, it would be interesting that the experience would be applied to different high-schools to compare the impact over students in different ages and origins in order to analyze their behavior through social networks. The project must expand the acquired knowledge to other subjects, not only theatre subject, for example, art, music or sciences.

Nevertheless, we must clarify that social networks and new technologies played an important role in the project, as demonstrated with analytics exposed and student's opinions, but it is not the only factor for project success. The work methodology, the motivation of group work, the interest in the content of the performance, and especially the degree of involvement of the teacher were also important. It would be possible to replicate the model, but it requires a full time involvement of the teacher or teachers, who needs to have knowledge about internet and social networks, being also a creative and artistic person.

### Acknowledgement

This paper is based on joint work from Universidad Politécnica de Madrid and I. E. S. Al-Satt with the aim of introducing new technologies in a secondary education institute, performed in the framework of project TEC2012-38402-C04-01 HORFI, which is partially funded by the Spanish Ministry of Science and Innovation.

The authors would like to acknowledge La Caixa Foundation for their support in Caixa Escena Encounters. Also, we thank the I. E. S. Al-Satt direction and teacher Silvia Eva Agosto, for their support and collaboration in the experimental project. Also, we acknowledge the members of local press "Crónica Norte" and "La Voz de Algete" for the diffusion of media contents, and the City Hall of Algete. And last but not least, thanks to all the students implied in the theatre project, whose effort made it possible.

For further information about the project on social networks. Twitter: @AlSattteatro, YouTube Channel: Alsattteatrodanza, Facebook: alsatt.teatrodanza, blog: alsattalgete.wordpress.com, email: alsattteatrodanza@gmail.com, website: www.alsattteatrodanza.tk

### References

- [1] Srivastava, S., "A study of multimedia & its impact on students' attitude," Technology Enhanced Education (ICTEE), 2012 IEEE International Conference on, vol., no., pp.1,5, 3-5 Jan. 2012
- [2] Srivastava, P., "Educational informatics: An era in education," Technology Enhanced Education (ICTEE), 2012 IEEE International Conference on, vol., no., pp.1,10, 3-5 Jan. 2012
- [3] Silius, K.; Miilumaki, T.; Huhtamaki, J.; Tebest, T.; Merilainen, J.; Pohjolainen, S., "Social media enhanced studying and learning in higher education," Education Engineering (EDUCON), 2010 IEEE, vol., no., pp.137,143, 14-16 April 2010
- [4] D. W. McMillan and D. M. Chavis, "Sense of community: A definition and theory," in Journal of Community Psychology, Vol. 14, No. 1, pp. 6-23, 1986.
- [5] Raetham, P.; Firpo, D., "Using Social Networking Technology to Enhance Learning in Higher Education: A Case Study Using Facebook," System Sciences (HICSS), 2011 44th Hawaii International Conference on, vol., no., pp.1,10, 4-7 Jan. 2011
- [6] Nielsen, "Social Media Report 2012"
- [7] Treepuech, W., "The application of using social networking Sites with available online tools for teaching and learning management," IT in Medicine and Education (ITME), 2011 International Symposium on, vol.1, no., pp.326,330, 9-11 Dec. 2011
- [8] Almadhoun, N.M.; Dominic, P.D.D.; Lai Fong Woon, "Social media as a promotional tool in higher education in Malaysia," National Postgraduate Conference (NPC), 2011, vol., no., pp.1,7, 19-20 Sept. 2011
- [9] Furuta, H.; Takahashi, K.; Ishibashi, K.; Usui, M.; Nakatsu, K., "A proposal of civil engineer education system applying social networking service," Soft Computing and Intelligent Systems (SCIS) and 13th International Symposium on Advanced Intelligent Systems
- [10] Chtouki, Y.; Harroud, H.; Khalidi, M.; Bennani, S., "The impact of YouTube videos on the student's learning," Information Technology Based Higher Education and Training (ITHET), 2012 International Conference on, vol., no., pp.1,4, 21-23 June 2012
- [11] Figl, K.; Kabicher, S.; and Toifl, K.; 2008 "Promoting social networks among Computer Science students," Proceedings of 38th IEEE Annual Frontiers in Education Conference, Saratoga, NY, October 2008, pp.S1C15-S1C20, 22-25.
- [12] Stollak, M. J.; Vandenberg, A.; Burklund, A.; Weiss, S.L. "Getting Social: The Impact of Social Networking Usage on Grades among College Students". St. Norbert College. ASBBS Annual Conference: Las Vegas 859 February 2011.
- [13] Tao Hu; Jian Fei, "The mental accounting and moderating role of habit in usage of online social media: Implications for higher education," IT in Medicine and Education (ITME), 2011 International Symposium on, vol.1, no., pp.403,407, 9-11 Dec. 2011
- [14] Lockyer, L.; Patterson, J. "Integrating social networking technologies in education: a case study of a formal learning environment" Eighth IEEE International Conference on Advanced Learning Technologies. IEEE, 2008.
- [15] Tulaboev, A.; Oxley, A., "A case study on using Web 2.0 social networking tools in higher education," Computer & Information Science (ICCIS), 2012 International Conference on, vol.1, no., pp.84,88, 12-14 June 2012
- [16] Oksman, V., "Daddy, daddy, my computer has a fever!" Children and communication technologies in everyday life," Technology and Society, 2002. (ISTAS'02). 2002 International Symposium on, vol., no., pp.186,189, 2002
- [17] Chelly, M.; Mataillet, H., "Social Media and the impact on education: Social media and home education," e-Learning and e-Technologies in Education (ICEEE), 2012 International Conference on, vol., no., pp.236,239, 24-26 Sept. 2012
- [18] Y. Oida, L. Marshall and P. Brook. "The Invisible Actor". Editorial: New Ed. February, 2002. ISBN-13: 978-0413696106.
- [19] N. S. Saeed, Suku, "Adoption of Twitter in higher education: a pilot study," presented at the 28th Annual Conference of the Australasian Society for Computers in Learning in Tertiary Education, Hobart, Tasmania, Australia, 2011.

## With the support of



[www.nem-summit.eu](http://www.nem-summit.eu)